How well do cardiovascular CPMs validate?

Benjamin Wessler MD, FACC
Assistant Professor of Medicine,
Associate Director, PACE Center,
Director, Valve Center, Tufts Medical Center

David M. Kent, MD, MSc
Professor of Medicine, Neurology, Clinical and Translational Science,
Director, Predictive Analytics and Comparative Effectiveness (PACE) Center,
Institute for Clinical Research and Health Policy Studies, Tufts Medical Center

Outline

- A Clinical Example of Prediction
- External Validations
  - Heart failure
  - Review of the Literature
- Fully Independent External Validations
- OHDSI– Pooled Cohort Equation results

A Problem with Predictions

There is no standard evaluation framework for CPMs and it is often merely assumed that predictions are trustworthy and accurate.

More importantly it is assumed that clinical decisions based on these predictions are superior to decisions made without these tools (i.e., lead to better outcomes).

Heart Failure

- 6.2 million people in United States have HF
- >650,000 new cases of HF diagnosed annually
- 50% morality within 5 years of diagnosis
- Total cost of HF in United States > $40 billion annually

Yancy et al. *Circulation* 2013, Virani et al. *Circulation* 2020

Heart Failure

Patient A
60 years old
SBP 110
HR 99
O2 98%
Cr 1.4
Troponin < 0.05
BNP 500

Patient B
85 years old
SBP 140
HR 84
O2 95%
Cr 0.9
Troponin 1.7
BNP 1400

ACCF/AHA Practice Guideline

2013 ACCF/AHA Guideline for the Management of Heart Failure

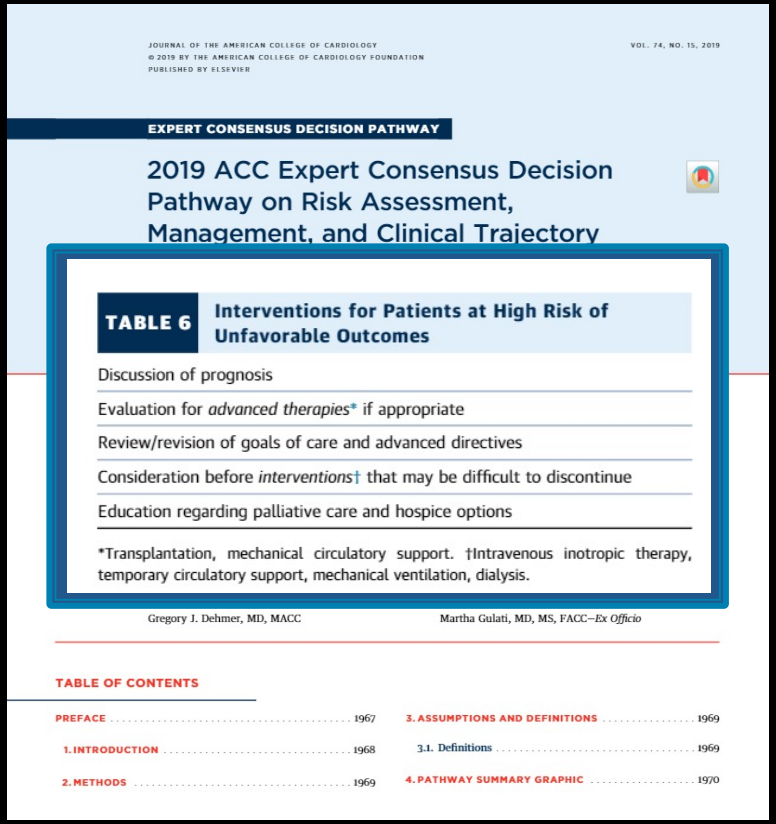
A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines

6.1.2. Risk Scoring: Recommendation

Class IIa

1. **Validated multivariable risk scores can be useful to estimate subsequent risk of mortality in ambulatory or hospitalized patients with HF.[199–207] *(Level of Evidence: B)***

Frederick A. Masoudi, MD, MSPH, FACC, FAHA†#; Patrick E. McBride, MD, MPH, FACC**;
John J.V. McMurray, MD, FACC*†; Judith E. Mitchell, MD, FACC, FAHA†;
Pamela N. Peterson, MD, MSPH, FACC, FAHA†; Barbara Riegel, DNSc, RN, FAHA†;
Flora Sam, MD, FACC, FAHA†; Lynne W. Stevenson, MD, FACC*†;
W.H. Wilson Tang, MD, FACC*†; Emily J. Tsai, MD, FACC†;
Bruce L. Wilkoff, MD, FACC, FHRS*††



JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2019 BY THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
PUBLISHED BY ELSEVIER

VOL. 74, NO. 15, 2019

EXPERT CONSENSUS DECISION PATHWAY

2019 ACC Expert Consensus Decision Pathway on Risk Assessment, Management, and Clinical Trajectory

TABLE 6 Interventions for Patients at High Risk of Unfavorable Outcomes

Discussion of prognosis
Evaluation for *advanced therapies** if appropriate
Review/revision of goals of care and advanced directives
Consideration before *interventions*† that may be difficult to discontinue
Education regarding palliative care and hospice options

*Transplantation, mechanical circulatory support. †Intravenous inotropic therapy, temporary circulatory support, mechanical ventilation, dialysis.

