ORIGINAL RESEARCH

Regional Validation and Recalibration of Clinical Predictive Models for Patients With Acute Heart Failure

Benjamin S. Wessler, MD, MS; Robin Ruthazer, MPH; James E. Udelson, MD; Mihai Gheorghiade, MD; Faiez Zannad, MD, PhD; Aldo Maggioni, MD; Marvin A. Konstam, MD; David M. Kent, MD, MSc

***Background*—**Heart failure clinical practice guidelines recommend applying validated clinical predictive models (CPMs) to support decision making. While CPMs are now widely available, the generalizability of heart failure CPMs is largely unknown.

***Methods and Results*—**We identified CPMs derived in North America that predict mortality for patients with acute heart failure and validated these models in different world regions to assess performance in a contemporary international clinical trial (N=4133) of patients with acute heart failure treated with guideline-directed medical therapy. We performed independent external validations of 3 CPMs predicting in-hospital mortality, 60-day mortality, and 1-year mortality, respectively. CPM discrimination decreased in all regional validation cohorts. The median change in area under the receiver operating curve was −0.09 (range −0.05 to −0.23). Regional calibration was highly variable (90th percentile of absolute difference between smoothed observed and predicted values range <1% to >50%). Calibration remained poor after global recalibrations; however, region-specific recalibration procedures significantly improved regional performance (recalibrated 90th percentile of absolute difference range <1% to 5% across all regions and all models).

***Conclusions*—**Acute heart failure CPM discrimination and calibration vary substantially across different world regions; region-specific (as opposed to global) recalibration techniques are needed to improve CPM calibration. **(*J Am Heart Assoc*. 2017;6: e006121. DOI: 10.1161/JAHA.117.006121.)**

Key Words: acute heart failure • cardiovascular disease risk factors • clinical predictive model • external validation • modeling • prediction • prognostic factor

It is increasingly recognized that patients with the same disease can differ from one another substantially with respect to their outcome risks, and the harms and benefits of treatment.[1,2] To aid physicians and patients in individualizing decisions, clinical predictive models (CPMs) are now widely available to estimate the likelihood of important outcomes (prognostic models) or diagnoses (diagnostic models) based on patient-specific characteristics.[3] In the case of heart failure, CPMs have been proposed to inform decisions for advanced therapies and palliative care[4] and also the common and costly admission decision for patients with acute heart failure (AHF) in the emergency department.[5] While many different CPMs exist for predicting mortality for HF,[6] CPM performance is often significantly better for the population on which the model was derived compared with similar yet distinct "validation" populations.[7]

Model performance across different world regions is largely unknown. Even within the restricted settings of randomized controlled trials for patients with HF, substantial regional heterogeneity in patient characteristics and in outcome rates have been observed.[8–10] Thus, an important but understudied concern is that CPMs may support appropriate decision making in 1 region, while yielding misleading information in another. Here we use data from the EVEREST (Efficacy of Vasopressin Antagonism in Heart Failure Outcome Study with Tolvaptan) trial[11] and perform regional independent external validations of previously published CPMs that

From the Tufts Cardiovascular Center, Tufts Medical Center, Boston, MA (B.S.W., J.E.U., M.A.K.); Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center/Tufts University School of Medicine, Boston, MA (B.S.W., R.R., D.M.K.); Northwestern University Feinberg School of Medicine, Chicago, IL (M.G.); Institut National de la Santé et de la Recherche Médicale (INSERM), Nancy, France (F.Z.); Associazione Nazionale Medici Cardioligi Ospedalieri Research Center, Florence, Italy (A.M.).

Accompanying Tables S1 through S4 and Figures S1, S2 are available at http://jaha.ahajournals.org/content/6/11/e006121/DC1/embed/inline-supplementary-material-1.pdf

Correspondence to: Benjamin S. Wessler, MD, MS, Tufts Cardiovascular Center, Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies (ICRHPS), Tufts Medical Center (TMC), 800 Washington St, Box 63, Boston, MA 02111. E-mail: bwessler@tuftsmedicalcenter.org

Received April 13, 2017; accepted September 1, 2017.

Clinical Perspective

What Is New?

- To assess the generalizability of acute heart failure clinical predictive models (CPMs), we validated and recalibrated a sample of acute heart failure CPMs predicting short- and long-term mortality in different world regions.

What Are the Clinical Implications?

- CPM discrimination and calibration vary substantially across different world regions, and regional (as opposed to global) recalibration techniques were needed to improve CPM calibration.
- Off-the-shelf acute heart failure CPMs may support appropriate decision making in 1 region, while yielding misleading information in another.
- Region-specific recalibrations can improve CPM calibration.

predict mortality following hospital admission for AHF. We evaluate CPMs for AHF derived on data from patients in 1 world region (here, North America) and determine whether these CPMs can generalize to patients in different world regions (Eastern Europe, Western Europe, and South America and whether global or regional recalibration procedures improve regional performance.

Methods

External validations explore CPM performance for patients not included in the derivation data set. The general approach requires matching CPMs to validation database(s) and assessing model performance. Here CPM performance was assessed in different world regions and recalibration techniques were evaluated.

Model Selection

Identifying CPMs that match the validation database is a process that involves evaluation of both the original CPM and the validation cohorts (Table 1). 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 an outcome captured in the validation cohort (here mortality), (3) all variables in the CPM were captured in the validation data sets 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 reviewing a recently published systematic review of CPMs for HF.[6] For this analysis, we present a sample of the compatible CPMs developed in North America that predict mortality at 3 different time points (in-hospital, 60 day, and 1 year) following hospitalization for HF.

Selected Models

Selected validated models are shown in Table 1 and Figure S1. Selected models were as follows: GWTG-HF[12] (The American Heart Association Get With the Guidelines-Heart Failure) model (7 variables, predicts in-hospital mortality), OPTIME-CHF[13] (Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure) (5 variables, predicts 60-day mortality after admission), and EFFECT[14] (Enhanced Feedback for Effective Cardiac Treatment) model (10 variables, predicts 1-year mortality after admission).

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.[12] 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. It has since been externally validated.[15]

The OPTIME-CHF study was a randomized clinical trial of 949 patients with HF with reduced ejection fraction hospitalized for worsening symptoms.[16] Patients were randomized to receive intravenous milrinone or placebo for 48 to 72 hours. The outcome of 60-day mortality did not differ significantly between the milrinone and placebo groups (10.3% versus 8.9%, P=0.41). Patients were enrolled from 78 centers across the United States from 1997 to 1999. A CPM based on a point score predicting 60-day mortality was derived from this data set using Cox proportional hazards analysis and internally validated in this database.[13]

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 prespecified definition of clinical HF. This CPM was created using logistic regression analysis and validated on 1407 patients from different hospitals in Ontario from a previous time period (1997–1999).

External Validation Cohort

The EVEREST trial has been previously reported.[17] This was a prospective, international, randomized, placebo-controlled study conducted in 359 sites worldwide from 2003 and 2006. The trial included 1251 patients from North America,

DOI: 10.1161/JAHA.117.006121

Table 1. Baseline Characteristics for Patients Among the Various Databases

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	27 850	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	†	†	†	†	†

Clinical predictive models derivation populations are presented on the left (bold border). Validation data sets (overall and regional) are shown on the right. Gray shading indicates variables that are included in the CPM derived from each database. BUN indicates blood urea nitrogen; BPM, beats per minute; CVA, cerebrovascular accident; COPD, chronic obstructive pulmonary disease; CPM, clinical predictive model; EE, Eastern Europe; EFFECT, Enhanced Feedback for Effective Cardiac Treatment study; EVEREST, Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptan; GWTG-HF, Get With The Guidelines-Heart Failure; NA, North American; NYHA, New York Heart Association; OPTIME-CHF, The Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure study; SA, South America; SBP, systolic blood pressure; WE, Western Europe.

