Leiden University Medical Center

UMC Utrecht

Towards Trustworthy Patient-level Predictions: A Multiverse of Uncertainty and Heterogeneity

May 2026, Tanger workshop

Ewout W. Steyerberg, PhD

Professor of Clinical Biostatistics

Julius Center

University Medical Center Utrecht

Dept of Biomedical Data Sciences

Leiden University Medical Center

Thanks to many for assistance and inspiration, specifically

Ben van Calster, Leuven

Trustworthiness of numbers

Prediction

“10% risk”

Clinical prediction models

Cancer Prognostic Resources

A Catalog of Interactive Cancer Prognostic Tools

Home | Resources | Submit a Tool | About Us

Need Help?

Tutorial FAQ

Glossary Contact Us

Google™ Custom S

This website was designed to help healthcare professionals choose among available interactive cancer prognostic tools. Interactive cancer prognostic tools use an algorithm to calculate likely cancer-related outcomes based on a patient's characteristics.

Use of these tools may support communication and understanding about cancer prognosis. Some of the tools can be used to support shared decision making with cancer patients. The website allows for the comparison of cancer site specific tools OR search of tools using your own criteria.

View All Tools

Review and choose among available interactive cancer prognostic tools.

Compare Tools by Cancer Site

See and compare tools designed for a specific cancer site.

Search Tools

Search tools using your own criteria.

Find models by title, author, specialty, MeSH, ...

Public models by Specialty

- Adolescent medicine (5)
- Aerospace medicine (0)
- Allergology (6)
- Anaesthesiology (2)
- Cardiology (76)
- Clinical chemistry (3)
- Clinical pharmacology (22)
- Dermatology (1)
- Emergency medicine (32)
- Endocrinology (1)
- Epidemiology (7)
- Gastroenterology (20)
- General practice (59)
- Geriatrics (37)
- Gerontology (1)
- Gynaecology (23)
- Health informatics (4)
- Hematology (6)
- Hepatology (4)
- Immunology (4)
- Infectious disease (11)
- Intensive care (32)
- Internal medicine (44)
- Microbiology (2)
- Neonatology (4)
- Nephrology (19)
- Neurology (14)
- Neurophysiology (0)
- Neuroradiology (0)
- Neurosurgery (0)
- Nuclear medicine (2)
- Obstetrics (8)
- Occupational therapy (1)
- Oncology (182)
- Ophthalmology (0)
- Orthodontics (0)
- Orthopaedics (13)
- Otorhinolaryngology (1)
- Paediatrics (6)
- Palliative medicine (1)
- Pathology (2)
- Physiatry (0)
- Physical therapy (7)
- Podiatry (1)
- Psychiatry (3)
- Psychotherapy (0)
- Public Health (27)
- Pulmonology (32)
- Radiology (5)
- Radiotherapy (2)
- Rheumatology (4)
- Sports medicine (0)
- Surgery (96)
- Traumatology (18)
- Unspecified (46)
- Urology (63)
- Vascular medicine (11)

PREDICT example

predict
breast cancer

NHS

Home | About Predict | Predict Tool | Contact | Legal | Change Language

What is Predict?

Predict is an online tool that helps patients and clinicians see how different treatments for early invasive breast cancer might improve survival rates after surgery.

It is endorsed by the American Joint Committee on Cancer (AJCC).

Start Predict | Change Language

Did you mean to visit Predict Prostate?

What does Predict do?

Predict asks for some details about the patient and the cancer. It then uses data about the survival of similar women in the past to show the likely proportion of such women expected to survive up to fifteen years after their surgery with different treatment combinations.

Who is Predict for?

Predict is for clinicians, patients and their families.

Patients should use it in consultation with a medical professional.

Where can I find out more?

To read more go to **About Predict**

PREDICT example: hypothetical patient, 1.4% benefit

DCIS or LCIS only? Yes | No

Age at diagnosis – 55 +

Age must be between 25 and 85

Post Menopausal? Yes | No | Unknown

ER status Positive | Negative

HER2/ERRB2 status Positive | Negative | Unknown

Ki-67 status Positive | Negative | Unknown

Positive means more than 10%

Invasive tumour size (mm) – 5 +

If there was more than one tumour, enter the size of the largest tumour. If neo-adjuvant therapy was undertaken, enter the size before neo-adjuvant therapy.