Gregory J. Dehmer, MD, MACC Martha Gulati, MD, MS, FACC—*Ex Officio*

TABLE OF CONTENTS

Yancy et al. *Circulation* 2013, Hollenberg et al. *JACC* 2019

CPM Performance: Discrimination

1.0 = perfect discrimination

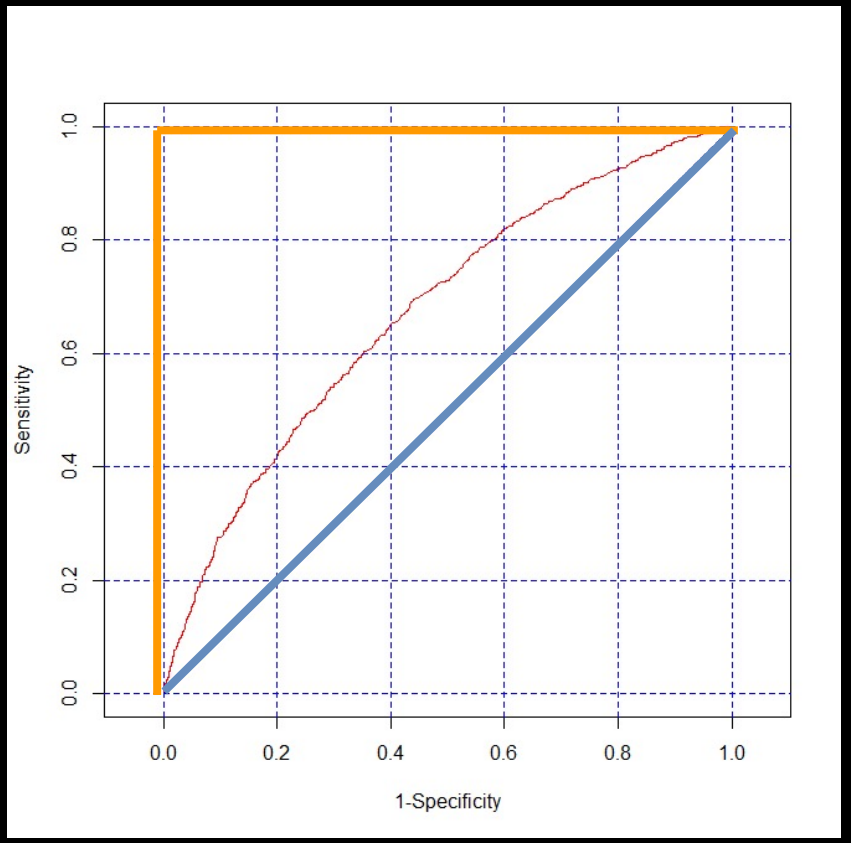
0.9

0.8 — Very good/ Excellent discrimination

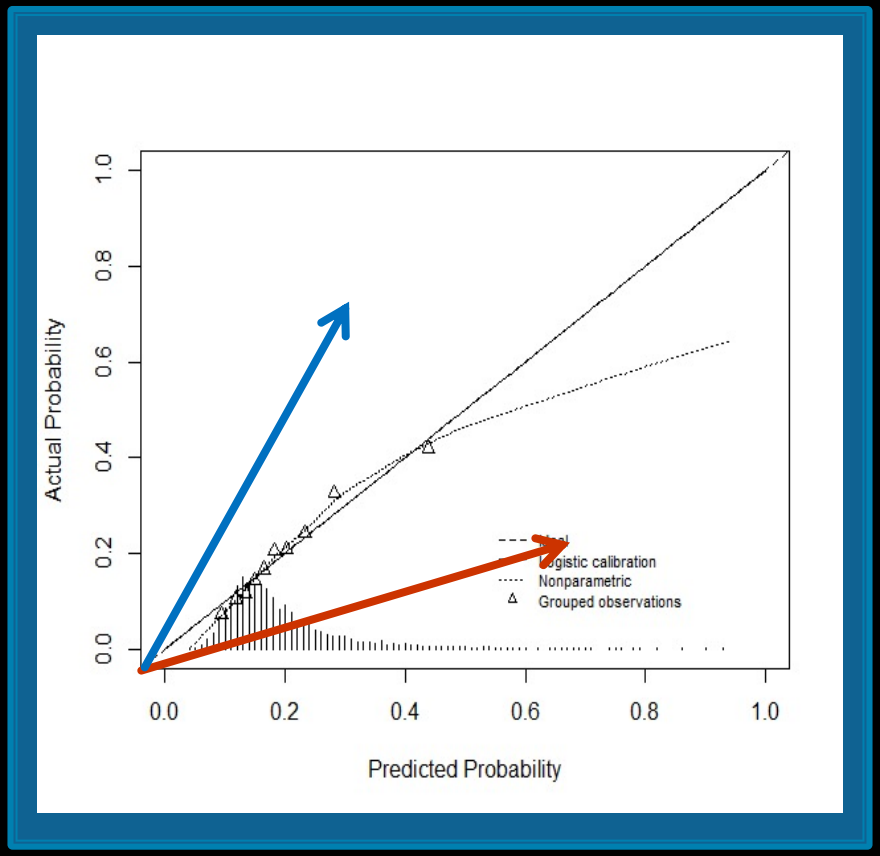
0.7

0.6

0.5 = coin flip



CPM Performance: Calibration

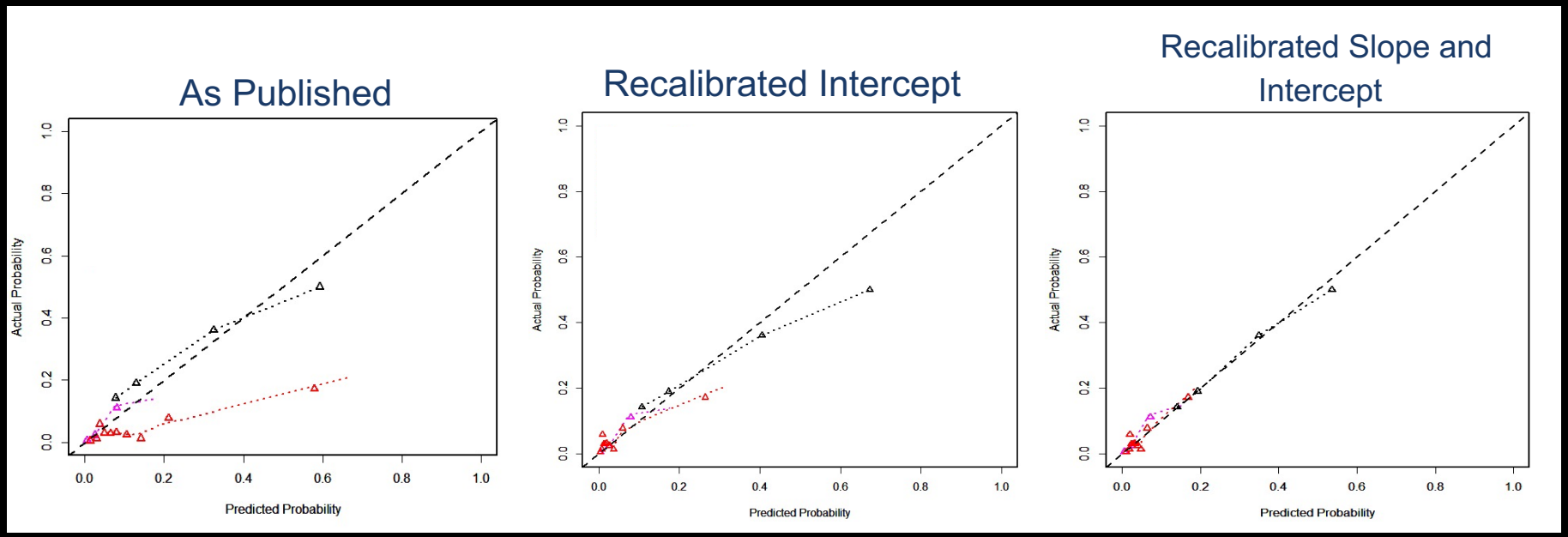


External Validations: Discrimination

CPM	Derivation AUC	Worldwide AUC	N. America AUC	S. America AUC	E. Europe AUC	W. Europe AUC
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%)