*Acute heart failure populations that include patients with both reduced and preserved ejection fractions.

†Variables that were exclusion criteria for a given database (these variables were coded as 0). NR indicates not reported. For the derivation populations, continuous variables are shown as means ($) or medians (&) as originally presented. For the validation populations, values are presented as median (interquartile range).

699 patients from South America, 564 patients from Western Europe, and 1619 patients from Eastern Europe (Figure 1). This study evaluated the addition of tolvaptan to standard medical therapy for AHF and reduced ejection fraction and enrolled patients within 48 hours of HF hospitalization. During a median follow-up of 9.9 months, 537 (26%) of the patients died and tolvaptan had no effect on long-term mortality for these patients (hazard ratio 0.98; 95% confidence interval, 0.87%–1.11%; *P*=0.68). The patients enrolled in this trial were treated with guideline-directed medical therapies for HF including angiotensin-converting enzyme inhibitors (84%), β-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.

Outcomes

All models were tested for their ability to predict all-cause mortality in the overall EVEREST cohort and separately in regional EVEREST cohorts using patient-level data. The GWTG-HF in-hospital mortality model was validated on in-hospital mortality in the EVEREST study; the OPTIME-CHF 60-day mortality model was validated on 60-day mortality in the EVEREST study; the EFFECT study 1-year mortality model was validated on 1-year mortality in the EVEREST study (Figure 1). Patients censored prior to 1 year were either dropped from the analysis (if last known alive and followed for <9 months, n=1471) or included as alive (if alive and followed for ≥9 months, n=2662). Sensitivity analyses to explore these assumptions are presented in Figure S2A through S2D.

Statistical Analysis and Model Recalibration

Our approach to validating these CPMs used patient-level data from EVEREST. For each patient and each CPM we calculated a point score based on covariate values. This point score was then converted into predicted event probabilities as described by the original CPM authors (Figure S1). When a range of probabilities was given, the midpoint probability was assigned for a given point score range. For various performance measures and both global and regional recalibration procedures, the estimated event probabilities were converted to the linear predictor using the equation [predicted value=$(1/(1+e^{-xbeta}))$] where xbeta is the linear predictor. We evaluated the loss in discrimination by assessing the change in Area under the Receiver Operating Curve (AUC). 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.

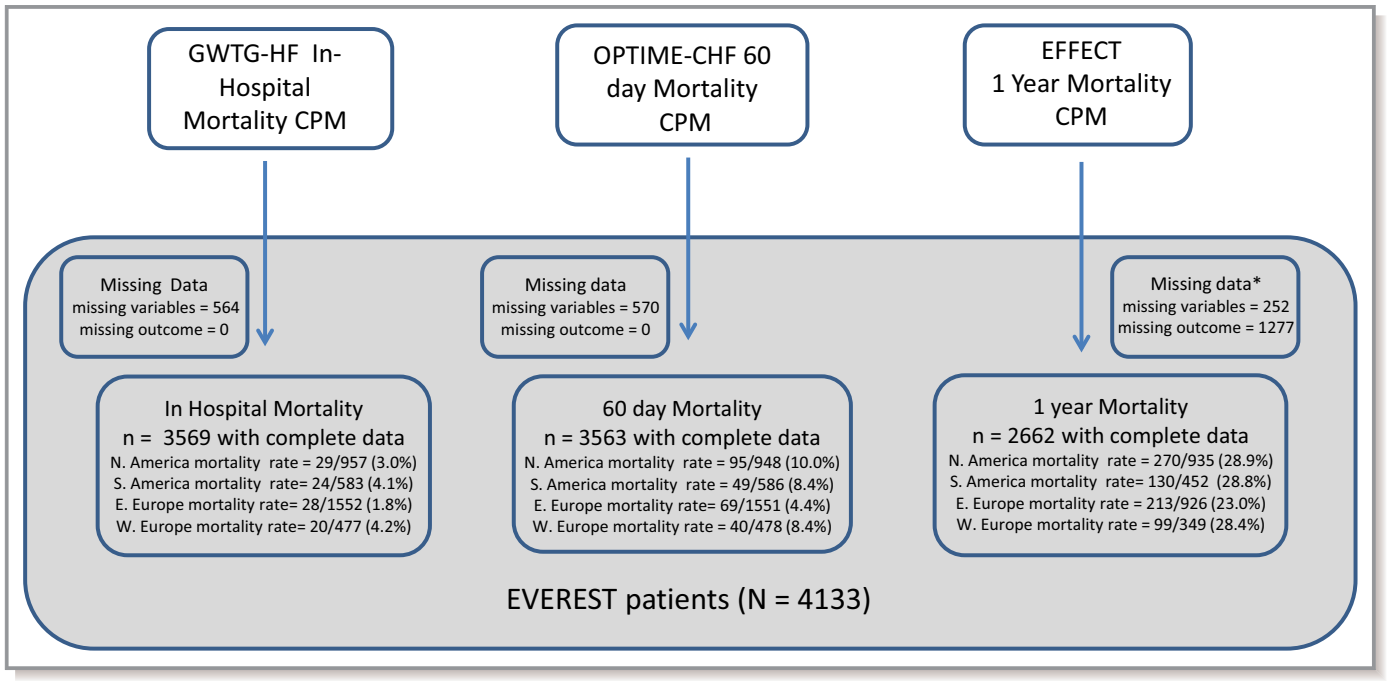


Figure 1. GWTG-HF is Get with the Guidelines-Heart Failure in-hospital mortality CPM. OPTIME-CHF is Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure 60-d mortality CPM. EFFECT is the Enhanced Feedback for Effective Cardiac Treatment 1-y mortality CPM. Validation exercises were done for patients with all variables available. *Indicates that for the 1-y mortality model, we considered patients to have missing data if they were last known alive with <9 mo of follow-up. CPM indicates clinical predictive models.

Measuring CPM Performance

Calibration-in-the-large is a measure of global fit. *Model discrimination* was represented here by the AUC. In this analysis, we assess percent decrement in discrimination, which is derived from the AUC for each region. *Model calibration* was assessed primarily through calibration plots. We also report Harrell's E statistic, which calculates a prediction error for each individual patient by using a lowess-estimated probability as the observed outcome rate.[18] We report E_{90} and E_{avg} statistics in this report. E_{avg} computes the average absolute calibration error (average absolute difference between the lowess-estimated calibration curve and the line of identity). E_{90} describes the 90th percentile of the absolute differences (ie, 90% of individuals have absolute prediction errors that are below this value).

Recalibration

CPM recalibration techniques have been previously described.[19] The simplest form of recalibration (technique 1) addresses calibration-in-the-large and 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. The next form of recalibration (technique 2) adjusts both the intercept and the slope (ie, applies a uniform correction factor to the regression coefficients of the independent variables to better fit the validation population). This recalibration technique corrects both for differences in prevalence unrelated to covariate effects (as in technique 1) and also can correct for overfitting in the derivation population. To assess whether global or region-specific recalibrations are needed to improve CPM performance, our recalibrations proceeded stepwise, first with global recalibrations on the entire EVEREST cohort (techniques 1 and 2) and next with region-specific recalibrations (techniques 1 and 2).

