Circulation: Cardiovascular Quality and Outcomes

ORIGINAL ARTICLE

Generalizability of Cardiovascular Disease Clinical Prediction Models: 158 Independent External Validations of 104 Unique Models

Gaurav Gulati, MD, MS*; Jenica Upshaw, MD, MS*; Benjamin S. Wessler, MD, MS*, Riley J. Brazil, MD, MPH; Jason Nelson, MPH; David van Klaveren, PhD; Christine M. Lundquist, MPH; Jinny G. Park, MPH; Hannah McGinnes, MPH; Ewout W. Steyerberg, PhD; Ben Van Calster, PhD; David M. Kent, MD, MS

BACKGROUND: While clinical prediction models (CPMs) are used increasingly commonly to guide patient care, the performance and clinical utility of these CPMs in new patient cohorts is poorly understood.

METHODS: We performed 158 external validations of 104 unique CPMs across 3 domains of cardiovascular disease (primary prevention, acute coronary syndrome, and heart failure). Validations were performed in publicly available clinical trial cohorts and model performance was assessed using measures of discrimination, calibration, and net benefit. To explore potential reasons for poor model performance, CPM-clinical trial cohort pairs were stratified based on relatedness, a domain-specific set of characteristics to qualitatively grade the similarity of derivation and validation patient populations. We also examined the model-based C-statistic to assess whether changes in discrimination were because of differences in case-mix between the derivation and validation samples. The impact of model updating on model performance was also assessed.

RESULTS: Discrimination decreased significantly between model derivation (0.76 [interquartile range 0.73–0.78]) and validation (0.64 [interquartile range 0.60–0.67], $P<0.001$), but approximately half of this decrease was because of narrower case-mix in the validation samples. CPMs had better discrimination when tested in related compared with distantly related trial cohorts. Calibration slope was also significantly higher in related trial cohorts (0.77 [interquartile range, 0.59–0.90]) than distantly related cohorts (0.59 [interquartile range 0.43–0.73], $P=0.001$). When considering the full range of possible decision thresholds between half and twice the outcome incidence, 91% of models had a risk of harm (net benefit below default strategy) at some threshold; this risk could be reduced substantially via updating model intercept, calibration slope, or complete re-estimation.

CONCLUSIONS: There are significant decreases in model performance when applying cardiovascular disease CPMs to new patient populations, resulting in substantial risk of harm. Model updating can mitigate these risks. Care should be taken when using CPMs to guide clinical decision-making.

Key Words: cardiovascular ■ cardiovascular diseases ■ clinical decision ■ decision support techniques ■ models ■ risk ■ validation study

See Editorial by Shah et al

Clinical prediction models (CPMs) are multivariable statistical algorithms that produce patient-specific estimates of clinically important outcome risks based on ascertainable clinical characteristics. They are designed to improve prognostication and thus clinical decision making. CPMs are increasingly common and

Correspondence to: David M. Kent, MD, MS, PACE, ICRHPS, Tufts Medical Center, 800 Washington St, Box No. 63, Boston, MA 02111. Email dkent1@tuftsmedicalcenter.org

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 Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

*G. Gulati, J. Upshaw, and B.S. Wessler contributed equally.

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCOUTCOMES.121.008487.

WHAT IS KNOWN

- Clinical prediction models (CPMs) are used routinely to guide clinical decision making, yet the majority of published CPMs have never been externally validated.
- How well these models perform on new populations, as well as how likely these models are to improve clinical decision making, is unknown.

WHAT THE STUDY ADDS

- In this collection of cardiovascular CPMs, discrimination and calibration decrease substantially when models are validated on external databases, with the largest decrease when derivation and validation cohorts are the most dissimilar.
- The majority of CPMs had the potential to motivate harmful clinical decisions, particularly when decision thresholds were far from the population average risk.
- Model updating can reduce the risk of harm and should be performed before widespread clinical use of a CPM.

important tools for patient-centered outcomes research and clinical care.

Recent reviews have demonstrated the abundance of CPMs in the literature but have also pointed at shortcomings.[1] Our own database, the Tufts Predictive Analytics and Comparative Effectiveness Center CPM Registry,[2] currently includes 1382 CPMs just for patients with cardiovascular disease (CVD), including 344 CPMs for patients with coronary artery disease, 195 for population-based samples (ie, predicting incident CVD), and 135 for patients with heart failure (HF).

How well these CPMs are likely to perform when tested on a new patient population is poorly understood. Large-scale evaluations of the model development methods have revealed that the vast majority of models do not follow best practice and are classified as having a high risk of bias.[3,4] The concern that prediction models may fail when disseminated into clinical practice has grown increasingly urgent, now that models are being broadly distributed by vendors and influencing care at a large scale. Examples of model failure of clinically influential and widely disseminated models have recently come to light.[5] Our prior literature review found that approximately 60% of published CPMs have never been externally validated. Most of those that have been externally validated have been evaluated only once.[6,7] Yet, our prior analysis also called into question the value of these single validations, since discriminatory performance typically varies tremendously when a single model is evaluated on multiple databases.[6]

However, there are inherent limitations in literature reviews in understanding how well models perform when evaluated on external data. For example, when discrimination in a new database is poor, it can be due to model invalidity, on the one hand, or because the case-mix in the external database is substantially restricted compared with the derivation database. The published literature does not distinguish between these possibilities. Further, when CPMs are validated, typically no clinically interpretable measure of calibration is reported, despite

Nonstandard Abbreviations and Acronyms

ACCORD	Action to Control Cardiovascular Risk in Diabetes trial
ACS	acute coronary syndrome
ALLHT-HTN	Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack
ALLHAT-LLT	Lipid-Lowering Therapy trial
AMIS	Aspirin-Myocardial Infarction Study
BEST	Beta Blocker Evaluation of Survival trial
BioLINCC	Biological Specimen and Data Repository Information Coordinating Center
CPM	clinical prediction model
CVD	cardiovascular disease
DIG	Digitalis Investigation Group trial
EAVG	Harrell's E-statistic measure of the mean
E90	Harrell's E-statistic measure of the 90th percentile
ENRICHD	Enhanced Recovery in Coronary Heart Disease trial
EVEREST	Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study with Tolvaptan trial
HEAAL	Heart Failure evaluation of Angiotensin II Antagonist Losartan trial
HF	heart failure
HF-ACTION	Heart Failure: A Controlled Trial Investigating Outcomes of Exercise Training trial
IQR	interquartile range
MAGIC	Magnesium in Coronaries trial
MB-c	model-based C-statistic
SCD-HeFT	Sudden Cardiac Death in Heart Failure trial
SOLVD	Studies of Left Ventricular Dysfunction trial
TIMI-II	Thrombolysis in Myocardial Infarction: phase II trial
TIMI-III	Thrombolysis in Myocardial Ischemia trial
TOPCAT	Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist trial
WHI	Women's Health Initiative trial

the fact that it is known that poor calibration can lead to harmful decision making. Finally, there is no widely accepted criteria by which one can claim that a model has been validated, since models are assessed for statistical accuracy, but scant attention is paid to whether models can improve the quality of decisions.

Given these limitations, it is difficult to understand the quality of CPMs reported in the literature, and how they might influence decision-making if widely disseminated. Thus, we sought to perform a large scale and systematic external validation on published CPMs, using both conventional and novel measures of model performance to address some of the above limitations. In particular, we sought to examine both discrimination and calibration, to examine the proportion of decreased performance that might be due to model invalidity versus case-mix, and to examine the influence that predictions might have on decisions through the use of decision curve analysis. We were especially interested in evaluating when CPMs might lead to harmful decision making. We also evaluated the effect of simple updating procedures.

METHODS

Source of Models

The Tufts Predictive Analytics and Comparative Effectiveness Center CPM Registry is a registry of CPMs published between January 1990 and December 2015 that predict outcomes in patients at risk for or with known cardiovascular disease. Detailed methods for development of the registry have been reported previously.[2,8] Briefly, for inclusion in the registry, articles must (1) develop a CPM as a primary aim, (2) contain at least 2 outcome predictors, and (3) present enough information to estimate the outcome probability for an individual patient. For this analysis, we selected from the registry all CPMs predicting outcomes for 3 index conditions: (1) acute coronary syndrome, (2) HF (both preserved and reduced ejection fraction), and (3) healthy patients at risk for CVD (primary prevention or population sample). Some data and materials for this analysis have been made publicly available and can be accessed at www.pacecpmregistry.org, and other data are available from the corresponding author upon reasonable request. The Tufts Health Sciences Institutional Review Board (IRB) approved this study.

