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

Benjamin S. Wessler[1,2], Jason Nelson[1], Jinny G. Park[1], Hannah McGinnes[1], Gaurav Gulati[1,2], Riley Brazil[1], Ben Van Calster[3], D. van Klaveren[1,4], Esmee Venema[5,6], Ewout Steyerberg[5,7], Jessica K. Paulus[1], David M. Kent[1]

1. Predictive Analytics and Comparative Effectiveness (PACE), Tufts Medical Center, United States of America
2. Division of Cardiology, Tufts Medical Center, Boston, MA
3. KU Leuven, Department of Development and Regeneration, Leuven, Belgium
4. Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands
5. Department of Public Health, Erasmus MC University Medical Center, Rotterdam, the Netherlands
6. Department of Neurology, Erasmus MC University Medical Center, Rotterdam, the Netherlands
7. Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

Abstract:

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 PACE 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: 2030 external validations of 1382 CPMs were identified. 807 (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 AUC was 0.73 (25^{th} -75^{th} percentile [IQR] 0.66, 0.79), representing a median percent decrease in discrimination of -11.1% (IQR -32.4%, +2.7%) compared to performance on derivation data. 81% (n = 1333) of validations reporting AUC 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, 3.1) for 'closely related' validations (n=123), -9.0 (IQR -27.6, 3.9) for 'related validations' (n=862) and -17.2% (IQR -42.3, 0) for 'distantly related' validations (n=717) (p<0.001).

Conclusion: 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.

Introduction:

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 (CVD) 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 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.

While 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] though 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 (i.e. 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 is available at www.pacecpmregistry.org and represents a field synopsis of

prediction models for patients at risk for and with known cardiovascular disease. 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 **Supplement Figure 1**. 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 two 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 non-overlapping 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 (AUC) change compared to the AUC 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 to predicted rates.

CPM performance

Consistent with prior work[12], percent 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 AUC - 0.5) – (Derivation AUC - 0.5) / (Derivation AUC - 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 **(Supplement Table 1)**. 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 (e.g., CPM validated on population with non-overlapping index condition or outcome). Generally, the relatedness rubrics were based on 5 domains: (1) recruitment setting (e.g., outpatient vs emergency room vs inpatient), (2) major inclusion/ exclusion criteria, (3) intervention type (e.g., 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. Non-random 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 (e.g., clinical trial vs. medical record), sample size, number of events, number of predictors, prediction time horizon (< 30 days, 30-265 days, >365 days), regression method (e.g., logistic regression vs 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 paper, 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 (GEE)[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, USA).

Results:

Overview of Validations

The Registry includes 1382 CPMs for CVD 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 papers. Only 575 (42%) of the CPMs in the Registry have ever been validated. On average there were 1.5 validations per de novo CPM, with a very skewed distribution. The Logistic EuroSCORE[16] has been externally validated 94 times. For this analysis, we included 1846 validations of 556 CPMs after exclusion of 19 decision trees and 156 validations performed on 'unrelated' (i.e. populations with different index conditions or non-overlapping outcomes) samples. The median external validation sample size was 861 (25^{th}-75^{th} percentile [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 area under the receiver operating characteristic curve (AUC). The median derivation AUC was 0.77 (IQR 0.73, 0.82). The median external validation AUC was 0.73 (IQR 0.66, 0.79) representing a median percent change in discrimination of -11.1% (IQR -32.4%, +2.7%) (**Table 2**). Of the validations with decreased performance (n = 795), 25% (n = 195) had less than 10% decrement in discrimination. Two percent (n = 35) had greater than 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%, 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.

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 CVD (population samples) and 286 validations of 52 CPMs for Cardiac Surgery. Only five index conditions had $\geq$ 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 **Supplemental Figure 2**).