This study was reviewed and approved via expedited review procedures by the Tufts Health Sciences IRB and informed consent requirement was waived.

Results

The covariates that are used to calculate probabilities with each CPM are shown in Table 1. Overall the patients in the derivation cohorts appear similar (related) to the patients in the validation cohorts (EVEREST database overall and region specific). The distribution of covariates is shown for each world region within the validation databases. The numbers of cases with complete data and the number of outcomes for each time point and each region are shown in Figure 1. Two CPMs (GWTG-HF and EFFECT) were derived from data sets including both patients with HF with reduced ejection fraction and those with preserved ejection fraction. GWTG-HF CPM was derived from registry data. The OPTIME-CHF 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 S1. The randomized controlled trials had more exclusion criteria than the registry database.

Independent External Validations

CPM discrimination was 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 2). Even within the North American EVEREST cohort, there was a substantial decrement in model discrimination, with percent decrement ranging from −19% for the EFFECT CPM predicting 1-year mortality to −30% for the OPTIME-CHF model predicting 60-day mortality. The median model percent decrement in discrimination across all world regions and all CPMs was −35%. The median percent decrement in discrimination for GWTG-HF CPM was −42% and in South America the CPM had essentially no ability to effectively rank event probabilities (AUC 0.54). The median percent decrement in discrimination for OPTIME-CHF CPM was 26% with the worst performance in Western Europe (AUC 0.66). The EFFECT CPM had a median percent decrement in discrimination of 43% and had the poorest discrimination in South America (AUC 0.58).

We assessed calibration-in-the-large for each mortality time point (in-hospital mortality, 60-day mortality, and 1-year mortality) for the validation databases (Table 3). The in-hospital mortality rate was 2.8% in the EVEREST trial. GWTG-HF CPM had excellent calibration-in-the-large for Eastern Europe and North America, while substantially underpredicting overall event rates in South America and Western Europe (difference in observed versus predicted event rates is −2.1% and −1.7%, respectively). The 60-day mortality rate in the EVEREST trial was 7.1%. OPTIME-CHF CPM predicted 60-day mortality rates were considerably higher than observed rates; the difference in observed versus predicted event rates ranged from 8.3% in Eastern Europe to 19.2% in North America. By 1 year, 26.7% of patients in the overall EVEREST trial had died. The EFFECT CPM systematically underpredicted overall 1-year event rates across the different world regions, particularly in Eastern Europe and South America (by −5.0% and −9.1%, respectively).

We assessed model calibration across ranges of predicted risk for different world regions. Regional calibration plots (without recalibration) are shown in Figure 2A through 2D. These curves demonstrate highly variable and generally poor calibration. For the GWTG-HF CPM without recalibration the E_{90} ranged from <1% in Eastern Europe and North America to

Table 2. Discrimination

CPM	Derivation AUC	Worldwide AUC [95% CI] (% Decrement)	North America AUC [95% CI] (% Decrement)	South America AUC [95% CI] (% Decrement)	Eastern Europe AUC [95% CI] (% Decrement)	Western Europe AUC [95% CI] (% Decrement)
GWTG-HF	0.75	0.64 [0.60–0.69] (−44%)	0.70 [0.62–0.77] (−20%)	0.54 [0.42–0.66] (−84%)	0.65 [0.58–0.73] (−40%)	0.64 [0.55–0.74] (−44%)
OPTIME-CHF	0.77	0.72 [0.68–0.75] (−19%)	0.69 [0.64–0.74] (−30%)	0.69 [0.61–0.77] (−30%)	0.71 [0.64–0.78] (−22%)	0.66 [0.57–0.74] (−41%)
EFFECT	0.77	0.66 [0.64–0.68] (−41%)	0.72 [0.68–0.75] (−19%)	0.58 [0.53–0.64] (−70%)	0.62 [0.58–0.66] (−56%)	0.69 [0.58–0.66] (−30%)

AUC indicates area under the receiver operator curve, % decrement is the percent decrease in discrimination and is calculated as [Derivation AUC−0.5]−[Regional AUC−0.5]/[Derivation AUC−0.5]×100; CI, confidence interval; CPM, clinical predictive models; EFFECT, Enhanced Feedback for Effective Cardiac Treatment study; GWTG-HF, Get With The Guidelines-Heart Failure; OPTIME-CHF, Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure study.

3.9% in South America. The OPTIME-CHF CPM demonstrated substantial miscalibration with the E_{90} ranging from 19% in Eastern Europe to 51% in Western Europe. For the EFFECT CPM, calibration varied significantly across different world regions where the E_{90} ranged from 3% in North America to 18% in South America. Tables S2 and S3 show a summary of CPM calibration across the different regional validation populations.

Model Recalibration (Global)

Our first set of recalibrations was based on global adjustments of the intercept (technique 1) and intercept and slope (technique 2), (Table S3). Despite global recalibration of the intercept, GWTG-HF CPM predicting in-hospital mortality E_{90} remained at 3.8% in South America, OPTIME-CHF CPM predicting 60-day mortality remained poorly calibrated in certain regions (eg, E_{90} was 13.7% in Western Europe) and the EFFECT CPM predicting 1-year mortality showed only minimal improvement from baseline performance (recalibrated E_{90} ranged from 4.4% to 16.1% across different world regions). Recalibrations based on global adjustment of the intercept and slope (technique 2) yielded similar results. GWTG-HF CPM E_{90} ranged from <1% to 3.7%, OPTIME-CHF CPM remained poorly calibrated (eg, E_{90} was 7.5% in South America), and EFFECT CPM predicting 1-year mortality also showed only minimal improvement from the base model performance (recalibrated E_{90} ranged from 1.1% to 12.9% across different world regions).

Model Recalibration (Regional)

Next we applied technique 1 using region-specific recalibrations (Figure 2A through 2D and Table S3). Despite region-specific updating of the intercept, the regional calibration of the GWTG-HF CPM predicting in-hospital mortality remained essentially unchanged (E_{90} ranged from <1% to 3.4% across different world regions). Technique 1 regional recalibration led to only modest improvements in regional calibration for the OPTIME-CHF CPM predicting 60-day mortality, and miscalibration for this CPM was most significant in South America where E_{90} remained at 13.5%. Following technique 1 recalibration, the regional calibration for the EFFECT CPM predicting 1-year mortality showed only minimal improvement (E_{90} was 12.9% in South America).

Regional recalibration of the CPM intercept and slope (technique 2) demonstrated significant improvements in calibration (Figure 2A through 2D and Table S3). Following technique 2 recalibration, E_{90} for the GWTG-HF CPM predicting in-hospital mortality decreased to ≤1.4% across all world regions. This regional recalibration technique lowered E_{90} for the OMPTIME-CHF CPM predicting 60-day mortality and the

Table 3. Calibration-in-the-Large

Model	Event Rate	EVEREST	N. America	S. America	E. Europe	W. Europe
GWTG-HF (in hospital)	Observed event rate	0.028	0.030	0.041	0.018	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.)	0.006	0.003	0.021	0.001	0.017
OPTIME-CHF (60 d)	Observed event rate	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
EFFECT (1 y)	Observed event rate	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

Observed and Predicted average event rates in the validation data sets. Average Pred. Rate indicates the mean predicted outcome rates in the validation data sets (SD); Diff. (Obs.−Pred.), the difference between the Observed event rate and the average predicted event rate; E. Europe, Eastern European patients in EVEREST; EVEREST, Efficacy of Vasopressin Antagonism in Heart Failure: Outcome Study with Tolvaptan; GWTG-HF, Get With The Guidelines-Heart Failure; N. America, North American patients in EVEREST; S. America, South American patients in EVEREST; W. Europe, Western European patients in EVEREST.