Source of Validation Cohorts

Deidentified patient-level data from clinical trials were obtained from the National Heart, Lung, and Blood Institute via application to the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC). For the Acute coronary syndrome index condition, we used the AMIS[9] (Aspirin-Myocardial Infarction Study), TIMI-II[10] (Thrombolysis in Myocardial Infarction: phase II), TIMI-III[11] (Thrombolysis in Myocardial Ischemia), MAGIC[12] (Magnesium in Coronaries), and ENRICHD[13] (Enhanced Recovery in Coronary Heart Disease) trials. For the HF index condition, we used the TOPCAT[14] (Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist), HEAAL[15] (Heart Failure evaluation of Angiotensin II Antagonist Losartan), HF-ACTION[16] (Heart Failure: A Controlled Trial Investigating Outcomes of Exercise Training), EVEREST[17] (Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study with Tolvaptan), SCD-HeFT[18] (Sudden Cardiac Death in Heart Failure), BEST[19] (Beta Blocker Evaluation of Survival), DIG[20] (Digitalis Investigation Group), and SOLVD[21] (Studies of Left Ventricular Dysfunction) trials. For the primary prevention index condition, we used the ACCORD[22] (Action to Control Cardiovascular Risk in Diabetes), ALLHAT-HTN[23] (Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack), ALLHAT-LLT[24] (Lipid-Lowering Therapy), and WHI[25] (Women's Health Initiative). Details of the trials have been reported previously and are summarized in Tables S1 through S3.

CPM-Dataset Matching Process

To identify which clinical trial dataset could be used to validate which CPMs, we employed a hierarchical matching procedure. First, each CPM was compared with each dataset by non-clinical research staff to identify pairs that had grossly similar inclusion criteria and outcomes, which were then reviewed for appropriateness by clinical experts. Potential pairs passing these screening steps were carefully reviewed at a granular level, and only pairs where sufficient patient-level data existed in the trial dataset such that the CPM could be used to generate a predicted outcome probability for each patient were included in the analysis. Observed outcomes in the patient-level data were defined using the CPM outcome definition and prediction time horizon. A list of CPMs included in this analysis is shown in Table S4.

Measuring CPM Performance

The performance of CPMs in external cohorts was evaluated with measures of discrimination, calibration, and net benefit when applied to external validation cohorts. For all model validations, observed outcome events that occurred after the prediction time horizon were censored. For time-to-event models, the Kaplan-Meier estimator was used for right-censored follow up times. For binary outcome models, unobserved outcomes (ie, due to loss-to-follow up before the prediction time horizon) were considered missing and excluded from analyses. For each CPM-database pair, the linear predictor was calculated for each patient in the dataset using the intercept and coefficients from the published CPM. Model discrimination was assessed using the C-statistic. The percent change in a CPM's discrimination from the derivation cohort to the validation cohort was calculated as [(Validation C-statistic−0.5)−(Derivation C-statistic−0.5)]/(Derivation C-statistic−0.5)×100.[26] If the C-statistic at model derivation was not reported, the model was excluded from the assessment of decrement in C-statistic relative to derivation.

Since changes in case-mix between derivation and validation population will affect the C-statistic in the validation population even without any change in measured effects, change in discrimination was also compared relative to the model-based C-statistic (MB-c). The MB-c is the C-statistic that would be obtained in the validation database under the assumption that the CPM is perfectly valid in the validation database and depends only on the distribution of the linear predictors in the

validation database.[27] For example, even a model with no invalidity would have both a C-statistic of 0.5 and an MB-c of 0.5 in a validation cohort if all patients in that cohort were identical with respect to their covariates. Thus, any difference between the derivation C-statistic and the validation MB-c reflects differences in case-mix, while the difference between the validation MB-c and the C-statistic in the validation cohort reflects model invalidity. Because calculation of the MB-c depends entirely on the validation cohort, MB-c could be calculated for all pairs.

Model calibration was assessed by converting the linear predictor to a predicted probability (including a specified time point if Cox proportional hazards modeling was used). From the predicted probabilities, calibration slope and Harrell's E_{AVG} and E_{90} statistics were calculated. Harrell's E_{AVG} and E_{90} statistics measure the mean and 90th percentile, respectively, of the absolute difference between the predicted and observed event probabilities, where observed probabilities are estimated nonparametrically using locally weighted scatterplot smoothing curves. For this analysis, E_{AVG} and E_{90} values were standardized by dividing by the outcome rate in the validation cohort to improve comparability between CPM-validation pairs. If point estimates of outcome incidences at similar time points in the CPM derivation cohort and paired validation cohort were not able to be calculated with published information, that pair was excluded from analysis of calibration.

Finally, decision curve analysis[28] was used to estimate the net benefit of each model in each paired validation dataset. Decision curve analysis presents a comprehensive assessment of the potential population-level clinical consequences of using CPMs to inform treatment decisions by examining misclassification of patients across a relevant range of decision thresholds, while weighting the relative utility of false-positive and false-negative predictions as implicitly determined by the threshold. As each model could be used to guide many different decisions, each with a unique threshold probability, we assessed whether each model resulted in a positive net benefit (above the best default strategy of treat all or treat none) or negative net benefit (below the best default strategy) first at 3 threshold probabilities spanning a broad range of plausible thresholds: half the outcome incidence, outcome incidence, and twice the outcome incidence, and then over the entire range of threshold probabilities from half the outcome incidence to twice the outcome incidence. A range centered around the outcome incidence was also chosen because model net benefit is most likely to differ from that of the default strategy at thresholds close to the outcome incidence. Models with negative net benefit were considered harmful, while models with positive net benefit were considered not harmful. We used standardized net benefit[29] to make results comparable across validations by controlling for variation in the incidence of the outcome.

Stratification by Relatedness

To explore sources of variability in model performance in external validation, we categorized each CPM-dataset pair based on the relatedness of the underlying study populations. Study populations were reviewed in detail by clinical experts on the basis of key clinical characteristics, such as inclusion/exclusion criteria, patient demographics, outcome, enrollment period, and follow-up duration. These characteristics were index condition-specific and are detailed in Tables S5 through S7). Pairs were categorized as related when there were no clinically relevant differences in inclusion criteria, exclusion criteria, recruitment setting, and baseline clinical characteristics. Any matches with clinically relevant differences in any criterion were categorized as distantly related. Clinical experts scoring relatedness were blinded to the derivation C-statistic of the CPM and outcome rates in the derivation and validation cohorts.

Model Updating

We assessed the impact of model updating on discrimination, calibration, and net benefit. Models were updated using data from each paired validation dataset in a sequential fashion: (1) by updating the model intercept using the observed outcome rate in the validation cohort (recalibration-in-the-large); (2) by updating the intercept and rescaling all the model coefficients by the calibration slope; and (3) by re-estimating all regression coefficients using data from the validation database (but maintaining the predictors from the original model).[30]

Statistical Analysis

Differences in various model performance measures were assessed using the Wilcoxon rank-sum test. All analyses were performed in R version 3.5.3 (R foundation for statistical computing, Vienna, Austria).

RESULTS

CPM-Validation Cohort Matching

From a set of 674 potential CPMs across all 3 index conditions, 548 (81%) were screened as potential matches based on title and abstract review and underwent granular review to assess for sufficient patient level variable and outcome data within the publicly available clinical trial databases. We matched 104 (15%) CPMs to at least one database, yielding 158 CPMs-database pairs across the 3 index conditions (Figure 1). The matching success frequency varied by index condition, from 6% (23 of 344) for acute coronary syndrome to 32% (59 of 195) for primary prevention. Details about the CPMs used in this analysis are summarized in Table S4.

CPM Discrimination in Independent External Validations

Of the 158 total CPM-database pairs, there were 111 pairs in which the CPM reported a C-statistic at model derivation. Among these, the median C-statistic in the derivation cohorts was 0.76 (interquartile range [IQR], 0.73–0.78) and the median C-statistic at model validation was 0.64 (IQR, 0.60–0.67, $P<0.001$; Table 1). Discriminative ability decreased by a median of 49% (IQR, 29%–64%). Approximately half the loss in discriminatory power was attributable to a decrease in case-mix heterogeneity, while half was attributable to model invalidity. When stratified by relatedness, 57 (36%) pairs were graded as related and 101 (64%) were graded

