# Regional Validation and Updating of Clinical Predictive Models for Patients with Acute Heart Failure

A thesis

Submitted by

Benjamin S. Wessler

In partial fulfillment of the requirements

for the degree of

Master of Science

In

Clinical and Translational Science

May 2016

Advisors:

Thesis Chair and Program Mentor: David Kent MD, MS

Project Mentor: James Udelson MD

Statistical Mentor: Robin Ruthazer MPH

Abstract:

It is increasingly common to incorporate risk-based analyses into clinical decisions for patients with cardiovascular disease. While many current cardiovascular guidelines, including those for heart failure, recommend integrating validated clinical predictive models (CPMs), these predictive instruments are rarely used. One major limitation is that heart failure is a heterogeneous phenotype and the generalizability of heart failure CPMs is largely unknown. We hypothesize that regional CPMs for patients with acute heart failure (AHF) do not generalize well to patients in other world regions and that regional model performance can be improved with simple updating procedures. We identified CPMs derived in North America that predict mortality for patients with AHF and validated these models in different world regions to assess performance in a contemporary international clinical trial (N = 4133) of patients with AHF treated with guideline directed medical therapy. We performed independent external validations of 3 compatible CPMs predicting in-hospital mortality, 60 day mortality, and 1 year mortality respectively. CPM performance varied substantially across different world regions (North America, South America, Eastern Europe, and Western Europe). The median percent decrease in discrimination across different world regions was - 40% (range -15% to -92%). Regional calibration was highly variable (Harrell's $E_{avg}$ range <1% to >19%) however simple updating procedures significantly improved local performance (recalibrated $E_{avg}$ < 3% across all regions and all models). Regional AHF CPM discrimination and

calibration vary substantially across different world regions and local performance must be understood and optimized before these tools can be leveraged to improve patient care.

**Acknowledgements**

Educational mentors: David Kent MD, MSc
Jessica Paulus ScD
Robin Ruthazer MPH
Farzad Noubary PhD
Angie Mae Rodday PhD
Robert Goldberg PhD

The Tufts Predictive Analytics and Comparative Effectiveness (PACE) Center

The Tufts Cardiovascular Center

T32 HL069770 [Richard Karas (PI)] Training Grant from the NIH

5 TL1 TR001062 [Harry Selker (PI)] Training Grant from the NIH-NCATS

**Table of Contents** **Page**

**List of Tables** **Page**

**Manuscript:**

## List of Figures

**Manuscript:**

## List of Abbreviations:

| | |
|---|---|
| AHF | acute heart failure |
| AUC | area under the curve |
| AuROC | area under the receiver operating curve |
| BPM | beats per minute |
| COPD | chronic obstructive pulmonary disease |
| CPM | clinical Predictive Model |
| CVD | cardiovascular disease |
| $E_{avg}$ | Harrell's E statistic |
| EE | Eastern Europe |
| EFFECT | Enhanced Feedback for Effective Cardiac Treatment |
| EVEREST | Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study with Tolvaptan |
| GWTG-HF | Get With the Guidelines- Heart Failure |
| HF | heart failure |
| Na | serum sodium |
| NA | North America |
| NYHA | New York Heart Association |
| OPTIME-CHF | Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure |
| PACE | predictive analytics and comparative effectiveness |
| SA | South America |
| SBP | systolic blood pressure |
| WE | Western Europe |

Introduction:

It is increasingly recognized that common clinical phenotypes encompass substantial heterogeneity with respect to patient values and preferences, outcome risks, and harms and benefits of treatment.[1,2] To aid physicians and patients, clinical predictive models (CPMs) are now widely available to estimate the likelihood of important outcomes (prognostic models) or diagnoses (diagnostic models) for individual patients. These tools hold the potential of allowing clinicians to align treatment decisions with patient risks and preferences and are increasingly included as part of cardiovascular disease (CVD) clinical practice guidelines.[3–7]

Despite this recognition and the availability of CPMs, until recently clinicians have largely ignored these tools and implementation lags far behind model development. In our own Tufts PACE CPM registry, we have described approximately 800 unique risk models in the cardiovascular literature[8], yet only a handful are applied routinely in clinical practice to guide important decisions. Potential reasons for this lack of translation into practice include marginal performance with respect to discrimination and calibration, the limited usefulness of the information with respect to clinical decision making, and limited usability at the point of care.[9–12] One major concern is that model performance is often significantly better for the population on which the model was derived (derivation population) compared to a similar yet distinct ‘real world’ populations (external validation population).[13] Efforts are now underway to better understand this

predictive modeling literature and to improve the reporting and overall utility of these tools.[14,15]

In the treatment of patients with heart failure (HF), current guidelines recommend using validated CPMs to individualize treatment decisions.[6] This recommendation emerges from the observation that patients with HF vary widely in their risk of death, as defined on the basis of easily obtainable clinical characterstics.[16] CPMs can help estimate this risk, however acceptable regional performance of HF CPMs and appropriate model updating have both been understudied. These are central questions for the care of patients with HF since significant regional heterogeneity is recognized even within the restricted settings of randomized controlled trials.[17–19] Models may support appropriate decision making in one region, while yielding misleading information in another. Here we use patient level data from The Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study with Tolvaptan (EVEREST)[20] to perform regional independent external validations of previously published CPMs that predict mortality at different time points following hospital admission for acute heart failure (AHF). Our hypothesis is that regional CPMs for AHF derived on data from patients in one world region (here North America) do not generalize well to patients in different world regions and that simple updating procedures can improve regional performance.

## Methods:

Model Selection: Identifying compatible CPMs is a process that involves evaluation of both the original CPM and the validation cohorts **(Table 1)**.

| Variable | GWTG-HF* | OPTIME-CHF | EFFECT* | EVEREST | NA EVEREST | SA EVEREST | EE EVEREST | WE EVEREST |
|---|---|---|---|---|---|---|---|---|
| Years | 2005-2007 | 1997-1999 | 1999-2001 | 2003-2006 | 2003-2006 | 2003-2006 | 2003-2006 | 2003-2006 |
| Data Source | Registry | Clinical Trial | Clinical Trial | Clinical Trial | Clinical Trial | Clinical Trial | Clinical Trial | Clinical Trial |
| N | 27850 | 949 | 2624 | 4133 | 957 | 586 | 1552 | 477 |
| Age | 72.5 | 68 | 76.3 | 67.0 (58.0-75.0) | 70.0 (60.0-78.0) | 63.0 (56.0-71.0) | 66.0 (58.0-73.0) | 70.0 (61.3-77.0) |
| SBP | 137 | 120 | 148 | 120.0 (105.0-131.0) | 112.0 (101.0-128.0) | 112.5 (100.0-117.1) | 122.0 (110.0-140.0) | 112.0 )100.0-130.0) |
| Na | 138 | 139 | 138 | 140.0 (137.0-142.0) | 139.0 (136.0-142.0) | 140.0 (137.0-142.0) | 140.0 (138.0-143.0) | 139.0 (137.0-142.0) |
| BUN (mg/dL) | 25 | 13 | 29.4 | 26.0 (20.0-35.0) | 30.0 (22.0-45.0) | 25.00 (19.0-32.0) | 23.0 (18.0-30.0) | 31.0 (22.0-45.0) |
| Heart Rate (BPM) | 82 | 84 | 94 | 78.0 (69.0-90.0) | 76.0 (68.0-86.0) | 78.0 (69.5- 90.0) | 80.0 (70.0-90.0) | 76.0 (68.0-88.0) |
| Respiratory Rate | NR | NR | 26 | 20.0 (18.0-22.0) | 20.0 (18.0-22.0) | 20.0 (18.75-22.0) | 20.0 (18.0-24.0) | 20.0 (18.0-23.0) |
| Prior CVA (%) | 14 | NR | 17 | 17 | 28 | 13 | 16 | 15 |
| COPD (%) | 28 | 23 | 21 | 10 | 18 | 6 | 5 | 9 |
| Black Race (%) | 18 | 33 | NR | 4 | 17 | 10 | 0 | 0 |
| Hemoglobin | 12.0 | NR | 12.4 | 13.2 (11.8-14.5) | 12.5 (11.2-13.9) | 13.5 (12.1-14.7) | 13.7 (12.5-14.9) | 13.0 (11.4-14.2) |
| NYHA class IV (%) | NR | 47 | NR | 42 | 44 | 46 | 43 | 34 |
| Dementia (%) | NR | # | 9 | # | # | # | # | # |
| Cancer (%) | NR | # | 9 | # | # | # | # | # |
| Liver Disease (%) | NR | # | 1 | # | # | # | # | # |

**Table 1**. **Baseline characteristics for patients the various databases**. CPM derivation populations are presented on the left (bold border). Validation datasets (overall and regional) are shown on the right. Blue shading indicates variables that are included in the CPM derived from each database. GWTG-HF is Get With The Guidelines-Heart Failure, OPTIME-HF is The Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure study, EFFECT is Enhanced Feedback for Effective Cardiac Treatment study. NA is North American, SA is South America, EE is Eastern Europe, WE is Western Europe. * indicates AHF populations that include patients with both reduced and preserved ejection fractions. # indicates variables that were exclusion criteria for a given database (these variables were coded as 0. NR indicates not reported. For the validation populations values are presented as median (IQR).

For this analysis compatible CPMs were defined by the following characteristics: 1) the index condition in the derivation cohort was similar to the index condition in the validation cohort (here AHF), 2) CPM predicts mortality, 3) all variables in the CPM were captured in the validation datasets (see below) and can be assigned a value, and 4) CPMs were derived in patient samples from a single world region (here North America). We identified compatible models by performing a query of our Tufts PACE CPM Registry. We complemented this search by reviewing a recently published systematic review of CPMs for HF.[21] For this analysis we present a sample of the CPMs developed in North America that predict mortality at three different time points following hospitalization for HF.