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 'closely related' populations, 862 (51%) were performed on 'related' populations, while 717 (42%) were performed on 'distantly related' populations **(Table 2)**. The median AUC for 'closely related' validations was 0.78 (IQR 0.719, 0.841). The median AUC for 'related population' validations was 0.75 (IQR 0.68, 0.803). The median AUC 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, 3.1) for 'closely related' validations, -9.0 (IQR -27.6, 3.9) for 'related validations' and -17.2% (-42.3, 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 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 EuroSCORE (validated 94 times) 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 **Supplemental Table 2**. 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 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 AUC decrease was -15.6% (95% CI -22.0, -9.1) compared to 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, -0.8). We also observed that validations demonstrated AUCs that were 9.8% (95% CI 5.4, 14.2) higher when reported in the same manuscript (with the same authors) as the de novo CPM report compared to validations reported in different manuscripts with non-overlapping authors.

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 have been validated only once. A small minority of models have been validated numerous times. The value of single validations is unclear, since 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 ten most validated CPMs have each been validated more than 20 times; all show substantial variation in discrimination across these validation studies, from virtually useless (i.e., 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, since 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[17] and calls for more robust stewardship of algorithms for health care[18].

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] 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.

In our review, 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 to performance on the derivation database).[6,8] There are several potential reasons why model performance might decrease, including model invalidity (e.g., due to 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 model performance when tested on distantly-related populations compared to 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[19] and outcome predictors[20] are different very early in the disease course. So too changes in treatments received (e.g., different ACS revascularization approaches,[21] stent types[22], or outcome definitions[23,24]) 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 (PCI) 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[25,26] should be accepted as trustworthy unless CPM performance is specifically known to be excellent on populations like those being treated. While 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 (e.g. in a cluster randomized trial[27]) is rarely performed prior to 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.

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 under-assessed. 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 (e.g., 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[28], may help compare performance across multiple validations. Decrements in calibration may be as serious as, or even more serious than, decrements in discrimination, since 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 to best "one-sized-fits-all" strategies.[4,29] Such evaluations would have required individual patient data, since 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[30]. 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 semi-automated approach heavily reliant on natural language processing.[31]

Conclusion:

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 to the model development phase. Calibration has been widely under-assessed 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 (e.g. through multiple validations across various practice settings) and how models are evaluated for clinical use.

Acknowledgements

Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1606-35555). The views in this work are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

The authors wish to acknowledge the contributions of Mr. Vandan Patel for his work on the relatedness effort.

BW is supported by K23AG055667 from NIH-NIA and R03AG056447 from NIH-NIA

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. *Diagnostic Progn Res*. 2017;1:20.

2. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, Briggs A, Udumyan R, Moons KGM, Steyerberg EW, Roberts I, Schroter S, Altman DG, Riley RD. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. *BMJ*. 2013;346:e5595.

3. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Circulation*. 2015;131:211–9.

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;21:128–38.

5. Vergouwe Y, Moons KGM, 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.

6. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68:25–34.

7. Adibi A, Sadatsafavi M, Ioannidis JPA. Validation and Utility Testing of Clinical Prediction Models. *JAMA*. 2020;

8. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, Lassale CM, Siontis GCM, Chiocchia V, Roberts C, Schlüssel MM, Gerry S, Black JA, Heus P, Van Der Schouw YT, Peelen LM, Moons KGM. Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*. 2016;353.

9. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17:230.

10. Van Calster B, Vickers AJ. Calibration of Risk Prediction Models. *Med Decis Mak*. 2015;35:162–169.

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;74:167–76.

12. Wessler BS, Lundquist CM, Koethe B, Park JG, Brown K, Williamson T, Ajlan M, Natto Z, Lutz JS, Paulus JK, Kent DM. Clinical Prediction Models for Valvular Heart Disease. *J Am Heart Assoc*. 2019;8:e011972.

13. Harrell , FE. Regression Modeling Strategies. Cham: Springer International Publishing; 2015.

14. Zeger SL, Liang K-Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*. 1986;42:121.

15. Zeger SL, Liang K-Y, Albert PS. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*. 1988;44:1049.

16. Roques F. The logistic EuroSCORE. *Eur Heart J*. 2003;24:882.

17. Adibi A, Sadatsafavi M, Ioannidis JPA. Validation and Utility Testing of Clinical Prediction Models. *JAMA*. 2020;324:235.