Wessler et al. *JAHA* 2017

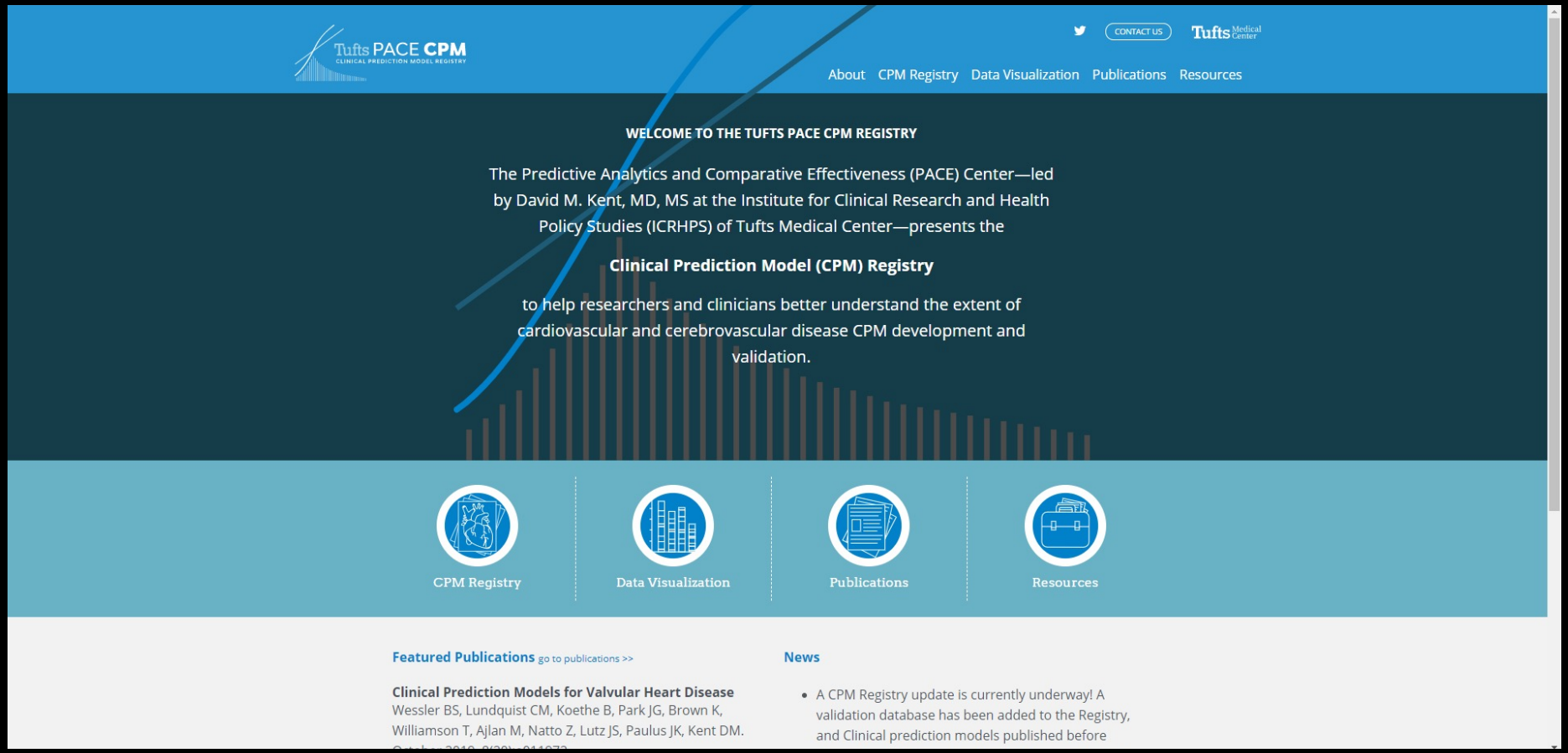
External Validations: Calibration



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

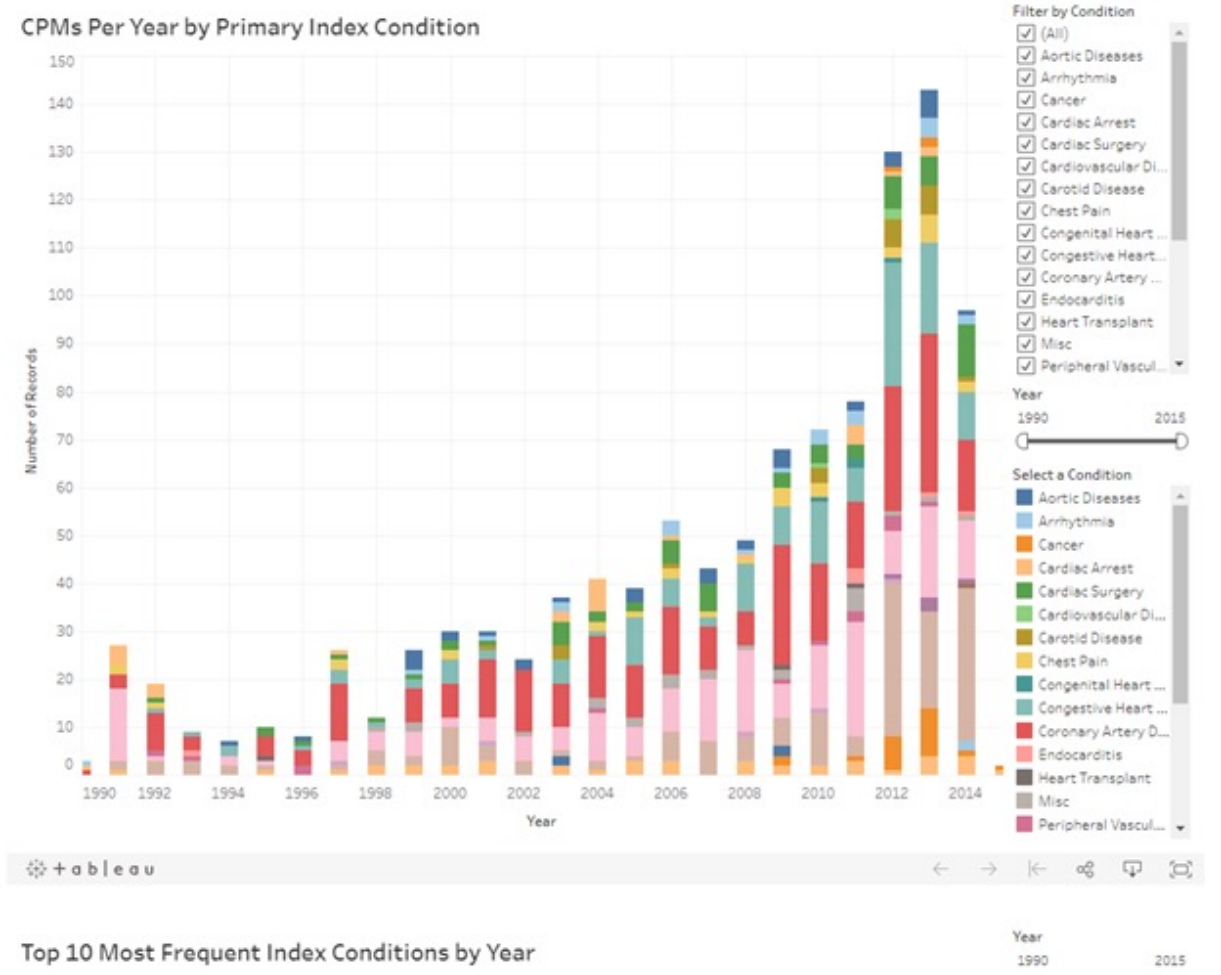
Wessler et al. *JAHA* 2017

Tufts PACE CPM Registry



Wessler et al. *Circ CQO* 2015

Tufts PACE CPM Registry



Cardiovascular CPMs broadly…

- 1382 clinical predictive models (CPMs)
  63% of de novo CPMs report a c-statistic

- We identified 2030 external validations of these CPMs

Wessler et al, CQO, in press

Cardiovascular CPMs broadly…

- Only 575 (42%) of the CPMs in the Registry have ever been externally validated.
- On average there were 1.5 validations per de novo CPM
- There was a very skewed distribution
  - The Logistic EuroSCORE has been validated 94 times

Roques et al. *Euro Heart J* 2003

Cardiovascular CPMs broadly…

Top 10 Most Validated CPMs				
Model Name	Index Condition	Number of validations	Median validation AUC (IQR)	Range in validation AUC
Logistic EuroSCORE	Cardiac Surgery	94	0.75 (0.67, 0.80)	0.48-0.90
Additive EuroSCORE	Cardiac Surgery	86	0.77 (0.72, 0.82)	0.58-0.90
EuroSCORE II	Valve Disease	65	0.76 (0.68, 0.81)	0.40-0.87
GRACE	CAD: ACS	53	0.80 (0.73, 0.84)	0.60-0.95
STS (valve) - Mortality	Cardiac Surgery	51	0.70 (0.64, 0.76)	0.45-0.85
CHA_2DS_2-VASc	Arrhythmia	45	0.66 (0.61, 0.69)	0.45-0.93
$CHADS_2$	Arrhythmia	37	0.65 (0.61, 0.68)	0.51-0.87
FRS - CHD	Population Sample	35	0.68 (0.63, 0.72)	0.54-0.80
ICH Score	Stroke	27	0.85 (0.75, 0.87)	0.69-0.94
ACEF Score	Cardiac Surgery	26	0.74 (0.68, 0.77)	0.54-0.87

Performance heterogeneity is the rule…

Wessler et al, CQO, in press

Cardiovascular CPMs broadly…

53% (n = 983) of the validations report some measure of CPM calibration.

- The Hosmer-Lemeshow test of goodness-of-fit was most commonly reported (30%), calibration-in-the-large (26%), and calibration plots (22%).

There is no external assessment of calibration for 86% (n = 1182) of Cardiovascular Predictive Models

Conclusions from Prelim Work

The tremendous proliferation and redundancy of CPMs is occurring without adequate—or even minimal—external evaluation.

Approximately 60% of published CPMs have never been externally validated. Approximately half of the CPMs that have been validated have been validated only once.

The value of single validations is unclear, since there is substantial performance heterogeneity and good (or poor) performance on a single validation does not appear to reliably forecast performance on subsequent validations.

Conclusions from Prelim Work

This work raises substantial concerns about the current approach to 'validating' cardiovascular CPMs.

There should be a major rethinking of how performance heterogeneity is explored and quantified and how cardiovascular CPMs are evaluated for clinical use.

Wessler et al, CQO, in press

Limitations of prelim work

- A major limitation of our literature review is that model performance is not generally presented in a way that makes it clear whether a given CPM is likely to improve or worsen decision making.
- Our main metric for model performance on external validation was the decrement in discrimination.
- The clinical significance of "change in discrimination" is unclear.

Outline

- A Clinical Example of Prediction
- External Validations
  - Heart failure
  - Review of the Literature
- Fully Independent External Validations
- OHDSI– Pooled Cohort Equation results

- We performed independent validations on a set of CPMs across 3 index conditions (acute coronary syndrome [ACS], heart failure [HF], and incident cardiovascular disease [CVD]) using publicly available clinical trial data and an evaluation framework.

Independent validations: including novel measures

- Model Based c-statistic
  - Standardizes for case mix
- Measures of calibration:
  - Harrell's E_{avg} and E_{90} (standardized)
- Measures of clinical utility:
  - Decision curve analysis

Use of Decision Curves

- Performance measures generally assess the quality of the predictions, not the quality of the decisions.
- ROC treats sensitivity and specific as equally important.

