AI and Machine Learning Have Significant Limitations – A Word of Caution

Mamas A. Mamas

Professor of Cardiology

University of Keele

@MMamas1973

Disclosure of Relevant Financial Relationships

I, Mamas Mamas DO NOT have any relevant financial relationships to disclose relevant to this talk.

Editorials

Looking Back on the Millennium in Medicine

Application of Statistics to Medicine

A natural starting point for a history of biostatistical thought in the past millennium is the work of Leonardo Fibonacci (c. 1170– after 1240), an Italian mathematician of the Middle Ages. By introducing Indian and Arabic mathematics and numbering to Europe in 1202, he freed Western thought from the limitations of the Roman-numeral system. This advance laid the foundation for modern computation and bookkeeping. Probability theory emerged only in the 16th and 17th centuries, when Pierre de Fermat (1601–1665) and Blaise Pascal (1623–1662) developed basic probabilistic calculations to analyze games of chance. Ideas of relative frequency were first applied to mortality statistics in 17th-century London at the time of the plague. John Graunt (1620–1674) introduced the notion of inference from a sample to an underlying population and described calculations of life expectancy that launched the insurance industry in the 17th and 18th centuries.

The German mathematician Karl Friedrich Gauss (1777–1855) played a central part in the development of modern statistical reasoning. His method of least-squares analysis, developed around 1794, underlies much of modern regression analysis. Thomas Bayes (1702–1761), the 18th-century English theologian and mathematician, was the first to show how probability can be used in inductive reasoning.

One of the earliest clinical trials took place in 1747, when James Lind treated 12 scorbutic ship passengers with cider, an elixir of vitriol, vinegar, sea water, oranges and lemons, or an electuary recommended by the ship's surgeon. The success of the citrus-containing treatment eventually led the British Admiralty to mandate the provision of lime juice to all sailors, thereby eliminating scurvy from the navy. The origin of modern epidemiology is often traced to 1854, when John Snow demonstrated the transmission of cholera from contaminated water by analyzing disease rates among citizens served by the Broad Street Pump in London's Golden Square. He arrested the further spread of the disease by removing the pump handle from the polluted well.

Biostatistical reasoning developed rapidly in Great Britain in the late 19th and early 20th centuries. Sir Ronald Fisher (1890–1962), the most important figure in modern statistics, developed the analysis of variance and multivariate analysis. He also introduced the principle of randomization as a method for avoiding bias in experimental studies. In the United States, Jerzy Neyman, a Russian immigrant, developed the theories of estimation and testing that shaped contemporary biostatistical practice.

A landmark of quantitative observational research as a tool for exploring the determinants of disease was Sir Richard Doll's study of smoking among British physicians. Randomized clinical trials emerged in England in the 1950s and were adopted by the National Institutes of Health in the United States in the early 1960s; there followed an explosion of clinical trials of treatment for cancer, heart disease, diabetes, and other diseases. Biostatistical methods expanded rapidly during this period. Sir David Cox's 1972 paper on proportional-hazards regression ignited the fields of survival analysis and semiparametric inference (using partial specification of the probability distribution of the outcomes under investigation). Rapid improvements in computer support were essential to the growing role of empirical investigation and statistical inference.

Are machine learning models really superior to traditional approaches?

Keele Cardiovascular Research Group

Table 4. Discrimination and Calibration of the Models for Predicting 10-Year Risk of Developing Heart Failure in the Validation Cohorts Among Black Adults and White Adults

	ARIC			MESA/DHS		
	C-index (95% CIs)	GND χ^2 (*P* value)	DeLong test (*P* value)*	C-index (95% CIs)	GND χ^2 (*P* value)	DeLong test (*P* value)*
Black adults						
ML risk score	0.80 (0.75–0.84)	10.1 (0.26)	Ref	0.83 (0.77–0.87)	11.7 (0.17)	Ref
ARIC-HF risk score	0.77 (0.73–0.80)	8.9 (0.35)	<0.001	0.80 (0.76–0.84)	29.8 (<0.001)	0.01
PCP-HF risk score	0.73 (0.69–0.77)	14.4 (0.07)	<0.001	0.75 (0.71–0.79)	16.1 (0.04)	<0.001
MESA-HF risk score	0.72 (0.67–0.75)	6.9 (0.55)	<0.001	0.78 (0.74–0.82)	6.0 (0.54)	0.006
White adults						
ML risk score	N/A			0.82 (0.78–0.86)	9.9 (0.27)	Ref
ARIC-HF risk score				0.79 (0.76–0.81)	25.5 (0.001)	0.008
PCP-HF risk score				0.75 (0.71–0.79)	19.1 (0.01)	<0.001
MESA-HF risk score				0.80 (0.76–0.83)	8.33 (0.40)	0.044

ARIC indicates Atherosclerosis Risk in Communities; DHS, Dallas Heart Study; GND, Greenwood-Nam-D'Agostino; HF, heart failure; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis; ML, machine learning; N/A, not applicable; PCP-HF, Pooled Cohort Equations-Heart Failure; and Ref, reference.

*The DeLong test of C-index compared with the ML risk score model.

