Circulation: Cardiovascular Quality and Outcomes

ORIGINAL ARTICLE

External Validations of Cardiovascular Clinical Prediction Models: A Large-Scale Review of the Literature

Benjamin S. Wessler, MD, MS; Jason Nelson, MPH; Jinny G. Park, MPH; Hannah McGinnes, MPH; Gaurav Gulati, MD; Riley Brazil, MD; Ben Van Calster, PhD; David van Klaveren, PhD; Esmee Venema, MD, PhD; Ewout Steyerberg, PhD; Jessica K. Paulus, PhD; David M. Kent, MD, MS

BACKGROUND: There are many clinical prediction models (CPMs) available to inform treatment decisions for patients with cardiovascular disease. However, the extent to which they have been externally tested, and how well they generally perform has not been broadly evaluated.

METHODS: A SCOPUS citation search was run on March 22, 2017 to identify external validations of cardiovascular CPMs in the Tufts Predictive Analytics and Comparative Effectiveness CPM Registry. We assessed the extent of external validation, performance heterogeneity across databases, and explored factors associated with model performance, including a global assessment of the clinical relatedness between the derivation and validation data.

RESULTS: We identified 2030 external validations of 1382 CPMs. Eight hundred seven (58%) of the CPMs in the Registry have never been externally validated. On average, there were 1.5 validations per CPM (range, 0–94). The median external validation area under the receiver operating characteristic curve was 0.73 (25th–75th percentile [interquartile range (IQR)], 0.66–0.79), representing a median percent decrease in discrimination of −11.1% (IQR, −32.4% to +2.7%) compared with performance on derivation data. 81% (n=1333) of validations reporting area under the receiver operating characteristic curve showed discrimination below that reported in the derivation dataset. 53% (n=983) of the validations report some measure of CPM calibration. For CPMs evaluated more than once, there was typically a large range of performance. Of 1702 validations classified by relatedness, the percent change in discrimination was −3.7% (IQR, −13.2 to 3.1) for closely related validations (n=123), −9.0 (IQR, −27.6 to 3.9) for related validations (n=862), and −17.2% (IQR, −42.3 to 0) for distantly related validations (n=717; $P<0.001$).

CONCLUSIONS: Many published cardiovascular CPMs have never been externally validated, and for those that have, apparent performance during development is often overly optimistic. A single external validation appears insufficient to broadly understand the performance heterogeneity across different settings.

Key Words: calibration ■ cardiovascular disease ■ decision making ■ literature review

Clinical prediction models (CPMs) are widely available to inform decisions in cardiovascular medicine. Our own database, the Tufts Predictive Analytics and Comparative Effectiveness (PACE) CPM Registry,[1] demonstrates continued growth of prediction models for patients with cardiovascular disease despite apparent substantial redundancy. The growth in the literature reflects the increasing ease with which these models can be developed, given the wide availability of both data and statistical software. Despite the publication

Correspondence to: Benjamin S. Wessler, MD, MS, Predictive Analytics and Comparative Effectiveness (PACE), Tufts Medical Center, 35 Kneeland St, Box No. 63, Boston, MA 02111. Email bwessler@tuftsmedicalcenter.org

The Data Supplement is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCOUTCOMES.121.007858.

For Sources of Funding and Disclosures, see page 910.

WHAT IS KNOWN

- There has been a proliferation of clinical prediction models (CPMs) to help risk-stratify patients at risk for cardiovascular disease. Clinically beneficial CPMs will yield accurate predictions for new patients and improve decision-making and clinical outcomes.

WHAT THE STUDY ADDS

- Here, we describe the extent to which CPMs have been validated and how performance varies across settings.
- We show that many CPMs have never been externally validated; for those that have, performance during model development is often overly optimistic and that isolated validations do not adequately capture CPM performance heterogeneity across different settings.

Nonstandard Abbreviations and Acronyms

AUROC	area under the receiver operating characteristic curve
CPM	clinical prediction model
EPV	events per included variable
IQR	interquartile range
PACE	Predictive Analytics and Comparative Effectiveness Center

of methodologic[2] and reporting guidelines[3] and a large set of potential performance metrics,[4] much remains unknown about the broad performance of these models, including the extent to which they have been validated, how well they validate, and how performance varies from one setting to another.

Although there are various ways to assess the performance of a statistical model,[4] clinically beneficial CPMs will yield accurate predictions on new cohorts (external validation)[5] and improve decision-making and subsequent clinical outcomes. Despite the increasing number of CPMs in the literature, how models perform generally during external validations and the determinants of that performance is largely unknown. Current reporting recommendations reinforce the need for external validation,[3] although recent analyses suggest that most CPMs either have not been externally validated[6] or have only been validated on a single external cohort.[7] CPM discriminatory performance cannot not be assumed to be stable (ie, equivalent to model performance at derivation) when tested in new settings.[8] Model calibration has been largely neglected and unless it is known to be excellent, CPMs may lead to harm if they are used to inform decisions at certain risk thresholds.[9,10]

Here, we perform a field synopsis of external validation studies of cardiovascular CPMs reported in a prior systematic review.[1] We aimed to describe the extent of external validation, variation in performance of models across databases, and to explore factors that are associated with worse model performance.

METHODS

Cardiovascular CPMs

The cardiovascular CPMs that form the basis of this review are found within the Tufts PACE CPM Registry. This registry represents a field synopsis of prediction models for patients at risk for and with known cardiovascular disease. All data and materials for this analysis have been made publicly available and can be accessed at www.pacecpmregistry.org. The search strategy and inclusion criteria have been previously reported.[1] Briefly, for inclusion in the Registry, an article must present the development of a cardiovascular CPM, contain a model predicting a binary clinical outcome, and the model must be presented in a way that allows prediction of outcome risk for a future patient. The search strategy for CPM identification was previously reported[1] and is presented in the Figure I in the Data Supplement. This analysis looked at cardiovascular CPMs published from 1990 through March 2015.

External Validation Search

A SCOPUS citation search of these cardiovascular CPMs was conducted on March 22, 2017. Citations were reviewed by 2 members of the study team to identify external validations of CPMs in the Registry. Discrepancies were reviewed by a third member of the research team. Consistent with prior work,[6] external validations were defined as any report that claimed to study the CPM for the same outcome as originally reported, but in a nonoverlapping population.