But false negatives are generally worse than a false positive

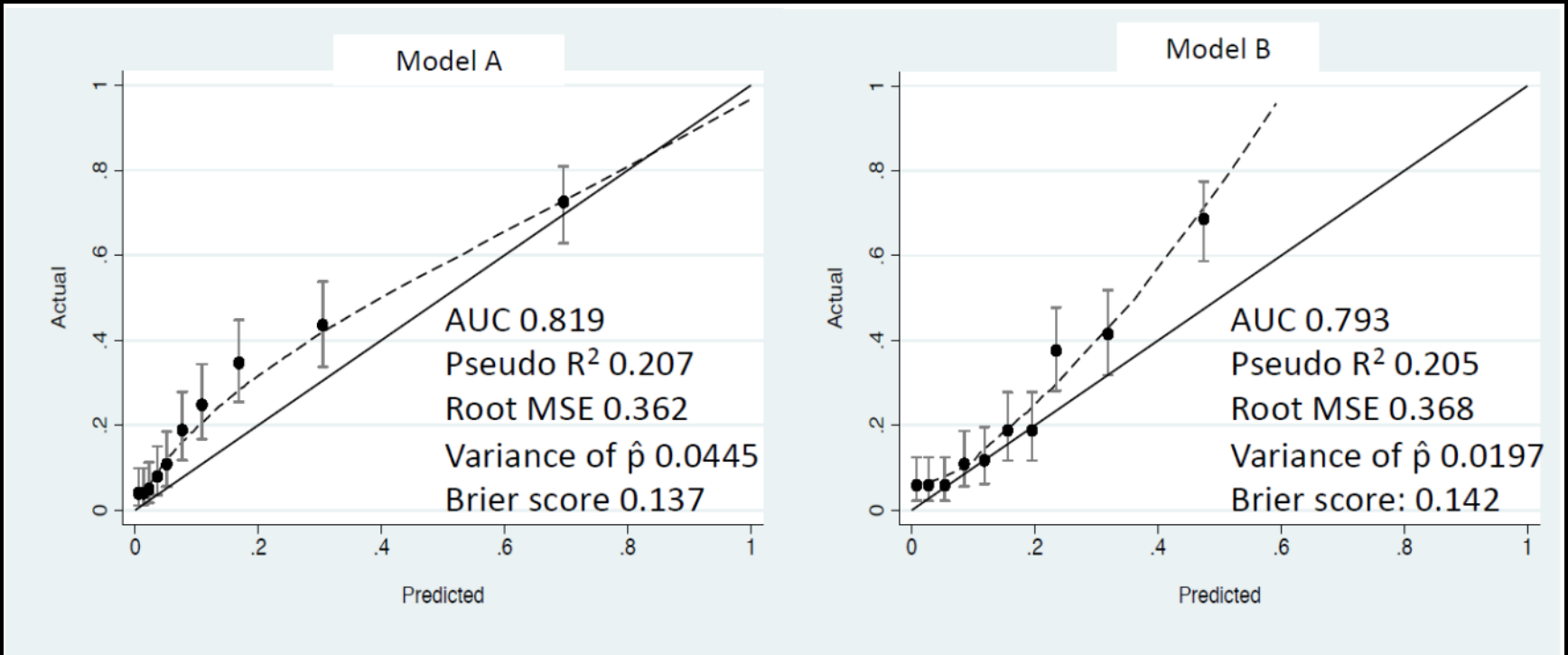
Decision curve analysis

1. Select a p_t
2. Positive test defined as $\hat{p} \geq p_t$
3. Calculate "Clinical Net Benefit" as:

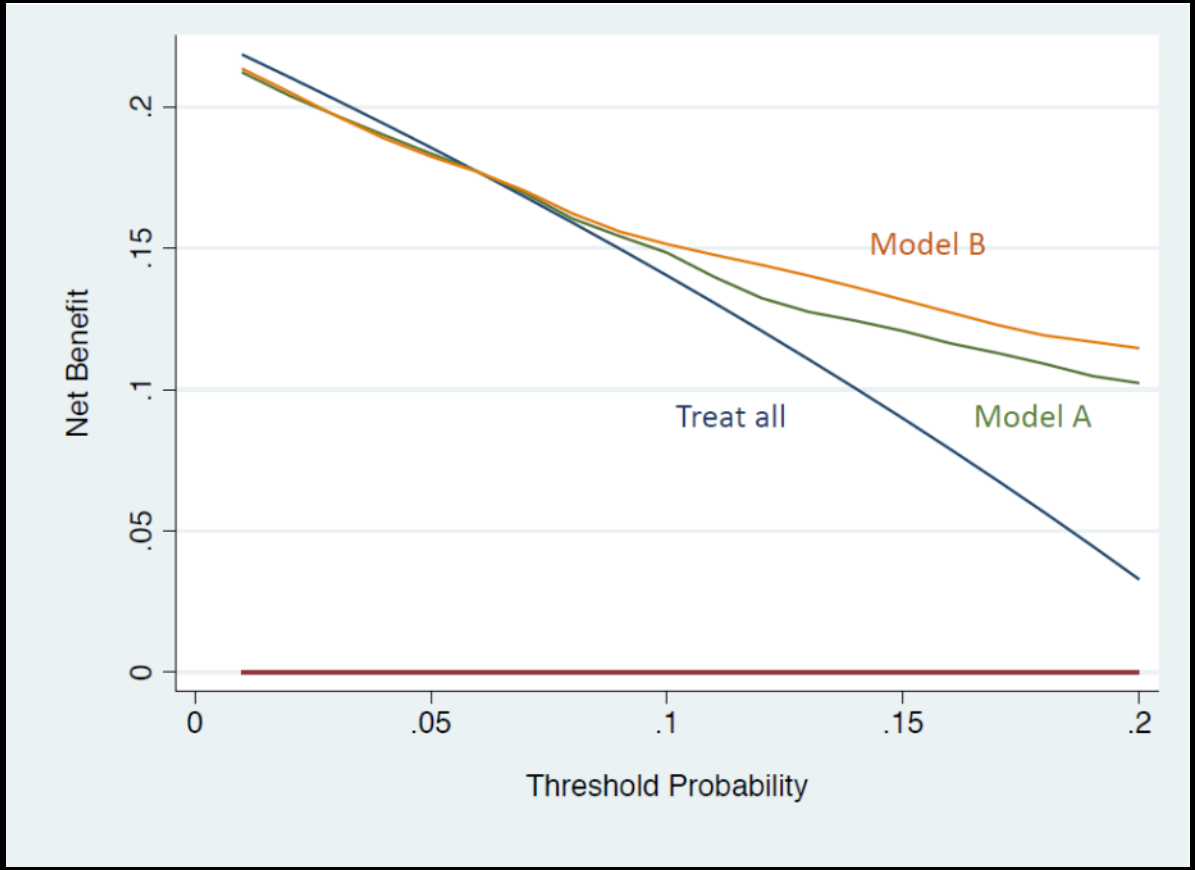
$$\frac{TruePositiveCount}{n} - \frac{FalsePositiveCount}{n}\left(\frac{p_t}{1-p_t}\right)$$

4. Vary p_t over an appropriate range

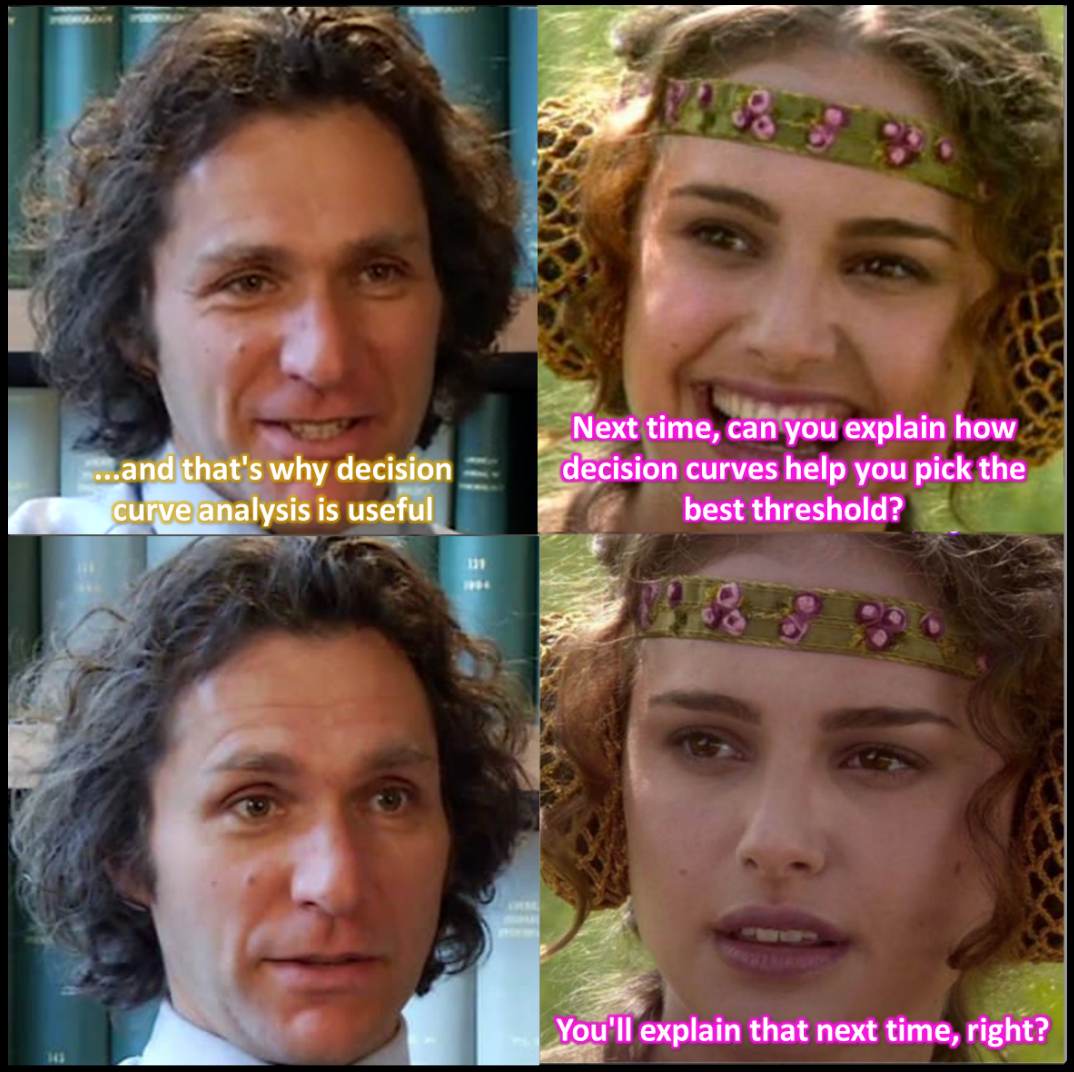
Scenario 1: Select patients for biopsy amongst men with elevated PSA