Circulation

ORIGINAL RESEARCH ARTICLE

Development and Validation of Machine Learning–Based Race-Specific Models to Predict 10-Year Risk of Heart Failure

A Multicohort Analysis

Matthew W. Segar, MD, MS; Byron C. Jaeger, PhD; Kershaw V. Patel, MD; Vijay Nambi, MD, PhD; Chiadi E. Ndumele, MD; Adolfo Correa, MD; Javed Butler, MD, MPH, MBA; Alvin Chandra, MD; Colby Ayers, MS; Shreya Rao, MD, MPH; Alana A. Lewis, MD; Laura M. Raffield, PhD; Carlos J. Rodriguez, MD, MPH; Erin D. Michos, MD, MHS; Christie M. Ballantyne, MD; Michael E. Hall, MD; Robert J. Mentz, MD; James A. de Lemos, MD; Ambarish Pandey, MD, MSCS

Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation

Glen P. Martin, MSc,[a] Matthew Sperrin, PhD,[a] Peter F. Ludman, MA, MD, FRCP, FESC,[b] Mark A. de Belder, MA, MD, FRCP,[c] Chris P. Gale, PhD, FRCP, FESC,[d] William D. Toff, MD, FRCP, FESC,[e,f] Neil E. Moat, MBBS, MS,[g] Uday Trivedi, MBBS,[h] Iain Buchan, MD, FFPH,[a] and Mamas A. Mamas, MA, DPhil, FRCP[a,i] *Manchester, Birmingham, Middlesbrough, Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds; Leicester, London, Brighton and Sussex University Hospitals, Brighton, and Stoke-on-Trent, United Kingdom*

Table III. Calibration, discrimination and Brier score for 30-day mortality in the whole cohort

Risk model	Calibration intercept (95% CI)*	Calibration slope (95% CI)	AUC (95% CI)	Brier score
LES	−1.75 (−1.86, −1.64)	0.35 (0.23, 0.48)	0.57 (0.54, 0.61)	0.093
ESII	−0.47 (−0.59, −0.36)	0.40 (0.28, 0.53)	0.59 (0.55, 0.62)	0.054
STS	**0.07 (−0.04, 0.18)**	0.56 (0.42, 0.71)	0.60 (0.57, 0.63)	0.051
German AV	−0.36 (−0.47, −0.25)	0.44 (0.32, 0.57)	0.59 (0.56, 0.62)	0.053
FRANCE-2	−0.60 (−0.71, −0.49)	0.69 (0.53, 0.86)	0.62 (0.59, 0.65)	0.053
OBSERVANT	−0.31 (−0.42, −0.20)	0.39 (0.25, 0.53)	0.57 (0.54, 0.60)	0.052
ACC TAVI	**0.04 (−0.07, 0.15)**	0.67 (0.52, 0.82)	0.64 (0.60, 0.67)	0.051

*The reported calibration intercept is that estimated assuming a slope of one; satisfactory calibration would occur if the 95% confidence intervals for the calibration intercept and slope span zero and one respectively. Bold items indicate that the 95% CI spans the corresponding reference value.

OPEN ACCESS

ORIGINAL RESEARCH ARTICLE

Novel United Kingdom prognostic model for 30-day mortality following transcatheter aortic valve implantation

Glen P Martin,[1] Matthew Sperrin,[1] Peter F Ludman,[2] Mark A de Belder,[3] Simon R Redwood,[4] Jonathan N Townend,[2] Mark Gunning,[5] Neil E Moat,[6] Adrian P Banning,[7] Iain Buchan,[1] Mamas A Mamas[1,5]

Keele Cardiovascular Research Group

Table 3 Variables and coefficients included in the final multivariable UK-TAVI CPM

Variable*	Coefficient (SE)	OR (95% CI)
Intercept	−3.6119 (0.1995)	NA
Mean-centred age	0.0115 (0.0085)	1.012 (0.995 to 1.028)
Female	0.1393 (0.1174)	1.150 (0.913 to 1.447)
Mean-centred BMI	−0.0257 (0.0119)	0.975 (0.952 to 0.998)
Mean-centred BMI squared	0.0011 (0.0007)	1.001 (1.000 to 1.002)
Glomerular filtration rate per 5 units increase	−0.0342 (0.0139)	0.966 (0.940 to 0.993)
Pulmonary disease	0.2140 (0.1266)	1.239 (0.966 to 1.588)
Extracardiac arteriopathy	0.1912 (0.1348)	1.211 (0.930 to 1.577)
Sinus preoperative heart rhythm	−0.1798 (0.1193)	0.835 (0.661 to 1.056)
Prior BAV	0.2469 (0.1633)	1.280 (0.930 to 1.763)
Critical preoperative status	0.5914 (0.2770)	1.807 (1.050 to 3.109)
Poor mobility	0.6302 (0.2052)	1.878 (1.256 to 2.808)
KATZ (per point drop from 6 points)	0.2362 (0.0689)	1.267 (1.107 to 1.450)
PA systolic pressure >60 mm Hg	0.1867 (0.1583)	1.205 (0.884 to 1.644)
Non-elective procedure	0.3719 (0.1554)	1.451 (1.070 to 1.967)
Non-transfemoral access	0.5436 (0.1268)	1.722 (1.343 to 2.208)

*Variable definitions are given in online supplementary table 1.
BAV, balloon aortic valvuloplasty; BMI, body mass index; CPM, clinical prediction model; NA, not applicable; PA, pulmonary artery; TAVI, transcatheter aortic valve implantation.

Table 4 Performance measures before (apparent) and after bootstrap-corrected optimism within the 2013–2014 data (n=2969)

Validation	Calibration intercept (95% CI)	Calibration slope (95% CI)	AUC (95% CI)
Apparent	0.00 (−0.18 to 0.18)	1.00 (0.76 to 1.24)	0.70 (0.65 to 0.75)
Internal*	0.02 (−0.17 to 0.20)	0.79 (0.55 to 1.03)	0.66 (0.61 to 0.71)

*Estimated as the apparent performance minus optimism, where optimism was obtained through bootstrap resampling.
AUC, area under the curve.