Data Extraction

Information about each CPM/validation pair was extracted, including sample size, continent of study, number of events, and reporting of measures of discrimination and calibration. CPM validation performance focused on discrimination (area under the receiver operating characteristic curve [AUROC]) change compared with the AUROC seen in the derivation population. We also document whether validations include any assessment of CPM calibration. There are many methods to assess model calibration and only recent consensus on best practices.[4,11] Given this lack of consistency and interpretability in the literature, we report whether or not this dimension of performance was assessed during external validation. Calibration assessment included any comparison of observed versus expected outcomes. Examples include a Hosmer-Lemeshow statistic or calibration plot. For this study, we also included measures of calibration-in-the-large, where overall observed event rates are compared with predicted rates.

CPM Performance

Consistent with prior work,[12] changes in CPM discrimination from derivation to validation are described on a scale of 0% (no change in discrimination) to −100% (complete loss

of discrimination) because it more intuitively reflects the true changes in discriminatory power.[13] Positive changes represent improvements in discrimination. The percent change in discrimination is calculated using the following equation ([validation AUROC−0.5]−[derivation AUROC−0.5]/(derivation AUROC−0.5]×100).

Population Relatedness

To explore potential explanations for decreased performance on validation data sets, we assessed the similarity between the derivation and validation populations by creating detailed relatedness rubrics for the 10 index conditions with the greatest number of CPMs (Table I in the Data Supplement). These rubrics were created by investigators with expertise in these clinical areas. Relatedness was assessed for each CPM/validation pair to divide validation databases into 3 categories—closely related, related, and distantly related. A fourth category no match was assigned to validations that were excluded from the analysis because they were not clinically appropriate matches (eg, CPM validated on population with nonoverlapping index condition or outcome). Generally, the relatedness rubrics were based on 5 domains: (1) recruitment setting (eg, outpatient versus emergency room versus inpatient), (2) major inclusion/ exclusion criteria, (3) intervention type (eg, percutaneous coronary intervention versus thrombolysis for acute myocardial infarction), (4) therapeutic era, (5) follow-up time. Two clinicians reviewed these domains for each CPM/validation match and assigned a relatedness category. Nonrandom split-sample validations were labeled as closely related validations. Discrepancies were reviewed by the study team to arrive at a consensus.

Factors Associated With CPM External Validation

We identified a set of study-level factors to evaluate associations with whether or not a CPM was externally validated. These factors were identified based on observed methodologic and reporting patterns as well as prior literature.[8] These factors included: Index clinical condition, internal validation performed, year of publication (divided here before 2004, 2004–2009, 2009–2012, after 2012), continent of origin, study design (eg, clinical trial versus medical record), sample size, number of events, number of predictors, prediction time horizon (<30 days, 30–265 days, >365 days), regression method (eg, logistic regression versus Cox regression), and reporting of discrimination or calibration. We analyzed unadjusted associations and used multivariable logistic regression to assess whether these variables were associated with CPM external validation.

Factors Associated With Poor Performance

A set of study-level factors defined a priori were evaluated for association with worse CPM performance (discrimination) during validation. These factors included: population relatedness (here, dichotomized as distantly related versus other), presence of overlapping authors, same or different article, CPM modeling method, CPM data source, validation data source, outcome rate difference between derivation and validation data (defined as > versus ≤40%), CPM events per included variable (EPV). We used generalized estimating equations[14,15] with robust covariance estimator to assess the multivariable association with the observed change in discrimination, taking into account the correlation between validations of the same CPM. Multiple imputation of 20 imputed data sets was used to account for missingness. These analyses estimated the absolute difference in the estimated percent change in the C statistic from derivation to validation populations, as calculated above. All statistical analyses were performed using SAS Enterprise Guide version 8.2 (SAS Institute, Inc, Cary, NC).

RESULTS

Overview of Validations

The Registry includes 1382 CPMs for cardiovascular disease and the citation search of these CPMs identified 54 086 citations that were screened (Figure 1). These citations identified 14 615 abstracts that were screened to identify 6039 full-text articles. A total of 2030 external validations were extracted from 413 articles. Only 575 (42%) of the CPMs in the Registry have ever been validated (Table 1). On average, there were 1.5 validations per de novo CPM, with a very skewed distribution. The Logistic European System for Cardiac Operative Risk Evaluation[16] has been externally validated 94×. For this analysis, we included 1846 validations of 556 CPMs after exclusion of 19 decision trees and 156 validations performed on unrelated (ie, populations with different index conditions or nonoverlapping outcomes) samples. The median external validation sample size was 861 (25th–75th percentile [interquartile range (IQR)] 326–3306), and the median number of outcome events was 68 (IQR, 29–192; Table 2).

CPM Validation Discrimination

Overall, 91.3% (n=1685) of the external validations report AUROC. The median derivation AUROC was 0.77 (IQR, 0.73–0.82). The median external validation AUROC was 0.73 (IQR, 0.66–0.79) representing a median percent change in discrimination of -11.1% (IQR, −32.4% to +2.7%; Table 2). Of the validations with decreased performance (n=795), 25% (n=195) had <10% decrement in discrimination. Two percent (n=35) had >80% drop in discrimination; 19% (n=352) of model validations showed CPM discrimination at or above the performance reported in the derivation dataset.

CPM Calibration

In total, 53% (n=983) of the validations report some measure of CPM calibration. The Hosmer-Lemeshow test of goodness-of-fit was most commonly reported (30%, n=555) followed by calibration-in-the-large (26%, n=488), and calibration plots (22%, n=399). (Table 2). Overall, there was no externally assessed calibration information available for 86% (n=1182) of the CPMs in the Registry.