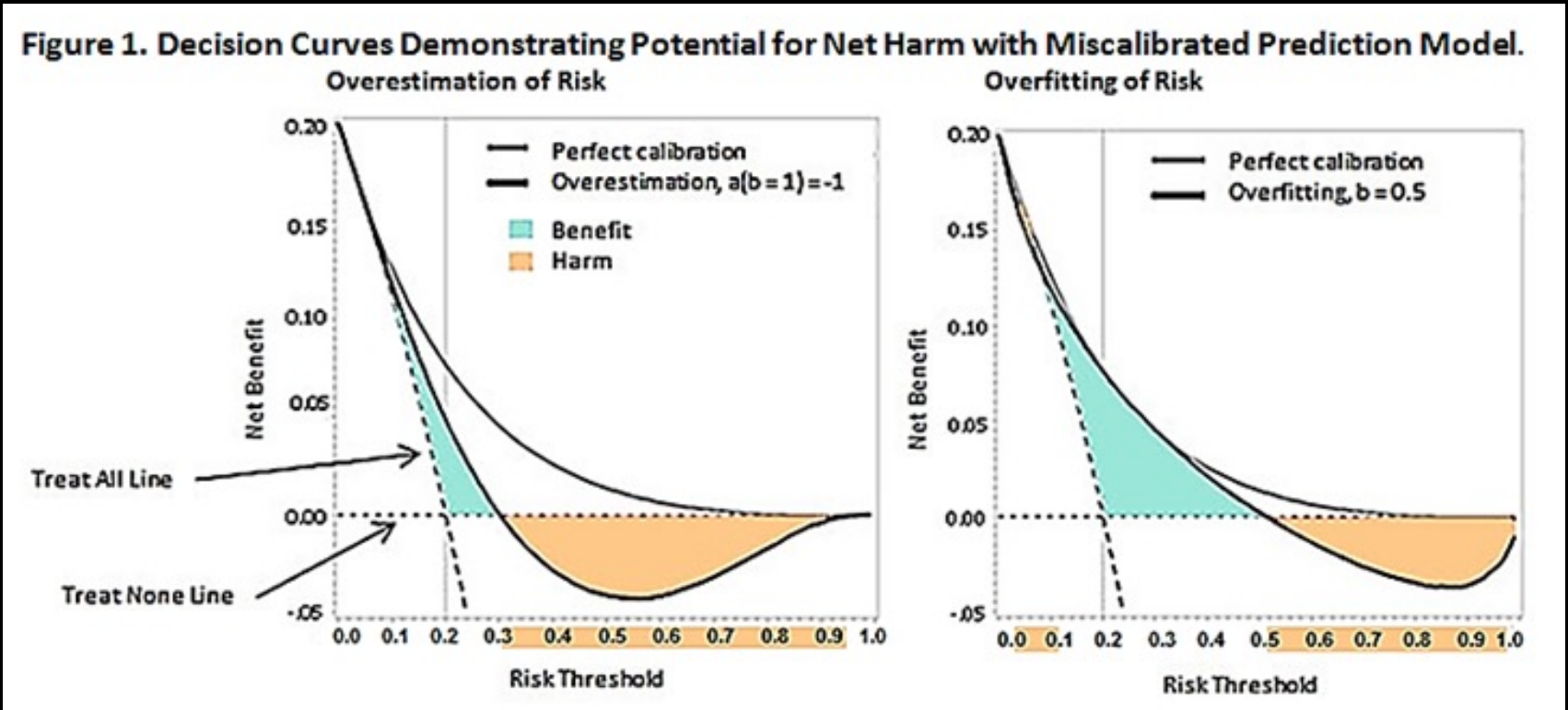
Scenario 1: Select patients for biopsy amongst men with elevated PSA



Decision curves do not help you pick the best decision threshold



Decision Curve: useful for detecting harm



Schematic showing how miscalibration leads to harm, and recalibration protects against harm.

Courtesy: Ben Van Calster

- We performed independent validations on a set of CPMs across 3 index conditions (acute coronary syndrome [ACS], heart failure [HF], and incident cardiovascular disease [CVD]) using publicly available clinical trial data and an evaluation framework.