Selected Models: Selected models are show in **Table 1** and **S1**. Selected models were: The American Heart Association Get With the Guidelines- Heart Failure (GWTG-HF)[28] model (7 variables, predicts in-hospital mortality), The Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure (OPTIME-CHF)[29] (5 variables, predicts 60 day mortality), and The Enhanced Feedback for Effective Cardiac Treatment (EFFECT)[30] model (10 variables, predicts 1 year mortality). Age, Systolic Blood Pressure, Serum Sodium, Serum Blood Urea Nitrogen are common variables in all three models.

The GWTG-HF program collected patient level data from patients hospitalized for HF at 287 hospitals in the United States between January 2005 and June 2007.[28] These data were used to build and validate a model predicting in-hospital mortality following admission for HF that was presented as a point score and online calculator in 2010. The model was built using logistic regression analysis from a final cohort of 27,850 patients (derivation cohort) and validated on 11,933 patients (validation cohort) from this program.

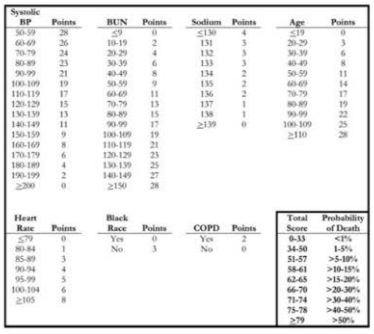
The OPTIME-CHF study was a randomized clinical trial of 949 patients with heart failure with reduced ejection fraction hospitalized for worsening symptoms.[31] Patients were randomized to receive intravenous milrinone or placebo for 48-72 hours. Patients were enrolled from 78 centers across the Unites States from 1997 to 1999. A CPM predicting 60 day mortality was derived from this dataset using Cox proportional hazards analysis and presented as a point score in 2004.[29]

The EFFECT study group presented a CPM derived from 2624 patients hospitalized in Ontario, Canada, from April 1999 to March 2001 for HF. Data for this model came from the Canadian Institutes of Health Information hospital discharge abstract and patients were included only if they met a pre-specified definition of clinical heart failure. This CPM was created using logistic regression analysis and a validation was performed on 1407 patients from different hospitals in Ontario from a previous time period (1997 to 1999).

Validation Cohort: The EVEREST trial has been previously reported.[22] This was a prospective, international, multicenter, randomized, placebo controlled study conducted in 359 sites worldwide from 2003 and 2006. The trial included 1,251 patients from North America, 699 patients from South America, 564 patients from Western Europe, and 1619 patients from Eastern Europe. This study evaluated the addition of tolvaptan to standard medical therapy for AHF and enrolled patients within 48 hours of HF hospitalization. Inclusion criteria were: 18 years or age or older, reduced left ventricular ejection fraction ($\leq$ 40%), clinical volume overload, New York Heart Association (NYHA) class III/ IV symptoms and hospitalized for ADHF. Exclusion criteria include: Cardiac surgery within 2 months of enrollment, intra-aortic balloon pump (AIBP) or other mechanical support, biventricular pacemaker placed within 2 months, other comorbidities with life expectancy less than 6 months, acute myocardial infarction at the time of hospitalization, significant uncorrected valvular disease, end stage heart failure, need for dialysis, supine systolic blood pressure < 90, creatinine > 3.5 mg/ dL, potassium > 5.5 mEq/L, hemoglobin < 9 g/dL. Patients were followed for the dual

primary end points of all-cause mortality and cardiovascular death or hospitalization for heart failure. During a median follow up of 9.9 months, 537 (26%) of the patients died. Tolvaptan had no effect on long-term mortality for these patients (hazard ratio 0.98; 95% confidence interval [CI], 0.87-1.11; P=0.68). The patients enrolled in this trial were treated with guideline directed medical therapies for HF including ACE inhibitors (84%), beta-blockers (70%), aldosterone blockers (54%), and diuretics (97%) and thus this trial provides an opportunity to evaluate the regional performance of previously published CPMs on an international population of patients with AHF treated with contemporary evidence based therapies.

Statistical Analysis and Model Updating: For each CPM we calculated a point score for each patient based on their covariate values. These point scores were then converted into predicted event probabilities as described by the original CPM authors **(Figure S1)**.



| Systolic BP | Points | BUN | Points | Sodium | Points | Age | Points |
|---|---|---|---|---|---|---|---|
| 50-59 | 28 | ≤9 | 0 | ≤130 | 4 | ≤19 | 0 |
| 60-69 | 26 | 10-19 | 2 | 131 | 3 | 20-29 | 3 |
| 70-79 | 24 | 20-29 | 4 | 132 | 3 | 30-39 | 6 |
| 80-89 | 23 | 30-39 | 6 | 133 | 3 | 40-49 | 8 |
| 90-99 | 21 | 40-49 | 8 | 134 | 2 | 50-59 | 11 |
| 100-109 | 19 | 50-59 | 9 | 135 | 2 | 60-69 | 14 |
| 110-119 | 17 | 60-69 | 11 | 136 | 2 | 70-79 | 17 |
| 120-129 | 15 | 70-79 | 13 | 137 | 1 | 80-89 | 19 |
| 130-139 | 13 | 80-89 | 15 | 138 | 1 | 90-99 | 22 |
| 140-149 | 11 | 90-99 | 17 | ≥139 | 0 | 100-109 | 25 |
| 150-159 | 9 | 100-109 | 19 | | | ≥110 | 28 |
| 160-169 | 8 | 110-119 | 21 | | | | |
| 170-179 | 6 | 120-129 | 23 | | | | |
| 180-189 | 4 | 130-139 | 25 | | | | |
| 190-199 | 2 | 140-149 | 27 | | | | |
| ≥200 | 0 | ≥150 | 28 | | | | |

| Heart Rate | Points | Black Race | Points | COPD | Points |
|---|---|---|---|---|---|
| ≤79 | 0 | Yes | 0 | Yes | 2 |
| 80-84 | 1 | No | 3 | No | 0 |
| 85-89 | 3 | | | | |
| 90-94 | 4 | | | | |
| 95-99 | 5 | | | | |
| 100-104 | 6 | | | | |
| ≥105 | 8 | | | | |

| Total Score | Probability of Death |
|---|---|
| 0-33 | <3% |
| 34-50 | 1-5% |
| 51-57 | >5-10% |
| 58-61 | >10-15% |
| 62-65 | >15-20% |
| 66-70 | >20-30% |
| 71-74 | >30-40% |
| 75-78 | >40-50% |
| ≥79 | >50% |

GWTG-HF from Peterson et al.

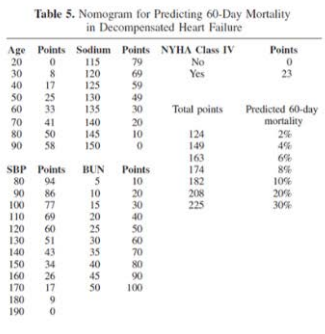


**Table 5.** Nomogram for Predicting 60-Day Mortality in Decompensated Heart Failure

| Age | Points | Sodium | Points | NYHA Class IV | Points |
|---|---|---|---|---|---|
| 20 | 0 | 115 | 79 | No | 0 |
| 30 | 8 | 120 | 69 | Yes | 23 |
| 40 | 17 | 125 | 59 | | |
| 50 | 25 | 130 | 49 | | |
| 60 | 33 | 135 | 30 | Total points | Predicted 60-day mortality |
| 70 | 41 | 140 | 20 | | |
| 80 | 50 | 145 | 10 | 124 | 2% |
| 90 | 58 | 150 | 0 | 149 | 4% |
| | | | | 163 | 6% |
| SBP | Points | BUN | Points | 174 | 8% |
| 80 | 94 | 5 | 10 | 182 | 10% |
| 90 | 86 | 10 | 20 | 208 | 20% |
| 100 | 77 | 15 | 30 | 225 | 30% |
| 110 | 69 | 20 | 40 | | |
| 120 | 60 | 25 | 50 | | |
| 130 | 51 | 30 | 60 | | |
| 140 | 43 | 35 | 70 | | |
| 150 | 34 | 40 | 80 | | |
| 160 | 26 | 45 | 90 | | |
| 170 | 17 | 50 | 100 | | |
| 180 | 9 | | | | |
| 190 | 0 | | | | |

OPTIME-CHF from Felker et al.

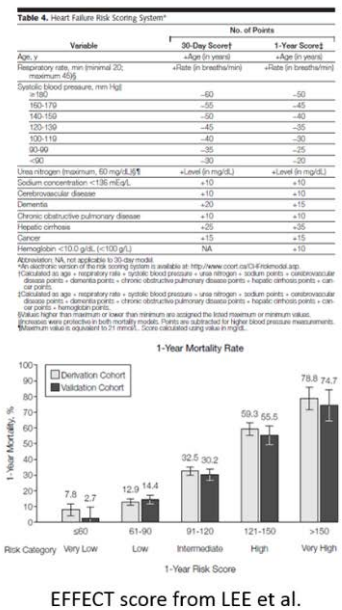


**Table 4.** Heart Failure Risk Scoring System*

| Variable | No. of Points: 30-Day Score† | No. of Points: 1-Year Score‡ |
|---|---|---|
| Age, y | +Age (in years) | +Age (in years) |
| Respiratory rate, min (minimal 20; maximum 45)§ | +Rate (in breaths/min) | +Rate (in breaths/min) |
| Systolic blood pressure, mm Hg§ | | |
| ≥180 | –60 | –50 |
| 160-179 | –55 | –45 |
| 140-159 | –50 | –40 |
| 120-139 | –45 | –35 |
| 100-119 | –40 | –30 |
| 90-99 | –35 | –25 |
| <90 | –30 | –20 |
| Urea nitrogen (maximum, 60 mg/dL)§¶ | +Level (in mg/dL) | +Level (in mg/dL) |
| Sodium concentration <136 mEq/L | +10 | +10 |
| Cerebrovascular disease | +10 | +10 |
| Dementia | +20 | +15 |
| Chronic obstructive pulmonary disease | +10 | +10 |
| Hepatic cirrhosis | +25 | +35 |
| Cancer | +15 | +15 |
| Hemoglobin <10.0 g/dL (<100 g/L) | NA | +10 |

Abbreviation: NA, not applicable to 30-day model.
*An electronic version of the risk scoring system is available at: http://www.coort.ca/O4Riskmodel.asp.
†Calculated as: age + respiratory rate + systolic blood pressure + urea nitrogen + sodium points + cerebrovascular disease points + dementia points + chronic obstructive pulmonary disease points + hepatic cirrhosis points + cancer points.
‡Calculated as: age + respiratory rate + systolic blood pressure + urea nitrogen + sodium points + cerebrovascular disease points + dementia points + chronic obstructive pulmonary disease points + hepatic cirrhosis points + cancer points + hemoglobin points.
§Values higher than maximum or lower than minimum are assigned the listed maximum or minimum values. Increases were protective in both mortality models. Points are subtracted for higher blood pressure measurements. ¶Maximum value is equivalent to 21 mmol/L. Score calculated using value in mg/dL.