Tumour grade 1 | 2 | 3

Detected by Screening | Symptoms | Unknown

Positive nodes – 0 +

Micrometastases only Yes | No | Unknown

Enabled when positive nodes is 1.
Why can't I enter micrometastases?

Treatment Options

Hormone Therapy No | 5 Years | 10 Years

Hormone (endocrine) therapy
Available when ER-status is positive

Chemotherapy None | 2nd gen | 3rd gen

Trastuzumab No | Yes

Available when HER2/ERRB2 status is positive

Bisphosphonates No | Yes

Available for post-menopausal women

Results

All treatments have side effects. Weigh up the benefits shown with the side effects in this website.

Table | Curves | Chart | Texts | Icons

Select number of years since surgery you wish to consider:

5 | 10 | 15

This table shows the percentage of women who survive at least 10 years after surgery.

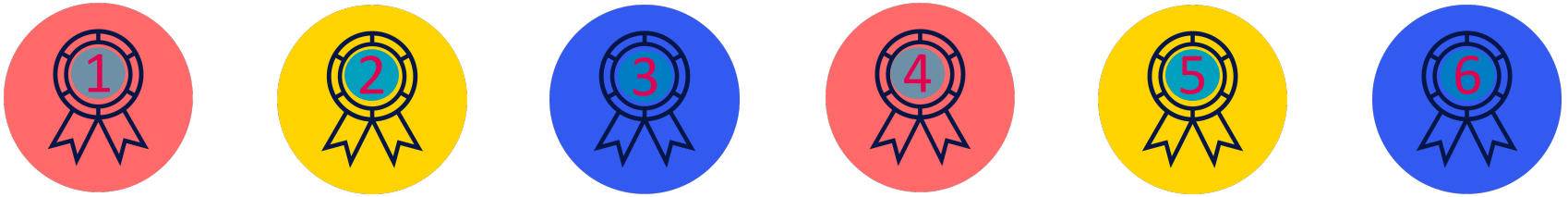
Treatment	Additional Benefit	Overall Survival %
Surgery only	-	90%
+ Chemotherapy	1.4% (1.0% – 1.7%)	92%

If death from breast cancer were excluded, 94% would survive at least 10 years, and

Can we trust these numbers?

- Predictions under care as usual?
  - On average
  - For this patient
- Treatment effect estimates?
  - On average
  - For this patient

Predictive algorithms: Medical AI



Phase 1	Phase 2	Phase 3	Phase 4	Phase 5	Phase 6
Data preparation	Development AI algorithm	Validation AI algorithm	Software environment	Impact assessment	Implementation in medical practice

Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review

Anne A. H. de Hond[1,2,3,8]✉, Artuur M. Leeuwenberg[4,8]✉, Lotty Hooft[4,5], Ilse M. J. Kant[1,2,3], Steven W. J. Nijman[4],

npj Digital Medicine (2022)5:2 ; https://doi.org/10.1038/s41746-021-00549-7

Int J Med Inform. 2021 Aug:152:104496. doi: 10.1016/j.ijmedinf.2021.104496. Epub 2021 May 15.

Machine learning for developing a prediction model of hospital admission of emergency department patients: Hype or hope?

Anne De Hond [1], Wouter Raven [2], Laurens Schinkelshoek [3], Menno Gaakeer [4], Ewoud Ter Avest [5], Ozcan Sir [6], Heleen Lameijer [7], Roger Apa Hessels [8], Resi Reijnen [9], Evert De Jonge [10], Ewout Steyerberg [11], Christian H Nickel [12], Bas De Groot [2]

Results: We included 172,104 ED patients of whom 66,782 (39 %) were hospitalized. The AUC of the multivariable logistic regression model was 0.82 (0.78-0.86) at triage, 0.84 (0.81-0.86) at ~30 min and 0.83 (0.75-0.92) after ~2 h. The best performing ML model over time was the gradient boosted decision trees model with an AUC of 0.84 (0.77-0.88) at triage, 0.86 (0.82-0.89) at ~30 min and 0.86 (0.74-0.93) after ~2 h.