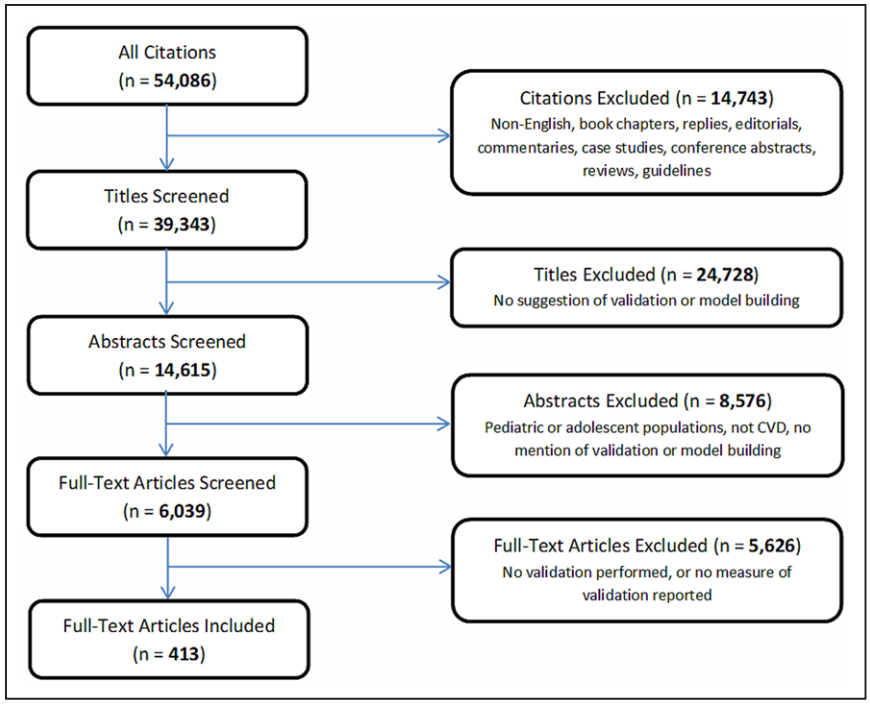

Figure 1. Flowchart of external validation review process.

Clinical Domains

The ten conditions with the most CPM validations comprised 92% (1702/1846) of the total validations included in this analysis (Table 3). The condition with the largest number of validations was stroke (299 validations performed on 104 CPMs). There were a total of 286 validations of 87 CPMs for populations at risk for developing cardiovascular disease (population samples) and 286 validations of 52 CPMs for Cardiac Surgery. Only 5 index conditions had ≥50% of available CPMs externally validated (arrhythmias [81%], valve disease [62%], venous thromboembolism [53%], cardiac surgery [51%], and aortic diseases [50%]). There is an extreme range of CPM performance and consistent loss of discriminatory performance during external validations (Figure 2, Table 3). These observations were apparent for all conditions that were studied (specific condition waterfall analyses shown in Figure II in the Data Supplement).

Relatedness

Relatedness was assigned to each of the 1702 of the CPM/validation pairs for the top 10 index conditions. Of these, 123 (7%) of the validations were performed on

Table 1. De Novo Models Summary

	Overall	Validated*	Never validated
Models	1382	556	807
Validations per model, mean (range)	1.5 (0–94)	3.3 (1–94)	0
Cohort size	1728 (509–6198)	2130 (688–8978)	1227 (405–5303)
Events	165 (71–456)	198 (88–649)	144 (63–328)
EPV final model†	22.3 (11.3–50.6)	26.1 (12.2–70.6)	20.4 (11.0–42.7)
C statistic‡	0.77 (0.725–0.821)	0.78 (0.73–0.83)	0.77 (0.72–0.816)
Multicenter, N (%)	830 (60)	380 (68)	438 (54)
Any calibration, N (%)	603 (44)	268 (48)	332 (41)
Hosmer-Lemeshow test, N (%)	415 (30)	189 (34)	225 (28)
Calibration plot, N (%)	254 (18)	112 (20)	141 (17)
Calibration-in-the-large, N (%)	82 (6)	44 (8)	37 (5)

Characteristics of unique CPMs in PACE CPM Registry in aggregate for all CPMs, CPMs that have ever been validated, and CPMs never validated. All values reported as median (IQR) unless otherwise noted. CPM, clinical prediction model; EPV, events per included variable; IQR, interquartile range; and PACE, Predictive Analytics and Comparative Effectiveness.

*Includes validations in sample set (Table 2).

†EPV refers to the calculation of events per included variable in the final model, not candidate variables.

‡C statistic reported in 91.3% of exercises.

Table 2. External Validations Summary

	Sample set*	Relatedness set†	Closely related‡	Related	Distantly related
Validation exercises	1846 (556 models)	1702 (483 models)	123 (117 models)	862 (301 models)	717 (216 models)
Cohort	861 (326 to 3306)	882 (330 to 3479)	1460 (494 to 4905)	960 (413 to 4492)	681 (256 to 2152)
Events	68 (29 to 192)	67 (28 to 188)	144 (48 to 275)	72 (31 to 201)	52 (25 to 158)
EPV final model§	6.9 (2.5 to 22.9)	6.3 (2.4 to 21.1)	14.1 (6.7 to 33.9)	6.5 (2.3 to 20.6)	5.6 (2.1 to 18.5)
C statistic‖	0.73 (0.66 to 0.794)	0.73 (0.664 to 0.796)	0.78 (0.719 to 0.841)	0.75 (0.68 to 0.803)	0.701 (0.64 to 0.77)
% Change in discrimination	−11.1 (−32.4 to 2.7)	−11.1 (−32 to 2.6)	−3.7 (−13.2 to 3.1)	−9.0 (−27.6 to 3.9)	−17.2 (−42.3 to 0)
Multicenter, N (%)	779 (42)	717 (42)	70 (57)	338 (39)	309 (43)
Any calibration, N (%)	983 (53)	930 (55)	66 (54)	542 (63)	322 (45)
Hosmer-Lemeshow test, N (%)	555 (30)	527 (31)	36 (29)	310 (36)	181 (25)
Calibration plot, N (%)	399 (22)	378 (22)	28 (23)	236 (27)	114 (16)
Calibration-in-the-large, N (%)	488 (26)	480 (28)	18 (15)	292 (34)	170 (24)

Characteristics of external validations of CPMs in PACE CPM Registry, stratified by inclusion in analysis sample, CPMs in top 10 most validated index conditions, and by relatedness category. All values reported as median (IQR) unless otherwise noted. CPM, clinical prediction model; EPV, events per included variable; IQR, interquartile range; and PACE, Predictive Analytics and Comparative Effectiveness.