EFFECT score from LEE et al.

When a range of probabilities was given the midpoint probability was assigned for a given point score range. Event probabilities were converted to a linear predictor using the equation [predicted value = (1 / (1+ $e^{-xbeta}$)] where xbeta is the linear predictor for a given probability. This transformation was used for the various performance measures (described below). Percent decrement in discrimination was calculated as [Derivation AUC-0.5]-[Regional AUC-0.5]/[Derivation AUC-0.5] * 100. All analyses were run in R Studio Version 0.99.489.

Measuring CPM Performance: *Calibration-in-the-large* is a measure of global fit that describes whether the mean observed death rate agrees with the mean predicted rate of death. *Model discrimination* refers to the ability of each model to separate those who experience the outcome of death from those who do not. This measure is often represented by the Area under the Receiver Operating Curve (AuROC). An AuROC curve of 0.5 indicates that the model does no better than a 'coin flip' in discriminating between those who ultimately die from those who do not. An AuROC of 1.0 indicated that the model perfectly discriminates between these groups. CPM performance is generally between these two bounds and values of 0.7-0.8 are sometimes described as representing benchmarks for good discrimination. In this analysis we assess percent decrement in discrimination which is derived (as shown above) from the AuROC for each region. *Model calibration* is a separate measure of performance and

reflects whether predicted outcome rates match observed outcome rates across a particular subgrouping scheme (e.g. when the sample is divided into quantiles of risk). This measure of performance is often evaluated by a Hosmer-Lemenshow statistic, which uses a chi square statistic to determine whether the predicted and observed outcome rates are significantly different across deciles. There are, however, important limitations to this commonly used statistic.[25] Instead we focus our analysis on a newer measure of calibration, Harrell's E statistic.[26] This measure of error computes the average absolute calibration error (average absolute difference between the lowess-estimated calibration curve and the line of identity, labeled $E_{avg}$.) We also evaluate $E_{max}$, a more conservative measure of calibration that describes that maximum absolute difference between predicted and calibrated probabilities.

While many updating techniques are available (**Table SA**, adapted from Steyerberg[23]), simple approaches are preferred for ease of use. The most direct

| Technique | Label | Notation |
|---|---|---|
| **1 (no updating)** | Apply original model | $\alpha_{orig} + \beta_{orig} * x_{1...8}$ |
| **2 (recalibration)** | Update intercept | $\alpha$ |
| **3 (recalibration)** | Re-calibration of intercept and slope | $\alpha + \beta_{overall}$ |
| 4 (recalibration) | Re-calibration + selective re-estimation | $\alpha + \beta_{overall} + \gamma_{1...8}$ |
| 5 (recalibration) | Re-estimation | $\alpha + \beta_{1...8}$ |
| 6 (model extension) | Re-calibration + selective re-estimation + selective extension | $\alpha + \beta_{overall} + \gamma_{1...8} + \beta_{9...16}$ |
| 7 (model extension) | Re-estimation + selective extension | $\alpha + \beta_{1...8} + \beta_{9...16}$ |
| 8 (model extension) | Re-estimation + extension | $\alpha + \beta_{1...16}$ |

**Table SA. Model updating techniques.** As described by Steyerberg et al. The simplest form of updating (technique 2) addresses calibration-in-the-large. This method considers the mean observed outcome rate in the derivation and validation cohorts and applies the difference between these rates to update the intercept ($\alpha$) of the CPM. Technique 3 takes the form of recalibrating both the intercept and applying a uniform correction factor to the regression coefficients of the independent variables to better fit the population ($\alpha$, as addressed in technique 2) and the regression coefficients, by multiplication by the calibration slope ($\beta_{overall}$). In technique 4 updating starts with updating $\alpha + \beta_{overall}$ as in technique 3 and then tests whether the effects of the included predictors are similar to the derivation cohort and also across regions. $\gamma$ indicates deviation from the recalibrated coefficients with additional adjustment of the $\beta$ only if there is significant deviation. In technique 5 the original model predictor variables would be fit to the validation cohorts with creation of new $\alpha + (\beta_{1...8})$. Techniques 6 through 8 depart substantially from the original model reports and more closely overlap with *de novo* model development, and are less useful as tools for understanding how currently available CPMs perform.

application of a published prediction model is to use the model as it was originally described (technique 1).

The simplest form of updating (technique 2) addresses calibration-in-the-large and is often considered of primary importance. This method considers the mean observed outcome rate in the derivation and validation cohorts and applies the difference between these rates to update the intercept (α) of the CPM. This approach is expected to work reasonably well if the variation in outcome rates is not directly attributable to changes in the effects of clinical variables contained in the model.[24]

The next form of updating (technique 3) takes the form of recalibrating the intercept and applying a uniform correction factor to the regression coefficients of the independent variables to better fit the validation population. This form of updating maintains the originally developed model while updating both the model intercept (α, as addressed in technique 2), and the remaining regression coefficients by multiplying each by a constant factor called the calibration slope ($\beta_{overall}$). The original model is maintained in this form of updating and is adjusted for overall fit. This updating technique corrects both for differences in prevalence unrelated to covariate effects (as in technique 2) and also can correct for overfitting in the original derivation dataset.

Data Adequacy:

The overall mortality rates and number of events for patients in the EVEREST trial, stratified by region, are shown in **Figure 1**.

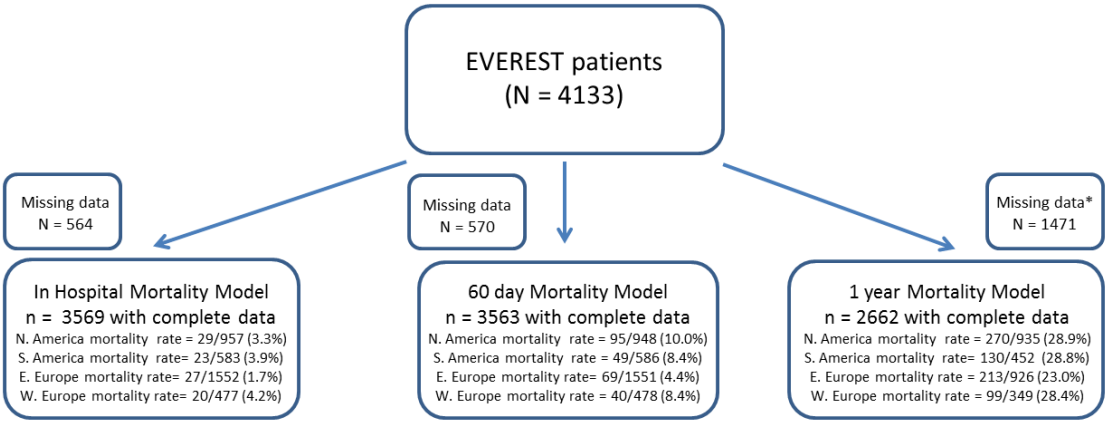



**Figure 1**. **Description of Independent External Validation Cohorts**. Validation exercises were done for patients with all variables available. * indicates that for the 1 year mortality model we considered patients to have missing data if they were last known alive with < 9 months of follow up.

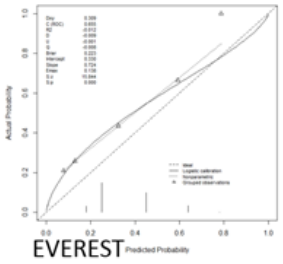
Event rates for in-hospital mortality are low and may negatively impact measures of model performance. Additionally there are certain validation datasets in specific regions with modest size (approximately 400 patients) and also low event rates (~2.5% for in-hospital mortality). These characteristics may adversely affect our ability to measure CPM performance.[27] Power for model re-calibration relates primarily to the ability to extend models beyond the published predictors or re-estimate individual coefficients (update techniques 4-8 in **Table SA**). We have decided *a priori* not to extend the predictive models by adding additional variables.

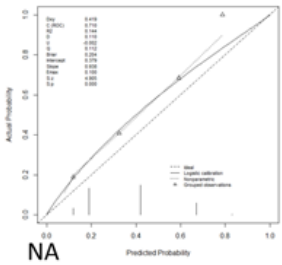
In order to maximize the number of patients available for the validations we relied on an assumption for the long-term mortality model. For the 1 year mortality outcome we included patients who had died (prior to or at 1 year) or were alive with at least 9 months of follow up. Patients without a death in the first year but who were confirmed alive at 9 months (or beyond) were considered alive at 12 months for the logistic regression models of 1 year mortality. We performed

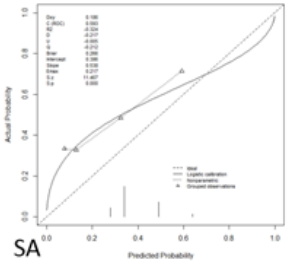
sensitivity analyses (**S2a-c)** to determine how this assumption affects model discrimination and calibration.