18. Eaneff S, Obermeyer Z, Butte AJ. The Case for Algorithmic Stewardship for Artificial Intelligence and Machine Learning Technologies. *JAMA*. 2020;324:1397.

19. Collins SP, Levy PD, Lindsell CJ, Pang PS, Storrow AB, Miller CD, Naftilan AJ, Thohan V, Abraham WT, Hiestand B, Filippatos G, Diercks DB, Hollander J, Nowak R, Peacock WF, Gheorghiade M. The Rationale for an Acute Heart Failure Syndromes Clinical Trials Network. *J Card Fail*. 2009;15:467–474.

20. Karam N, Bataille S, Marijon E, Tafflet M, Benamer H, Caussin C, Garot P, Juliard J-M, Pires V, Boche T, Dupas F, Le Bail G, Lamhaut L, Simon B, Allonneau A, Mapouata M, Loyeau A, Empana J-P, Lapostolle F, Spaulding C, Jouven X, Lambert Y, 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.

21. Mehta SR, Wood DA, Storey RF, Mehran R, Bainey KR, Nguyen H, Meeks B, Di Pasquale G, López-Sendón J, Faxon DP, Mauri L, Rao S V., Feldman L, Steg PG, Avezum Á, Sheth T, Pinilla-Echeverri N, Moreno R, Campo G, Wrigley B, Kedev S, Sutton A, Oliver R, Rodés-Cabau J, Stanković G, Welsh R, Lavi S, Cantor WJ, Wang J, Nakamya J, Bangdiwala SI, Cairns JA. Complete Revascularization with Multivessel PCI for Myocardial Infarction. *N Engl J Med*. 2019;381:1411–1421.

22. Piccolo R, Bonaa KH, Efthimiou O, Varenne O, Baldo A, Urban P, Kaiser C, Remkes W, Räber L, de Belder A, van 't Hof AWJ, Stankovic G, Lemos PA, Wilsgaard T, Reifart J, Rodriguez AE, Ribeiro EE, Serruys PWJC, Abizaid A, Sabaté M, Byrne RA, de la Torre Hernandez JM, Wijns W, Jüni P, Windecker S, Valgimigli M, Piccolo R, Bonaa KH, Efthimiou O, Varenne O, Baldo A, Urban P, Kaiser C, Remkes W, Räber L, de Belder A, van't Hof AWJ, Stankovic G, Lemos PA, Wilsgaard T, Reifart J, Rodriguez AE, Ribeiro EE, Serruys PWJC, Abizaid A, Sabaté M, Byrne RA, de la Torre Hernandez JM, Wijns W, Jüni P, Windecker S, Valgimigli M. 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.

23. 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;51:701–7.

24. Mehran R, Rao S V., Bhatt DL, Gibson CM, Caixeta A, Eikelboom J, Kaul S, Wiviott SD, Menon V, Nikolsky E, Serebruany V, Valgimigli M, Vranckx P, Taggart D, Sabik JF, Cutlip DE, Krucoff MW, Ohman EM, Steg PG, White H. Standardized bleeding definitions for cardiovascular clinical trials: A consensus report from the bleeding academic research consortium. *Circulation*. 2011;123:2736–2747.

25. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, Robinson JG, Schwartz JS, Shero ST, Smith SC, Sorlie P, Stone NJ, Wilson PWF. 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;63:2935–2959.

26. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, Fonarow GC, Geraci SA, Horwich T, Januzzi JL, Johnson MR, Kasper EK, Levy WC, Masoudi FA, McBride PE, McMurray JJ V, Mitchell JE, Peterson PN, Riegel B, Sam F, Stevenson LW, Tang WHW, Tsai EJ, Wilkoff BL. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. *Circulation*. 2013;128:e240-327.