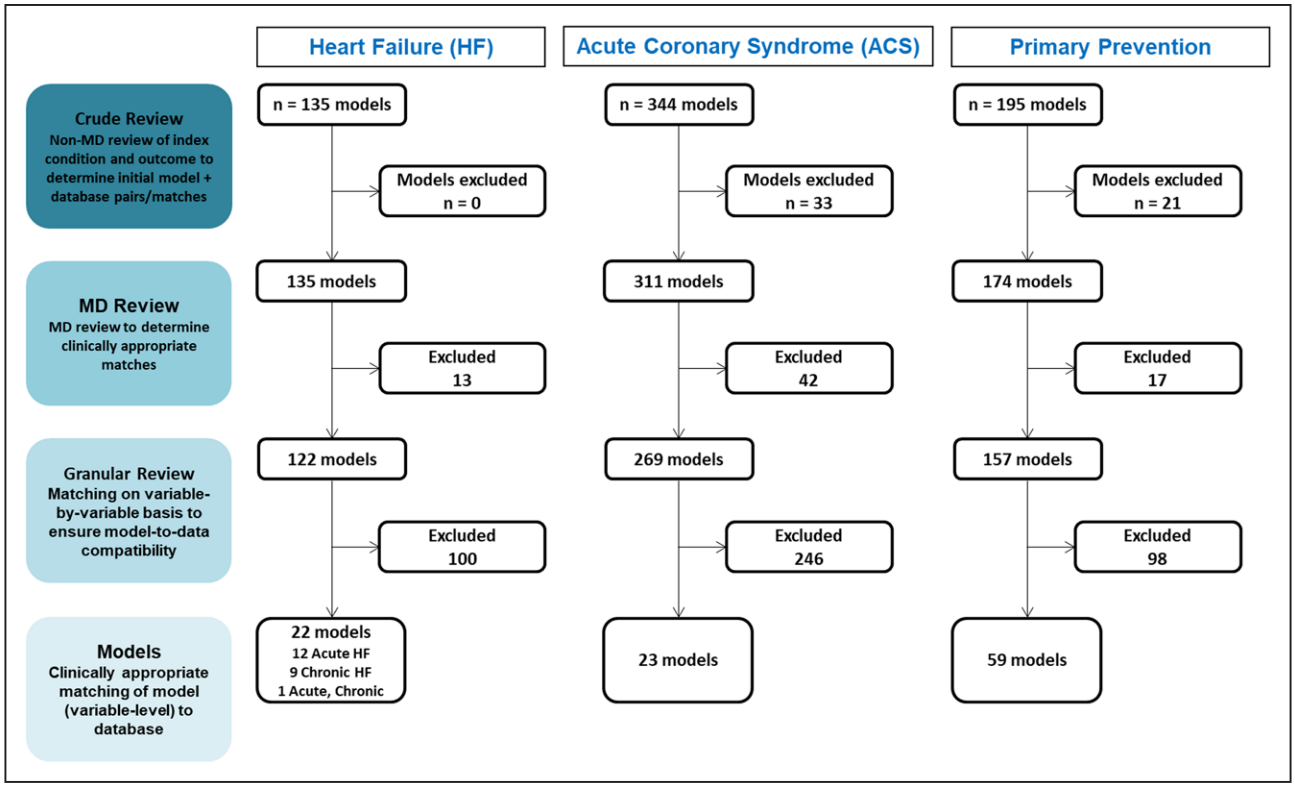

Figure 1. Flowchart of clinical prediction model-database matching process.
ACS indicates acute coronary syndrome; and HF, heart failure.

as distantly related. CPM-trial pairs that were related had significantly higher MB-c and validation C statistics than pairs that were distantly related (Table 1). Median percentage decrement in discrimination among related pairs was 30% (IQR, 16%–45%), of which approximately two-thirds was due to a decrease in case-mix heterogeneity and one-third due to model invalidity. In contrast, CPM-trial pairs that were distantly related had a median percentage decrement in discrimination of 55% (IQR, 40%–68%, $P<0.001$ versus related pairs), approximately half of which was due to case-mix heterogeneity and half due to model invalidity.

CPM Calibration in Independent External Validations

Of the 158 total CPM-database pairs, there were 132 pairs in which the validation was assessed for calibration. The median calibration slope in the external validations was 0.64 (IQR, 0.48–0.84). Median calibration slope among related pairs was 0.77 (IQR, 0.59–0.90), significantly higher than the median calibration slope among distantly related pairs (0.59, IQR, 0.43–0.73, $P=0.001$). Median E_{AVG} and median E_{90} standardized to the outcome incidence among all pairs was 0.53 (IQR,

Table 1. Discriminative Ability of Clinical Prediction Models in Derivation and Validation Cohorts Stratified by Cohort Relatedness

		Cohort relatedness		
	All validations	Related	Distantly related	Related vs distantly related, *P* Value
	N=76 models	N=30 models	N=53 models	
	N=158 validations	N=57 validations	N=101 validations	
Derivation c-statistic*	0.76 (0.73–0.78)	0.76 (0.71–0.78)	0.77 (0.74–0.79)	NA
Validation MB-c	0.68 (0.61–0.71)†	0.69 (0.65–0.75)†	0.68 (0.66–0.70)†	0.2
Validation c-statistic	0.64 (0.60–0.67)†‡	0.67 (0.64–0.71)†‡	0.62 (0.59–0.66)†‡	<0.01

Derivation and validation cohorts were classified as related or distantly related using an index condition-specific rubric (see Methods and Tables S5 through S7 for more details). Values are presented as median (IQR). MB-c indicates model-based c-statistic; and NA, not applicable.

*Twenty-eight of the 104 unique models did not report c-statistic. Sum of related and distantly related models is > 76 as some models were validated in related and distantly related pairs.

†$P<0.01$ vs derivation.

‡$P<0.01$ vs MB-c.

0.38–0.72) and 0.95 (IQR, 0.62–1.25), respectively and did not differ significantly between related and distantly related pairs (Table 2).

Net Benefit

When we assessed net benefit at 3 thresholds (half the outcome incidence, outcome incidence, and twice the outcome incidence), the vast majority of models (110 of 132, 83%) were harmful when used in their paired database at one or more threshold. Even more models (120 of 132, 91%) were harmful at some point within the range of thresholds between half and twice the outcome incidence. However, when we considered each threshold individually, models were much less likely to be harmful at a threshold equal to the outcome incidence than at either of the more extreme thresholds. In particular, only 10 of 132 (8%) of models were harmful at a threshold equal to the outcome incidence, while 72 (55%) were harmful at half the outcome incidence and 58 (44%) were harmful at twice the outcome incidence (Table 3). When exploring the likelihood of a model being harmful at any point within the range of thresholds explored, we found that 97% (103/106) of models with a validation C-statistic below 0.7 were potentially harmful at some threshold, while only 65% (17/26) of those with a validation C-statistic above 0.7 were potentially harmful. Similarly, we found that 96% (106/111) of models with a standardized E_{AVG} above 0.3 were potentially harmful, while only 67% (14/21) of models with a standardized E_{AVG} below 0.3 were potentially harmful (Figure 2).

Effects of Updating

E_{AVG} improved by a median of 56% (IQR, 23%–79%) across all the CPM-trial pairs with updating of the intercept and by a median of 93% (IQR, 80%–99%) after updating the intercept and slope (Table 4). Similar results were seen for E_{90}. No further improvement in calibration error was seen with re-estimation. Plots of harmful and nonharmful models across the range of validation C-statistic and standardized E_{AVG} showed the points moving progressively downward (and rightward after re-estimation), reflecting improved calibration (when the intercept and slope are adjusted) and discrimination (when the coefficients are re-estimated, Figure 2). Similar results were seen when net benefit was assessed only at the 3 thresholds (Figure S1). Significant improvement in net benefit was seen with sequential model updating (Table 3). Updating the intercept alone reduced the likelihood of model harm at any threshold in the full range of thresholds considered from 91% to 73% (97/132). Updating the intercept and slope reduced this likelihood further, to 53% (70/132), and complete re-estimation reduced the likelihood to 48% (63/132).

DISCUSSION

The major finding of this analysis is that off-the-shelf CPMs often perform poorly in new populations, and this very frequently results in potential for net harm. Indeed, only 12/132 (9%) of the unique evaluations we performed were either beneficial or neutral in the full range of thresholds examined, and only 22/132 (17%) were either beneficial or neutral at each of the 3 thresholds. In contrast to what is often assumed, use of an explicit data-driven CPM is often not likely to be better than nothing. Model re-updating substantially reduced the risk of harm, although half the evaluations showed potential harm at least at some threshold within the range considered even after re-estimation. These findings emphasize the need for close oversight, governance and regulation of CPMs as they are more broadly deployed in clinical practice.[31,32]

The risk of harm of using CPMs in clinical practice is most salient when decision thresholds depart substantially from the average risk in the patient population of interest. For example, risk of harm would be substantial when trying to deselect a very low risk population for a test or treatment that is clearly beneficial on average, or when trying to select a very high risk population for a

Table 2. Calibration Performance of Clinical Prediction Models on External Validation Stratified by Cohort Relatedness

		Cohort relatedness		
	All validations	Related	Distantly related	
	N=158	N=57	N=101	*P* Value
Calibration slope*	0.64 (0.48–0.84)	0.77 (0.59–0.90)	0.59 (0.43–0.73)	<0.001
Standardized E_{AVG}	0.53 (0.38–0.72)	0.53 (0.37–0.67)	0.52 (0.40–0.81)	0.5
Standardized E_{90}	0.95 (0.62–1.25)	0.82 (0.67–1.19)	1.04 (0.56–1.28)	0.5

Derivation and validation cohorts were classified as related or distantly related using an index condition-specific rubric (see Methods and Tables S5 through S7 for more details). Calibration error was measured using Harrell's E_{AVG} and E_{90} statistics, standardized to the outcome incidence. For example, if the outcome incidence in a validation population was 5% and E_{AVG} was 0.05, standardized E_{AVG}=1.0. Values are presented as median (IQR). E_{AVG} indicates Harrell's E-statistic measure of the mean; and E_{90}, Harrell's E-statistic measure of the 90th percentile.