Conclusions: Our study showed that machine learning models had an excellent but similar predictive performance as the logistic regression model for predicting hospital admission. I

JMIR Hum Factors. 2023 Jan 5:10:e39114. doi: 10.2196/39114.

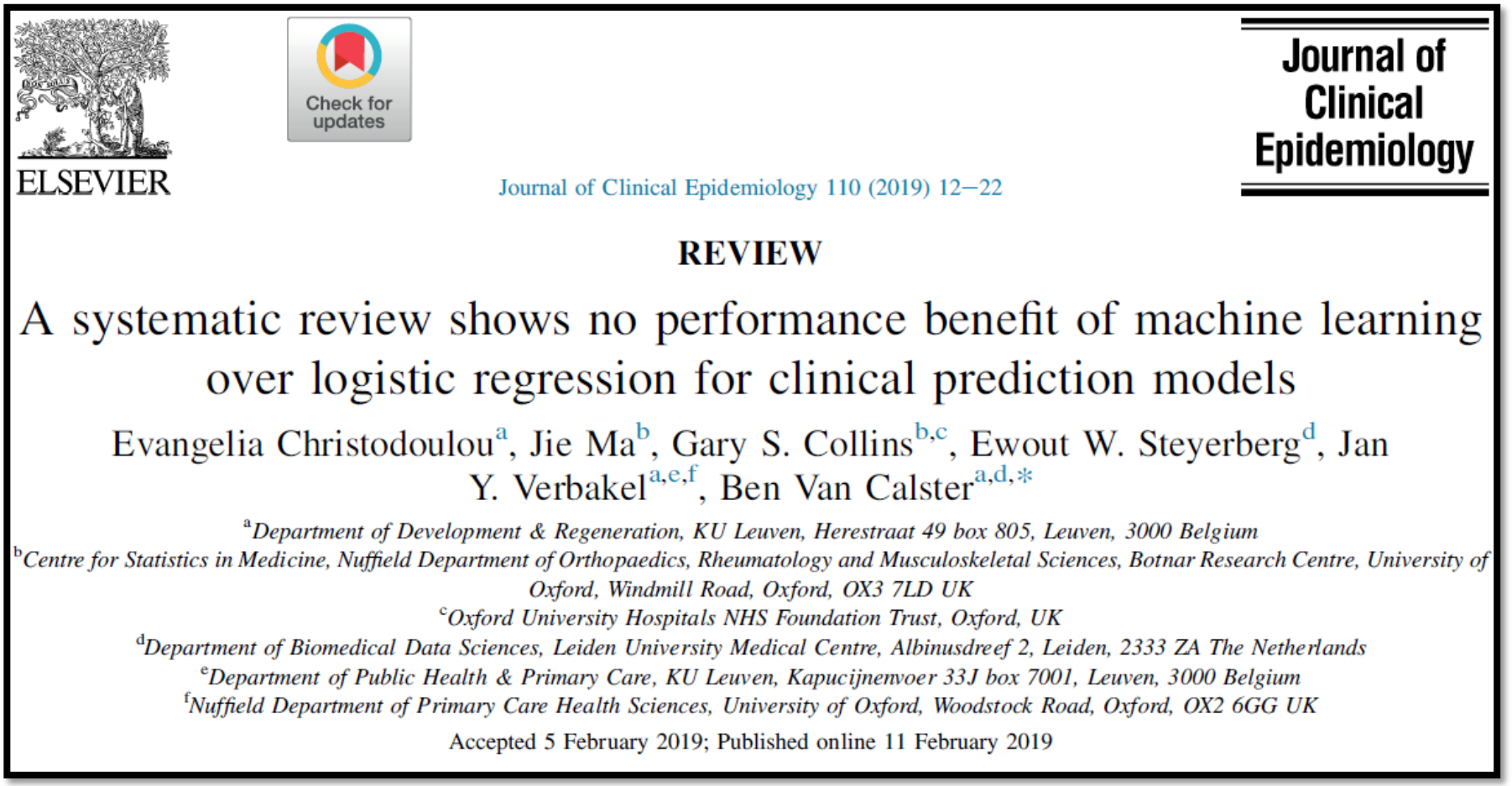
Intensive Care Unit Physicians' Perspectives on Artificial Intelligence-Based Clinical Decision Support Tools: Preimplementation Survey Study