EFFECT CPM predicting 1-year mortality across all world regions to ≤2.2% and ≤5.1%, respectively. The region-specific intercept and slope corrections that optimize calibration are shown in Table S2. In general, the OPTIME-CHF CPM and the EFFECT CPM had recalibrated slopes that were <1 across all world regions, suggesting that the original models were substantially overfit. Notably, the major decrements in discrimination that we observed remain unchanged despite the various recalibration procedures.

Discussion

Here a series of independent external validations demonstrate that published CPMs for AHF frequently perform poorly (with respect to discrimination and calibration) and have limited generalizability. Further, performance can vary substantially across different world regions even in the same clinical trial with uniform inclusion criteria. Finally, performance (specifically calibration) can be improved significantly with simple recalibration procedures, but only when recalibration is performed using region-specific corrections. Since different adjustments (to intercept and slope) are necessary to optimize performance across various world regions, it appears unrealistic to expect a single "off-the-shelf" CPM to perform well across all settings.

Consistent with a recent report limited only to North America,[15] The GWTG-HF CPM showed a moderate drop in discrimination in our North American validation cohort. CPM discrimination across different world regions was generally considerably worse for each of the 3 models compared with performance reported in the initial derivation samples and the decrement in discrimination varied substantially across different world regions. This may reflect (1) overfitting in the derivation population; (2) differences in case-mix/disease severity across regions; and (3) phenotype heterogeneity across regions (ie, the effects of the independent variables may be different across the different populations). Techniques that minimize the risks of overfitting include avoiding data-driven variable selection procedures and ensuring a large number (*often between 10 and 20*) events per considered variable.[20,21] An example of this heterogeneity is noted in South America where the causes of HF are different and also use of certain therapies (such as implantable cardioverter-defibrillators and β-blockers) are less common.[8] While the percent decrement in discrimination in different world regions is often large, we acknowledge uncertainty surrounding these point estimates. Unfortunately, the simple recalibration techniques done here (in the absence of adding variables or recalculating individual beta coefficients) do nothing to improve this loss of discrimination.

A similarly important (and often neglected[22]) measure of performance is calibration. Calibration of the originally published CPMs varies across world regions and is often poor. The reasons for poor regional calibration include regional differences in HF causes, severity, and treatment.[8,23,24] Additionally, certain variables such as New York Heart Association class[25] and various vital signs[26] 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 postdischarge 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—such as discharging low-risk patients from the hospital or considering advanced therapies for high-risk patients—holds the potential to do harm when compared with "treat all" or "treat none" approaches. Good calibration protects models from motivating harmful changes in decisions regardless of model discrimination.[27,28]

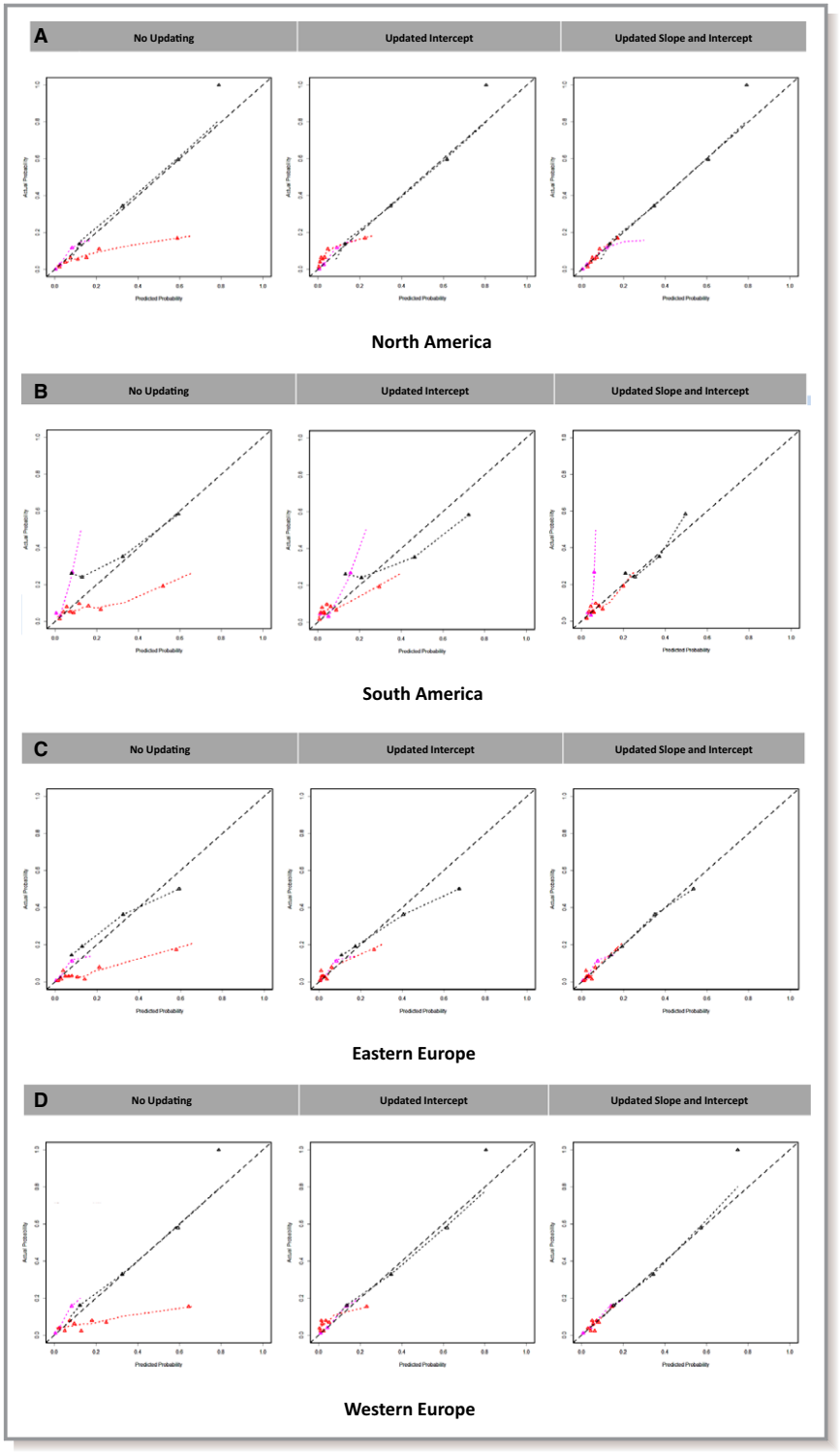


Figure 2. GWTG-HF is Get With the Guidelines–Heart Failure in-hospital mortality CPM. OPTIME-CHF is Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure 60-d mortality CPM. EFFECT is the Enhanced Feedback for Effective Cardiac Treatment 1-y mortality CPM. No updating is the original CPM applied to the validation population. Updated intercept is technique 1 with regional updating, Updated Intercept and Slope is technique 2 with regional updating (described in the text). A, North American calibration plots, (B) South American calibration plots, (C) Eastern European calibration plots, (D) Western European calibration plots. Calibration plots are presented according to deciles of predicted probabilities. CPM indicates clinical predictive models.