*Twenty-six of 158 validation pairs were not assessed for calibration. Sample size for calibration is 132 (57 related, 75 distantly related).

Table 3. Net Benefit Analysis of Models at 3 Representative Decision Thresholds Before and After Sequential Model Updating

	Net benefit relative to default strategy (N=132)	Half outcome incidence	Outcome incidence	Twice outcome incidence	Any point within range
Original model	Positive	38 (29%)	113 (86%)	35 (26%)	NA
	Neutral	22 (17%)	9 (7%)	39 (30%)	NA
	Negative	72 (55%)	10 (8%)	58 (44%)	120 (91%)
Sequential model updating					
Updated intercept	Positive	52 (39%)	132 (100%)	65 (49%)	NA
	Neutral	16 (12%)	0 (0%)	14 (11%)	NA
	Negative	64 (49%)	0 (0%)	53 (40%)	97 (73%)
Updated intercept and slope	Positive	69 (52%)	132 (100%)	74 (56%)	NA
	Neutral	38 (29%)	0 (0%)	32 (24%)	NA
	Negative	25 (19%)	0 (0%)	26 (20%)	70 (53%)
Re-estimated	Positive	90 (68%)	132 (100%)	100 (76%)	NA
	Neutral	19 (14%)	0 (0%)	3 (2%)	NA
	Negative	23 (17%)	0 (0%)	29 (22%)	63 (48%)

The 3 threshold probabilities at which net benefit was quantified were: outcome incidence in the validation cohort, half the outcome incidence, and twice the outcome incidence. Net benefit was assessed as positive if it was above the default strategy at that threshold, negative if it was below the default strategy at that threshold, and neutral if it was equivalent to the default strategy at that threshold. We also assessed how many models had net benefit below the default strategy at any point within the range of half the outcome incidence to twice the outcome incidence. NA indicates not applicable.

test or treatment that is clearly not indicated for those at average risk. However, when the point of indifference lies closer to the average risk (ie, decision is unclear for a typical patient), CPMs seem to be more likely to yield net clinical benefit and to be tolerant of some miscalibration. These findings were consistent across the 3 index conditions we tested.

That the decision threshold emerged as a very important determinant of the utility of applying the CPMs in this sample emphasizes the importance of selecting the right decision context for CPM application—an often neglected issue. Based on our results, CPMs yielding typical (ie, nonexcellent) performance should generally be reserved for applications where the decision threshold is near the population average risk, particularly when model updating is not feasible (as it often is not). Intuitively, the value of risk information is the highest when the decision threshold is near the average risk, since even relatively small shifts from the average risk due to using a CPM can reclassify patients into more appropriate decisions.

Our prior literature review[6] was unable to examine calibration because it is frequently unreported and, when reported, the metrics used vary from study to study and are largely uninformative with regard to the magnitude of miscalibration (eg, Hosmer Lemeshow, which yields only a *P*, which tends to be large in small samples and small in large samples). The validations we performed ourselves revealed that CPM-predicted outcome rates frequently deviate from observed outcome rates even when discrimination was good. The typical standardized E_{AVG} was 0.5 (IQR, 0.4–0.7), which means that the absolute error is half the average risk. In exploratory analysis, we found that when the standardized E_{AVG} was > 0.3 (average prediction was off by at least 30%), models were generally found to yield harmful decisions at least at one threshold within the range examined (half the outcome rate to twice the outcome rate). The importance of good calibration in guarding against harmful decision making has recently been emphasized.[33–35] Similarly, it was very unusual to find models that were consistently nonharmful at all examined thresholds when the validation C-statistic dropped below 0.7.

We found that the risk of harm can be substantially mitigated often simply by adjusting the intercept alone. Indeed, updating the intercept alone resulted in 100% of the models yielding positive net benefit when the decision threshold was set at the average risk. Yet for the more extreme thresholds, there was still substantial risk of harm; 60% (79/132) of CPMs tested yielded harmful predictions at one or more of the extreme thresholds, even after intercept updating. When both the slope and the intercept were updated, 62/132 (47%) of models were consistently beneficial or nonharmful across all examined thresholds. This underscores the importance of calibration in determining the risk of harm—and also the importance of clear and consistent reporting of calibration, which is largely absent from the literature. Unfortunately, in many clinical settings, recalibration may not be possible.

Among other notable findings, we discovered that the vast majority of CPMs were impossible to validate on publicly available patient-level trial databases. The most common reason was a mismatch between the variables in the models and those collected in the publicly available databases. Among the CPMs that we were able

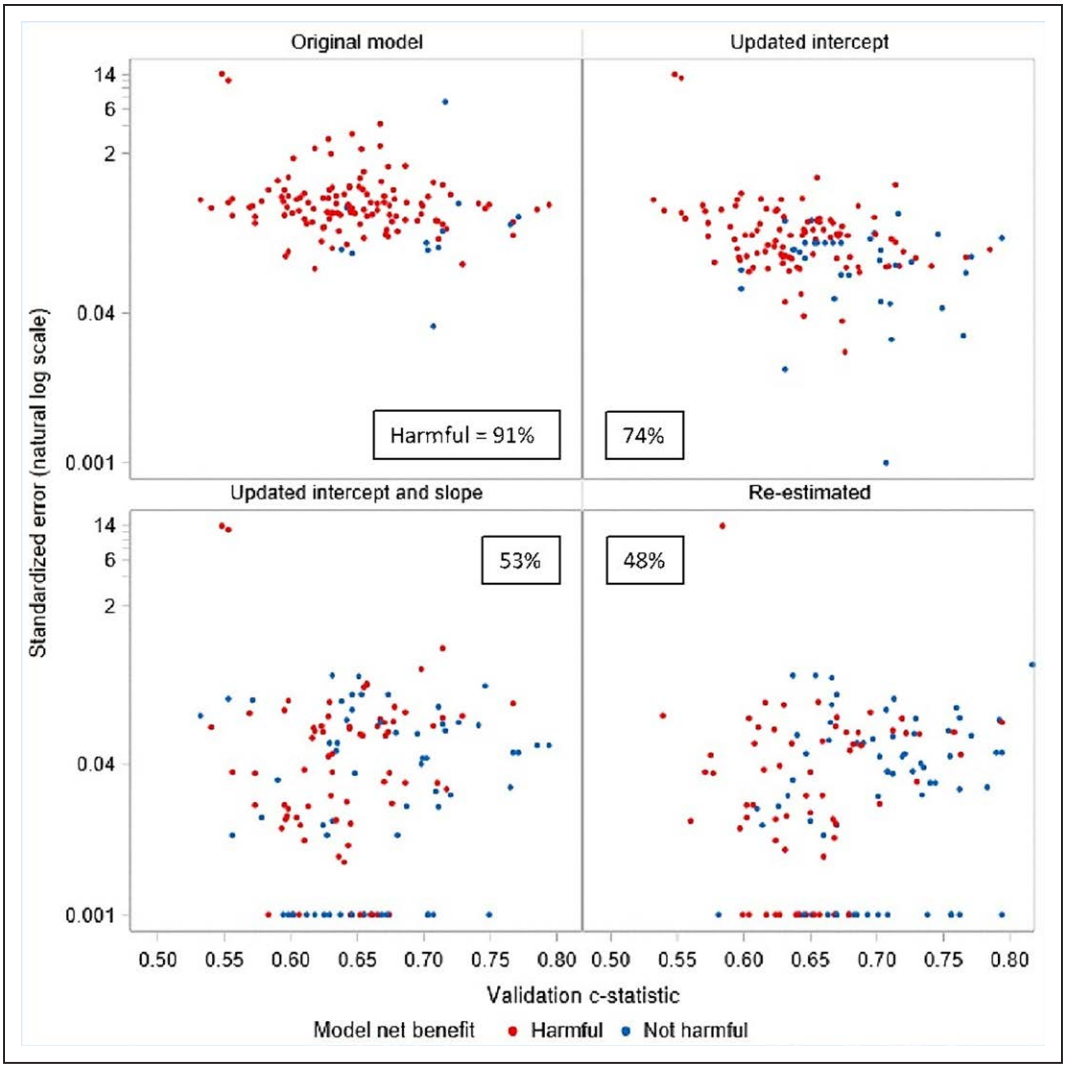

Figure 2. Model harm status before and after sequential model updating based on validation c-statistic and standardized calibration error.