*Excluded decision tree, classification and regression tree, and mismatched index condition validations.

†CPMs comprising top 10 most validated index conditions in CPM Registry (acute coronary syndrome, aortic disease, arrhythmia, cardiac surgery, chronic heart failure, population sample, revascularization, stroke, valve disease, venous thromboembolism).

‡Validation is split-sample external validation, as defined by Steyerberg.[17]

§EPV refers to the calculation of events in the validation exercise per included variable in the final model, not candidate variables.

‖C statistic reported in 91% of exercises.

closely related populations, 862 (51%) were performed on related populations, whereas 717 (42%) were performed on distantly related populations (Table 2). The median AUROC for closely related validations was 0.78 (IQR, 0.719–0.841). The median AUROC for related population validations was 0.75 (IQR, 0.68–0.803). The median AUROC for distantly related validations was 0.70 (IQR, 0.64–0.77; $P<0.001$). Overall, the median percent change in discrimination was -3.7% (IQR, −13.2 to 3.1) for closely related validations, −9.0% (IQR, −27.6 to 3.9) for related validations, and −17.2% (−42.3 to 0) for distantly related validations ($P<0.001$).

Range of Performance for Individual CPMs

Table 4 shows the variation in performance across the 10 CPMs[16,18–26] that were validated most frequently. Uniformly, there was a substantial range in performance of each CPM across datasets, from virtually useless to excellent. For example, discrimination for the Logistic European System for Cardiac Operative Risk Evaluation (validated 94×) ranged from 0.48 to 0.90 across different databases. None of these highly cited (and validated) CPMs had consistently good discrimination across validation databases.

Predictors of External Validation

Study features that are associated with CPM external validation (yes/no) are shown in Table II in the Data Supplement. The index condition was strongly associated with subsequent external validation. Models that were internally validated and models that were published more recently were less likely to be externally validated. Sample size, number of predictors, and reporting of discrimination or calibration were positively associated with external validation. On multivariable analysis, these predictors remained associated with CPM external validation. Study design, prediction time horizon, and regression method were not apparently associated with a model being externally validated.

Predictors of Poor Performance

Predictors of CPM validation performance are shown in Table 5. On univariate analysis, population relatedness was significantly associated with CPM discrimination in validations. When CPMs were tested on distantly related cohorts, the AUROC decrease was −15.6% (95% CI, −22.0 to −9.1) compared with the reference (validations done on closely related cohorts). When evaluated in a multivariable model, population relatedness remained significantly associated with CPM discrimination in validations (−9.8% [95% CI, −18.8 to −0.8). We also observed that validations demonstrated AUROCs that were 9.8% (95% CI, 5.4–14.2) higher when reported in the same article (with the same authors) as the de novo CPM report compared with validations reported in different articles with nonoverlapping authors. There was a trend toward higher AUROC (+7.3% [95% CI, −1.2 to 15.8], $P=0.09$) when validations were reported by overlapping authors in a subsequent publication (compared with reports by nonoverlapping authors).

Table 3. Conditions With the Most External Validations (Top 10)

							Delta C, median (Q1, Q3)		
Index condition	Validated CPMs (% of total)	Validations	Closely related	Related	Distantly related	N missing	Closely related	Related	Distantly related
Stroke	104 (48)	299	5 (1.7)	127 (42.5)	167 (55.9)	69	−7.1 (−12.8 to 2.8)	−6.9 (−17.7 to 3.5)	−12.9 (−33.3 to 0.4)
Cardiac surgery	52 (51)	286	19 (6.6)	216 (75.5)	51 (17.8)	141	5.9 (−26.9 to 8.9)	−10.3 (−27.6 to 6.9)	−17.2 (−43.4 to 2.9)
Population sample	87 (38)	286	7 (2.5)	162 (56.6)	117 (40.9)	162	−8.7 (−13.3 to −3.6)	−15.2 (−38.2 to −1.1)	−16.1 (−48.6 to 1.3)
ACS	57 (45)	209	20 (9.6)	85 (40.7)	104 (49.8)	59	−2.3 (−10.3 to 4.0)	−3.2 (−17.8 to 7.8)	−11.7 (−39.1 to 2.2)
Valve disease	37 (62)	202	12 (5.9)	71 (35.2)	119 (58.9)	56	−8.9 (−17.9 to 0)	−6.3 (−32.5 to 2.3)	−31.8 (−52.3 to −11.5)
Arrhythmia	17 (81)	98	3 (3.1)	55 (56.1)	40 (40.8)	11	−12.7 (−13.6 to −11.8)	−17.3 (−50.0 to 44.8)	−31.8 (−70.1 to 46.2)
CHF	47 (32)	92	18 (19.6)	47 (51.1)	27 (29.4)	30	−3.7 (−13.5 to 3.5)	−21.3 (−29.4 to −5.5)	−17.2 (−28.9 to −5.2)
Revascularization	47 (36)	92	27 (29.4)	37 (40.2)	28 (30.4)	21	−1.1 (−14.9 to 1.6)	−1.7 (−16.7 to 2.6)	−15.1 (−25.0 to −5.9)
Aortic disease	31 (50)	72	7 (9.7)	32 (44.4)	33 (45.8)	65	NA	−7.9 (−10.3 to −0.9)	39.4 (39.4 to 39.4)
VTE	27 (53)	66	5 (7.6)	30 (45.5)	31 (47.0)	34	−10 (−17.2 to −5.0)	−3.4 (−19.2 to 28.6)	−7.4 (−25.0 to 6.2)
Total	506 (44)	1702	123 (7.2%)	862 (50.6%)	717 (42.1%)	648	−3.7 (−13.3 to 3.5)	−9 (−27.6 to 3.9)	−17.2 (−42.4 to 0)

Discrimination and characteristics of validations for CPMs with top 10 most validated index conditions in PACE CPM Registry. N missing refers to either derivation AUC or validation AUC missing (delta C not available). ACS indicates acute coronary syndrome; AUC, area under curve; CHF, chronic heart failure; CPMs, clinical prediction models; NA, not applicable; PACE, Predictive Analytics and Comparative Effectiveness; and VTE, venous thromboembolism.