Simple recalibration techniques can significantly improve calibration, and the recalibration procedures needed to optimize performance are region specific. As CPMs are used to aid clinical decisions, it is important to understand model performance within local care systems. If models are used for administrative purposes, differences between observed and predicted event rates related to processes of care (and not poor CPM performance) may be informative and potentially actionable. Without these independent external measures of performance, our assessment of CPMs (and the information they yield) is incomplete (at best) and potentially harmful.

Our study had several limitations. First, our sample of AHF models did not comprehensively explore all published AHF CPMs and may not be representative of models generally or HF models in particular. We believe that these models are representative of AHF CPMs generally since they were created from contemporary clinical trial and registry data, have been variably incorporated into guidelines, and have been previously validated by the original investigators. There are certain validation data sets in specific regions with modest size (≈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] The GWTG-HF and EFFECT were derived on patients with AHF and preserved and reduced ejection fraction while the EVEREST database included only a subset of these patients (with reduced ejection fraction). If the effects of covariates are different across these unique HF subtypes or if there is less relatedness between these populations, then we should anticipate worse model performance across the EVEREST databases. Also, the CPMs examined here were point scores with predications based on *observed* outcome rates in point score strata rather than model-based probability estimates. Using these observed rates may have increased the error in prediction. Nevertheless, these observed outcome rates are presented in the original CPM articles as substitutes for risk predictions, and so are appropriate to use in our analysis. Finally, we used complete case analyses in these validations, which may bias our results if the included cases are not representative of the larger population of patients with AHF. This is unlikely to be a major concern since the patients included in the complete case analyses of these CPMs appear very similar across the different analytic timeframes (Table S4).

Performance of these North American CPMs for AHF is generally poor and varies substantially across different world regions. Simple recalibration procedures improve the calibration (but not discrimination) of previously published CPMs for regional populations with AHF, but only when region-specific recalibrations are applied. This analysis shows the importance of independent external validations, especially when clinical decisions 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.

Sources of Funding

This work was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Methods Award (ME-1606-35555), as well as by the National Institutes of Health (T32 HL069770 Training Grant from the NIH, 5 TL1 TR001062 Training Grant from the NIH-NCATS, 4U01NS086294-04). All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the PCORI, its Board of Governors, or Methodology Committee.

Disclosures

Drs Udelson, Konstam, Zannad, and Gheorghiade received research support from Otsuka for participating in the original EVEREST trial. The current analysis was not funded by Otsuka.

References

1. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA*. 2007;298:1209–1212.
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;66:818–825.
3. Wessler BS, Lai YH L, Kramer W, Cangelosi M, Raman G, Lutz JS, Kent DM. Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circ Cardiovasc Qual Outcomes*. 2015;8:368–375.
4. Allen LA, Stevenson LW, Grady KL, Goldstein NE, Matlock DD, Arnold RM, Cook NR, Felker GM, Francis GS, Hauptman PJ, Havranek EP, Krumholz HM, Mancini D, Riegel B, Spertus JA. Decision making in advanced heart failure: a scientific statement from the American Heart Association. *Circulation*. 2012;125:1928–1952.
5. Collins SP, Pang PS, Fonarow GC, Yancy CW, Bonow RO, Gheorghiade M. Is hospital admission for heart failure really necessary? The role of the emergency department and observation unit in preventing hospitalization and rehospitalization. *J Am Coll Cardiol*. 2013;61:121–126.
6. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, Woodward M, Patel A, McMurray J, MacMahon S. Risk prediction in patients with heart failure: a systematic review and analysis. *JACC Heart Fail*. 2014;2:440–446.
7. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, Moons KG. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol*. 2003;56:826–832.
8. Blair JE, Zannad F, Konstam MA, Cook T, Traver B, Burnett JC Jr, Grinfeld L, Krasa H, Maggioni AP, Orlandi C, Swedberg K, Udelson JE, Zimmer C, Gheorghiade M; EVEREST Investigators. 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;52:1640–1648.
9. Pfeffer MA, Claggett B, Assmann SF, Boineau R, Anand IS, Clausell N, Desai AS, Diaz R, Fleg JL, Gordeev I, Heitner JF, Lewis EF, O'Meara E, Rouleau J-L, Probstfield JL, Shaburishvili T, Shah SJ, Solomon SD, Sweitzer NK, McKinlay SM, Pitt B. Regional variation in patients and outcomes in the Treatment of Preserved Cardiac Function Heart Failure With an Aldosterone Antagonist (TOPCAT) trial. *Circulation*. 2015;131:34–42.
10. Greene SJ, Fonarow GC, Solomon SD, Subacius H, Maggioni AP, Böhm M, Lewis EF, Zannad F, Gheorghiade M; ASTRONAUT Investigators and Coordinators. 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;17:591–600.

11. Gheorghiade M, Orlandi C, Burnett JC, Demets D, Grinfeld L, Maggioni A, Swedberg K, Udelson JE, Zannad F, Zimmer C, Konstam MA. 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;11:260–269.
12. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcomes*. 2010;3:25–32.
13. Felker GM, Leimberger JD, Califf RM, Cuffe MS, Massie BM, Adams KF, Gheorghiade M, O'Connor CM. Risk stratification after hospitalization for decompensated heart failure. *J Card Fail*. 2004;10:460–466.
14. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *JAMA*. 2003;290:2581–2587.
15. Lagu T, Pekow PS, Shieh M-S, Stefan M, Pack QR, Kashef MA, Atreya AR, Valania G, Slawsky MT, Lindenauer PK. Validation and comparison of seven mortality prediction models for hospitalized patients with acute decompensated heart failure. *Circ Heart Fail*. 2016;9:e002912.
16. Cuffe MS, Califf RM, Adams KF, Benza R, Bourge R, Colucci WS, Massie BM, O'Connor CM, Pina I, Quigg R, Silver MA, Gheorghiade M; Outcomes of a Prospective Trial of Intravenous Milrinone for Exacerbations of Chronic Heart Failure (OPTIME-CHF) Investigators. Short-term intravenous milrinone for acute exacerbation of chronic heart failure: a randomized controlled trial. *JAMA*. 2002;287:1541–1547.
17. Konstam MA, Gheorghiade M, Burnett JC, Grinfeld L, Maggioni AP, Swedberg K, Udelson JE, Zannad F, Cook T, Ouyang J, Zimmer C, Orlandi C. Effects of oral tolvaptan in patients hospitalized for worsening heart failure: the EVEREST Outcome Trial. *JAMA*. 2007;297:1319–1331.
18. Harrell FE. *Regression Modeling Strategies*. Cham, Switzerland: Springer International Publishing; 2015.
19. Steyerberg EW. *Clinical Prediction Models*. New York, NY: Springer New York; 2009.
20. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstem AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49:1373–1379.
21. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making*. 2001;21:45–56.
22. 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.
23. Ferreira JP, Girerd N, Rossignol P, Zannad F. Geographic differences in heart failure trials. *Eur J Heart Fail*. 2015;17:893–905.
24. Kristensen SL, Martinez F, Jhund PS, Arango JL, Bělohlávek J, Boytsov S, Cabrera W, Gomez E, Hagège AA, Huang J, Kiatchoosakun S, Kim K-S, Mendoza I, Senni M, Squire IB, Vinereanu D, Wong RC-C, Gong J, Lefkowitz MP, Rizkala AR, Rouleau JL, Shi VC, Solomon SD, Swedberg K, Zile MR, Packer M, McMurray JJ. Geographic variations in the PARADIGM-HF heart failure trial. *Eur Heart J*. 2016;37:3167–3174.
25. 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*. 2002;31:262–270.
26. Edmonds ZV, Mower WR, Lovato LM, Lomeli R. The reliability of vital sign measurements. *Ann Emerg Med*. 2002;39:233–237.
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;74:167–176.
28. Cook NR, Ridker PM. Calibration of the Pooled Cohort Equations for atherosclerotic cardiovascular disease: an update. *Ann Intern Med*. 2016;165:786–794.