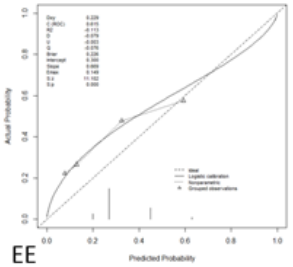
**Supplement S2a: Sensitivity Analysis of EFFECT CPM (Including only patients dead or alive with ≥ 12 months of follow up**

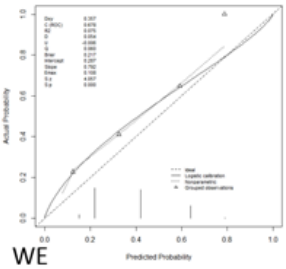
| Region | AUC |
|---|---|
| EVEREST | 0.66 |
| North America | 0.71 |
| South America | 0.59 |
| Eastern Europe | 0.62 |
| Western Europe | 0.68 |


EVEREST


NA

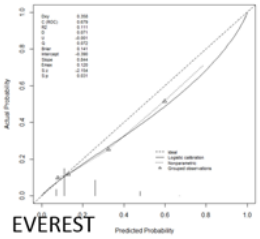

SA

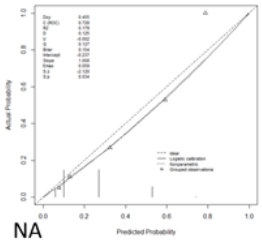

EE

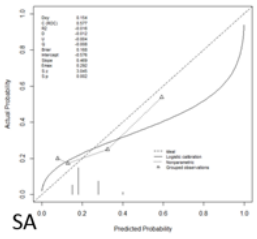

WE

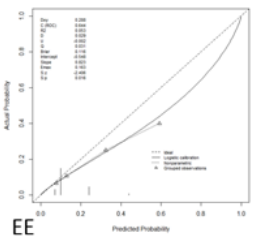
**Supplement S2b: Sensitivity Analysis of EFFECT CPM (Including only patients dead or alive with ≥ 6 months of follow up**

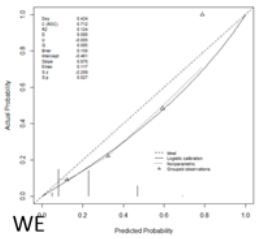
| Region | AUC |
|---|---|
| EVEREST | 0.68 |
| North America | 0.73 |
| South America | 0.58 |
| Eastern Europe | 0.64 |
| Western Europe | 0.71 |


EVEREST


NA

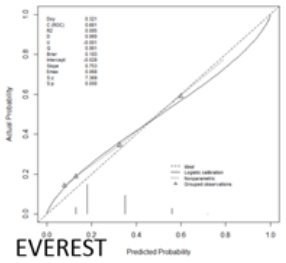

SA

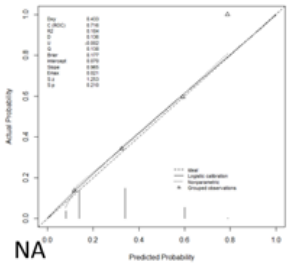

EE

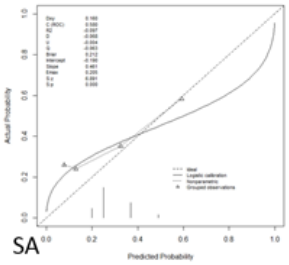

WE

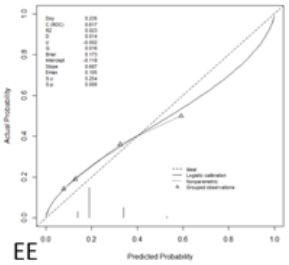
**Supplement S2c: Sensitivity Analysis of EFFECT CPM (Including only patients dead or alive with ≥ 9 months of follow up**

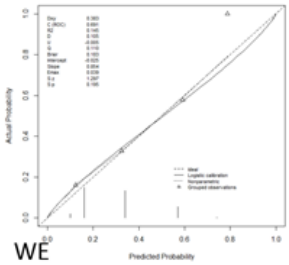
| Region | AUC |
|---|---|
| EVEREST | 0.66 |
| North America | 0.72 |
| South America | 0.58 |
| Eastern Europe | 0.62 |
| Western Europe | 0.69 |




Results:

Database Comparisons:

Overall the patients in the derivation cohorts were similar to but distinct from the patients in the validation cohorts (EVEREST database overall and region specific); **Table 1**. The distribution of covariates is shown for each world region within the validation databases **(Table 1)**. Median age ranged from 63 years in South America to 70 years in North America and Western Europe databases. Median systolic blood pressure (SBP) ranged from 122.0 in Eastern Europe to 112.0 in the other world regions. The percentage of Black patients ranged from 0% in Eastern Europe to 16.9% in North America. Two CPMs (GWTG-HF and EFFECT) were derived from patients with heart failure with reduced ejection fraction (HFrEF) and also heart failure with preserved ejection fraction (HFpEF).

GWTG-HF CPM was derived from registry data. The OPTIME-HF CPM was derived from data collected between 5 and 7 years before the EVEREST study was conducted. Exclusion criteria for these databases are shown in **Table S4**. The randomized controlled trials had more exclusion criteria than the registry database.

Supplement S4: Database Exclusion Criteria

| Database | Exclusion Criteria |
|---|---|
| OPTIME CHF | 1. Patient requires IV vasopressor or inotropic support. 2. Patient requires admission primarily for concurrent morbidity. Left ventricular failure primarily from uncorrected obstructive valvular disease, hypertrophic obstructive cardiomyopathy, uncorrected thyroid disease, known acute myocarditis, known amyloid cardiomyopathy, or known malfunctioning artificial heart valve. 4. Patient is scheduled for heart surgery. 5. There is evidence of unstable angina, active myocardial ischemia, or myocardial infarction within 3 months. 6. Patient has atrial fibrillation with a sustained ventricular response rate >110 beats/min. 7. Patient has sustained ventricular tachycardia or fibrillation. 8. Patient has systolic blood pressure <80 or >150 mm Hg. 9. Patient has severe renal impairment with a creatinine level >3.0 mg/dL or requires dialysis. 10. Patient has suspected digitalis intoxication. 11. Patient has known hypersensitivity to milrinone. |
| EFFECT | Patients who developed heart failure after admission (ie, in-hospital complication), patients transferred from another acute care facility, those aged 105 years or older, nonresidents, and those with an invalid health card |
| GWTG-HF | Patients were excluded from analysis if they did not have a diagnosis of HF, if they were transferred to a different acute care facility, if the discharge date was invalid, or if data were missing for their discharge status , or left ventricular ejection fraction (LVEF). |
| EVEREST | Cardiac surgery within 60 d of potential study enrollment, excluding percutaneous coronary interventions. Planned revascularization procedures, electrophysiologic device implantation, cardiac mechanical support implantation, cardiac transplantation, or other cardiac surgery within 30 days after study enrollment. Subjects who are on cardiac mechanical support. History of biventricular pacer placement within the last 60 d. Comorbid condition with an expected survival < 6 mo. Subjects with acute ST segment elevation myocardial infarction at the time of hospitalization. History of sustained ventricular tachycardia or ventricular fibrillation within 30 days, unless in the presence of an automatic implantable cardioverter defibrillator. History of a cerebrovascular accident within the last 30 d. Hemodynamically significant uncorrected primary cardiac valvular disease. Hypertrophic cardiomyopathy (obstructive or nonobstructive) Congestive heart failure from uncorrected thyroid disease, active myocarditis, or known amyloid cardiomyopathy. Subjects with refractory, end-stage, heart failure defined as subjects who are appropriate candidates for specialized treatment strategies, such as ventricular assist devices, continuous positive intravenous inotropic therapy, or hospice care Progressive or episodic neurologic disease such as multiple sclerosis or history of multiple Strokes. History of primary significant liver disease or acute hepatic failure. Chronic uncontrolled diabetes mellitus as determined by the investigator. Subjects currently treated with hemofiltration or dialysis. Morbid obesity, defined as 159 kg (or 350 lb) or body mass index 42. Supine systolic arterial blood pressure 90 mm Hg. Serum creatinine 3.5 mg/dL or 309.4 µmol/L. Serum potassium 5.5 mEq/L or 5.5 mmol/L. Hemoglobin 9 g/dL or 90 g/L or 5.586 mmol/L. History of hypersensitivity or idiosyncratic reaction to benzazepine derivatives (such as benazepril). Women who will not adhere to the reproductive precautions as outlined in the informed consent form. Positive urine pregnancy test. Inability to provide written informed consent. History of drug or medication abuse within the past year, or current alcohol abuse. Previous participation in this or any other tolvaptan clinical trial. Inability to take oral medications. Participation in another clinical drug or device trial in which the last dose of drug was within the past 30 d or an investigation medical device is implanted |

Table S4: Exclusion criteria as written in the original reports (Peterson et al., Lee et al.) or in the Design and Rationale reports (Cuffe et al. and Gheorghiade et al.)