Siri L van der Meijden [1 2 3], Anne A H de Hond [2 4], Patrick J Thoral [5], Ewout W Steyerberg [4], Ilse M J Kant [2 4], Giovanni Cinà [6 7 8], M Sesmu Arbous [1]

Conclusions: ICU physicians showed a favorable attitude toward the integration of AI-CDS tools into the ICU setting in general, and in particular toward a tool to predict a patient's risk of readmission and mortality within 7 days after discharge. The findings of this questionnaire will be used to improve the implementation process and training of end users.

Trust in AI: Machine Learning vs Regression

ML vs LR: systematic review



ELSEVIER

Check for updates

Journal of Clinical Epidemiology

Journal of Clinical Epidemiology 110 (2019) 12–22

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*

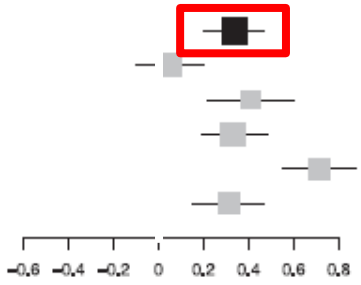
[e]*Department of Public Health & Primary Care, KU Leuven, Kapucijnenvoer 33J box 7001, Leuven, 3000 Belgium*

[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

Random effects meta-regression

	Diff logit(AUC) (95% CI)	N
Overall		
– Any ML vs LR	0.25 (0.12;0.38)	282
– Tree vs LR	0.00 (–0.15;0.15)	42
– RF vs LR	0.33 (0.18;0.49)	59
– SVM vs LR	0.24 (0.10;0.39)	43
– ANN vs LR	0.47 (0.32;0.62)	52
– Other ML vs LR	0.22 (0.07;0.37)	86
Low risk of bias		
– Any ML vs LR	0.00 (–0.18;0.18)	145
– Tree vs LR	–0.34 (–0.65;–0.04)	16
– RF vs LR	0.06 (–0.15;0.26)	39
– SVM vs LR	0.03 (–0.20;0.26)	17
– ANN vs LR	–0.12 (–0.35;0.12)	27
– Other ML vs LR	–0.09 (–0.30;0.12)	46
High risk of bias		
– Any ML vs LR	0.34 (0.20;0.47)	137
– Tree vs LR	0.05 (–0.10;0.20)	26
– RF vs LR	0.41 (0.22;0.60)	20
– SVM vs LR	0.33 (0.19;0.48)	26
– ANN vs LR	0.71 (0.55;0.88)	25
– Other ML vs LR	0.31 (0.15;0.47)	40



Conclusion 1

- AI / Machine Learning vs regression at group level:
  similar performance for simple prediction tasks
  - Different from image; tekst analysis

- What do we mean with performance?
  - Many measures:
    discrimination; calibration; overall; decision-analytic

- Is there a difference in trustworthiness at the patient level?