JAMA Network | Open

Original Investigation | Cardiology

Comparison of Machine Learning Methods With National Cardiovascular Data Registry Models for Prediction of Risk of Bleeding After Percutaneous Coronary Intervention

Bobak J. Mortazavi, PhD; Emily M. Bucholz, MD, PhD, MPH; Nihar R. Desai, MD, MPH; Chenxi Huang, PhD; Jeptha P. Curtis, MD; Frederick A. Masoudi, MD, MSPH; Richard E. Shaw, MA, PhD; Sahand N. Negahban, PhD; Harlan M. Krumholz, MD, SM

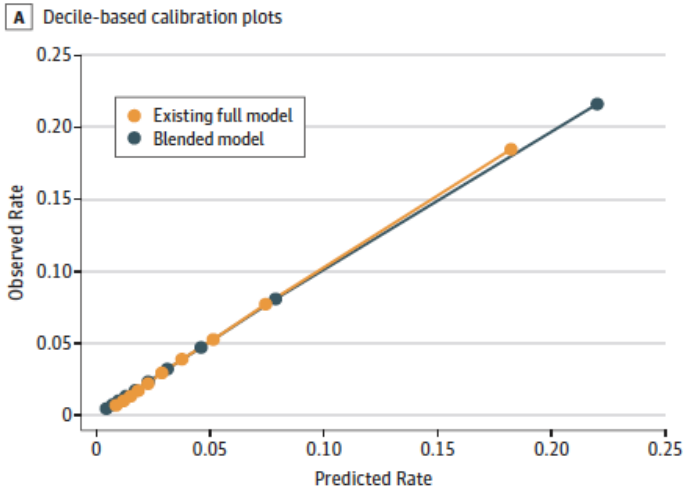
Keele Cardiovascular Research Group

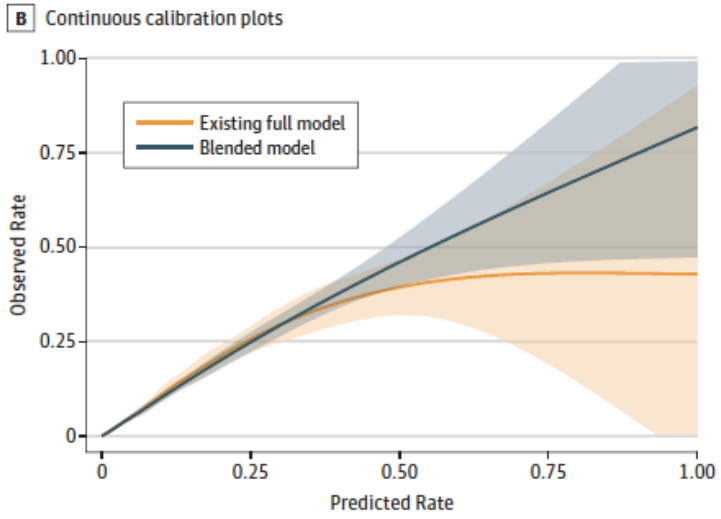
Table 3. C Statistics of 5-Fold Cross-validation Results for the Existing Simplified Risk Score and the Blended Model

Timing	Variable Set	Mean (95% CI) C Statistic
Existing simplified risk score	Existing simplified risk score	0.77 (0.77-0.77)
	Existing simplified risk score with lasso regularization	0.77 (0.77-0.77)
	Existing simplified risk score with gradient descent boosting	0.81 (0.80-0.81)
Blended model	Existing full model	0.78 (0.78-0.78)
	Existing full model with lasso regularization	0.78 (0.78-0.78)
	Existing full model with gradient descent boosting	0.78 (0.78-0.78)
	Blended model with lasso regularization	0.78 (0.78-0.78)
	Blended model with gradient descent boosting	0.82 (0.82-0.82)

Model identifies additional 168 bleeding cases per 100 000 PCI cases.

ML blended model -59 variables





ELSEVIER
Journal of Clinical Epidemiology 110 (2019) 12–22
Journal of Clinical Epidemiology

REVIEW

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

Evangelia Christodoulou[a], Jie Ma[b], Gary S. Collins[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

[a]Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium
[b]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK
[c]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[d]Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands
[e]Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium
[f]Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK

Accepted 5 February 2019; Published online 11 February 2019

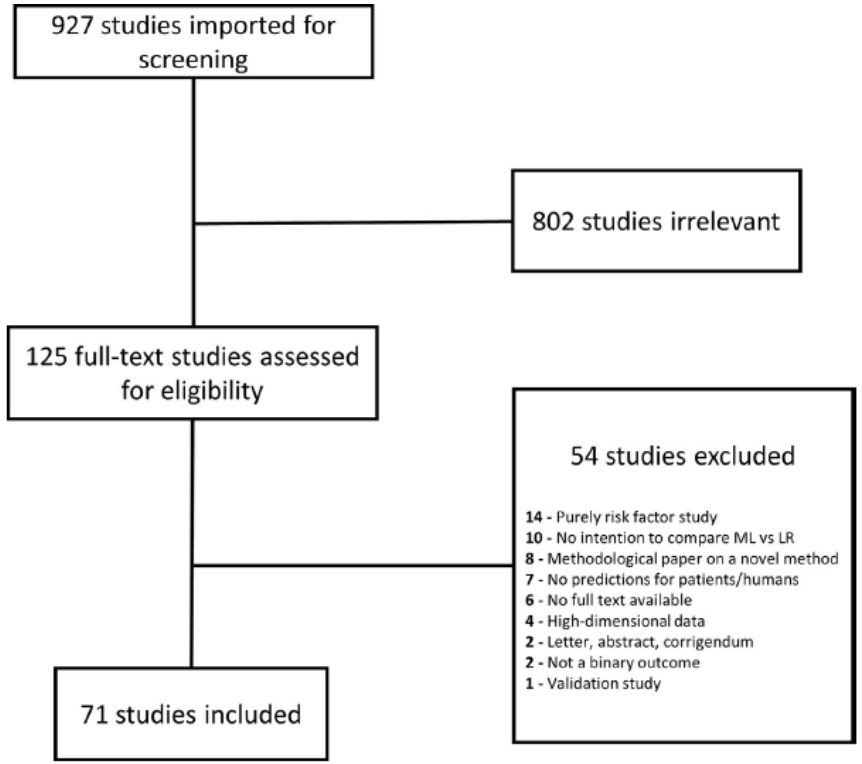
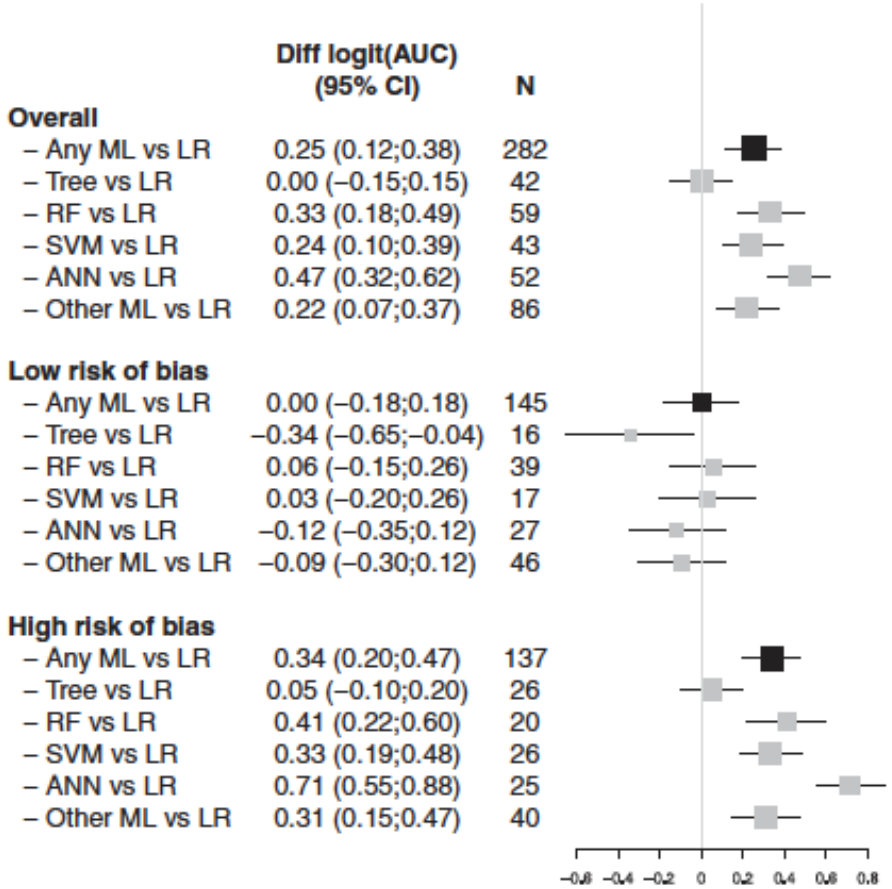


Fig. 1. PRISMA flowchart. PRISMA, preferred reporting items for systematic reviews and meta-analysis.



ELSEVIER
Journal of Clinical Epidemiology 110 (2019) 12–22
Journal of Clinical Epidemiology