Independent External Validations:

We assessed calibration-in-the large for each mortality time point (in-hospital mortality, 60 day mortality, and 1 year mortality) for the derivation databases and the validation databases (**Table 2)**.

| In-Hospital Mortality | GWTG-HF | EVEREST | N. America | S. America | E. Europe | W. Europe |
|---|---|---|---|---|---|---|
| Observed Event rate | 0.029 | 0.027 | 0.030 | 0.039 | 0.017 | 0.042 |
| Average Pred. Rate | | 0.022 (0.016) | 0.027 (0.021) | 0.020 (0.014) | 0.017 (0.012) | 0.025 (0.018) |
| Diff. (Obs.- Pred.) | | 0.005 | 0.003 | 0.019 | 0 | 0.017 |

| 60 day Mortality | OPTIME-CHF | EVEREST | N. America | S. America | E. Europe | W. Europe |
|---|---|---|---|---|---|---|
| Observed Event rate | 0.096 | 0.071 | 0.100 | 0.084 | 0.045 | 0.084 |
| Average Pred. Rate | | 0.198 (0.223) | 0.292 (0.258) | 0.172 (0.192) | 0.128 (0.166) | 0.271 (0.25) |
| Diff. (Obs.- Pred.) | | -0.127 | -0.192 | -0.088 | -0.083 | -0.187 |

| 1 year Mortality | EFFECT | EVEREST | N. America | S. America | E. Europe | W. Europe |
|---|---|---|---|---|---|---|
| Observed Event rate | 0.329 | 0.267 | 0.289 | 0.288 | 0.230 | 0.283 |
| Average Pred. rate | | 0.227 (0.152) | 0.271 (0.169) | 0.197 (0.131) | 0.180 (0.115) | 0.274 (0.170) |
| Diff. (Obs.- Pred.) | | 0.040 | 0.018 | 0.091 | 0.050 | 0.009 |

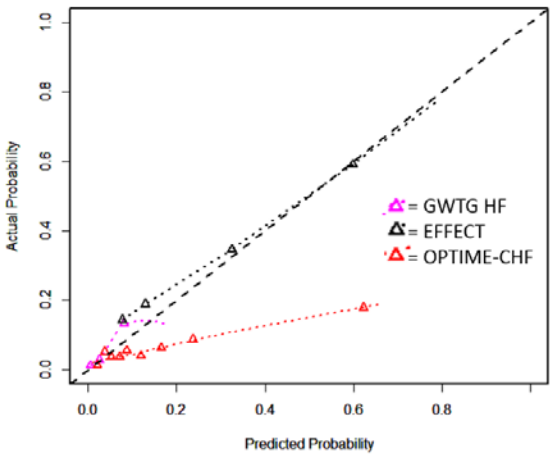
**Table 2. Calibration-in-the large**. Observed average event rates in the derivation datasets and validation datasets. Average Pred. Rate is the mean predicted outcome rates in the validation datasets (standard deviation). Diff. (Obs.- Pred.) is the difference between the Observed event rate and the average predicted event rate. N. America is North American patients in EVEREST, S. American is South American patients in EVEREST, E. Europe is Eastern European patients in EVEREST, W. Europe is Western European patients in EVEREST.

In-hospital mortality (predicted by the GWTG-HF CPM) occurred with an average rate of 2.7% across all regions (observed rates ranged 1.7% in Eastern Europe to 4.2% in Western Europe). Predicted rates for the regional populations ranged from 1.7% in Eastern Europe to 2.7% in North America. There was no difference in the observed and predicted in-hospital event rates for Eastern Europe while the largest difference between the observed and predicted event rates was seen in South America (observed mortality rate of 3.9%, predicted mortality rate of 2.0%, difference of 1.9%). 60 day mortality (predicted by the OPTIME-CHF CPM) occurred at a rate of 7.1% across all world regions. The observed event rates ranged from 4.5% in Eastern Europe to 10.0% in North America. Predicted average event rates ranged from 12.8% in Eastern Europe to 27% in Western Europe. There were large differences between the observed and average predicted 60 day mortality rates across all world regions (differences ranged from 8.3% in Eastern Europe to 19.2% in North America). 1 year mortality

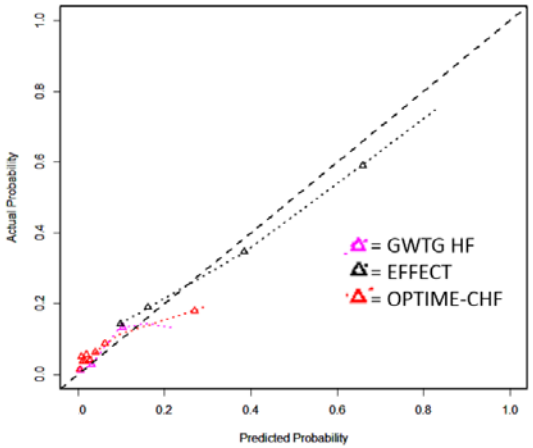
(predicted by the EFFECT model) had an observed event rate of 32.9% across all regions. The regional observed rates ranged from 23.0% in Eastern Europe to 28.9% in North America. Predicted average rates ranged from 18.0% in Eastern Europe to 27.4% in Western Europe. There were many regional differences between the observed and average predicted 1 year mortality rates and differences ranged from 9.1% in South America to < 1% in Western Europe.

Another dimension of CPM performance that we assessed was model calibration across ranges of predicted risk for different world regions. Regional calibration plots (with updating) are shown in the **Figures S5 to S7.**

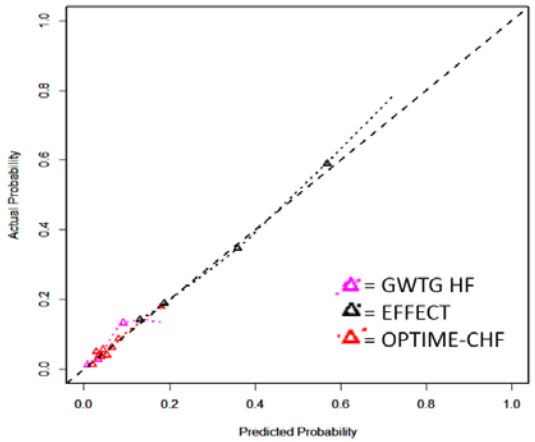
**Worldwide: No updating**




**Worldwide: Updated Intercept**

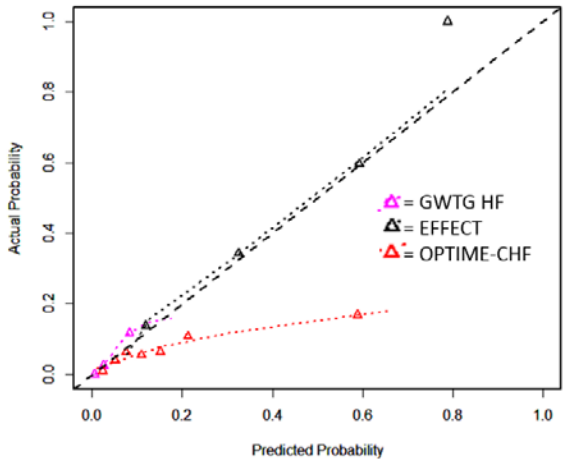



**Worldwide: Updated Slope and Intercept**

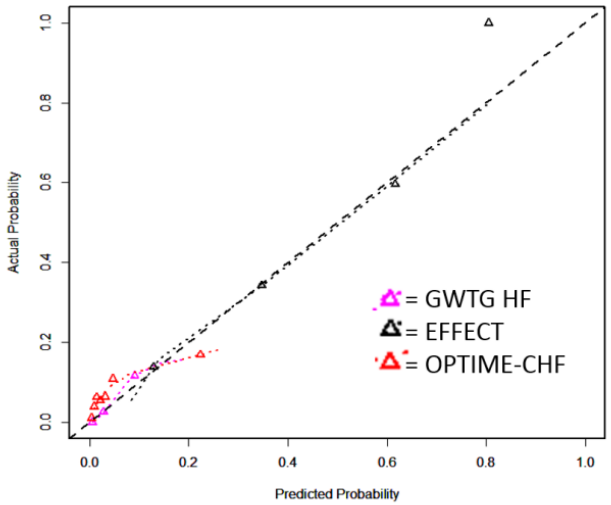



**S5a, S6a, S7a:** Calibration Plots for EVEREST (worldwide) (from left to right: No update (original model applied), Updated intercept, Updated Intercept and Slope). X-axis is predicted probabilities, Y-axis is observed probabilities. GWTG HF is Get With The Guidelines Heart Failure Model, EFFECT is The Enhanced Feedback for Effective Cardiac Treatment Model, OPTIME-CHF is the Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure.