27. Chew DP, Hyun K, Morton E, Horsfall M, Hillis GS, Chow CK, Quinn S, D'Souza M, Yan AT, Gale CP, Goodman SG, Fox K, Brieger D. Objective Risk Assessment vs Standard Care for Acute Coronary Syndromes. *JAMA Cardiol*. 2020;

28. 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.

29. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 26:565–74.

30. 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.

31. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8:163.

32. Steyerberg EW. Clinical Prediction Models. New York, NY: Springer New York; 2009.

Tables:

Table 1. De Novo Models Summary. Characteristics of unique CPMs in PACE CPM Registry in aggregate for all CPMs, CPMs that have ever been validated, and CPMs never validated.

	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)

All values reported as median (IQR) unless otherwise noted; EPV indicates events per variable.
* 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.
‡ Includes validations in Sample Set (Table 2).

Table 2. External Validations Summary. 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.

	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, 3306)	882 (330, 3479)	1460 (494, 4905)	960 (413, 4492)	681 (256, 2152)
Events	68 (29, 192)	67 (28, 188)	144 (48, 275)	72 (31, 201)	52 (25, 158)
EPV Final Model*	6.9 (2.5, 22.9)	6.3 (2.4, 21.1)	14.1 (6.7, 33.9)	6.5 (2.3, 20.6)	5.6 (2.1, 18.5)
C statistic†	0.73 (0.66, 0.794)	0.73 (0.664, 0.796)	0.78 (0.719, 0.841)	0.75 (0.68, 0.803)	0.701 (0.64, 0.77)
% Change in Discrimination	-11.1 (-32.4, 2.7)	-11.1 (-32, 2.6)	-3.7 (-13.2, 3.1)	-9.0 (-27.6, 3.9)	-17.2 (-42.3, 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)

All values reported as median (IQR) unless otherwise noted; EPV indicates events per variable.

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

† C-statistic reported in 91% of exercises.

‡ Excluded decision tree, CART, 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.[32]

Table 3: Conditions with the Most External Validations (Top 10). Discrimination and characteristics of validations for CPMs with top 10 most validated index conditions in PACE CPM Registry.

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

CPMs indicates clinical prediction models; ACS, acute coronary syndrome; CHF, chronic heart failure; NA, not applicable; VTE, venous thromboembolism. N missing refers to the number of CPMs not reporting a baseline AUC. N missing refers to either derivation AUC or validation AUC missing (delta C not available).

Table 4. Top 10 Most Validated CPMs. Description of top 10 most validated CPMs in PACE CPM Registry and validation performance.

Model Name	Index Condition	Number of validations	Development AUC	Median validation AUC (IQR)	Range in validation AUC
Logistic EuroSCORE	Cardiac Surgery	94	NR	0.75 (0.67, 0.80)	0.48-0.90
Additive EuroSCORE	Cardiac Surgery	86	0.79	0.77 (0.72, 0.82)	0.58-0.90
EuroSCORE II	Valve Disease	65	0.81	0.76 (0.68, 0.81)	0.40-0.87
GRACE	CAD: ACS	53	0.83	0.80 (0.73, 0.84)	0.60-0.95
STS (valve) - Mortality	Cardiac Surgery	51	0.81	0.70 (0.64, 0.76)	0.45-0.85
CHA_2DS_2-VASc	Arrhythmia	45	0.61	0.66 (0.61, 0.69)	0.45-0.93
$CHADS_2$	Arrhythmia	37	0.82	0.65 (0.61, 0.68)	0.51-0.87
FRS - CHD	Population Sample	35	NR	0.68 (0.63, 0.72)	0.54-0.80
ICH Score	Stroke	27	0.92	0.85 (0.75, 0.87)	0.69-0.94
ACEF Score	Cardiac Surgery	26	0.74	0.74 (0.68, 0.77)	0.54-0.87

CPMs indicates clinical prediction models; AUC, area under curve; IQR, interquartile range; EuroSCORE, European System for Cardiac Operative Risk Evaluation; NR, not reported; CAD, coronary artery disease; ACS, acute coronary syndrome; GRACE, Global Registry of Acute Coronary Events; STS, Society of Thoracic Surgeons; CHADS-VASc, Congestive heart failure/LV dysfunction, Hypertension, Age, Diabetes mellitus, Stroke/TIA/TE, Vascular disease, Age, Sex category; CHADS, CHF, Hypertension, Age, DM, History of Stroke or TIA; FRS – CHD, Framingham Risk Score for Coronary Heart Disease; ICH, IntraCerebral Hemorrhage; ACEF, Age, Creatinine, and left ventricular Ejection Fraction.