Trustworthiness for individuals

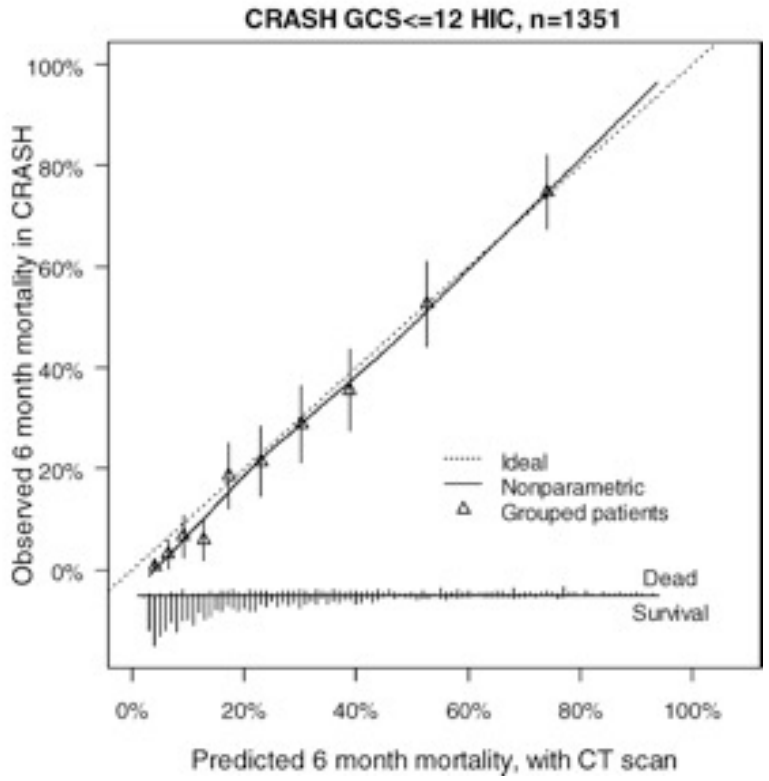
- Calibration = reliability of predictions

Calibration: the Achilles heel of predictive analytics

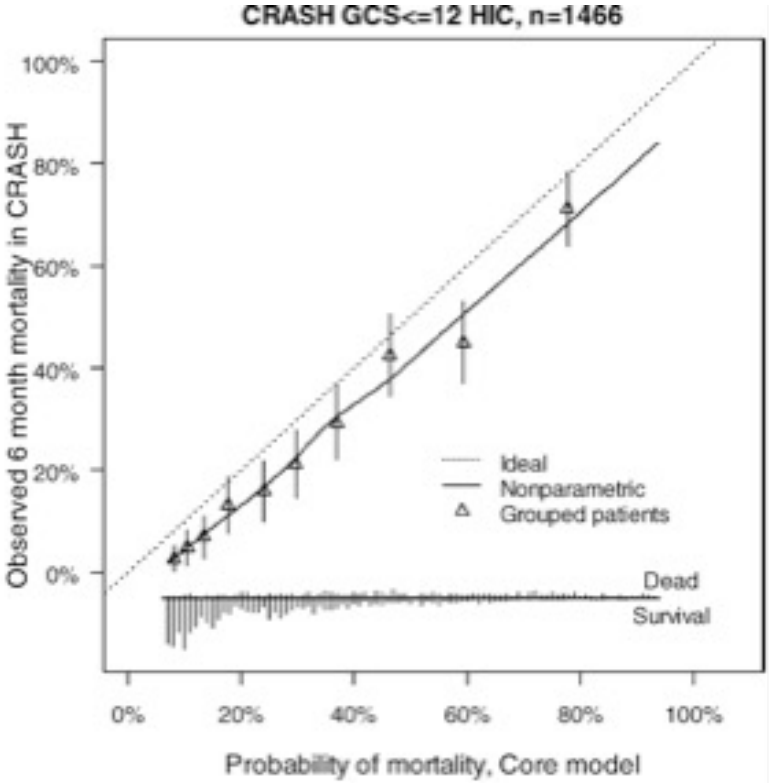
Ben Van Calster[1,2,6*], David J. McLernon[3,6], Maarten van Smeden[2,4,6], Laure Wynants[1,5], Ewout W. Steyerberg[2,6]
On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative[6]