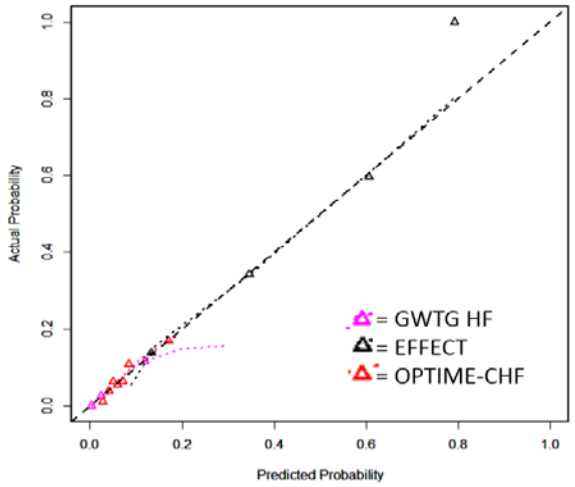
**North America: No updating**




**North America: Updated Intercept**

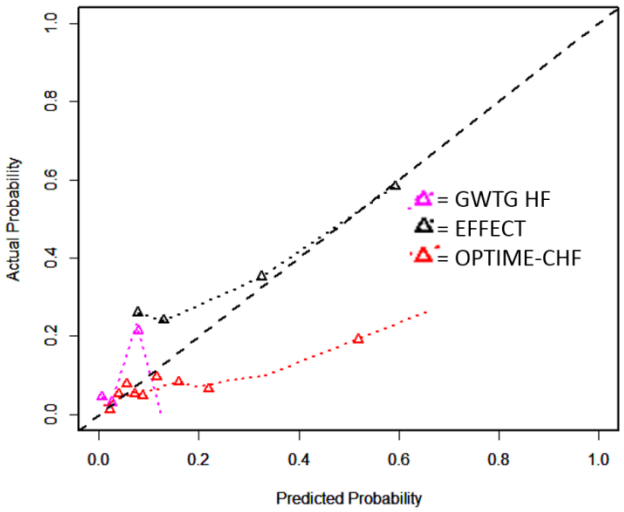



**North America: Updated Slope and Intercept**

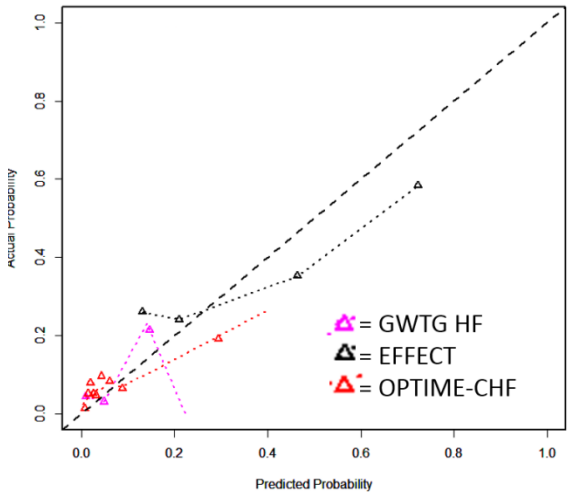



**S5b, S6b, S7b:** Calibration Plots for EVEREST North America (from left to right: No update (original model applied), Updated intercept, Updated Intercept and Slope). X-axis is predicted probabilities, Y-axis is observed probabilities. GWTG HF is Get With The Guidelines Heart Failure Model, EFFECT is The Enhanced Feedback for Effective Cardiac Treatment Model, OPTIME-CHF is the Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure.

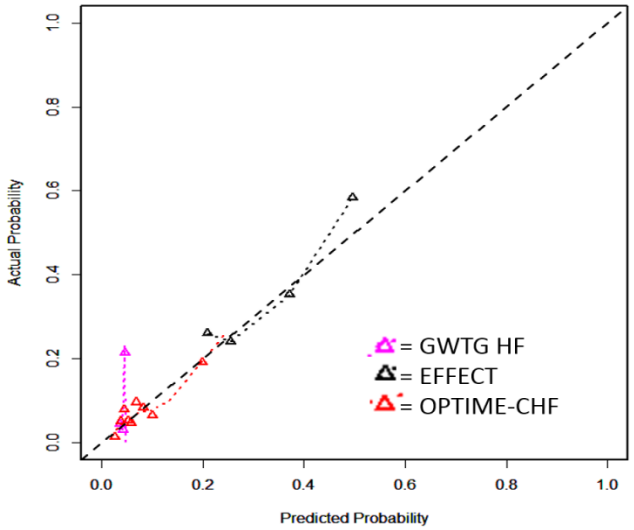
**South America: No updating**




**South America: Updated Intercept**

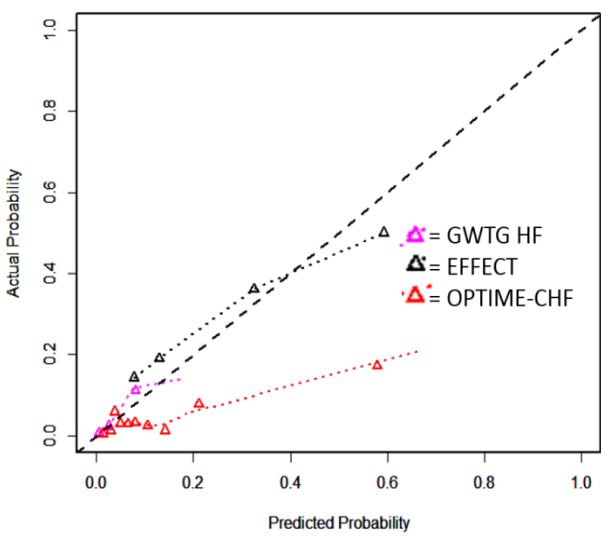



**South America: Updated Slope and Intercept**

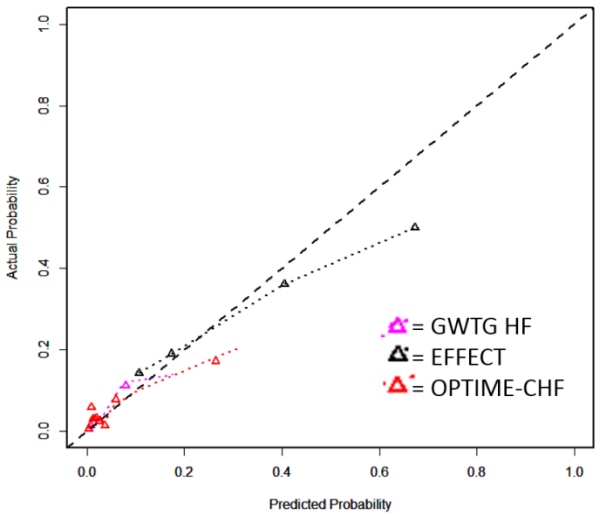



**S5c, S6c, S7c:** Calibration Plots for EVEREST South America (from left to right: No update (original model applied), Updated intercept, Updated Intercept and Slope). X-axis is predicted probabilities, Y-axis is observed probabilities. GWTG HF is Get With The Guidelines Heart Failure Model, EFFECT is The Enhanced Feedback for Effective Cardiac Treatment Model, OPTIME-CHF is the Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure.

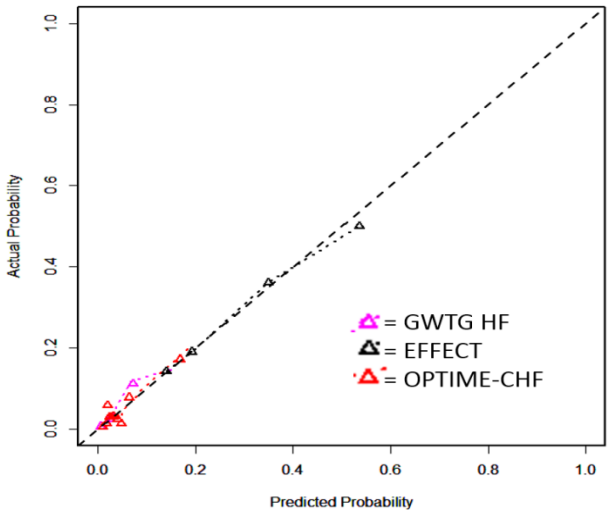
**Eastern Europe: No updating**




**Eastern Europe: Updated Intercept**

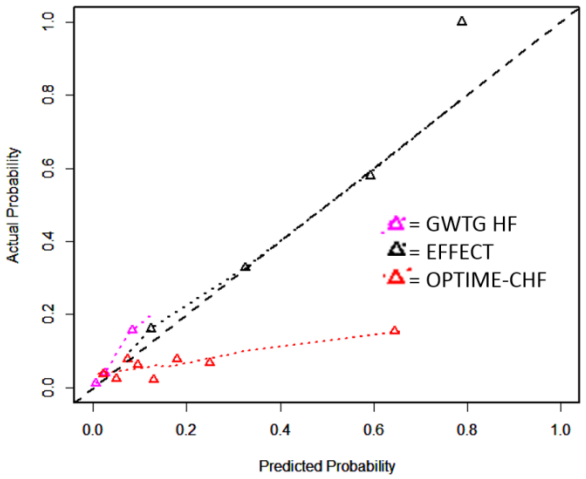



**Eastern Europe: Updated Slope and Intercept**

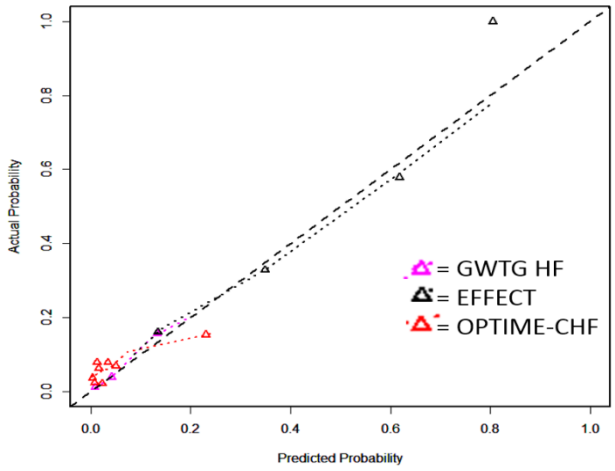



**S5d, S6d, S7d:** Calibration Plots for EVEREST Eastern Europe (from left to right: No update (original model applied), Updated intercept, Updated Intercept and Slope). X-axis is predicted probabilities, Y-axis is observed probabilities. GWTG HF is Get With The Guidelines Heart Failure Model, EFFECT is The Enhanced Feedback for Effective Cardiac Treatment Model, OPTIME-CHF is the Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure.