DISCUSSION

Our Tufts PACE CPM Registry documents the tremendous proliferation and redundancy of CPMs being developed and published. The review reported here underscores that this proliferation is occurring without adequate—or even minimal—external evaluation. Approximately 60% of published CPMs have never been externally validated. Approximately half of the CPMs that have been validated only once. A small minority of models have been validated numerous times. The value of single validations is unclear because there is substantial performance heterogeneity and good (or poor) performance on a single validation does not appear to reliably forecast performance on subsequent validations. No CPM showed consistently good discrimination across multiple validation databases. For example, the 10 most validated CPMs have each been validated >20×; all show substantial variation in discrimination across these validation studies, from virtually useless (ie, C statistic=≈0.5) to very good (C statistic=~0.8 or higher). This demonstrates the difficulty of defining the quality of a model generically because performance greatly depends on characteristics of the database on which a model is tested. These findings underscore recent calls for a fundamental paradigm shift in how models are assessed for validity and utility[7] and calls for more robust stewardship of algorithms for health care.[27]

The majority of cardiovascular CPMs in our Registry have never been externally validated. This finding mirrors an observation made in previous assessment of primary prevention models[8,28] and broadly suggests that cardiovascular clinicians should be skeptical about the accuracy of individual risk estimates. In our registry, model level predictors associated with subsequent external validation include the disease being studied and also larger sample size, higher outcome rates, and whether discrimination or calibration were reported in the original presentation. Older CPMs were generally more likely to be externally validated—an observation that may relate to insufficient time to allow for validation of more recently published CPMs. Given the extreme redundancy of CPMs and the relative scarcity of external validations, it seems reasonable to prioritize the study of existing cardiovascular CPMs (as opposed to developing new ones), and how these might be optimized for clinical use.

Although this review focuses on external validations, this emphasis does not imply that internal validation is not important. Internal and external validation provide different information. Internal validation is especially important when the sample size and the number of outcomes are relatively small for the complexity of the

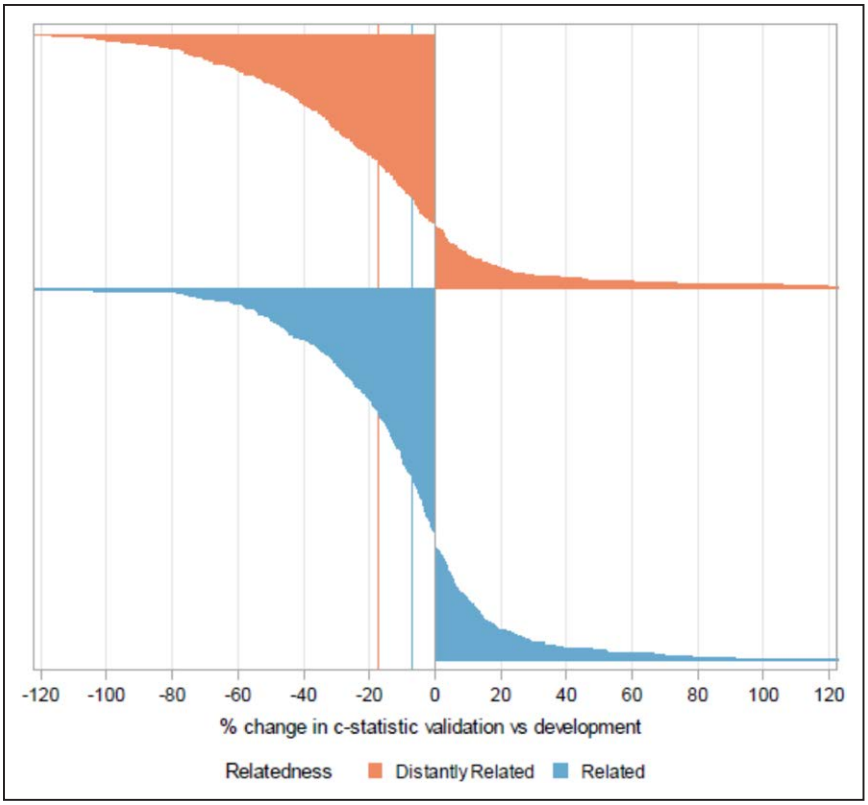

Figure 2. Waterfall plot depicting the percent change in the C statistic in related (related and closely related) validations (in blue) and distantly validations (in orange).
Plots comprise horizontal lines representing a total of 1701 validations that present C statistic that can be compared with the development C statistic. Vertical lines show that the median decrement in discrimination was more pronounced in the distantly related models than the related models.

model-building procedure. In such cases, reporting the apparent performance is likely over-optimistic. External validation provides information about the transportability of the model to other settings and across time, and how robust predictions are to distributional shifts in the data. Combination of internal-external validation procedures to assess CPM performance may represent best practice to broadly understand CPM performance.[29] Yet for those charged with deciding whether a given model is deployed in clinical practice, understanding how a model performs in the local setting may be most important. Our work suggests that this may be difficult to understand from the literature, especially if the target of inference is a setting other than where a model was validated.

It was common to observe substantial decrements in discrimination during validations. This finding is consistent with prior reports that have shown CPM validation discriminatory ability that is highly variable and often worse than anticipated (when compared with performance on the derivation database).[6,8] There are several potential reasons why model performance might decrease, including model invalidity (eg, due to

Table 4. Top 10 Most Validated CPMs

Model Name	Index condition	No. of validations	Development AUC	Median validation AUC (IQR)	Range in validation AUC
Logistic EuroSCORE[16]	Cardiac surgery	94	NR	0.75 (0.67–0.80)	0.48–0.90
Additive EuroSCORE[18]	Cardiac surgery	86	0.79	0.77 (0.72–0.82)	0.58–0.90
EuroSCORE II[19]	Valve disease	65	0.81	0.76 (0.68–0.81)	0.40–0.87
GRACE[20]	CAD: ACS	53	0.83	0.80 (0.73–0.84)	0.60–0.95
STS (valve)–mortality[21]	Cardiac surgery	51	0.81	0.70 (0.64–0.76)	0.45–0.85
CHA_2DS_2–VASc[22]	Arrhythmia	45	0.61	0.66 (0.61–0.69)	0.45–0.93
$CHADS_2$[23]	Arrhythmia	37	0.82	0.65 (0.61–0.68)	0.51–0.87
FRS–CHD[24]	Population sample	35	NR	0.68 (0.63–0.72)	0.54–0.80
ICH score[25]	Stroke	27	0.92	0.85 (0.75–0.87)	0.69–0.94
ACEF score[26]	Cardiac surgery	26	0.74	0.74 (0.68–0.77)	0.54–0.87