REVIEW

A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models

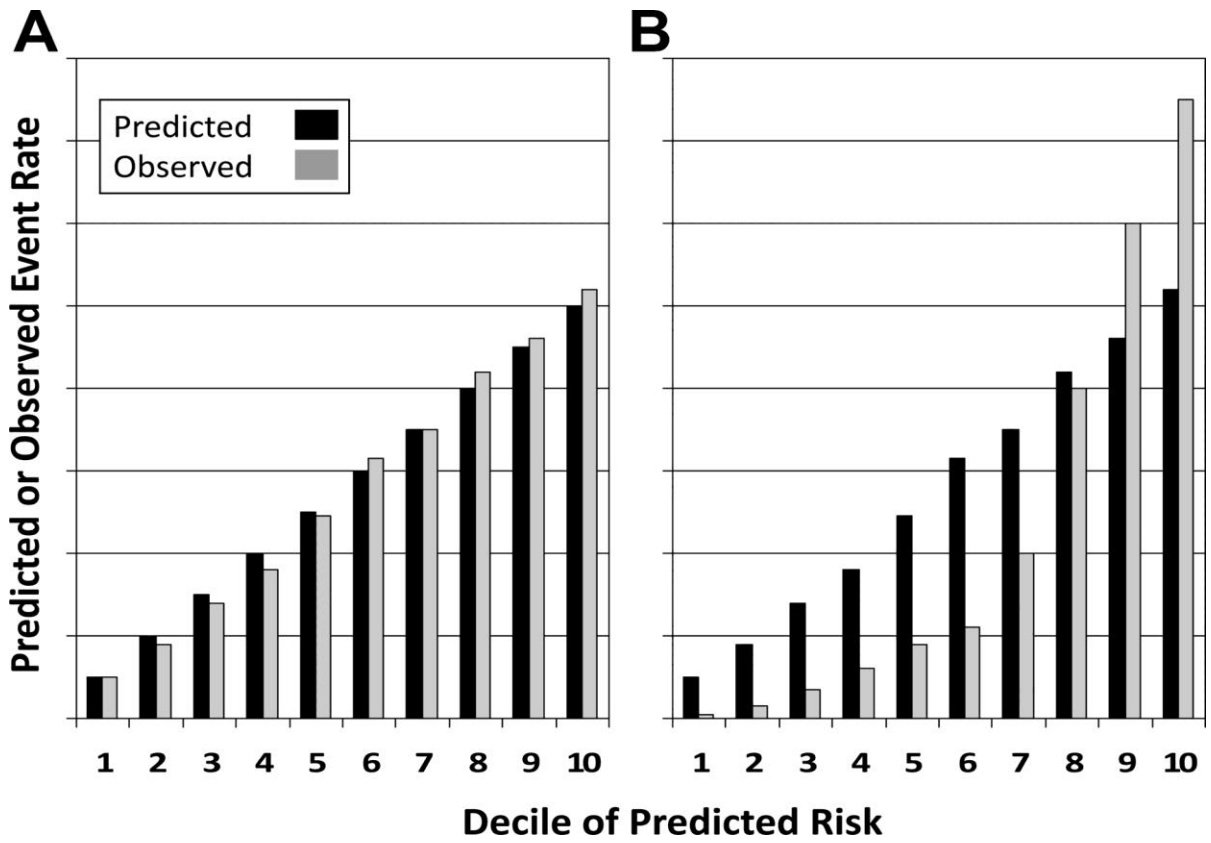
Evangelia Christodoulou[a], Jie Ma[b], Gary S. Collins[b,c], Ewout W. Steyerberg[d], Jan Y. Verbakel[a,e,f], Ben Van Calster[a,d,*]

[a]Department of Development & Regeneration, KU Leuven, Herestraat 49 box 805, Leuven, 3000 Belgium
[b]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford, OX3 7LD UK
[c]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[d]Department of Biomedical Data Sciences, Leiden University Medical Centre, Albinusdreef 2, Leiden, 2333 ZA The Netherlands
[e]Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium
[f]Nuffield Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, OX2 6GG UK

Accepted 5 February 2019; Published online 11 February 2019

Key findings

- Applied studies comparing clinical prediction models based on logistic regression and machine learning algorithms suffered from poor methodology and reporting, in particular, with respect to the validation procedure.
- The studies rarely assessed whether risk predictions are reliable (calibration), but the area under the receiver operating characteristic curve (AUC) was almost always provided.
- The AUC of logistic regression and machine learning models for clinical risk prediction were similar when comparisons were at low risk of bias; machine learning (ML) performance was higher in comparisons that were at high risk of bias.



Usability

CHA_2DS_2-VASc Score for Atrial Fibrillation Stroke Risk

Calculates stroke risk for patients with atrial fibrillation, possibly better than the $CHADS_2$ Score.

When to Use | Pearls/Pitfalls | Why Use

Age	<65 0	65-74 +1	≥75 +2
Sex	Female +1	Male 0	
CHF history	**No 0**	Yes +1	
Hypertension history	**No 0**	Yes +1	
Stroke/TIA/thromboembolism history	**No 0**	Yes +2	
Vascular disease history (prior MI, peripheral artery disease, or aortic plaque)	**No 0**	Yes +1	
Diabetes history	**No 0**	Yes +1	

Usability

Contents lists available at ScienceDirect

International Journal of Cardiology

journal homepage: www.elsevier.com/locate/ijcard

Prediction of clinical outcomes after percutaneous coronary intervention: Machine-learning analysis of the National Inpatient Sample

Akhmetzhan Galimzhanov [a,b,*], Andrija Matetic [b,c], Erhan Tenekecioglu [d,e], Mamas A. Mamas [b]

[a] Department of Propedeutics of Internal Disease, Semey Medical University, Semey, Kazakhstan
[b] Keele Cardiovascular Research Group, Keele University, Keele, UK
[c] Department of Cardiology, University Hospital of Split, Split 21000, Croatia
[d] Department of Cardiology, Bursa Education and Research Hospital, Health Sciences University, Bursa,Turkey
[e] Department of Cardiology, Thoraxcenter, Erasmus MC, Erasmus University, Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Machine learning
Percutaneous coronary intervention
Thrombosis
Bleeding
Prognosis
Precision medicine