Example in neurotrauma, external validation

well calibrated



over prediction

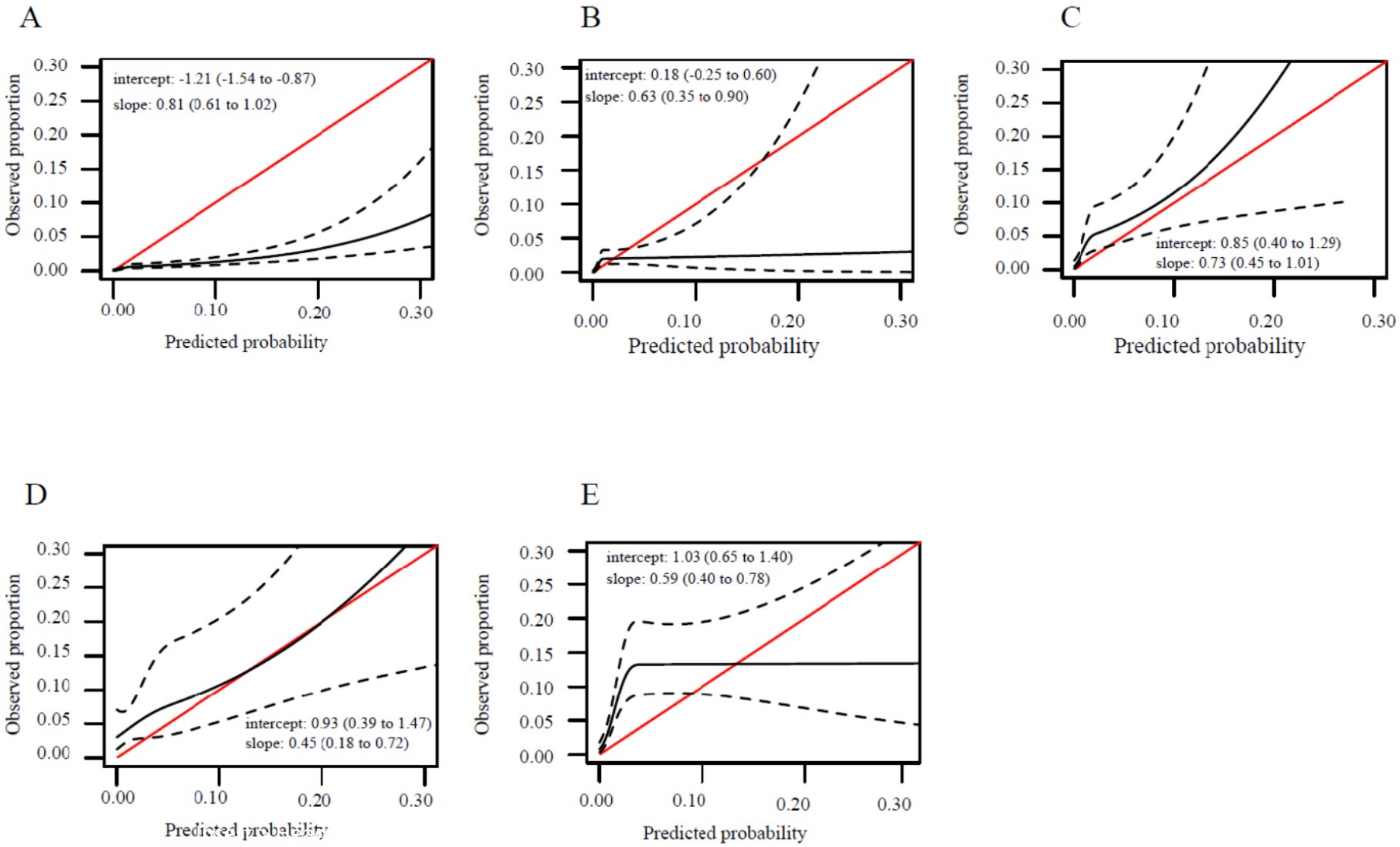


Modification of `val.prob()` in rms; **`val.prob.ci.2()`**

Steyerberg et al, PLoS Med 2008

Very heterogenous validations

Appendix 8: Calibration plot: observed proportion vs predicted probability of the clinical prediction model for 5 internal-external cross-validations.