A model was considered harmful if net benefit was below default strategy at any point across a range of thresholds from half the outcome incidence to twice the outcome incidence.

to validate, we found that discrimination and calibration deteriorated substantially when compared with the derivation cohorts. Interestingly, much of the decrease in discrimination was due to a narrower case-mix in the validation cohorts as well as model invalidity—although this varied somewhat across the different index conditions. For example, in our examination of acute coronary syndrome models, the median derivation C-statistic was 0.76. This was found to decrease on validation to 0.70. Almost all of this decrease, however, was due to changes in case-mix, not model performance (median MB-c=0.71). In heart failure and population models, the decrement in discriminatory performance appeared more evenly due to case-mix and model invalidity.

Our analysis also showed the potential usefulness of the MB-c. This is the first large-scale evaluation to apply this rarely utilized tool. By permitting an estimation of the C-statistic based on the variation of predictions only (ie, independent of the actual outcomes), the MB-c permits comparison of the actual c-statistic to a more appropriate baseline determined by the case-mix in the validation sample, rather than to that in the derivation population. This was particularly germane for our study since we used publicly available clinical trials to evaluate the CPMs. These databases are generally assumed to have a narrower case-mix than registry or real world populations derived from electronic health records—an assumption supported by our results.

Our analysis also showed a larger decrement in discrimination when externally validating a CPM on a distantly related cohort than if the cohort were more closely related, a result that confirms findings from our literature review.[6] Furthermore, the proportion of decrement in model discrimination attributable to model invalidity was somewhat higher when the cohorts were distantly related. Relatedness often hinged on subtle

Table 4. Impact of Model Updating on Calibration Error and Risk of Harm

	E_{AVG}			E_{90}			Potentially harmful models
							All validations
	All validations	Related	Distantly related	All validations	Related	Distantly related	N=132
Original model	NA	NA	NA	NA	NA	NA	120 (91%)
Updated intercept	−56 (−79 to −23)	−61 (−81 to −37)	−47 (−78 to −13)	−65 (−83 to −27)	−67 (−84 to −41)	−64 (−82 to −7)	97 (73%)
Updated intercept and slope	−93 (−99 to −80)	−92 (−99 to −79)	−93 (−99 to −82)	−93 (−99 to −81)	−90 (−99 to −75)	−94 (−99 to −82)	70 (53%)
Complete re−estimation	−94 (−99 to −86)	−91 (−99 to −85)	−95 (−99 to −86)	−94 (−99 to −87)	−93 (−99 to −83)	−95 (−99 to −88)	63 (48%)

For E_{AVG} and E_{90}, percent change in statistic relative to original model is shown. A model was considered potentially harmful if net benefit was below default strategy at any point across a range of thresholds from half the outcome incidence to twice the outcome incidence. Values reported in table are median (IQR); 26 of 158 validation pairs were not assessed for calibration. Sample size for calibration is 132 (57 related, 75 distantly related). E_{AVG} indicates Harrell's E-statistic measure of the mean; E90, Harrell's E-statistic measure of the 90th percentile; and NA, not applicable.

but clinically relevant differences between cohorts, such as years of enrollment or the distribution of baseline comorbidities, which required careful review from expert clinicians to identify.

Limitations

Our analysis has several limitations. Models published after 2015 were not included, so our results may not generalize to more recent models in these clinical domains. The sample of databases used was a convenience sample and this sample determined the CPMs selected, since many models were not compatible with the available databases usually because of incongruence between variables required for prediction and those collected in the trial. The validation databases generally represent older therapeutic eras, as this work reflects databases that are currently available through BioLINCC. Using derivation and validation databases from different errors, however, might be thought to simulate the kind of calibration problems that models are likely to confront from data shifts over time and in different settings.[36] Since the database are randomized trials, we anticipate poorer discrimination in these samples just on the basis of the restricted case-mix.

Many potential CPM-validation database matches were not possible because of missing or differently defined variables in the validation databases. Given the small number of CPM-validation database matches, we were seldom able to match a CPM to > 1 validation database. A given CPM may perform differently when validated against different cohorts, and more research is required to understand the sources of this variation before validation performance can be used to grade the quality of a model. Our relatedness categorization was one such attempt, but it requires content area expertise, is inherently subjective, and is difficult to generalize to CPMs for other clinical domains.

Further, our net benefit analyses used a range of decision thresholds that may be considered clinically arbitrary in that they were not informed by the relative cost of over-treatment versus under-treatment in the specific clinical context. However, we would anticipate that most clinically relevant thresholds would fall within this range, since risk prediction is much less likely to be useful for decision thresholds that are even more extreme. Nevertheless, considering any negative net benefit within this range as indicative of a potentially harmful model may provide an unduly pessimistic view, since many models that are labeled harmful may be beneficial at most thresholds, including the clinically most relevant ones.

CONCLUSIONS

Discrimination and calibration often decrease substantially when CPMs for cardiovascular disease are tested in external populations, especially when validation cohorts are only distantly related to model derivation cohorts. This leads to substantial risk of net harm, particularly when decision thresholds are not near the population average risk. Model updating can reduce this risk substantially and will likely be needed to realize the full potential of risk-based decision making. Our findings underscore the need for more thorough model evaluation, including the use of novel measures assessing utility, and better model oversight and stewardship.

ARTICLE INFORMATION

Received August 5, 2021; accepted February 23, 2022.

Affiliations

Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center, Boston, MA (G.G., J.U., B.S.W., R.J.B., J.N., D.v.K., C.M.L., J.G.P., H.M., D.M.K.). Division of Cardiology, Tufts Medical Center, Boston, MA (G.G., J.U., B.S.W.). Department of Biomedical Data Sciences, Leiden University Medical Centre, Netherlands (D.v.K., E.W.S., B.V.C.). KU Leuven, Department of Development and Regeneration, Belgium (B.V.C.). EPI-Center, KU Leuven, Belgium (B.V.C.).

Sources of Funding

Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1606-35555).

Disclosures

None.