36 Clinical Trials

Acute Coronary Syndrome

AMIS
ENRICHD
MAGIC
TIMI-II
TIMI-III

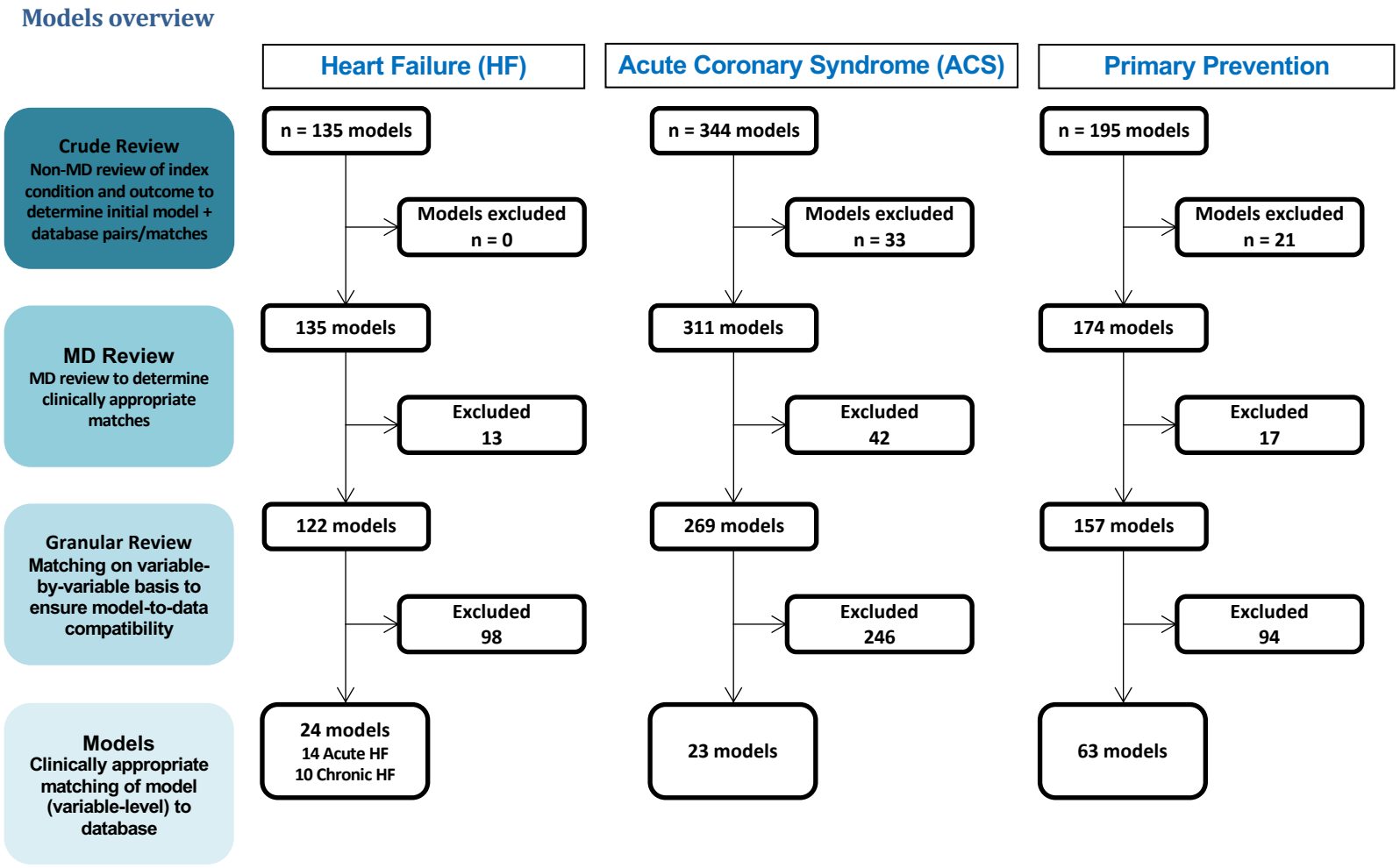
Heart Failure

BEST
DIG
EVEREST
TOPCAT
HEAAL
HF-ACTION
SCD-HeFT
SOLVD

Population Sample

ACCORD
ALLHAT-HTN
ALLHAT-LLT
WHI

108 unique CPM tested 158 times



Validation Performance

All Matches

(n = 158)	Mean (SD)	Median (IQR)	Range
Discrimination			
Development c-statistic	0.76 (0.05)	0.76 (0.73, 0.78)	0.63, 0.9
Validation model-based c-statistic (MBc)	0.68 (0.06)	0.68 (0.66, 0.71)	0.52, 0.84
Validation c-statistic	0.64 (0.06)	0.64 (0.6, 0.67)	0.44, 0.79
% Change in discrimination due to…			
Total (val.c vs. dev.c)	-46 (28)	-49 (-64, -29)	-138, 50
Case mix heterogeneity (MBc vs. dev.c)	-27 (22)	-28 (-39, -13)	-88, 55
Model validity (val.c vs. MBc)	-20 (60)	-24 (-43, -3)	-400, 400
Calibration (12.4% observed outcome rate)			
Slope	0.69 (0.33)	0.64 (0.48, 0.84)	0.17, 2.5
standardized E	0.9 (1.7)	0.5 (0.4, 0.7)	0, 14.2
standardized E90	1.5 (2.3)	1 (0.6, 1.3)	0, 14.6

*26 distantly related validations (population CPMs) are not assessed for calibration

Validation Performance

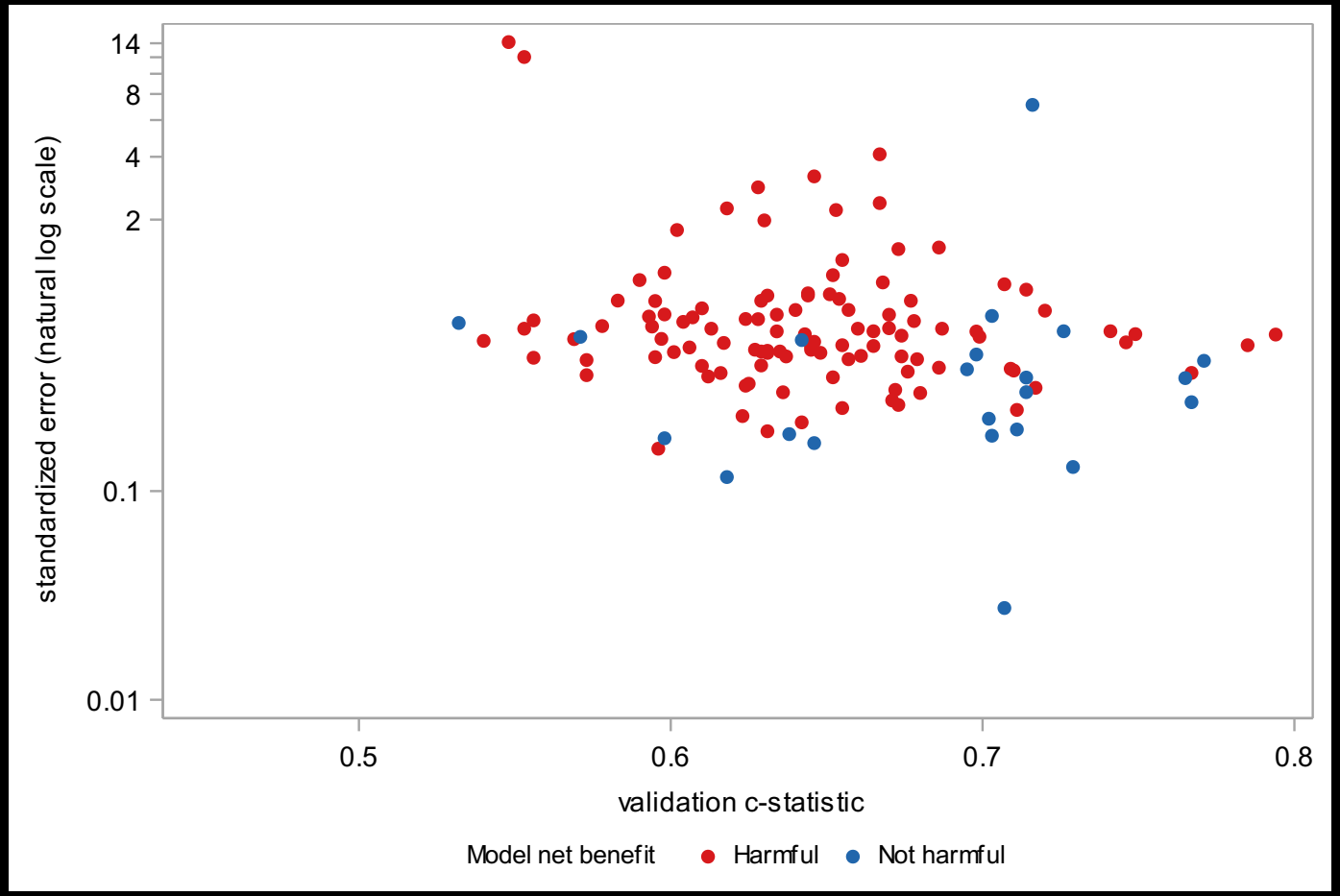
Net Benefit Above Default Strategy (All Matches)

Validation	Threshold	N	Compared to default strategy		
			% Above	% Neutral*	% Below
Original model	Prev./2	132	28.8	16.7	54.5
	Prevalence	132	85.6	6.8	7.6
	Prev.*2	132	26.5	29.5	43.9
Updated intercept	Prev./2	132	39.4	12.1	48.5
	Prevalence	132	100.0	0.0	0.0
	Prev.*2	132	49.2	10.6	40.2
Updated intercept and slope	Prev./2	132	52.3	28.8	18.9
	Prevalence	132	100.0	0.0	0.0
	Prev.*2	132	56.1	24.2	19.7
Re-estimated	Prev./2	132	68.2	14.4	17.4
	Prevalence	132	100.0	0.0	0.0
	Prev.*2	132	75.8	2.3	22.0

*Neutral defined as model NB equal to default strategy

Validation Performance

Harmful vs Not harmful (original model)



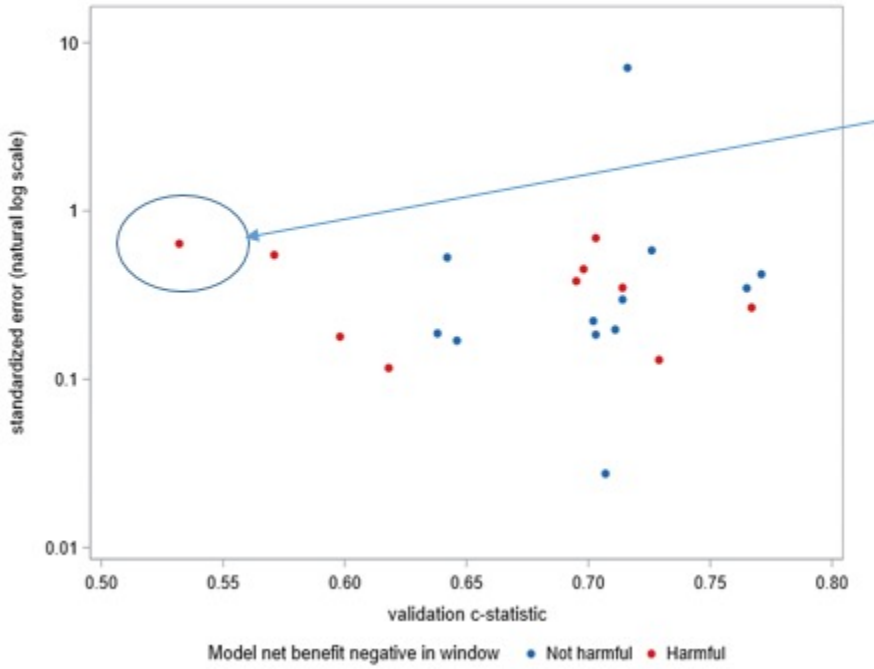
Net benefit	Harmful	**110**	NB below default at any of the 3 thresholds
	Not harmful	**22**	NB above default AND/OR neutral at all 3 of the thresholds