SUPPLEMENTAL MATERIAL

Table S1. 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

Exclusion criteria as written in the original reports (Person et al, Lee et al.) or the Design and Rationale resports (Cuffe et al. and Gheorghiade et al)

Table S2. Regional Intercept and Slope Corrections

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

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

Table S3. Calibration with Various Recalibration Techniques

	Model	Recalibration method	Eavg (E90) North America	Eavg (E90) South America	Eavg (E90) Eastern Europe	Eavg (E90) Western Europe
*Regional Calibration without updating	GWTG-HF	None	0.004 (0.005)	0.021 (0.039)	0.001 (0.001)	0.017 (0.014)
	OPTIME-CHF	None	0.193 (0.478)	0.092 (0.395)	0.084 (0.185)	0.192 (0.505)
	EFFECT	None	0.022 (0.030)	0.095 (0.182)	0.058 (0.065)	0.020 (0.040)
# Regional Calibration with various Global Recalibration techniques	GWTG-HF	Intercept	0.008 (0.007)	0.017 (0.038)	0.005 (0.008)	0.009 (0.006)
		Slope and Intercept	0.009 (0.008)	0.017 (0.037)	0.006 (0.008)	0.009 (0.006)
	OPTIME	Intercept	0.055 (0.110)	0.031 (0.042)	0.017 (0.018)	0.058 (0.137)
		Slope and Intercept	0.010 (0.019)	0.018 (0.075)	0.011 (0.024)	0.015 (0.035)
	EFFECT	Intercept	0.028 (0.047)	0.079 (0.161)	0.034 (0.044)	0.034 (0.063)
		Slope and Intercept	0.031 (0.066)	0.051 (0.129)	0.006 (0.011)	0.025 (0.031)
¥Calibration with various Regional Recalibration techniques.	GWTG-HF	Intercept	0.005 (0.006)	0.027 (0.034)	0.002 (0.001)	0.004 (0.003)
		Slope and Intercept	0.002 (0.003)	0.019 (0.014)	0.001 (0.001)	0.004 (0.004)
	OPTIME-CHF	Intercept	0.049 (0.079)	0.037 (0.135)	0.018 (0.016)	0.048 (0.084)
		Slope and Intercept	0.007 (0.012)	0.009 (0.022)	0.006 (0.015)	0.005 (0.006)
	EFFECT	Intercept	0.013 (0.019)	0.073 (0.129)	0.031 (0.044)	0.024 (0.028)
		Slope and Intercept	0.010 (0.014)	0.025 (0.051)	0.006 (0.012)	0.012 (0.016)

*represents regional calibration without recalibration. # represents regional calibration with Global recalibrations.¥represents regional calibration with region specific recalibrations. GWTG-HF predicts in-hospital mortality. OPTIME-CHF predicts 60 day mortality, EFFECT predicts 1 year mortality. Recalibration method is the technique of model updating. Intercept is update of the intercept to the overall database for the global recalibrations and to the specific region for the regional recalibrations. , Slope and Intercept is update of the slope and intercept to the overall database for the global recalibrations and to the specific region for the regional recalibrations. Eavg is Harrell's E statistic and represents the average difference between observed and predicted values. E90 represents the 90th percentile of absolute difference between observed and predicted values.

Table S4. Comparison Included vs. Excluded

Data source and Variable	Pooled	Include	Exclude	p-value	test
I. GWTIn (In-Hospital Outcome Model)	**N=4133**	**N=3568 (86%)**	**N=565 (13%)**		
Age	65.8 +/- 11.9 (4133)	65.8 +/- 11.8 (3568)	65.2 +/- 12.0 (565)	0.2486	(ttest)
Systolic blood pressure	120.5 +/- 19.7 (4091)	120.7 +/- 19.7 (3568)	118.8 +/- 19.1 (523)	0.0392	(ttest)
Sodium	139.6 +/- 4.6 (4030)	139.7 +/- 4.7 (3568)	139.2 +/- 4.0 (462)	0.0520	(ttest)
Blood urea nitrogen	30.2 +/- 16.3 (3960)	30.2 +/- 16.1 (3568)	30.3 +/- 18.4 (392)	0.9044	(ttest)
Death outcome_in hosp	2.6% (109/4129)	2.8% (101/3568)	1.4% (8/561)	0.0537	
region	N=4133	N=3568	N=565	<.0001	(chisq) df=3
EASTERN EUROPE	39.2% (1619)	43.5% (1552)	11.9% (67)		
NORTH AMERICA	30.3% (1251)	26.8% (956)	52.2% (295)		
SOUTH AMERICA	16.9% (699)	16.3% (583)	20.5% (116)		
WESTERN EUROPE	13.6% (564)	13.4% (477)	15.4% (87)		
II. Optime60 (60day Outcome Model)	**N=4133**	**N=3569 (86%)**	**N=564 (13%)**		
Age , mean +/- stdev	65.8 +/- 11.9 (4133)	65.8 +/- 11.8 (3563)	65.3 +/- 12.2 (570)	0.3302	(ttest)
Systolic blood pressure	120.5 +/- 19.7 (4091)	120.7 +/- 19.7 (3563)	118.9 +/- 19.1 (528)	0.0489	(ttest)
Sodium	139.6 +/- 4.6 (4030)	139.7 +/- 4.7 (3563)	139.2 +/- 4.0 (467)	0.0326	(ttest)
Blood urea nitrogen	30.2 +/- 16.3 (3960)	30.1 +/- 16.1 (3563)	30.5 +/- 18.6 (397)	0.6340	(ttest)
Death outcome_60d	7.1% (295/4133)	7.1% (253/3563)	7.4% (42/570)	0.8177	(chisq) df=1
region	N=4133	N=3563	N=570	<.0001	(chisq) df=3
EASTERN EUROPE	39.2% (1619)	43.5% (1551)	11.9% (68)		
NORTH AMERICA	30.3% (1251)	26.6% (948)	53.2% (303)		
SOUTH AMERICA	16.9% (699)	16.4% (586)	19.8% (113)		
WESTERN EUROPE	13.6% (564)	13.4% (478)	15.1% (86)		
IIIb. Effect365 (365 day Outcome Model)	**N=4133**	**N=2662 (64%)**	**N=1471 (36%)**		
Age	65.8 +/- 11.9 (4133)	65.8 +/- 12.2 (2662)	65.6 +/- 11.3 (1471)	0.6335	(ttest)
Systolic blood pressure	120.5 +/- 19.7 (4091)	119.4 +/- 19.6 (2662)	122.5 +/- 19.6 (1429)	<.0001	(ttest)
Sodium	139.6 +/- 4.6 (4030)	139.4 +/- 4.8 (2658)	140.0 +/- 4.3 (1372)	0.0001	(ttest)
Blood urea nitrogen	30.2 +/- 16.3 (3960)	31.1 +/- 17.1 (2662)	28.2 +/- 14.4 (1298)	<.0001	(ttest)
Death outcome_ 1 year	26.8% (765/2856)	26.7% (712/2662)	27.3% (53/194)	0.8619	
region	N=4133	N=2662	N=1471	<.0001	(chisq) df=3
EASTERN EUROPE	39.2% (1619)	34.8% (926)	47.1% (693)		
NORTH AMERICA	30.3% (1251)	35.1% (935)	21.5% (316)		
SOUTH AMERICA	16.9% (699)	17.0% (452)	16.8% (247)		
WESTERN EUROPE	13.6% (564)	13.1% (349)	14.6% (215)		

Figure S1. Originally Presented Point Scores described by the authors. These Predictive Models allow for calculation of individual event rates based on clinical variables.