Supplemental Material

Figure S1
Tables S1–S7
References 9–25,37–107

REFERENCES

1. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, Altman DG, Moons KG. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med*. 2012;9:1–12. doi: 10.1371/journal.pmed.1001221
2. Wessler BS, Yh LL, Kramer W, Cangelosi M, Raman G, Lutz JS, Kent DM. Clinical prediction models for cardiovascular disease: Tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circ Cardiovasc Qual Outcomes*. 2015;8: 8–75. doi: 10.1161/CIRCOUTCOMES.115.001693
3. Venema E, Wessler BS, Paulus JK, Salah R, Raman G, Leung LY, Koethe BC, Nelson J, Park JG, van Klaveren D, et al. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. *J Clin Epidemiol*. 2021;138:32–39. doi: 10.1016/j.jclinepi.2021.06.017
4. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Bonten MMJ, Dahly DL, Damen JAA, Debray TPA, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ*. 2020;369:m1328. doi: 10.1136/bmj.m1328
5. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, Pestrue J, Phillips M, Konye J, Penoza C, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181:1065–1070. doi: 10.1001/jamainternmed.2021.2626
6. Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, Van Calster B, van Klaveren D, Venema E, Steyerberg E, et al. External validations of cardiovascular clinical prediction models: a large-scale review of the literature. *Circ Cardiovasc Qual Outcomes*. 2021;14:e007858. doi: 10.1161/CIRCOUTCOMES.121.007858
7. 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
8. Wessler BS, Paulus JK, 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.
9. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA*. 1980;243:661–669.
10. Cannon CP, Weintraub WS, Demopoulos LA, Vicari R, Frey MJ, Lakkis N, Neumann FJ, Robertson DH, DeLucca PT, DiBattiste PM, et al; TACTICS (Treat Angina with Aggrastat and Determine Cost of Therapy with an Invasive or Conservative Strategy)–Thrombolysis in Myocardial Infarction 18 Investigators. Comparison of early invasive and conservative strategies in patients with unstable coronary syndromes treated with the glycoprotein IIb/IIIa inhibitor tirofiban. *N Engl J Med*. 2001;344:1879–1887. doi: 10.1056/NEJM200106213442501
11. Effects of tissue plasminogen activator and a comparison of early invasive and conservative strategies in unstable angina and non-Q-wave myocardial infarction. Results of the TIMI IIIB Trial. Thrombolysis in Myocardial Ischemia. *Circulation*. 1994;89:1545–1556. doi: 10.1161/01.cir.89.4.1545
12. Magnesium in Coronaries (MAGIC) Trial Investigators. Early administration of intravenous magnesium to high-risk patients with acute myocardial infarction in the Magnesium in Coronaries (MAGIC) Trial: a randomised controlled trial. *Lancet*. 2002;360:1189–1196. doi: 10.1016/s0140-6736(02)11278-5
13. Berkman LF, Blumenthal J, Burg M, Carney RM, Catellier D, Cowan MJ, Czajkowski SM, DeBusk R, Hosking J, Jaffe A, et al; Enhancing Recovery in Coronary Heart Disease Patients Investigators (ENRICHD). Effects of treating depression and low perceived social support on clinical events after myocardial infarction: the Enhancing Recovery in Coronary Heart Disease Patients (ENRICHD) Randomized Trial. *JAMA*. 2003;289:3106–3116. doi: 10.1001/jama.289.23.3106
14. Pitt B, Pfeffer MA, Assmann SF, Boineau R, Anand IS, Claggett B, Clausell N, Desai AS, Diaz R, Fleg JL, et al; TOPCAT Investigators. Spironolactone for heart failure with preserved ejection fraction. *N Engl J Med*. 2014;370:1383–1392. doi: 10.1056/NEJMoa1313731
15. Konstam MA, Neaton JD, Dickstein K, Drexler H, Komajda M, Martinez FA, Riegger GA, Malbecq W, Smith RD, Guptha S, et al; HEAAL Investigators. Effects of high-dose versus low-dose losartan on clinical outcomes in patients with heart failure (HEAAL study): a randomised, double-blind trial. *Lancet*. 2009;374:1840–1848. doi: 10.1016/S0140-6736(09)61913-9
16. O'Connor CM, Whellan DJ, Lee KL, Keteyian SJ, Cooper LS, Ellis SJ, Leifer ES, Kraus WE, Kitzman DW, Blumenthal JA, et al; HF-ACTION Investigators. Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *JAMA*. 2009;301:1439–1450. doi: 10.1001/jama.2009.454
17. Konstam MA, Gheorghiade M, Burnett JC Jr, Grinfeld L, Maggioni AP, Swedberg K, Udelson JE, Zannad F, Cook T, Ouyang J, et al; Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study With Tolvaptan (EVEREST) Investigators. Effects of oral tolvaptan in patients hospitalized for worsening heart failure: the EVEREST Outcome Trial. *JAMA*. 2007;297:1319–1331. doi: 10.1001/jama.297.12.1319
18. Bardy GH, Lee KL, Mark DB, Poole JE, Packer DL, Boineau R, Domanski M, Troutman C, Anderson J, Johnson G, et al; Sudden Cardiac Death in Heart Failure Trial (SCD-HeFT) Investigators. Amiodarone or an implantable cardioverter-defibrillator for congestive heart failure. *N Engl J Med*. 2005;352:225–237. doi: 10.1056/NEJMoa043399
19. Beta-Blocker Evaluation of Survival Trial Investigators, Eichhorn EJ, Domanski MJ, Krause-Steinrauf H, Bristow MR, Lavori PW. A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *N Engl J Med*. 2001;344:1659–1667. doi: 10.1056/NEJM200105313442202
20. Digitalis Investigation Group. The effect of digoxin on mortality and morbidity in patients with heart failure. *N Engl J Med*. 1997;336:525–533. doi: 10.1056/NEJM199702203360801
21. SOLVD Investigators, Yusuf S, Pitt B, Davis CE, Hood WB, Cohn JN. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med*. 1991;325:293–302. doi: 10.1056/NEJM199108013250501
22. Action to Control Cardiovascular Risk in Diabetes Study Group, Gerstein HC, Miller ME, Byington RP, Goff DC, Jr., Bigger JT, Buse JB, Cushman WC, Genuth S, Ismail-Beigi F, et al. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med*. 2008;358:2545–2559. doi: 10.1056/NEJMoa0802743
23. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA*. 2002;288:2981–2997. doi: 10.1001/jama.288.23.2981
24. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial. Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT-LLT). *JAMA*. 2002;288:2998–3007. doi: 10.1001/jama.288.23.2998
25. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*. 2002;288:321–333. doi: 10.1001/jama.288.3.321
26. Wessler BS, Ruthazer R, Udelson JE, Gheorghiade M, Zannad F, Maggioni A, Konstam MA, Kent DM. Regional validation and recalibration of clinical predictive models for patients with acute heart failure. *J Am Heart Assoc*. 2017;6:e006121. doi: 10.1161/JAHA.117.006121
27. 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
28. 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
29. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *J Clin Oncol*. 2016;34:2534–2540. doi: 10.1200/JCO.2015.65.5654
30. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. 2004;23:2567–2586. doi: 10.1002/sim.1844

31. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science*. 2019;33:810–812. doi: 10.1126/science.aaw0029
32. Shah N, Steyerberg E, Kent D. Big data and predictive analytics: recalibrating expectations. *JAMA*. 2018;320:27–28. doi: 10.1001/jama.2018.5602
33. 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
34. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making*. 2015;35:162–169. doi: 10.1177/0272989X14547233
35. Olchanski N, Cohen JT, Neumann PJ, Wong JB, Kent DM. Understanding the value of individualized information: the impact of poor calibration or discrimination in outcome prediction models. *Med Decis Making*. 2017;37:790–801. doi: 10.1177/0272989X17704855
36. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S. The Clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385:283–286. doi: 10.1056/NEJMc2104626
37. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J*. 1991;121(1 Pt 2):293–298. doi: 10.1016/0002-8703(91)90861-b
38. Arima H, Yonemoto K, Doi Y, Ninomiya T, Hata J, Tanizaki Y, Fukuhara M, Matsumura K, Iida M, Kiyohara Y. Development and validation of a cardiovascular risk prediction model for Japanese: the Hisayama study. *Hypertens Res*. 2009;32:1119–1122. doi: 10.1038/hr.2009.161
39. Assmann G, Schulte H, Cullen P, Seedorf U. Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Münster (PROCAM) study. *Eur J Clin Invest*. 2007;37:925–932. doi: 10.1111/j.1365-2362.2007.01888.x
40. Balkau B, Hu G, Qiao Q, Tuomilehto J, Borch-Johnsen K, Pyörälä K; DECODE Study Group; European Diabetes Epidemiology Group. Prediction of the risk of cardiovascular mortality using a score that includes glucose as a risk factor. The DECODE Study. *Diabetologia*. 2004;47:2118–2128. doi: 10.1007/s00125-004-1574-5
41. Cederholm J, Eeg-Olofsson K, Eliasson B, Zethelius B, Nilsson PM, Gudbjörnsdottir S; Swedish National Diabetes Register. Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register. *Diabetes Care*. 2008;31:2038–2043. doi: 10.2337/dc08-0662
42. Chien KL, Su TC, Hsu HC, Chang WT, Chen PC, Sung FC, Chen MF, Lee YT. Constructing the prediction model for the risk of stroke in a Chinese population: report from a cohort study in Taiwan. *Stroke*. 2010;41:1858–1864. doi: 10.1161/STROKEAHA.110.586222
43. Chien KL, Hsu HC, Su TC, Chang WT, Chen PC, Sung FC, Lin HJ, Chen MF, Lee YT. Constructing a point-based prediction model for the risk of coronary artery disease in a Chinese community: a report from a cohort study in Taiwan. *Int J Cardiol*. 2012;157:263–268. doi: 10.1016/j.ijcard.2012.03.017
44. D'Agostino RB, Wolf PA, Belanger AJ, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. The Framingham Study. *Stroke*. 1994;25:40–43. doi: 10.1161/01.str.25.1.40
45. D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117:743–753. doi: 10.1161/CIRCULATIONAHA.107.699579
46. Donnan PT, Donnelly L, New JP, Morris AD. Derivation and validation of a prediction score for major coronary heart disease events in a U.K. type 2 diabetic population. *Diabetes Care*. 2006;29:1231–1236. doi: 10.2337/dc05-1911
47. Faeh D, Rohrmann S, Braun J. Better risk assessment with glycated hemoglobin instead of cholesterol in CVD risk prediction charts. *Eur J Epidemiol*. 2013;28:551–555. doi: 10.1007/s10654-013-9827-6
48. Ferrario M, Chiodini P, Chambless LE, Cesana G, Vanuzzo D, Panico S, Sega R, Pilotto L, Palmieri L, Giampaoli S; CUORE Project Research Group. Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE Cohort Study prediction equation. *Int J Epidemiol*. 2005;34:413–421. doi: 10.1093/ije/dyh405
49. Folsom AR, Chambless LE, Duncan BB, Gilbert AC, Pankow JS; Atherosclerosis Risk in Communities Study Investigators. Prediction of coronary heart disease in middle-aged adults with diabetes. *Diabetes Care*. 2003;26:2777–2784. doi: 10.2337/diacare.26.10.2777
50. 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
51. Jørgensen PG, Jensen JS, Marott JL, Jensen GB, Appleyard M, Mogelvang R. Electrocardiographic changes improve risk prediction in asymptomatic persons age 65 years or above without cardiovascular disease. *J Am Coll Cardiol*. 2014;64:898–906. doi: 10.1016/j.jacc.2014.05.050
52. Kang GD, Guo L, Guo ZR, Hu XS, Wu M, Yang HT. Continuous metabolic syndrome risk score for predicting cardiovascular disease in the Chinese population. *Asia Pac J Clin Nutr*. 2012;21:88–96.
53. Kim HC, Greenland P, Rossouw JE, Manson JE, Cochrane BB, Lasser NL, Limacher MC, Lloyd-Jones DM, Margolis KL, Robinson JG. Multimarker prediction of coronary heart disease risk: the Women's Health Initiative. *J Am Coll Cardiol*. 2010;55:2080–2091. doi: 10.1016/j.jacc.2009.12.047
54. Knuiman MW, Vu HT. Prediction of coronary heart disease mortality in Busselton, Western Australia: an evaluation of the Framingham, national health epidemiologic follow up study, and WHO ERICA risk scores. *J Epidemiol Community Health*. 1997;51:515–519. doi: 10.1136/jech.51.5.515
55. Liu J, Hong Y, D'Agostino RB Sr, Wu Z, Wang W, Sun J, Wilson PW, Kannel WB, Zhao D. Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study. *JAMA*. 2004;291:2591–2599. doi: 10.1001/jama.291.21.2591
56. Mainous AG 3rd, Koopman RJ, Diaz VA, Everett CJ, Wilson PW, Tilley BC. A coronary heart disease risk score based on patient-reported information. *Am J Cardiol*. 2007;99:1236–1241. doi: 10.1016/j.amjcard.2006.12.035
57. Noda H, Maruyama K, Iso H, Dohi S, Terai T, Fujioka S, Goto K, Horie S, Nakano S, Hirobe K; 3M Study Project Committee of the Japan Association of Occupational Physicians "San-yu-kai". Prediction of myocardial infarction using coronary risk scores among Japanese male workers: 3M Study. *J Atheroscler Thromb*. 2010;17:452–459. doi: 10.5551/jat.3277
58. Prieto-Merino D, Dobson J, Gupta AK, Chang CL, Sever PS, Dahlöf B, Wedel H, Pocock S, Poulter N; ASCOT-BPLA Investigators. ASCORE: an up-to-date cardiovascular risk score for hypertensive patients reflecting contemporary clinical practice developed using the (ASCOT-BPLA) trial data. *J Hum Hypertens*. 2013;27:492–496. doi: 10.1038/jhh.2013.3
59. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. *Circulation*. 2008;118:2243–51, 4p following 2251. doi: 10.1161/CIRCULATIONAHA.108.814251
60. Stevens RJ, Kothari V, Adler AI, Stratton IM; United Kingdom Prospective Diabetes Study (UKPDS) Group. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clin Sci (Lond)*. 2001;101:671–679.
61. Støvring H, Harmsen CG, Wisløff T, Jarbøl DE, Nexøe J, Nielsen JB, Kristiansen IS. A competing risk approach for the European Heart SCORE model based on cause-specific and all-cause mortality. *Eur J Prev Cardiol*. 2013;20:827–836. doi: 10.1177/2047487312445425
62. Teramoto T, Ohashi Y, Nakaya N, Yokoyama S, Mizuno K, Nakamura H; MEGA Study Group. Practical risk prediction tools for coronary heart disease in mild to moderate hypercholesterolemia in Japan: originated from the MEGA study data. *Circ J*. 2008;72:1569–1575. doi: 10.1253/circj.cj-08-0191
63. 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
64. Wilson PW, Bozeman SR, Burton TM, Hoaglin DC, Ben-Joseph R, Pashos CL. Prediction of first events of coronary heart disease and stroke with consideration of adiposity. *Circulation*. 2008;118:124–130. doi: 10.1161/CIRCULATIONAHA.108.772962
65. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke*. 1991;22:312–318. doi: 10.1161/01.str.22.3.312
66. Wu Y, Liu X, Li X, Li Y, Zhao L, Chen Z, Li Y, Rao X, Zhou B, Detrano R, et al; USA-PRC Collaborative Study of Cardiovascular and Cardiopulmonary Epidemiology Research Group; China Multicenter Collaborative Study of Cardiovascular Epidemiology Research Group. Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. *Circulation*. 2006;114:2217–2225. doi: 10.1161/CIRCULATIONAHA.105.607499
67. Yang X, So WY, Kong AP, Ho CS, Lam CW, Stevens RJ, Lyu RR, Yin DD, Cockram CS, Tong PC, et al. Development and validation of stroke risk equation for Hong Kong Chinese patients with type 2 diabetes: the Hong Kong Diabetes Registry. *Diabetes Care*. 2007;30:65–70. doi: 10.2337/dc06-1273