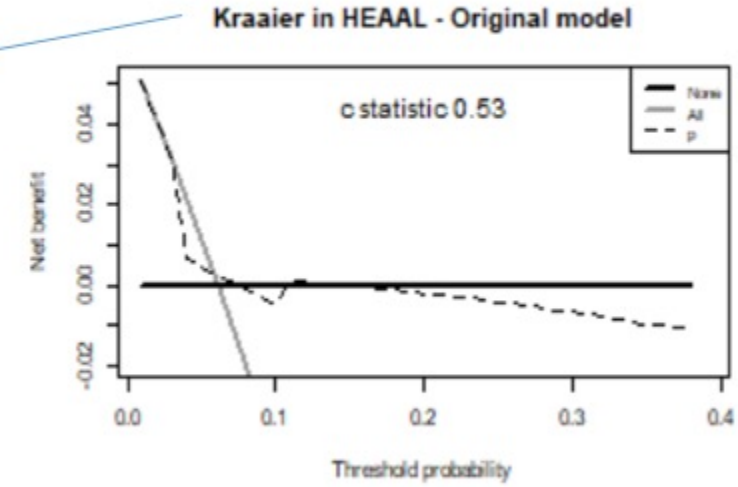
Validation Performance

Harmful vs Not harmful (original model)

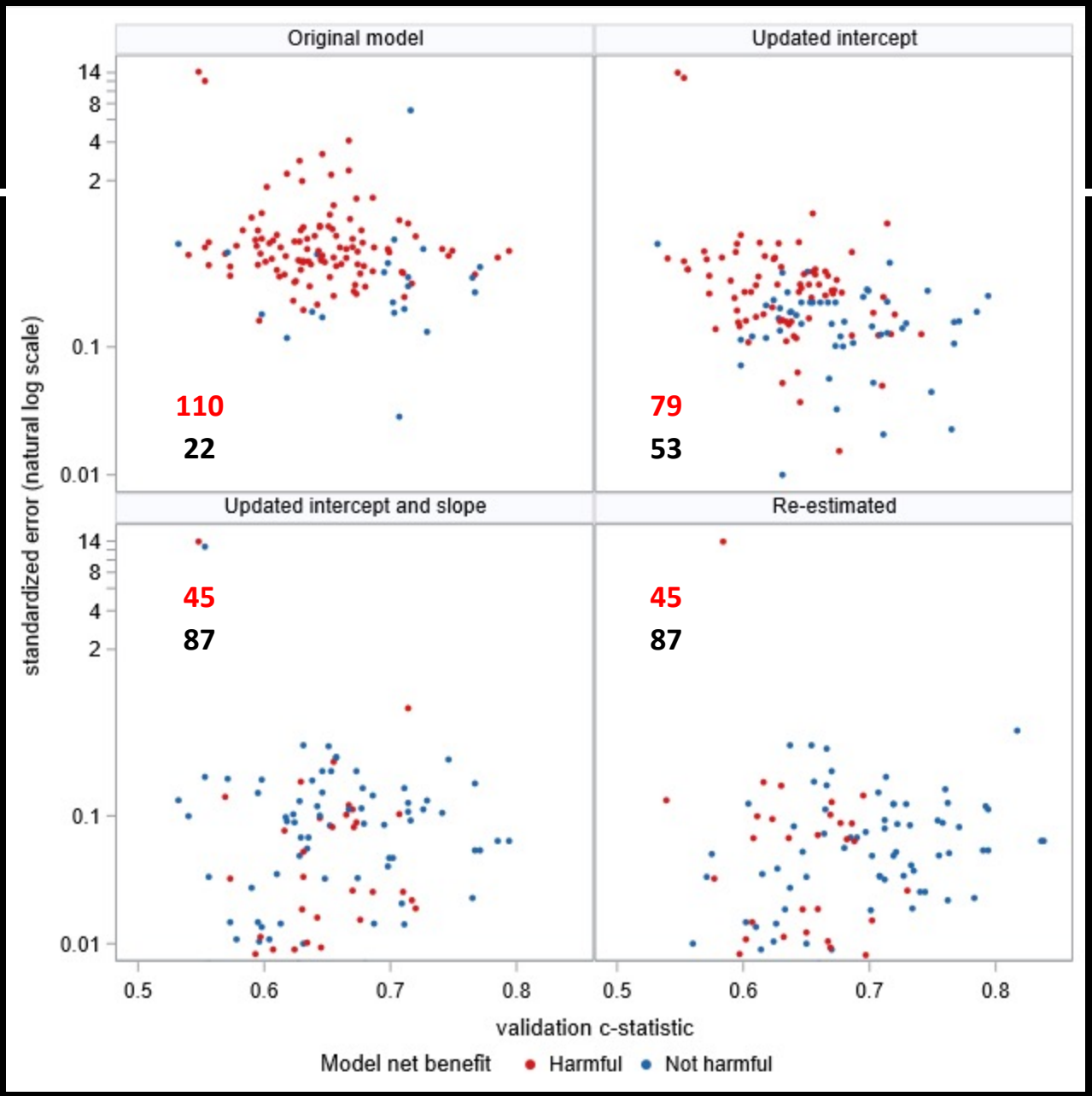
N=22 defined as "not harmful" (blue dots from previous figure)

"harmful" = any NB in the range from half to twice prevalence that is below default strategy





Net benefit	Not harmful	**12**	NB above default in threshold range
	Harmful	**10**	NB **below** default in threshold range

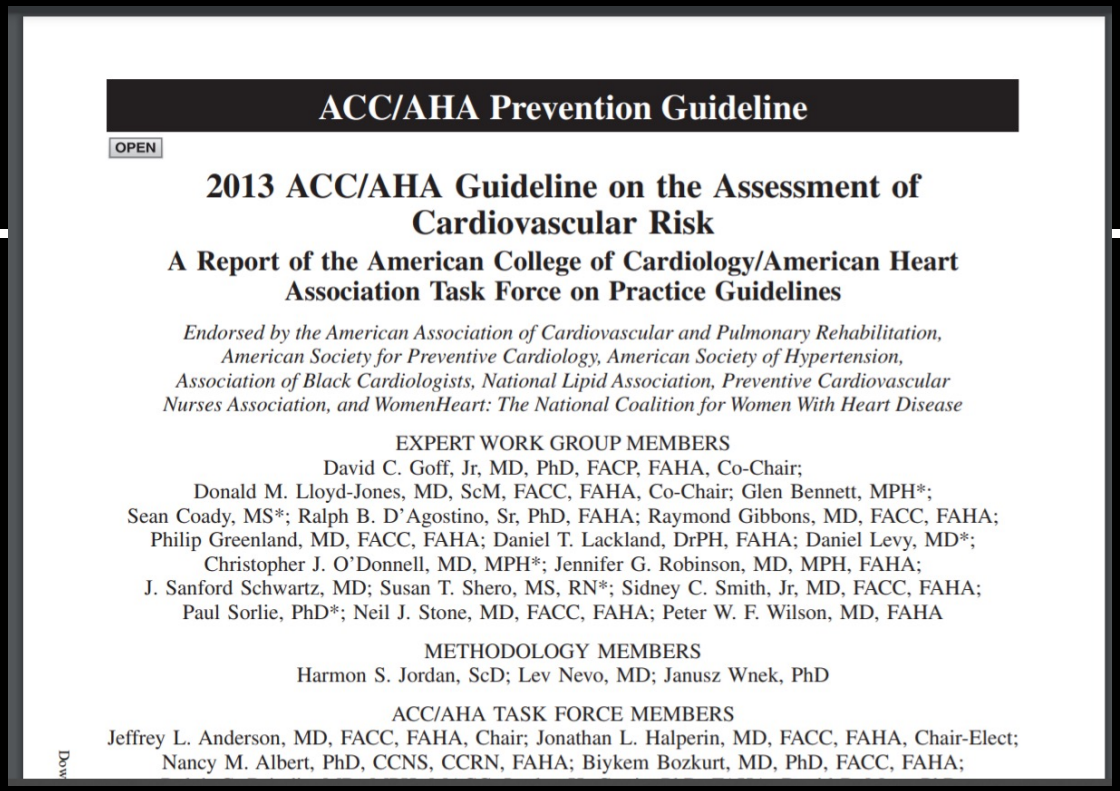

Original model
Updated intercept
Updated intercept and slope
Re-estimated
standardized error (natural log scale)
14
8
4
2
0.1
0.01
110
22
79
53
45
87
45
87
0.5
0.6
0.7
0.8
validation c-statistic
Model net benefit
Harmful
Not harmful

Outline

- A Clinical Example of Prediction
- External Validations
  - Heart failure
  - Review of the Literature
- Fully Independent External Validations
- OHDSI– Pooled Cohort Equation results

The Clinical Focus of PROTEUS

- Cholesterol is a major modifiable risk factor for experiencing MI or stroke.
- Statins are widely available to decrease cholesterol levels and reduce rates of MI and stroke
- Individual predicted risk of MI or stroke is used to inform treatment with statins



ACC/AHA Prevention Guideline

OPEN

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

Endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation, American Society for Preventive Cardiology, American Society of Hypertension, Association of Black Cardiologists, National Lipid Association, Preventive Cardiovascular Nurses Association, and WomenHeart: The National Coalition for Women With Heart Disease