Description of top 10 most validated CPMs in PACE CPM Registry and validation performance. ACEF indicates age, creatinine, and left ventricular ejection fraction; ACS, acute coronary syndrome; AUC, area under curve; CAD, coronary artery disease; CPMs, clinical prediction models; EuroSCORE, European System for Cardiac Operative Risk Evaluation; FRS–CHD, Framingham Risk Score for Coronary Heart Disease; GRACE, Global Registry of Acute Coronary Events; ICH, intracerebral hemorrhage; IQR, interquartile range; NR, not reported; PACE, Predictive Analytics and Comparative Effectiveness; and STS, Society of Thoracic Surgeons.

Table 5. Predictors of Worse Discrimination: Variable Distributions and GEE Model Results

		Univariate			Multivariable (n=1054)*	
		N	Delta AUC difference (95% CI)	*P* value	Delta AUC difference (95% CI)	*P* value
Relatedness, n (%)	Frequency N/A=93 (8.1%)	1054				
Closely related	79 (7.5)		Reference		Reference	
Related	544 (51.6)		−5.9 (−10.5 to −1.4)	0.011	−1.3 (−7.1 to 4.5)	0.660
Distantly related	431 (40.9)		−15.6 (−22.0 to −9.1)	<0.001	−9.8 (−18.8 to −0.8)†	0.033†
CPM authors, n (%)		1147				
Diff article, author overlap	94 (8.2)		7.3 (−1.2 to 15.8)	0.092	5.1 (−4.2 to 14.4)	0.283
Diff article, no author overlap	849 (74.0)		Reference		Reference	
Same article	204 (17.8)		9.8 (5.4 to 14.2)	<0.001	5.5 (−0.8 to 11.9)	0.088
CPM method, n (%)	Frequency missing=20 (1.7%)	1127				
Logistic regression	859 (76.2)		Reference		Reference	
Other	7 (0.6)		−0.1 (−12.4 to 12.2)	0.985	−1.1 (−13.7 to 11.6)	0.870
Time-to-event regression	261 (23.2)		2.6 (−5.8 to 11)	0.541	−1.4 (−11 to 8.2)	0.768
CPM data source, n (%)	Frequency missing=4 (0.3%)	1143				
Clinical trial	118 (10.3)		5.8 (−8.9 to 20.5)	0.437	3.7 (−14.1 to 21.5)	0.684
Medical record	614 (53.7)		Reference		Reference	
Other	114 (10.0)		−1.3 (−11.1 to 8.6)	0.803	−2.5 (−14.0 to 9.0)	0.669
Registry	297 (26.0)		2.3 (−5.5 to 10)	0.569	2.1 (−6.3 to 10.5)	0.616
Validation data source, n (%)	Frequency missing=47 (4.1%)	1100				
Clinical trial	99 (9.0)		−5.9 (−12.5 to 0.6)	0.077	−7.3 (−15.4 to 0.8)	0.076
Medical record	606 (55.1)		Reference		Reference	
Other	58 (5.3)		3.9 (−5.2 to 13)	0.402	3.3 (−5.1 to 11.7)	0.440
Registry	337 (30.6)		1.5 (−3.5 to 6.5)	0.560	1.2 (−3.5 to 6.0)	0.606
Relative outcome rate difference > 40%, n (%)	Frequency missing=402 (35.0%)	745				
Yes	384 (51.5)		−4.4 (−9.4 to 0.7)	0.091	−1.5 (−6.3 to 3.3)	0.540
No	361 (48.5)		Reference		Reference	
CPM EPV, median (IQR)	Frequency missing=214 (18.7%) 23.4 (16.3 to 58.8)	933	0.4 (−1.9 to 2.7)‡	0.718	1.8 (−1.1 to 4.6)	0.218

Results of regression analysis to detect predictors of change in discrimination performance from derivation to validation. EPM: events per included variable in the final model, not candidate variables. AUC indicates area under curve; CPM, clinical prediction model; Diff, different; EPV, events per included variable; GEE, generalized estimating equation; IQR, interquartile range; and N/A, not assessed.

*Multiple imputation for missing data (20 imputed data sets).

†significant with $p<0.05$

‡Natural log-transformed.

over-fitting on the derivation population) and a change in case mix.[5] Model invalidity might be expected to be more pronounced when models are evaluated in populations that are dissimilar to the derivation population. We found that models had a substantially larger decrease in discriminatory performance when tested on distantly related populations compared with either related or closely related populations. However, judging the relatedness of the populations is laborious and requires substantial clinical expertise. Differences that may appear subtle can be very influential. For example, a CPM developed on patients in the emergency room might not be expected to have similar discriminatory performance if the validation cohort includes only patients admitted to the hospital since—as in the case of many acute cardiac syndromes—care[30] and outcome predictors[31] are different very early in the disease course. So too changes in treatments received (eg, different ACS revascularization approaches,[32] stent types,[33] or outcome definitions[34,35]) likely impact model validation performance. If the model was derived on patients receiving lytic therapy and validated using data from a more contemporary percutaneous coronary intervention trial, it should not be surprising that model performance appears worse than expected. Other study-level characteristics we examined apart from relatedness did not appear to greatly influence model performance.