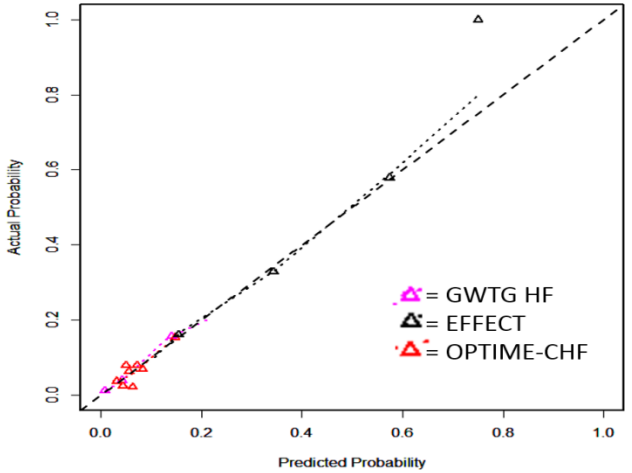
**Western Europe: No updating**




**Western Europe: Updated Intercept**




**Western Europe: Updated Slope and Intercept** e




**S5e, S6e, S7e:** Calibration Plots for EVEREST Western Europe (from left to right: No update (original model applied), Updated intercept, Updated Intercept and Slope). X-axis is predicted probabilities, Y-axis is observed probabilities. GWTG HF is Get With The Guidelines Heart Failure Model, EFFECT is The Enhanced Feedback for Effective Cardiac Treatment Model, OPTIME-CHF is the Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure.

| Model | Recalibration method | Eavg Worldwide | Eavg North America | Eavg South America | Eavg Eastern Europe | Eavg Western Europe |
|---|---|---|---|---|---|---|
| GWTG-HF | None | 0.006 | 0.004 | 0.020 | 0.002 | 0.017 |
| | Intercept | 0.006 | 0.005 | 0.025 | 0.002 | 0.004 |
| | Slope and Intercept | 0.006 | 0.002 | 0.014 | 0.002 | 0.004 |
| OPTIME-CHF | None | 0.130 | 0.193 | 0.092 | 0.084 | 0.192 |
| | Intercept | 0.035 | 0.048 | 0.037 | 0.018 | 0.047 |
| | Slope and Intercept | 0.001 | 0.007 | 0.008 | 0.006 | 0.005 |
| EFFECT | None | 0.04 | 0.022 | 0.095 | 0.058 | 0.020 |
| | Intercept | 0.036 | 0.013 | 0.073 | 0.031 | 0.024 |
| | Slope and Intercept | 0.008 | 0.010 | 0.025 | 0.006 | 0.012 |

**Table 3: Calibration with various Recalibration techniques.** Recalibration method is the technique of model updating. None indicates the model as originally applied without any updating, Intercept is update of the intercept, Slope and Intercept is update of the slope and intercept. Eavg represents the Harrell's E statistic, a measure of calibration. This measure of error computes the average absolute calibration error (average absolute difference between the lowess-estimated calibration curve and the line of identity).

These curves demonstrate highly variable calibration. **Table 3** shows CPM calibration across the regional validation populations.

Generally model calibration appears more favorable for the CPM predicting in-hospital mortality (GWTG-HF). For the GWTG-HF CPM the $E_{avg}$ ranged from 0.2% in Eastern Europe to 2.0% in South America. For the OPTIME-CHF CPM the $E_{avg}$ ranged from 8.4% in Eastern Europe to 19.3% in North America. For the EFFECT CPM the $E_{avg}$ ranged from 2.0% in Western Europe to 9.5% in South America. $E_{max}$ is shown in **Figure S8.**

| Model | Recalibration method | Emax Worldwide | Emax North America | Emax South America | Emax Eastern Europe | Emax Western Europe |
|---|---|---|---|---|---|---|
| GWTG-HF | None | 0.057 | 0.271 | 0.901 | 0.088 | 0.181 |
| | Intercept | 0.112 | 0.244 | 0.907 | 0.088 | 0.047 |
| | Slope and Intercept | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| OPTIME-CHF | None | 0.559 | 0.580 | 0.476 | 0.532 | 0.664 |
| | Intercept | 0.364 | 0.393 | 0.331 | 0.319 | 0.517 |
| | Slope and Intercept | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| EFFECT | None | 0.068 | 0.021 | 0.205 | 0.105 | 0.039 |
| | Intercept | 0.106 | 0.012 | 0.267 | 0.155 | 0.055 |
| | Slope and Intercept | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

**Table S8: Calibration with various Recalibration techniques.** Recalibration method is the technique of model updating. None indicates the model as originally applied without any updating, Intercept is update of the intercept, Slope and Intercept is update of the slope and intercept. EMax represents the Harrell's E statistic, a measure of calibration. This measure of error computes the maximum absolute difference in predicted and calibrated probabilities.

CPM discrimination was also assessed across different world regions and we observed major decrements in the ability of the CPMs to discriminate between those who died from those who did not **(Table 4).** The median model decrement in discrimination across all world regions and all CPMs was 40%. The median decrement for GWTG-HF CPM was 40% with the poorest performance in South America (AUC 0.52). The median decrement in discrimination for OPTIME-CHF CPM was 22% with the worst performance in Western Europe (AUC 0.66). The ability of the EFFECT CPM to discriminate those who died at one year from those who did not varied substantially across different world regions. This model had the poorest discrimination in South America (AUC 0.58) while the AUC in Western Europe and North America was 0.69 to 0.72 respectively.

Model Updating

We applied model updating technique 2 (updated intercept, **Table SA**) and demonstrated generally significantly improved calibration **(Table 4, Figure S6a-e)**. With this approach the GWTG-HF CPM showed a recalibrated $E_{avg}$ that ranged from 0.2% in Eastern Europe to 2.5% in South America. The OPTIME-CHF CPM demonstrated a recalibrated $E_{avg}$ that ranged from 1.8% in Eastern Europe to 4.8% in North America. The EFFECT CPM showed a recalibrated $E_{avg}$ that ranged from 1.3% in Western Europe to 7.3% in South America ($E_{max}$ is shown in **Figure S8).**

Following regional updating of the CPM intercept and slope (technique 3) we observed additional improvements in calibration **(Table 4, Figure S7a-e)**.

| CPM | Derivation AUC | Worldwide AUC (% decrement) | North America AUC (% decrement) | South America AUC (% decrement) | Eastern Europe AUC (% decrement) | Western Europe AUC (% decrement) |
|---|---|---|---|---|---|---|
| GWTG-HF | 0.75 | 0.64 (-44%) | 0.70 (-20%) | 0.52 (-92%) | 0.65 (-40%) | 0.65 (-40%) |
| OPTIME-CHF | 0.77 | 0.72 (-19%) | 0.69 (-30%) | 0.71 (-22%) | 0.71 (-22%) | 0.66 (-41%) |
| EFFECT | 0.77 | 0.66 (-41%) | 0.72 (-19%) | 0.58 (-70%) | 0.62 (-56%) | 0.69 (-30%) |

**Table 4: Discrimination**. AUC is area under the receiver operator curve, % decrement is the percent decrease in discrimination and is calculated as [Derivati AUC-0.5]-[Regional AUC-0.5]/[Derivation AUC-0.5] * 100.

With this approach the GWTG-HF CPM showed a recalibrated $E_{avg}$ that ranged from 0.2% in Eastern Europe and North America to 1.4% in South America. The OPTIME-CHF CPM demonstrated a recalibrated $E_{avg}$ with this technique that ranged from 0.5% in Western Europe to 0.8% in South America. The EFFECT CPM showed a recalibrated $E_{avg}$ that ranged from 1.0% in North America to 2.5% in South America ($E_{max}$ is shown in **Figure S8.)**

Slope and intercept values that optimize regional CPM performance (minimizing

| Model | Timeframe | Intercept, slope (Worldwide) | Intercept, slope (North America) | Intercept, slope (South America) | Intercept, slope (Eastern Europe) | Intercept, slope (Western Europe) |
|---|---|---|---|---|---|---|
| GWTG-HF | In hospital | -0.159, 0.883 | 1.21, 1.335 | -2.783, 0.099 | -0.318, 0.917 | 0.748, 1.061 |
| OPTIME-CHF | 60 days | -1.806, 0.532 | -1.777, 0.468 | -1.482, 0.558 | -1.849, 0.626 | -1.983, 0.375 |
| EFFECT | 1 year | -0.028, 0.753 | 0.070, 0.965 | -0.190, 0.461 | -0.118, 0.687 | -0.025, 0.854 |

**Table 5. Corrections.** Optimized regional intercept and slope corrections that optimize calibration so that predicted outcome rates match observed outcome rates.

$E_{avg}$ ) are shown in **Table 5** and demonstrate that distinct corrections are needed for each world region. 13 of the 15 (86%) of the slope corrections are < 1 indicating that the original CPMs are generally overfit to the derivation populations.

Generalizability by Region:

EFFECT model and GWTG-HF predicted probabilities generally best matched observed probabilities in North America,. Calibration of the OPTIME-CHF CPM is improved significantly with updating of the intercept only. For South America the baseline unadjusted CPMs were not well calibrated and following updating of the intercept only substantial differences between predicted and observed rates remain (for all models). After updating both the intercept and slope the OPTIME-CHF and EFFECT models demonstrated improved calibration though the GWTG-HF model remained poorly calibrated. Eastern European validation showed that the EFFECT model generally under predicted outcomes in the absence of any recalibration. Substantial improvements in calibration were seen for all models following updating of the intercept only for all CPMs in this region. In Western Europe calibration of the original EFFECT model is excellent without the need for recalibration. Calibration of the remaining models improved significantly following updating of the intercept only. Notably, the major decrements in discrimination that we observed remain unchanged following the various updating procedures.

**Discussion:**

There has been significant attention paid to the development and publication of CPMs (generally) and for cardiovascular diseases (in particular).[8] At the same time there has been relatively little focus on the performance of these models on

populations separate from the derivation cohort and more similar to patients seen in clinic. Here we show that published CPMs frequently perform extremely poorly and have limited generalizability in new populations both with respect to discrimination and calibration. We demonstrate that performance can vary substantially across different world regions even in the same clinical trial with uniform inclusion criteria (up to a 3-fold difference in observed vs. predicted event rates and > 90% decrease in discrimination) and also that performance (specifically calibration) can be improved significantly with simple updating procedures. Different adjustments (to intercept and slope) are necessary to optimize performance across various world regions such that, in the case of heart failure, it appears unrealistic to expect a single 'off-the-shelf' CPM to perform well across all settings.