Table 5. Predictors of Worse Discrimination: Variable distributions and GEE model results. Results of regression analysis to detect predictors of change in discrimination performance from derivation to validation.

		Univariate			Multivariate (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%)*					
Closely Related	79 (7.5)	1054	*Reference*		*Reference*	
Related	544 (51.6)		-5.9 (-10.5, -1.4)	0.011	-1.3 (-7.1, 4.5)	0.660
Distantly Related	431 (40.9)		-15.6 (-22.0, -9.1)	<.001	**-9.8 (-18.8, -0.8)**	**0.033**
CPM Authors, n (%)						
Diff Paper, Author Overlap	94 (8.2)	1147	7.3 (-1.2, 15.8)	0.092	5.1 (-4.2, 14.4)	0.283
Diff Paper, No Author Overlap	849 (74.0)		*Reference*		*Reference*	
Same Paper	204 (17.8)		9.8 (5.4, 14.2)	<.001	5.5 (-0.8, 11.9)	0.088
CPM Method, n (%)	*Frequency Missing = 20 (1.7%)*					
Logistic Regression	859 (76.2)	1127	*Reference*		*Reference*	
Other	7 (0.6)		-0.1 (-12.4, 12.2)	0.985	-1.1 (-13.7, 11.6)	0.870
Time-to-Event Regression	261 (23.2)		2.6 (-5.8, 11)	0.541	-1.4 (-11, 8.2)	0.768
CPM data source, n (%)	*Frequency Missing = 4 (0.3%)*					
Clinical Trial	118 (10.3)	1143	5.8 (-8.9, 20.5)	0.437	3.7 (-14.1, 21.5)	0.684
Medical Record	614 (53.7)		*Reference*		*Reference*	
Other	114 (10.0)		-1.3 (-11.1, 8.6)	0.803	-2.5 (-14.0, 9.0)	0.669
Registry	297 (26.0)		2.3 (-5.5, 10)	0.569	2.1 (-6.3, 10.5)	0.616
Validation data source, n (%)	*Frequency Missing = 47 (4.1%)*					
Clinical Trial	99 (9.0)	1100	-5.9 (-12.5, 0.6)	0.077	-7.3 (-15.4, 0.8)	0.076
Medical Record	606 (55.1)		*Reference*		*Reference*	
Other	58 (5.3)		3.9 (-5.2, 13)	0.402	3.3 (-5.1, 11.7)	0.440
Registry	337 (30.6)		1.5 (-3.5, 6.5)	0.560	1.2 (-3.5, 6.0)	0.606
Relative outcome rate difference > 40%, n (%)	*Frequency Missing = 402 (35.0%)*	745				
Yes	384 (51.5)		-4.4 (-9.4, 0.7)	0.091	-1.5 (-6.3, 3.3)	0.540
No	361 (48.5)		*Reference*		*Reference*	

CPM EPV, median (IQR)	*Frequency Missing = 214 (18.7%)*	933				
	23.4 (16.3, 58.8)		0.4 (-1.9, 2.7)**	0.718	1.8 (-1.1, 4.6)	0.218

GEE, generalized estimating equation; AUC, area under curve; CPM, clinical prediction model; Diff, different; EPV, events per variable (events per included variable in the final model, not candidate variables); IQR, interquartile range.

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

**Natural log transformed

Figure 1. Flowchart of clinical prediction model external validation review process.