One of the most striking observations of this work is that isolated validations appear insufficient to understand the performance of CPMs when tested in new

populations. There was often an extreme range in performance for CPMs evaluated in multiple databases—an observation that calls into question the generalizability of any one validation result. These data challenge the current approach in which a model might be evaluated on a single external population and then declared to be a validated prediction model that is ready for use. Even when a model performs well using statistical criteria, it is unclear whether such a model improves decision-making when used on a closely related population. Further, good statistical performance on one external database does not guarantee good statistical performance in another setting—such as where a CPM is eventually used to support care. There is no evidence from our analysis that so-called validated CPMs that have been integrated into clinical practice guidelines[36,37] should be accepted as trustworthy unless CPM performance is specifically known to be excellent on populations like those being treated. Although having a single CPM that is accepted by the clinical community and promoted in guidelines is appealing as a means of standardizing practice across a range of different settings, the degree of variation seen in our review suggests that this paradigm may result in substantial variation of performance across different settings and poor performance in some settings. Testing CPMs for improved decision-making and better clinical outcomes (eg, in a cluster-randomized trial[38]) is rarely performed before dissemination into practice. Novel paradigms, emphasizing increasing the accuracy of model performance on local populations, through continual recalibration and updating, are an appealing approach that deserves further consideration.

There are several potential reasons why external validations of prediction models are so rare. First, model developers typically exhaust their data deriving (and sometimes internally validating) their model and may not have additional data sources. Second, informally, there appear to be much stronger academic incentives for the development of new models, rather than the validation of previously published models. Third, there is limited understanding that it is informative to test and retest a validated model on new data to understand how robust predictions are to distributional shifts over time and settings. This is supported indirectly by the observation that internally validated models appear to be less likely to be externally validated than other models. Finally, as predictive modeling methodologic and reporting standards have been published and adopted,[2,3] there remain few standards for how best to conduct and report on validations of existing models.

Our review has several limitations. First, the review was limited by the information collected and presented in the original articles. We relied on changes in discrimination largely because CPM calibration is woefully underassessed. Only 62% of models in the CPM Registry have had calibration formally assessed in an external population; even among the models that were validated only 48% report any calibration. Finally, even when calibration is reported, it is usually reported in a form that is not clinically interpretable (eg, as a Hosmer-Lemeshow statistic[4,13]) or graphically (easy to summarize according to calibration slope [ideal: 1] and systematic under or overestimation [intercept ideally 0]). Some less frequently used metrics, such as the integrated calibration index,[39] may help compare performance across multiple validations. Decrements in calibration may be as serious as, or even more serious than, decrements in discrimination because miscalibrated models yield misinformation which may cause harmful decision-making.[9] Ideally, we would be able to evaluate the net benefit of model use, which integrates discrimination, calibration, and relative utility to compare the value of prediction-based decision-making compared with best one-sized-fits-all strategies.[4,40] Such evaluations would have required individual patient data because these approaches are so rarely used in the published literature. Similarly, we could not assess how much of the decrement in discrimination was due to differences in case mix, rather than invalidity, which would have also required evaluation of patient-level data.[41] Finally, our systematic review does not include more recent validations after 2017, due to the enormous scope of this literature, the lack of efficient search strategies, and the laborious nature of comprehensive data extraction and evaluation of relatedness. We do not anticipate the more recent literature would substantially change our findings. Maintenance and continual updating data of this registry will registry will require a semiautomated approach heavily reliant on natural language processing.[42]

CONCLUSIONS

Many published cardiovascular CPMs have never been externally validated, and for those that have, it is common to see significant performance heterogeneity and marked decreases in the discriminatory performance compared with the model development phase. Calibration has been widely underassessed, and single validations do not sufficiently capture CPM performance. Granular information about population relatedness is associated with CPM performance in external validations, and when CPMs are tested on distantly related populations, model performance is often substantially worse than expected. This review raises substantial concerns about the current approach to validating cardiovascular CPMs and underscores the need for a radical rethinking for how performance heterogeneity is explored and quantified (eg, through multiple validations across various practice settings) and how models are evaluated for clinical use.

ARTICLE INFORMATION

Received January 13, 2021; accepted June 8, 2021.

Affiliations

Predictive Analytics and Comparative Effectiveness (PACE) (B.S.W., J.N., J.G.P., H.G., G.G., R.B., D.v.K., J.K.P., D.M.K.) and Division of Cardiology (B.S.W., G.G.), Tufts Medical Center, Boston, MA. KU Leuven, Department of Development and Regeneration, Belgium (B.V.C.). Department of Biomedical Data Sciences (D.v.K.) and Department of Biomedical Data Sciences (E.S.), Leiden University Medical Centre, Netherlands. Department of Public Health (E.V., E.S.) and Department of Neurology (E.V.), Erasmus MC University Medical Center, Rotterdam, the Netherlands.

Acknowledgments

We wish to acknowledge the contributions of Vandan Patel for his work on the relatedness effort.

Sources of Funding

Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1606-35555). The views, statements, opinions presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of the PCORI, its Board of Governors, or Methodology Committee. Dr Wessler is supported by K23AG055667 from National Institutes of Health (NIH)–National Institute on Aging (NIA) and R03AG056447 from NIH-NIA.

Disclosures

None.