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	<1%
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%

Reproduced with permission from: Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi F a. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. *Circ Cardiovasc Qual Outcomes*. 2010;3:25–32.

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

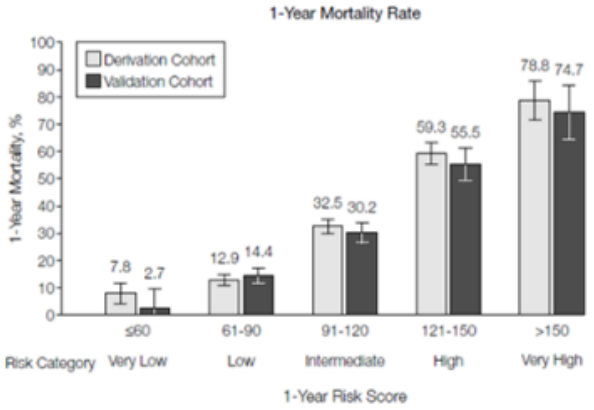
Age	Points	Sodium	Points	NYHA Class IV	Points
20	0	115	79	No	0
30	8	120	69	Yes	23
40	17	125	59		
50	25	130	49		
60	33	135	30	Total points	Predicted 60-day mortality
70	41	140	20		
80	50	145	10	124	2%
90	58	150	0	149	4%
				163	6%
SBP	Points	BUN	Points	174	8%
80	94	5	10	182	10%
90	86	10	20	208	20%
100	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				

Reproduced with permission from: Felker GM, Leimberger JD, Califf RM, Cuffe MS, Massie BM, Adams KF, Gheorghiade M, O'Connor CM. Risk stratification after hospitalization for decompensated heart failure. *J Card Fail*. 2004;10:460–466.

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

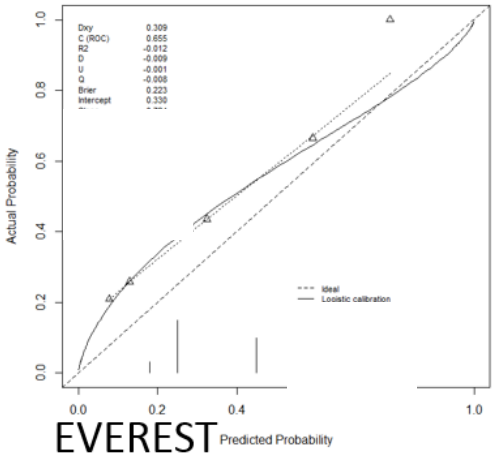
Abbreviation: NA, not applicable to 30-day model.
*An electronic version of the risk scoring system is available at: http://www.coort.ca/CI-Finkemodal.asp.
†Calculated as age + respiratory rate + systolic blood pressure + urea nitrogen + sodium points + cerebrovascular disease points + dementia points + 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.

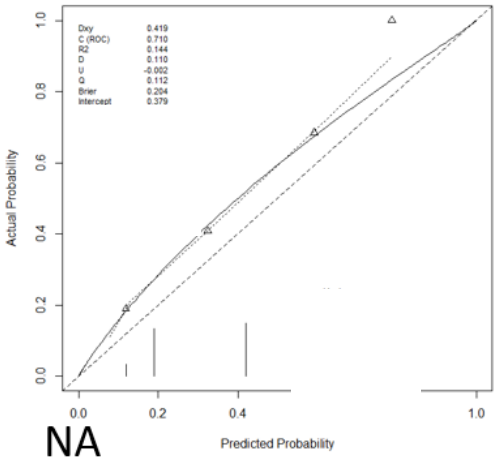


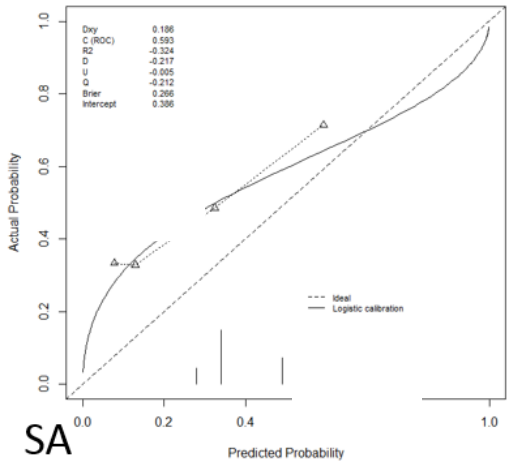
Reproduced with permission from: Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu J V. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *JAMA*. 2003;290:2581–7.

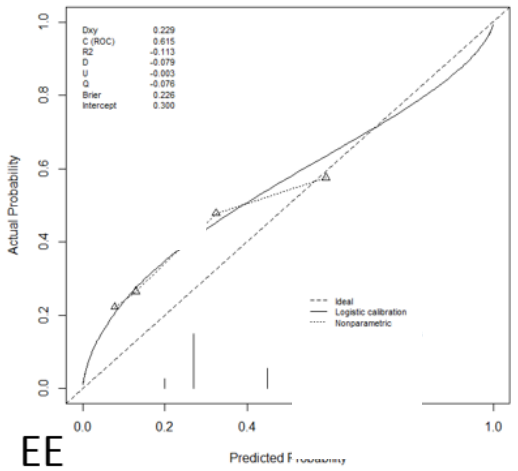
Figure S2a. Sensitivity Analysis of EFFECT CPM
Including only patients dead or alive with > 12 months of follow up

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









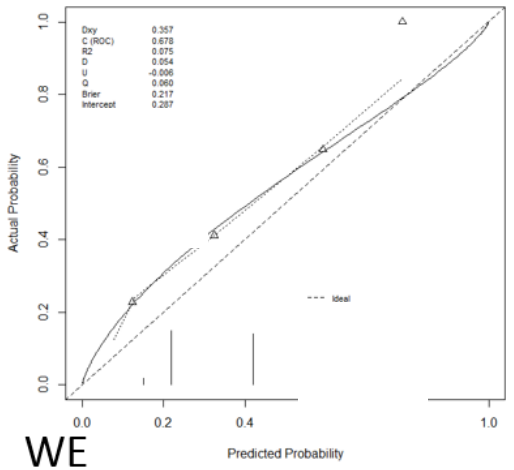
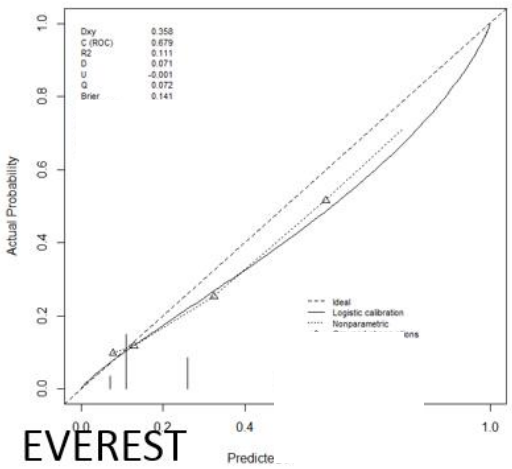
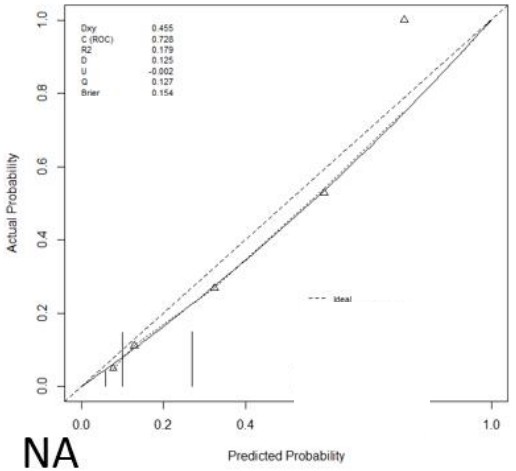