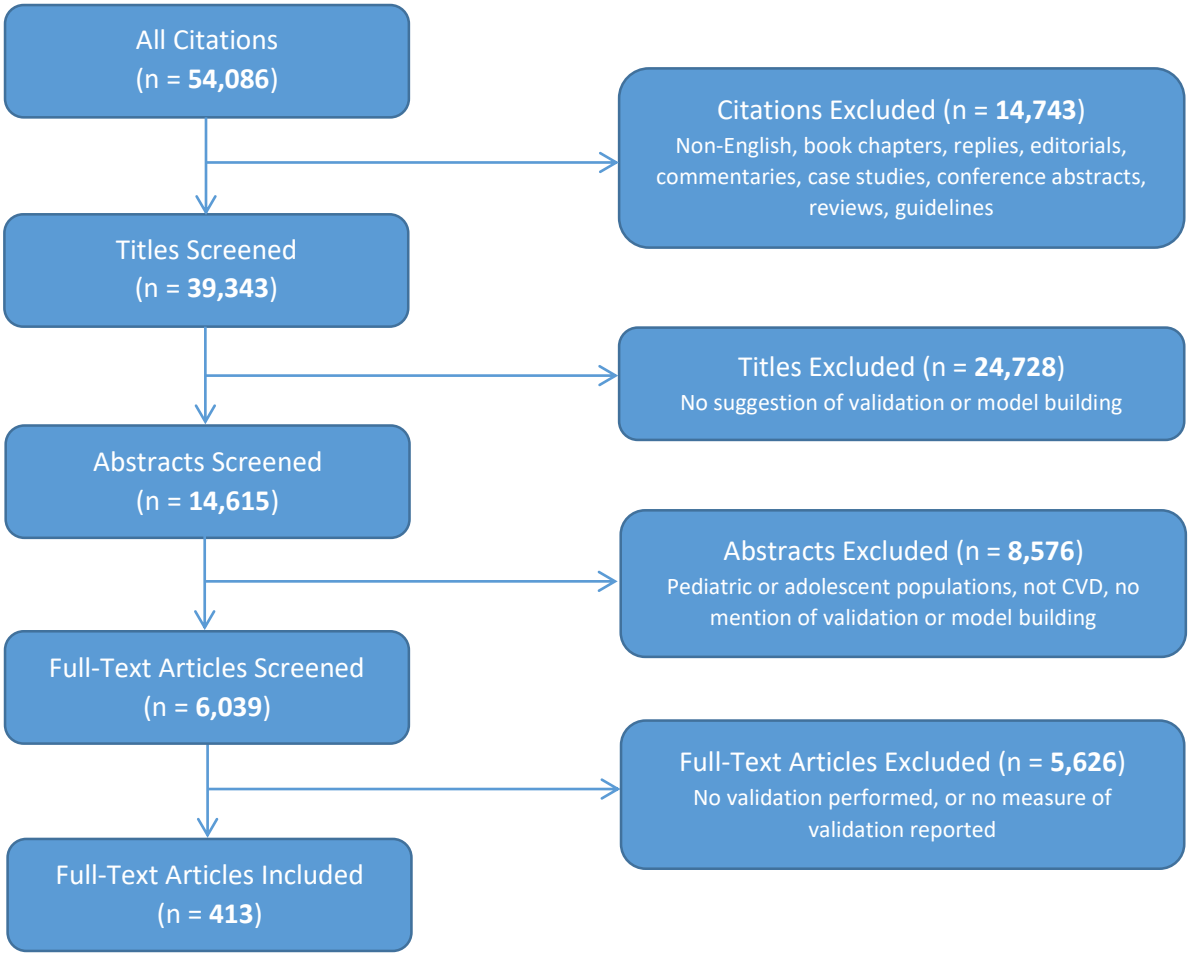
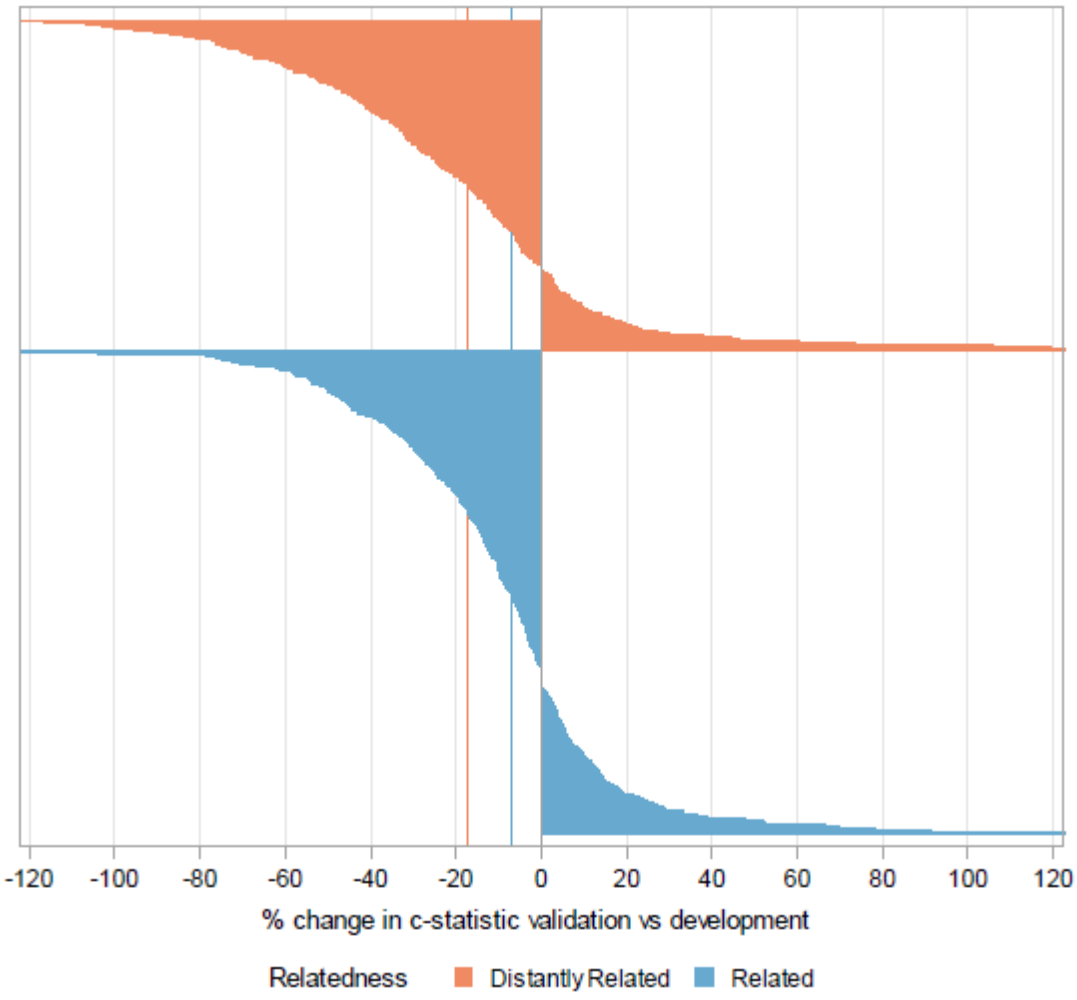


Figure 2. Percent Change in Validation C-statistic Performance versus Derivation C-statistic Performance.



All Index Conditions delta AUC								
Relatedness	N Obs	N	N Miss	Median	Lower Quartile	Upper Quartile	Minimum	Maximum
Distantly Related	717	430	287	-17.2	-42.4	0.0	-212.0	190.6
Related	984	627	357	-7.1	-25.0	3.8	-128.6	157.5

Waterfall plot depicting the percent change in the c-statistic in related [related and closely related] validations (in blue) and distantly validations (in orange). Vertical lines show that the median decrement in discrimination was more pronounced in the distantly related models than the related models.