ABSTRACT

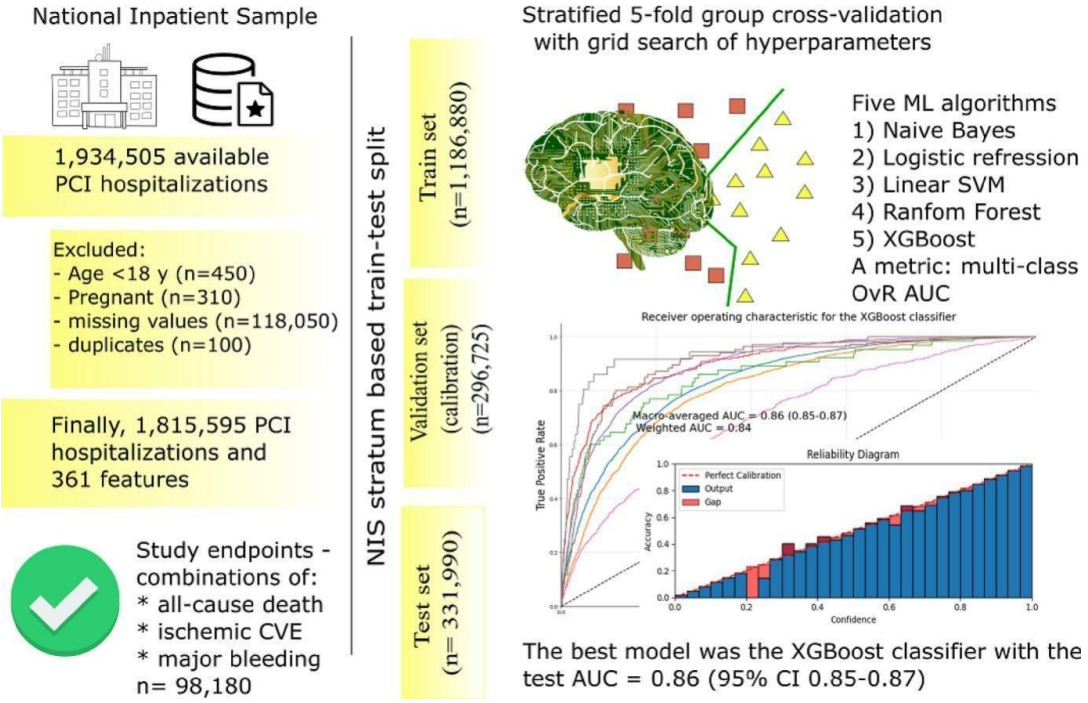
Background: This study aimed to develop a multiclass machine-learning (ML) model to predict all-cause mortality, ischemic and hemorrhagic events in unselected hospitalized patients undergoing percutaneous coronary intervention (PCI).

Methods: This retrospective study included 1,815,595 unselected weighted hospitalizations undergoing PCI from the National Inpatient Sample (2016–2019). Five most common ML algorithms (logistic regression, support vector machine (SVM), naive Bayes, random forest (RF), and extreme gradient boosting (XGBoost)) were trained and tested with 101 input features. The study endpoints were different combinations of all-cause mortality, ischemic cerebrovascular events (CVE) and major bleeding. An area under the curve (AUC) with 95% confidence interval (95% CI) was selected as a performance metric.

Results: The study population was split to a training cohort of 1,186,880 PCI discharges, validation cohort (for calibration) of 296,725 hospitalizations and a test cohort of 331,990 PCI discharges. A total of 98,180 (5.4%) hospital entries included study outcomes. Logistic regression, SVM, naive Bayes, and RF model demonstrated AUCs of 0.83 (95% CI 0.82–0.84), 0.84 (95% CI 0.83–0.86), 0.81 (95% CI 0.80–0.82), and 0.83 (95% CI 0.81–0.84), retrospectively. The XGBoost classifier performed the best with an AUC of 0.86 (95% CI 0.85–0.87) with excellent calibration. We then built a web-based application that provides predictions based on the XGBoost model.

Conclusion: We derived the multi-task XGBoost classifier based on 101 features to predict different combinations of all-cause death, ischemic CVE and major bleeding. Such models may be useful in benchmarking and risk prediction using routinely collected administrative data.