EXPERT WORK GROUP MEMBERS

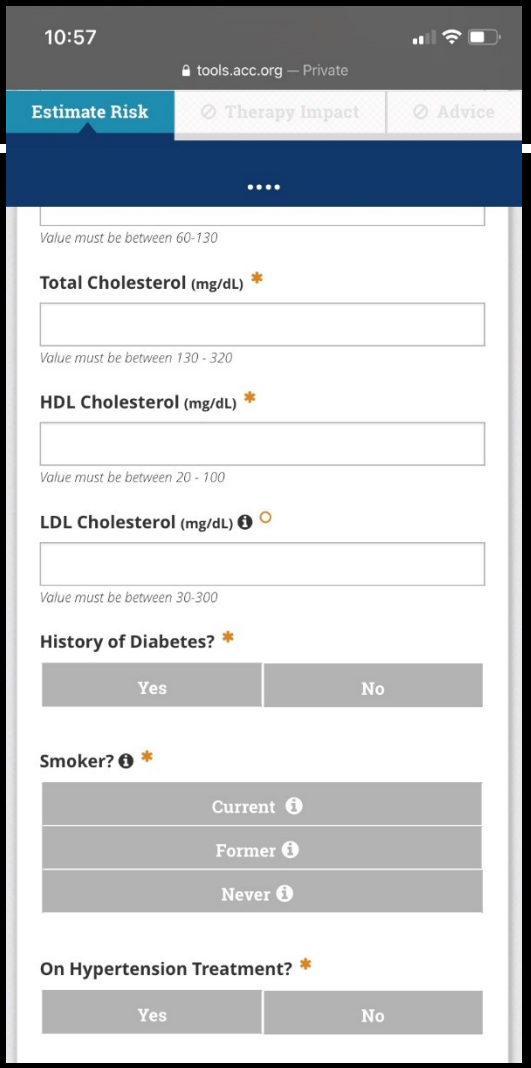
David C. Goff, Jr, MD, PhD, FACP, FAHA, Co-Chair; Donald M. Lloyd-Jones, MD, ScM, FACC, FAHA, Co-Chair; Glen Bennett, MPH*; Sean Coady, MS*; Ralph B. D'Agostino, Sr, PhD, FAHA; Raymond Gibbons, MD, FACC, FAHA; Philip Greenland, MD, FACC, FAHA; Daniel T. Lackland, DrPH, FAHA; Daniel Levy, MD*; Christopher J. O'Donnell, MD, MPH*; Jennifer G. Robinson, MD, MPH, FAHA; J. Sanford Schwartz, MD; Susan T. Shero, MS, RN*; Sidney C. Smith, Jr, MD, FACC, FAHA; Paul Sorlie, PhD*; Neil J. Stone, MD, FACC, FAHA; Peter W. F. Wilson, MD, FAHA

METHODOLOGY MEMBERS

Harmon S. Jordan, ScD; Lev Nevo, MD; Janusz Wnek, PhD

ACC/AHA TASK FORCE MEMBERS

Jeffrey L. Anderson, MD, FACC, FAHA, Chair; Jonathan L. Halperin, MD, FACC, FAHA, Chair-Elect; Nancy M. Albert, PhD, CCNS, CCRN, FAHA; Biykem Bozkurt, MD, PhD, FACC, FAHA;



Adults 40–75 years of age…with an estimated 10-year ASCVD risk ≥7.5% should be treated with moderate- to high-intensity statin therapy.

ACC/AHA Guideline on The Treatment of Blood Cholesterol

OHDSI Databases

- Asia
  - Ajou University School of Medicine Database (AUSOM)
  - Japan Medical Data Center (JMDC)
- Europe
  - Clinical Practice Research Datalink (CPRD)
  - Integrated Primary Care Information (IPCI)
- US
  - Columbia University Irving Medical Center Data Warehouse (CUIMC)
  - IBM MarketScan® Commercial Database (CCAE)
  - Optum® De-identified Clinformatic Data Mart Database – Date of Death (Optum DOD)
  - Optum® De-identified Electronic Health Record Dataset (Optum EHR)
  - The Stanford Medicine Research Data Repository (STARR-OMOP)
  - Tufts Research Data Warehouse (TRDW)

OHDSI Databases

Database	Treated: Systolic BP, mmHg (mean [SD])	Untreated: Systolic BP, mmHg (mean [SD])	Age (mean [SD])	Smoking (%)	HDL-C, mg/dL (mean [SD])	Male (%)	Total Cholesterol, mg/dL (mean [SD])
AUSOM	134.6 (21)	125.5 (15.3)	50.2 (9.5)	0	53.6 (13.3)	55.8	192.7 (33.1)
CPRD	143.9 (19.6)	134.6 (18.7)	55.2 (10.3)	93.1	56.3 (14.3)	46.1	214.5 (37.1)
CCAE	125.8 (12.9)	118.9 (12.4)	50 (6.9)	20.1	57.2 (14.8)	42	194 (31.7)
CUIMC	132.8 (18.3)	121.8 (15.4)	54.6 (10.7)	8.2	55.4 (14.9)	34.6	216.5 (35.7)
IPCI	146.3 (20.1)	137.5 (19.4)	57.1 (9.9)	4.4	55.6 (13.5)	44.1	216.5 (35.7)
JMDC	131.5 (15.7)	118.5 (15.3)	48.7 (7.5)	0.7	64.2 (14.6)	55.7	206.1 (32.3)
Optum DOD	129 (15.2)	121.7 (14.3)	50.3 (8)	63.4	55.8 (14)	47.2	196.1 (32.8)
Optum EHR	130.7 (16.5)	120.9 (13.9)	52.7 (10)	10.9	54.1 (14.6)	37.2	192.9 (31.4)
STARR-OMOP	133.4 (18)	123.1 (15.3)	55.1 (10.6)	13.9	56.2 (14.8)	39.8	195.1 (33.8)
TRDW	134.1 (18.3)	121.7 (14.5)	51.1 (9.9)	12.1	49.9 (13.9)	43.5	198.4 (34.5)

BP indicates blood pressure; SD, standard deviation.

Validation Results: Non-Black, Non-Female

Database	Outcomes	N	AUC	DAUC	Observed 3-year event rate (%)	standardized E (E_{avg}/event rate)	standardized E90 (E_{90}/event rate)	NB model @ threshold*
AUSOM	278	64997	0.816	-28.4	0.56	0.971	2.394	0.0003
CCAE	386	37321	0.668	31.9	1.47	0.249	0.307	0.0011
CPRD	9427	794858	0.732	5.9	1.44	1.544	2.78	-0.0013
CUIMC	179	9554	0.681	26.6	2.51	0.231	0.258	0.0063
IPCI	903	82028	0.67	31	1.49	1.012	2.626	-0.0002
JMDC	2232	390292	0.74	2.4	0.85	0.07	0.11	0.0008
OPTUMDOD	22	7248	0.789	-17.4	2.72	0.135	0.177	0.0158
OPTUMEHR	688	49606	0.785	-15.9	1.85	0.082	0.083	0.007
STARR-OMOP	670	44537	0.749	-1.2	1.96	0.067	0.204	0.0065
TUFTS	188	10867	0.752	-2.5	2.18	0.232	0.479	

*Net Benefit decision threshold 2.25% (3 year follow up)

Validation Results

Database	Outcomes	N	AUC	DAUC	Observed 3-year event rate (%)	standardized E (E_{avg}/event rate)	standardized E90 (E_{90}/event rate)	NB model @ threshold*
AUSOM	278	64997	0.816	-28.4	0.56	0.971	2.394	+
CCAE	386	37321	0.668	31.9	1.47	0.249	0.307	+
CPRD	9427	794858	0.732	5.9	1.44	1.544	2.78	-
CUIMC	179	9554	0.681	26.6	2.51	0.231	0.258	+
IPCI	903	82028	0.67	31	1.49	1.012	2.626	-
JMDC	2232	390292	0.74	2.4	0.85	0.07	0.11	+
OPTUMDOD	22	7248	0.789	-17.4	2.72	0.135	0.177	+
OPTUMEHR	688	49606	0.785	-15.9	1.85	0.082	0.083	+
STARR-OMOP	670	44537	0.749	-1.2	1.96	0.067	0.204	+
TUFTS	188	10867	0.752	-2.5	2.18	0.232	0.479	

*Net Benefit decision threshold 2.25% (3 year follow up)

A Note about Model Performance

- There are several potential reasons why model performance might decrease
  - Overfitting
  - Changes in case mix

Vergouwe Y. et al, Am J Epidemiol 2010

Conclusions

- All measures of performance are highly variable
- For databases where PCE was highly miscalibrated, model use to support decision making would lead to net harm
- Re-calibration guards against harm

Acknowledgements

- 262 OHDSI Symposium participants
- Evan Minty
- Jenna Reps
- Andrew Williams
- Patrick Ryan
- Jason Nelson
- And many others…