68. Yatsuya H, Iso H, Yamagishi K, Kokubo Y, Saito I, Suzuki K, Sawada N, Inoue M, Tsugane S. Development of a point-based prediction model for the incidence of total stroke: Japan public health center study. *Stroke*. 2013;44:1295–1302. doi: 10.1161/STROKEAHA.111.677534
69. Zhang XF, Attia J, D'Este C, Yu XH, Wu XG. A risk score predicted coronary heart disease and stroke in a Chinese cohort. *J Clin Epidemiol*. 2005;58:951–958. doi: 10.1016/j.jclinepi.2005.01.013
70. Abraham WT, Fonarow GC, Albert NM, Stough WG, Gheorghiade M, Greenberg BH, O'Connor CM, Sun JL, Yancy CW, Young JB; OPTIMIZE-HF Investigators and Coordinators. Predictors of in-hospital mortality in patients hospitalized for heart failure: insights from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). *J Am Coll Cardiol*. 2008;52:347–356. doi: 10.1016/j.jacc.2008.04.028
71. Alehagen U, Lindstedt G, Levin LA, Dahlström U. The risk of cardiovascular death in elderly patients with possible heart failure. Results from a 6-year follow-up of a Swedish primary care population. *Int J Cardiol*. 2005;100:17–27. doi: 10.1016/j.ijcard.2004.03.031
72. Alla F, Briançon S, Juillière Y, Mertes PM, Villemot JP, Zannad F. Differential clinical prognostic classifications in dilated and ischemic advanced heart failure: the EPICAL study. *Am Heart J*. 2000;139:895–904. doi: 10.1016/s0002-8703(00)90023-1
73. Bilchick KC, Stukenborg GJ, Kamath S, Cheng A. Prediction of mortality in clinical practice for medicare patients undergoing defibrillator implantation for primary prevention of sudden cardiac death. *J Am Coll Cardiol*. 2012;60:1647–1655. doi: 10.1016/j.jacc.2012.07.028
74. Borleffs CJ, van Welsenes GH, van Bommel RJ, van der Velde ET, Bax JJ, van Erven L, Putter H, van der Bom JG, Rosendaal FR, Schalij MJ. Mortality risk score in primary prevention implantable cardioverter defibrillator recipients with non-ischaemic or ischaemic heart disease. *Eur Heart J*. 2010;31:712–718. doi: 10.1093/eurheartj/ehp497
75. Bouvy ML, Heerdink ER, Leufkens HG, Hoes AW. Predicting mortality in patients with heart failure: a pragmatic approach. *Heart*. 2003;89:605–609. doi: 10.1136/heart.89.6.605
76. Felker GM, Leimberger JD, Califf RM, Cuffe MS, Massie BM, Adams KF Jr, Gheorghiade M, O'Connor CM. Risk stratification after hospitalization for decompensated heart failure. *J Card Fail*. 2004;10:460–466. doi: 10.1016/j.cardfail.2004.02.011
77. Fonarow GC, Adams KF Jr, Abraham WT, Yancy CW, Boscardin WJ; ADHERE Scientific Advisory Committee, Study Group, and Investigators. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA*. 2005;293:572–580. doi: 10.1001/jama.293.5.572
78. Goldenberg I, Vyas AK, Hall WJ, Moss AJ, Wang H, He H, Zareba W, McNitt S, Andrews ML; MADIT-II Investigators. Risk stratification for primary implantation of a cardioverter-defibrillator in patients with ischemic left ventricular dysfunction. *J Am Coll Cardiol*. 2008;51:288–296. doi: 10.1016/j.jacc.2007.08.058
79. Huynh BC, Rovner A, Rich MW. Identification of older patients with heart failure who may be candidates for hospice care: development of a simple four-item risk score. *J Am Geriatr Soc*. 2008;56:1111–1115. doi: 10.1111/j.1532-5415.2008.01756.x
80. Kinugasa Y, Kato M, Sugihara S, Hirai M, Kotani K, Ishida K, Yanagihara K, Kato Y, Ogino K, Igawa O, et al. A simple risk score to predict in-hospital death of elderly patients with acute decompensated heart failure–hypoalbuminemia as an additional prognostic factor. *Circ J*. 2009;73:2276–2281. doi: 10.1253/circj.cj-09-0498
81. Kraaier K, Scholten MF, Tijssen JG, Theuns DA, Jordaens LJ, Wilde AA, van Dessel PF. Early mortality in prophylactic implantable cardioverter defibrillator recipients: development and validation of a clinical risk score. *Europace*. 2014;16:40–46. doi: 10.1093/europace/eut223
82. Kramer DB, Friedman PA, Kallinen LM, Morrison TB, Crusan DJ, Hodge DO, Reynolds MR, Hauser RG. Development and validation of a risk score to predict early mortality in recipients of implantable cardioverter-defibrillators. *Heart Rhythm*. 2012;9:42–46. doi: 10.1016/j.hrthm.2011.08.031
83. Martínez-Sellés M, Martínez E, Cortés M, Prieto R, Gallego L, Fernández-Avilés F. Determinants of long-term survival in patients hospitalized for heart failure. *J Cardiovasc Med (Hagerstown)*. 2010;11:164–169. doi: 10.2459/JCM.0b013e328332ea96
84. O'Connor CM, Abraham WT, Albert NM, Clare R, Gattis Stough W, Gheorghiade M, Greenberg BH, Yancy CW, Young JB, Fonarow GC. Predictors of mortality after discharge in patients hospitalized with heart failure: an analysis from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). *Am Heart J*. 2008;156:662–673. doi: 10.1016/j.ahj.2008.04.030
85. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA; American Heart Association Get With the Guidelines-Heart Failure Program. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcomes*. 2010;3:25–32. doi: 10.1161/CIRCOUTCOMES.109.854877
86. Pocock SJ, Ariti CA, McMurray JJ, Maggioni A, Køber L, Squire IB, Swedberg K, Dobson J, Poppe KK, Whalley GA, et al; Meta-Analysis Global Group in Chronic Heart Failure. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J*. 2013;34:1404–1413. doi: 10.1093/eurheartj/ehs337
87. Scrutinio D, Guida P, Passantino A, Lagioia R, Pepe S, Catanzaro R, Santoro D. Amino-terminal pro-B-type natriuretic peptide for risk prediction in acute decompensated heart failure. *Congest Heart Fail*. 2012;18:308–314. doi: 10.1111/j.1751-7133.2012.00301.x
88. Subramanian D, Subramanian V, Deswal A, Mann DL. New predictive models of heart failure mortality using time-series measurements and ensemble models. *Circ Heart Fail*. 2011;4:456–462. doi: 10.1161/CIRCHEARTFAILURE.110.958496
89. van Rees JB, Borleffs CJ, van Welsenes GH, van der Velde ET, Bax JJ, van Erven L, Putter H, van der Bom JG, Schalij MJ. Clinical prediction model for death prior to appropriate therapy in primary prevention implantable cardioverter defibrillator patients with ischaemic heart disease: the FADES risk score. *Heart*. 2012;98:872–877. doi: 10.1136/heartjnl-2011-300632
90. Velavan P, Khan NK, Goode K, Rigby AS, Loh PH, Komajda M, Follath F, Swedberg K, Madeira H, Cleland JG. Predictors of short term mortality in heart failure - insights from the Euro Heart Failure survey. *Int J Cardiol*. 2010;138:63–69. doi: 10.1016/j.ijcard.2008.08.004
91. Addala S, Grines CL, Dixon SR, Stone GW, Boura JA, Ochoa AB, Pellizzon G, O'Neill WW, Kahn JK. Predicting mortality in patients with ST-elevation myocardial infarction treated with primary percutaneous coronary intervention (PAMI risk score). *Am J Cardiol*. 2004;93:629–632. doi: 10.1016/j.amjcard.2003.11.036
92. Amin ST, Morrow DA, Braunwald E, Sloan S, Contant C, Murphy S, Antman EM. Dynamic TIMI risk score for STEMI. *J Am Heart Assoc*. 2013;2:e003269. doi: 10.1161/JAHA.112.003269
93. Antman EM, Cohen M, Bernink PJ, McCabe CH, Horacek T, Papuchis G, Mautner B, Corbalan R, Radley D, Braunwald E. The TIMI risk score for unstable angina/non-ST elevation MI: A method for prognostication and therapeutic decision making. *JAMA*. 2000;284:835–842. doi: 10.1001/jama.284.7.835
94. Califf RM, Woodlief LH, Harrell FE Jr, Lee KL, White HD, Guerci A, Barbash GI, Simes RJ, Weaver WD, Simoons ML, et al. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. *Am Heart J*. 1997;133:630–639. doi: 10.1016/s0002-8703(97)70164-9
95. de Boer SP, Barnes EH, Westerhout CM, Simes RJ, Granger CB, Kastrati A, Widimsky P, de Boer MJ, Zijlstra F, Boersma E. High-risk patients with ST-elevation myocardial infarction derive greatest absolute benefit from primary percutaneous coronary intervention: results from the Primary Coronary Angioplasty Trialist versus thrombolysis (PCAT)-2 collaboration. *Am Heart J*. 2011;161:500–507.e1. doi: 10.1016/j.ahj.2010.11.022
96. Dorsch MF, Lawrance RA, Sapsford RJ, Oldham J, Greenwood DC, Jackson BM, Morrell C, Ball SG, Robinson MB, Hall AS. A simple benchmark for evaluating quality of care of patients following acute myocardial infarction. *Heart*. 2001;86:150–154. doi: 10.1136/heart.86.2.150
97. Giraldez RR, Sabatine MS, Morrow DA, Mohanavelu S, McCabe CH, Antman EM, Braunwald E. Baseline hemoglobin concentration and creatinine clearance composite laboratory index improves risk stratification in ST-elevation myocardial infarction. *Am Heart J*. 2009;157:517–524. doi: 10.1016/j.ahj.2008.10.021
98. Huynh T, Kouz S, Yan AT, Yan A, Danchin N, O'Loughlin J, Loughlin JO, Schampaert E, Yan RT, Yan R, et al. Canada Acute Coronary Syndrome Risk Score: a new risk score for early prognostication in acute coronary syndromes. *Am Heart J*. 2013;166:58–63. doi: 10.1016/j.ahj.2013.03.023
99. Krumholz HM, Chen J, Wang Y, Radford MJ, Chen YT, Marciniak TA. Comparing AMI mortality among hospitals in patients 65 years of age and older: evaluating methods of risk adjustment. *Circulation*. 1999;99:2986–2992. doi: 10.1161/01.cir.99.23.2986