Supplemental Materials

Figures I and II
Tables I and II

REFERENCES

1. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, Jethmalani N, Raman G, Lutz JS, Kent DM. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. *Diagn Progn Res*. 2017;1:20. doi: 10.1186/s41512-017-0021-2
2. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG; PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. 2013;10:e1001381. doi: 10.1371/journal.pmed.1001381
3. Collins GS, Reitsma JB, Altman DG, Moons KG; TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. *Circulation*. 2015;131:211–219. doi: 10.1161/CIRCULATIONAHA.114.014508
4. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138. doi: 10.1097/EDE.0b013e3181c30fb2
5. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*. 2010;172:971–980. doi: 10.1093/aje/kwq223
6. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68:25–34. doi: 10.1016/j.jclinepi.2014.09.007
7. Adibi A, Sadatsafavi M, Ioannidis JPA. Validation and utility testing of clinical prediction models: time to change the approach. *JAMA*. 2020;324:235–236. doi: 10.1001/jama.2020.1230
8. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416. doi: 10.1136/bmj.i2416
9. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230. doi: 10.1186/s12916-019-1466-7
10. Van Calster B, Vickers AJ. Calibration of risk prediction models. *Med Decis Making*. 2015;35:162–169. doi: 10.1177/0272989X14547233
11. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167–176. doi: 10.1016/j.jclinepi.2015.12.005
12. Wessler BS, Lundquist CM, Koethe B, Park JG, Brown K, Williamson T, Ajlan M, Natto Z, Lutz JS, Paulus JK, et al. Clinical prediction models for valvular heart disease. *J Am Heart Assoc*. 2019;8:e011972. doi: 10.1161/JAHA.119.011972
13. Harrell FE. *Regression Modeling Strategies*. Springer International Publishing; 2015.
14. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42:121–130.
15. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988;44:1049–1060.
16. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. *Eur Heart J*. 2003;24:881–882. doi: 10.1016/s0195-668x(02)00799-6
17. Steyerberg EW. *Clinical Prediction Models*. Springer New York; 2009.
18. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg*. 1999;16:9–13. doi: 10.1016/s1010-7940(99)00134-7
19. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, Lockowandt U. EuroSCORE II. *Eur J Cardiothorac Surg*. 2012;41:734–44; discussion 744. doi: 10.1093/ejcts/ezs043
20. Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, Van De Werf F, Avezum A, Goodman SG, Flather MD, et al; Global Registry of Acute Coronary Events Investigators. Predictors of hospital mortality in the global registry of acute coronary events. *Arch Intern Med*. 2003;163:2345–2353. doi: 10.1001/archinte.163.19.2345
21. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, Normand SL, DeLong ER, Shewan CM, Dokholyan RS, et al; Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2–isolated valve surgery. *Ann Thorac Surg*. 2009;88(1 Suppl):S23–S42. doi: 10.1016/j.athoracsur.2009.05.056
22. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137:263–272. doi: 10.1378/chest.09-1584
23. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*. 2001;285:2864–2870. doi: 10.1001/jama.285.22.2864
24. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. 1998;97:1837–1847. doi: 10.1161/01.cir.97.18.1837
25. Hemphill JC 3rd, Bonovich DC, Besmertis L, Manley GT, Johnston SC. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage. *Stroke*. 2001;32:891–897. doi: 10.1161/01.str.32.4.891
26. Ranucci M, Castelvecchio S, Menicanti L, Frigiola A, Pelissero G. Risk of assessing mortality risk in elective cardiac operations: age, creatinine, ejection fraction, and the law of parsimony. *Circulation*. 2009;119:3053–3061. doi: 10.1161/CIRCULATIONAHA.108.842393
27. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA*. 2020;324:1397–1398. doi: 10.1001/jama.2020.9371
28. Ban JW, Stevens R, Perera R. Predictors for independent external validation of cardiovascular risk clinical prediction rules: Cox proportional hazards regression analyses. *Diagn Progn Res*. 2018;2:3. doi: 10.1186/s41512-018-0025-6
29. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–247. doi: 10.1016/j.jclinepi.2015.04.005
30. Collins SP, Levy PD, Lindsell CJ, Pang PS, Storrow AB, Miller CD, Naftilan AJ, Thohan V, Abraham WT, Hiestand B, et al. The rationale for an acute heart failure syndromes clinical trials network. *J Card Fail*. 2009;15:467–474. doi: 10.1016/j.cardfail.2008.12.013
31. Karam N, Bataille S, Marijon E, Tafflet M, Benamer H, Caussin C, Garot P, Juliard JM, Pires V, Boche T, et al; e-MUST Study Investigators. Incidence, mortality, and outcome-predictors of sudden cardiac arrest complicating myocardial infarction prior to hospital admission. *Circ Cardiovasc Interv*. 2019;12:e007081. doi: 10.1161/CIRCINTERVENTIONS.118.007081
32. Mehta SR, Wood DA, Storey RF, Mehran R, Bainey KR, Nguyen H, Meeks B, Di Pasquale G, López-Sendón J, Faxon DP, et al; COMPLETE Trial Steering Committee and Investigators. Complete revascularization with multivessel PCI for myocardial infarction. *N Engl J Med*. 2019;381:1411–1421. doi: 10.1056/NEJMoa1907775

33. Piccolo R, Bonaa KH, Efthimiou O, Varenne O, Baldo A, Urban P, Kaiser C, Remkes W, Räber L, de Belder A, et al; Coronary Stent Trialists' Collaboration. Drug-eluting or bare-metal stents for percutaneous coronary intervention: a systematic review and individual patient data meta-analysis of randomised clinical trials. *Lancet*. 2019;393:2503–2510. doi: 10.1016/S0140-6736(19)30474-X
34. Kip KE, Hollabaugh K, Marroquin OC, Williams DO. The problem with composite end points in cardiovascular studies: the story of major adverse cardiac events and percutaneous coronary intervention. *J Am Coll Cardiol*. 2008;51:701–707. doi: 10.1016/j.jacc.2007.10.034
35. Mehran R, Rao SV, Bhatt DL, Gibson CM, Caixeta A, Eikelboom J, Kaul S, Wiviott SD, Menon V, Nikolsky E, et al. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. *Circulation*. 2011;123:2736–2747. doi: 10.1161/CIRCULATIONAHA.110.009449
36. Goff DC Jr, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB Sr, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2014;63(25 pt B):2935–2959. doi: 10.1016/j.jacc.2013.11.005
37. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Drazner MH, Fonarow GC, Geraci SA, Horwich T, Januzzi JL, et al. 2013 ACCF/AHA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. *Circulation*. 2013;128:1810–1852. doi: 10.1161/CIR.0b013e31829e8807
38. Chew DP, Hyun K, Morton E, Horsfall M, Hillis GS, Chow CK, Quinn S, D'Souza M, Yan AT, Gale CP, et al. Objective risk assessment vs standard care for acute coronary syndromes: a randomized clinical trial. *JAMA Cardiol*. 2021;6:304–313. doi: 10.1001/jamacardio.2020.6314
39. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med*. 2019;38:4051–4065. doi: 10.1002/sim.8281
40. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565–574. doi: 10.1177/0272989X06295361
41. van Klaveren D, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. *Stat Med*. 2016;35:4136–4152. doi: 10.1002/sim.6997
42. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8:163. doi: 10.1186/s13643-019-1074-9