- 101 features used in model
- ICD-10 based codes



Algorithms not transparent

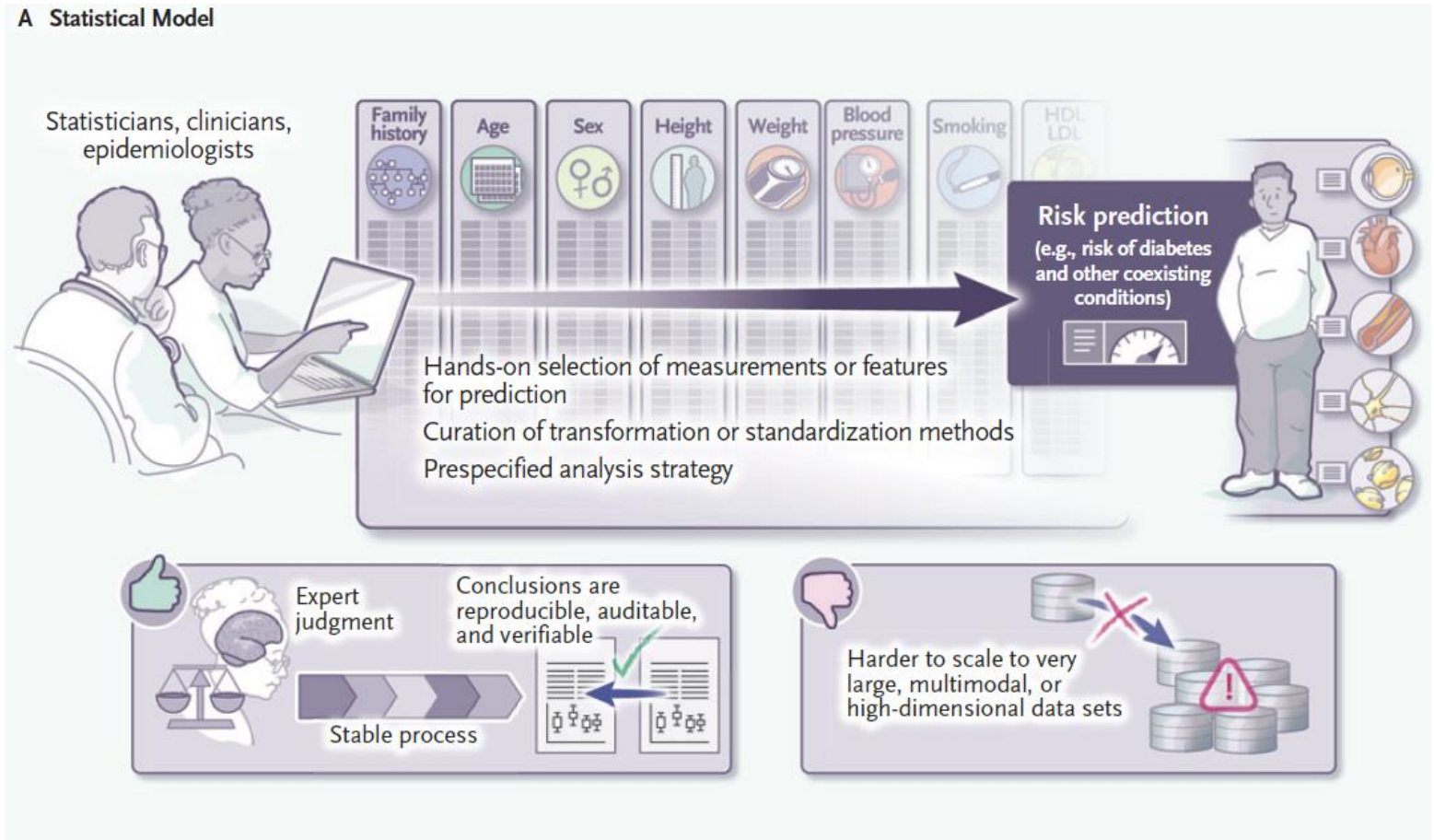
The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

AI IN MEDICINE
Jeffrey M. Drazen, M.D., *Editor*, Isaac S. Kohane, M.D., Ph.D., *Guest Editor*, and Tze-Yun Leong, Ph.D., *Guest Editor*

Where Medical Statistics Meets Artificial Intelligence

David J. Hunter, M.B., B.S., and Christopher Holmes, Ph.D.



Keele Cardiovascular Research Group

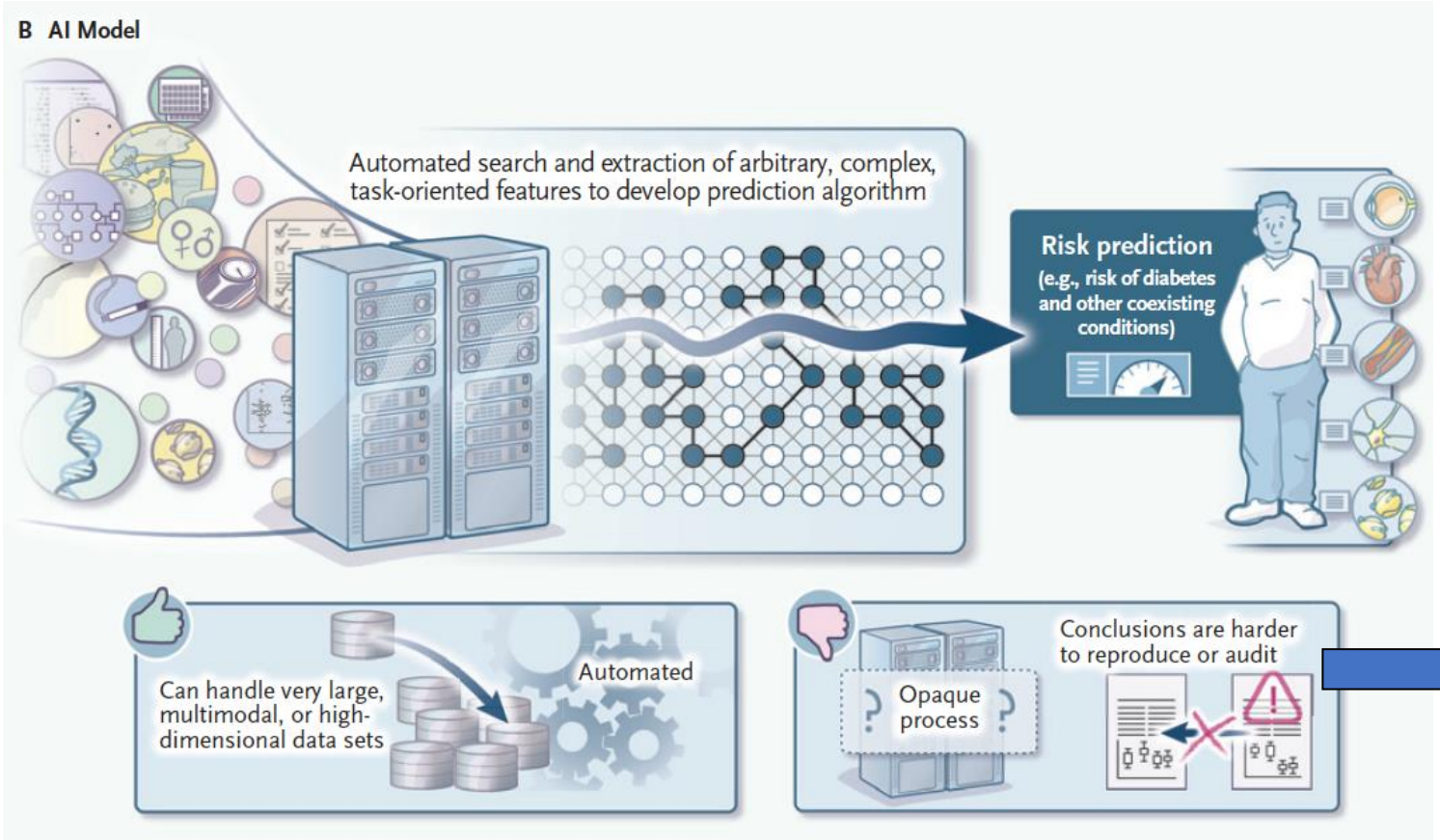
The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