100. Morrow DA, Antman EM, Charlesworth A, Cairns R, Murphy SA, de Lemos JA, Giugliano RP, McCabe CH, Braunwald E. TIMI risk score for ST-elevation myocardial infarction: a convenient, bedside, clinical score for risk assessment at presentation: an intravenous nPA for treatment of infarcting myocardium early II trial substudy. *Circulation*. 2000;102:2031–2037. doi: 10.1161/01.cir.102.17.2031
101. Morrow DA, Antman EM, Giugliano RP, Cairns R, Charlesworth A, Murphy SA, de Lemos JA, McCabe CH, Braunwald E. A simple risk index for rapid initial triage of patients with ST-elevation myocardial infarction: an InTIME II substudy. *Lancet*. 2001;358:1571–1575. doi: 10.1016/S0140-6736(01)06649-1
102. Moscucci M, Kline-Rogers E, Share D, O'Donnell M, Maxwell-Eward A, Meengs WL, Kraft P, DeFranco AC, Chambers JL, Patel K, et al. Simple bedside additive tool for prediction of in-hospital mortality after percutaneous coronary interventions. *Circulation*. 2001;104:263–268. doi: 10.1161/01.cir.104.3.263
103. Negassa A, Monrad ES, Bang JY, Srinivas VS. Tree-structured risk stratification of in-hospital mortality after percutaneous coronary intervention for acute myocardial infarction: a report from the New York State percutaneous coronary intervention database. *Am Heart J*. 2007;154:322–329. doi: 10.1016/j.ahj.2007.03.052
104. Negassa A, Monrad ES, Srinivas VS. A simple prognostic classification model for postprocedural complications after percutaneous coronary intervention for acute myocardial infarction (from the New York State percutaneous coronary intervention database). *Am J Cardiol*. 2009;103:937–942. doi: 10.1016/j.amjcard.2008.11.055
105. Qureshi MA, Safian RD, Grines CL, Goldstein JA, Westveer DC, Glazier S, Balasubramanian M, O'Neill WW. Simplified scoring system for predicting mortality after percutaneous coronary intervention. *J Am Coll Cardiol*. 2003;42:1890–1895. doi: 10.1016/j.jacc.2003.06.014
106. Singh M, Lennon RJ, Holmes DR Jr, Bell MR, Rihal CS. Correlates of procedural complications and a simple integer risk score for percutaneous coronary intervention. *J Am Coll Cardiol*. 2002;40:387–393. doi: 10.1016/s0735-1097(02)01980-0
107. Yap YG, Duong T, Bland M, Malik M, Torp-Pedersen C, Køber L, Connolly SJ, Gallagher MM, Camm AJ. Potential demographic and baselines variables for risk stratification of high-risk post-myocardial infarction patients in the era of implantable cardioverter defibrillator–a prognostic indicator. *Int J Cardiol*. 2008;126:101–107. doi: 10.1016/j.ijcard.2007.03.122