An important (and often neglected) measure of performance is calibration, specifically how well predicted probabilities correlate with observed outcome rates. Here we show that overall calibration of the originally published CPMs varies across world regions and is often poor. The reasons for poor regional (as presented) model include regional differences in HF etiology, severity, and treatment.[17] So too, certain variables such as NYHA class[32] and various vital signs[33], are likely captured with varying fidelity across different databases and regions. It is also likely that the threshold to admit patients for AHF, local systems for post-discharge care and follow up are all highly variable across the globe and relate to prognosis. Reasonable local calibration is essential since applying poorly calibrated models to inform clinical decisions holds the potential to do

harm when compared to ‘treat all’ or ‘treat none’ approaches, whereas good calibration protects models from motivating harmful changes in decisions regardless of model discrimination.[34]

We note major and variable decrement in CPM discrimination across these different world regions. There are likely two major drivers of this observation; 1) modeling techniques and 2) phenotype heterogeneity. It is now well recognized that CPM discrimination is often substantially better on the data that were used to derive the model as compared to new patients that are similar but distinct from the derivation cohort.[36] This optimistic performance often emerges from the statistical methods used to select predictors (here univariate and multivariable selection methods) since these methods suffer from the same limitations associated with other forms of multiple testing, are known to underestimate P-values, and lead to overfitting of the final model.[13] In addition to this concern, if major outcome predictors are not included in the CPMs and differentially distributed throughout the world then this measure of performance will suffer (and be different across world regions). An example of this heterogeneity is noted in South America where the etiology of HF is much less likely to be from ischemic heart disease (as compared to North America) and also use of certain therapies (such as implantable cardioverter-defibrillators and beta blockers) is less common.[17] Unfortunately the simple updating techniques done here (in the absence of adding variables or recalculating individual beta coefficients) do nothing to improve this loss of discrimination.

Simple updating techniques can significantly improve calibration. We demonstrate that the updating procedures needed to optimize performance are region specific and that generally the most significant improvements are seen after updating the intercept, a technique that improves calibration-in-the large. Ideally clinicians would understand (and optimize) performance for local populations similar to those seen in clinic. This level of performance (not assessed here) would account for many of the unmeasured local determinants of outcomes.[35] As CPMs are used to aid clinical decisions it is important to understand model performance within local care systems. If models are employed for administrative purposes to compare the performance of various sin differentregions, accurate (and updated) calibration becomes of central importance. Without this measure of performance our assessment of CPMs (and the information they yield) is incomplete (at best) and potentially harmful.

While it is appealing to separate individuals who are likely experience an event from those who are not, such information may be of little value; reporting to patients an event probability is not inherently helpful (and may be harmful if the prediction is poorly calibrated). Too often there is no clinical decision that is tied to such information. In fact there are only isolated examples from across the cardiovascular subdisclipines where specific clinical decisions are tied to predicted outcome risk. Well-known examples include statin therapy in the primary prevention setting[37] and warfarin use for patients with non-valvular atrial fibrillation.[38] More often patients with common clinical presentations and disease states are treated uniformly based on general phenotypes and this is especially

common when the number needed to treat is reasonably low and the harm associated with therapies is minimal (as for many heart failure medical therapies). When there are significant risks associated with treatments such as surgical and procedural interventions *accurate* predictions may indeed be very beneficial.

There are analytic limitations that should be discussed. The size of the validation cohorts (and regional event rates) varied across different world regions and in some regions (Western Europe and South America) the smaller sample sizes produce more uncertainty about observed performance (both discrimination and calibration. This was especially the case for the GWTG model, since in hospital mortality outcome rates were very low. Also, our chosen measure of calibration, $E_{avg}$, is related to the absolute event rate (observed and predicted) and thus appears better in populations with low event rates. We used complete case analyses in these validations. If data are not missing completely at random than this may bias our results. For example, if missingness occurred more frequently for patients with higher risk and there was an association with unmeasured risk variables, there is a chance that we are preferentially eliminating some high risk patients from our analysis and our calibration in-the-large (average event rate) would be an underestimate of the true event rate.

Our results show that the performance of North American CPMs varies across different world regions and that in certain regions published CPMs are not helpful. We demonstrate that simple updating procedures improve the calibration (but not discrimination) of previously published CPMs for regional populations

with AHF. This analysis shows the importance of independent external validations, especially when clinical decisions or administrative comparisons might be leveraged by the output. Poorly calibrated models hold the potential for harm and there should be renewed emphasis on *local* performance of CPMs and also on what decisions might be improved through use of these models. Clinicians and the research community should call for systematic independent external validations of proposed (and previously published CPMs) to more fully evaluate this important dimension of performance. Ultimately, when considering potential application of CPMs to inform clinical decisions or assess outcomes the most important performance measures are not those described during model development but instead are those that describe local performance

References:


1. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. JAMA 2007;298(10):1209–12.
2. Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 2013;66(8):818–25.
3. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338(june):b606.
4. Goff DC, Lloyd-Jones DM, Bennett G, 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. Circulation 2014;129(25 Suppl 2):S49–73.
5. January CT, Wann LS, Alpert JS, et al. 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: A report of the American college of cardiology/American heart association task force on practice guidelines and the heart rhythm society. J. Am. Coll. Cardiol. 2014;64(21):e1–76.
6. Yancy CW, Jessup M, Bozkurt B, et al. 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(16):e240–327.
7. Pearson T a. AHA Guidelines for Primary Prevention of Cardiovascular Disease and Stroke: 2002 Update: Consensus Panel Guide to Comprehensive Risk Reduction for Adult Patients Without Coronary or Other Atherosclerotic Vascular Diseases. Circulation 2002;106(3):388–91.
8. Wessler BS, Lai Yh L, Kramer W, et al. Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. Circ Cardiovasc Qual Outcomes 2015;
9. Kent DM, Shah ND. Risk models and patient-centered evidence: should physicians expect one right answer? JAMA 2012;307(15):1585–6.
10. Alba AC, Agoritsas T, Jankowski M, et al. Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review. Circ Heart Fail 2013;6(5):881–9.
11. Allan GM, Nouri F, Korownyk C, Kolber MR, Vandermeer B, McCormack J. Agreement among cardiovascular disease risk calculators. Circulation 2013;127(19):1948–56.
12. Gabbay E, Calvo-Broce J, Meyer KB, Trikalinos TA, Cohen J, Kent DM.

The empirical basis for determinations of medical futility. J Gen Intern Med 2010;25(10):1083–9.

13. Bleeker SE, Moll H a., Steyerberg EW, et al. External validation is necessary in prediction research: A clinical example. J Clin Epidemiol 2003;56:826–32.

14. Hemingway H, Croft P, Perel P, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ 2013;346:e5595.

15. 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(2):211–9.

16. Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. JAMA 2005;293(5):572–80.

17. Blair JE a, Zannad F, Konstam M a, et al. Continental differences in clinical characteristics, management, and outcomes in patients hospitalized with worsening heart failure results from the EVEREST (Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptan) program. J Am Coll Cardiol 2008;52(20):1640–8.

18. Pfeffer MA, Claggett B, Assmann SF, et al. Regional Variation in Patients and Outcomes in the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) Trial. [A1-15,202] Circ 2014;

19. Greene SJ, Fonarow GC, Solomon SD, et al. Global variation in clinical profile, management, and post-discharge outcomes among patients hospitalized for worsening chronic heart failure: findings from the ASTRONAUT trial. Eur J Heart Fail 2015;n/a – n/a.

20. Gheorghiade M, Orlandi C, Burnett JC, et al. Rationale and design of the multicenter, randomized, double-blind, placebo-controlled study to evaluate the Efficacy of Vasopressin antagonism in Heart Failure: Outcome Study with Tolvaptan (EVEREST). J Card Fail 2005;11(4):260–9.

21. Rahimi K, Bennett D, Conrad N, et al. Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail 2014;2(5):440–6.

22. Konstam MA, Gheorghiade M, Burnett JC, et al. Effects of oral tolvaptan in patients hospitalized for worsening heart failure: the EVEREST Outcome Trial. JAMA 2007;297(12):1319–31.

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

24. Morise AP, Diamond GA, Detrano R, Bobbio M, Gunel E. The effect of disease-prevalence adjustments on the accuracy of a logistic prediction model. Med Decis Making 16(2):133–42.

25. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16(May 1995):965–80.

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

27. 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;

28. Peterson PN, Rumsfeld JS, Liang L, et al. 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(1):25–32.

29. Felker GM, Leimberger JD, Califf RM, et al. Risk stratification after hospitalization for decompensated heart failure. J Card Fail 2004;10(6):460–6.

30. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu J V. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. JAMA 2003;290(19):2581–7.

31. Cuffe MS, Califf RM, Adams KF, et al. Short-term intravenous milrinone for acute exacerbation of chronic heart failure: a randomized controlled trial. JAMA 2002;287(12):1541–7.

32. Bennett JA, Riegel B, Bittner V, Nichols J. Validity and reliability of the NYHA classes for measuring research outcomes in patients with cardiac disease. Heart Lung 31(4):262–70.

33. Edmonds Z V, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. Ann Emerg Med 2002;39(3):233–7.

34. 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;

35. Hawkins NM, Jhund PS, McMurray JJ V, Capewell S. Heart failure and socioeconomic status: accumulating evidence of inequality. Eur J Heart Fail 2012;14(2):138–46.

36. Jr FEH, Lee KL, Mark DB. TUTORIAL IN BIOSTATISTICS MULTIVARIABLE PROGNOSTIC MODELS : ISSUES IN DEVELOPING MODELS , EVALUATING ASSUMPTIONS AND ADEQUACY , AND MEASURING AND REDUCING ERRORS. Stat Med 1996;15:361–87.

37. Stone NJ, Robinson JG, Lichtenstein AH, et al. 2013 ACC/AHA Guideline

on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: 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):2889–934.

38. Lip GYH, Nieuwlaat R, Pisters R, Lane D a, Crijns HJGM. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest 2010;137(2):263–72.