Figure S2b. Sensitivity Analysis of EFFECT CPM. Including only patients dead or alive with ≥ 6 months of follow up

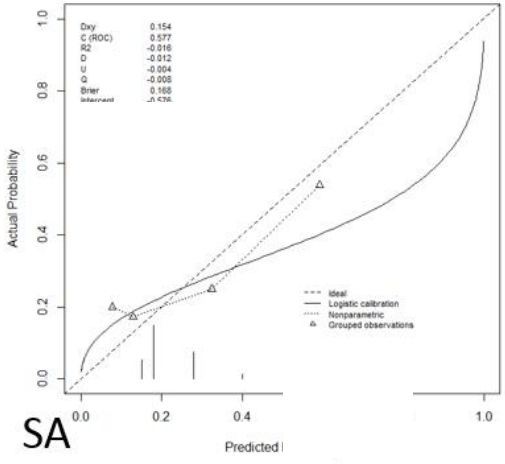
Region	AUC
EVEREST	0.68
North America	0.73
South America	0.58
Eastern Europe	0.64
Western Europe	0.71



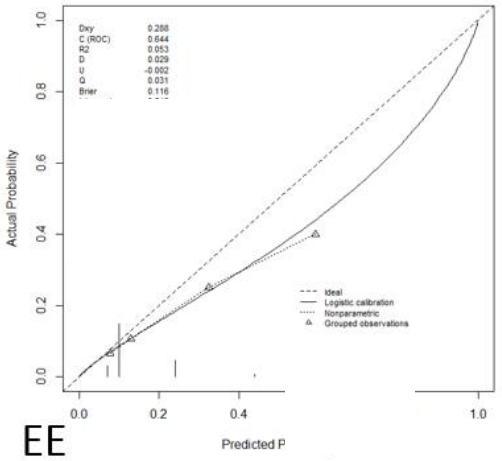
EVEREST



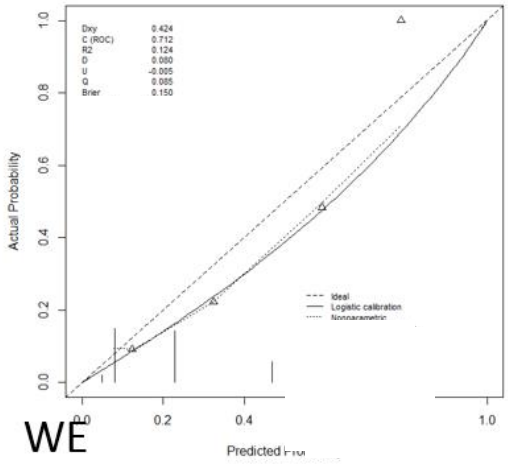
NA



SA



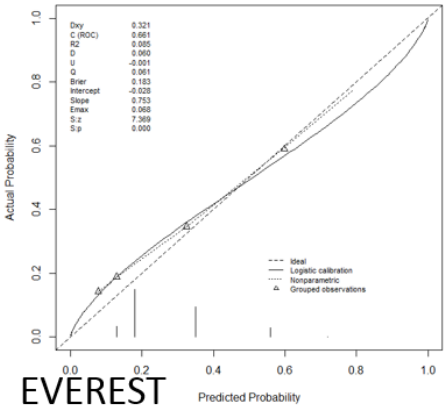
EE



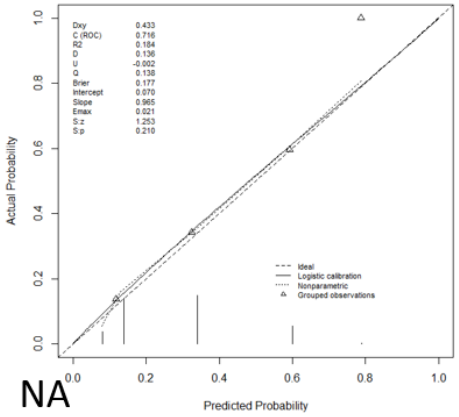
WE

Figure S2c. Sensitivity Analysis of EFFECT CPM (Including only patients dead or alive with > 9 months of follow up

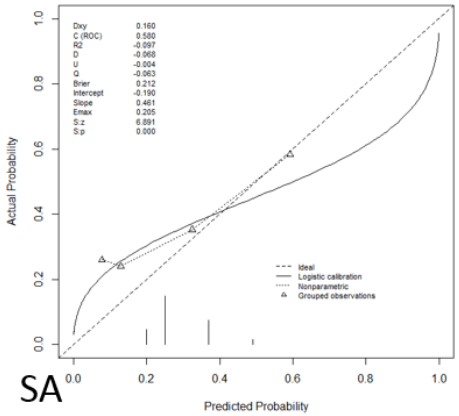
Region	AUC
EVEREST	0.66
North America	0.72
South America	0.58
Eastern Europe	0.62
Western Europe	0.69



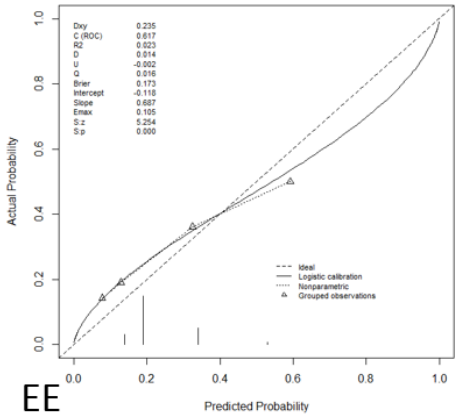
EVEREST



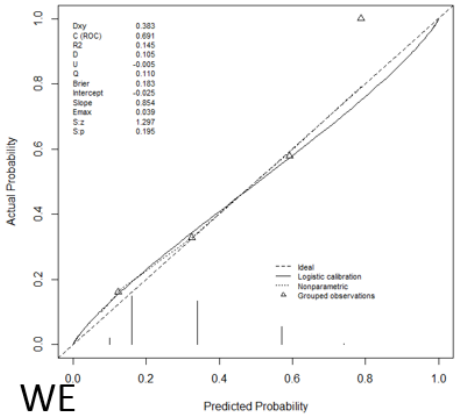
NA



SA



EE



WE

Figure S2d. Sensitivity Analysis of EFFECT CPM
Patient's status alive or dead imputed according to survival probability at last follow up n = 3881