AI IN MEDICINE
Jeffrey M. Drazen, M.D., *Editor*, Isaac S. Kohane, M.D., Ph.D., *Guest Editor*, and Tze-Yun Leong, Ph.D., *Guest Editor*

Where Medical Statistics Meets Artificial Intelligence

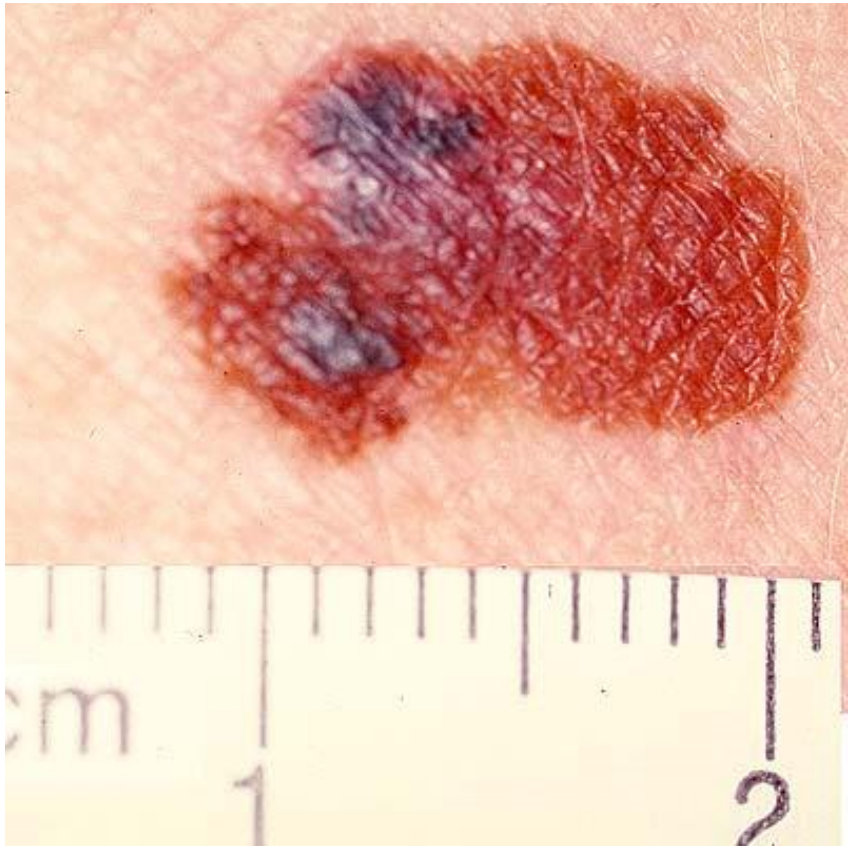
David J. Hunter, M.B., B.S., and Christopher Holmes, Ph.D.



Accidently fitting confounders

Accidently fitting confounders

Keele Cardiovascular Research Group



Amplify disparities

1) Model bias (i.e. models selected to best represent majority and not underrepresented groups)
2) Model variance (due to inadequate data from minorities)

Can ML / AI prognosis models be implemented ? Do they change outcomes?

Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial

*The INFANT Collaborative Group**

Keele Cardiovascular Research Group

	Decision support (n=23 263)	No decision support (n=23 351)	Adjusted risk ratio (CI)
Composite neonatal primary outcome			
Composite primary outcome*	172 (0·7%)	171 (0·7%)	1·01 (95% CI 0·82–1·25)
Intrapartum stillbirths†	1 (0)	2 (0)	0·50 (95% CI 0·05–5·53)
Neonatal deaths up to 28 days after birth‡	6 (0)	4 (0)	1·51 (95% CI 0·42–5·33)
Moderate or severe neonatal encephalopathy (requiring cooling)	18 (0·1%)	21 (0·1%)	0·86 (95% CI 0·46–1·61)
Admission to neonatal unit within 48 h of birth for ≥48 h because of feeding difficulties, respiratory illness or symptoms, or encephalopathy and evidence of compromise at birth	147 (0·6%)	144 (0·6%)	1·02 (95% CI 0·81–1·29)

Conclusions

- ML algorithms can be useful particularly for heterogenous data sources ie EHR / Imaging / Biology
- ML has not been shown to be superior traditional approaches for prognosis models
- Issues around lack of reproducibility, black box algorithms, model instability, potentiate bias
- Lack of data around whether improve clinical outcomes