Example with low N

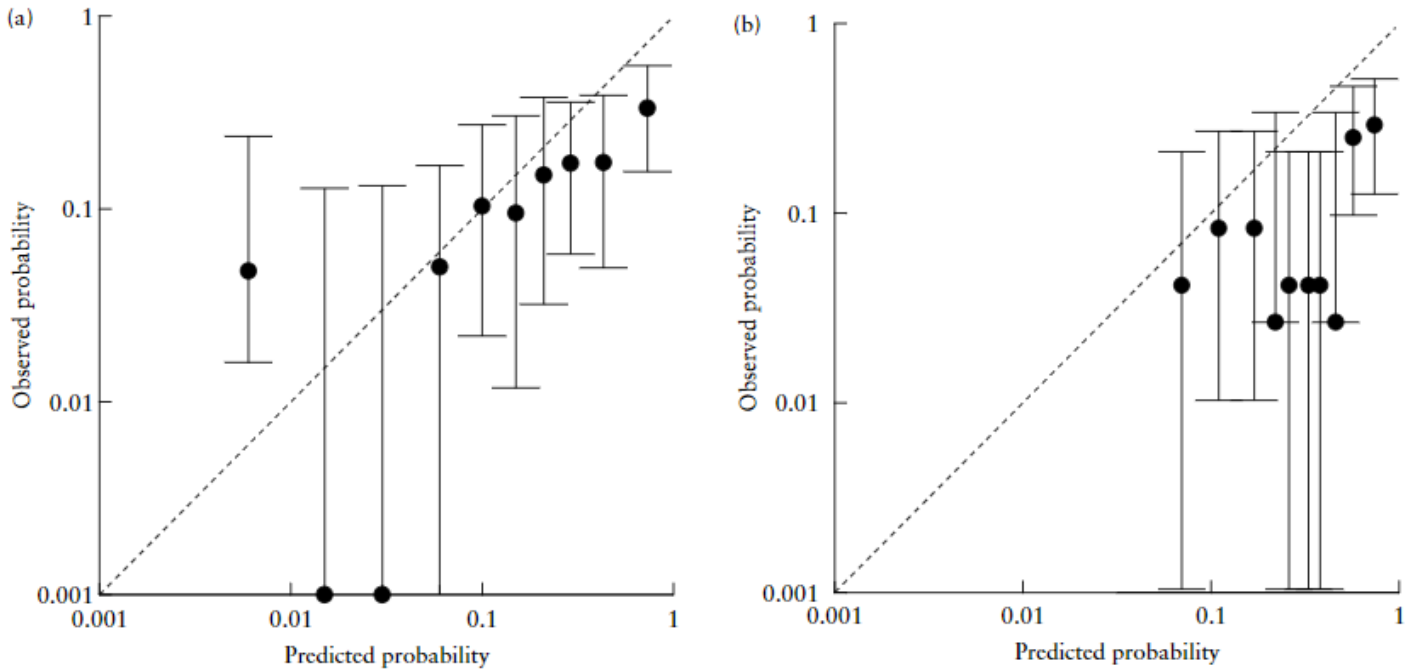


Figure 1 Calibration plots of the Peregrine *et al.* (a) and Rane *et al.* (b) prediction models for Cesarean delivery after induction of labor. Bars indicate 95% CIs of observed probability.

"Calibration of the model on the right was not as good as the calibration of the model on the left"

Levels of calibration

Journal of Clinical Epidemiology 74 (2016) 167–176

A calibration hierarchy for risk models was defined: from utopia to empirical data

Ben Van Calster[a,b,*], Daan Nieboer[b], Yvonne Vergouwe[b], Bavo De Cock[a], Michael J. Pencina[c,d], Ewout W. Steyerberg[b]

1. **Mean** calibration / calibration-in-the-large: average correct?
2. **Weak** calibration: slope correct?
3. **Moderate** calibration: pattern correct?
4. **Strong** calibration: model correct → **utopia**

Work motivated by a thought provoking paper from Werner Vach (JCE 2013;66:1296-1301)

1. Mean calibration

The average estimated risk is accurate

Compare average risk with outcome prevalence/incidence

Quantify and test:

a) In a logistic regression model with logit(p) as offset

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = a + offset(L)$$

with L the linear predictor or logit(p)

b) As observed:expected (O:E) ratio

2. Weak calibration

On average, the model does not overestimate or underestimate risk, and does not give too extreme or too modest risks

'Logistic recalibration' framework:

Evaluate **calibration slope** b: $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = a + bL$

$b < 1$ means too extreme risks, $b > 1$ means too modest risks

3. Moderate calibration

Observed proportion of events correspond to estimated risk

Construct a flexible calibration curve based on $\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = a + f(L)$.

$f(.)$ e.g. a loess fit, or splines.

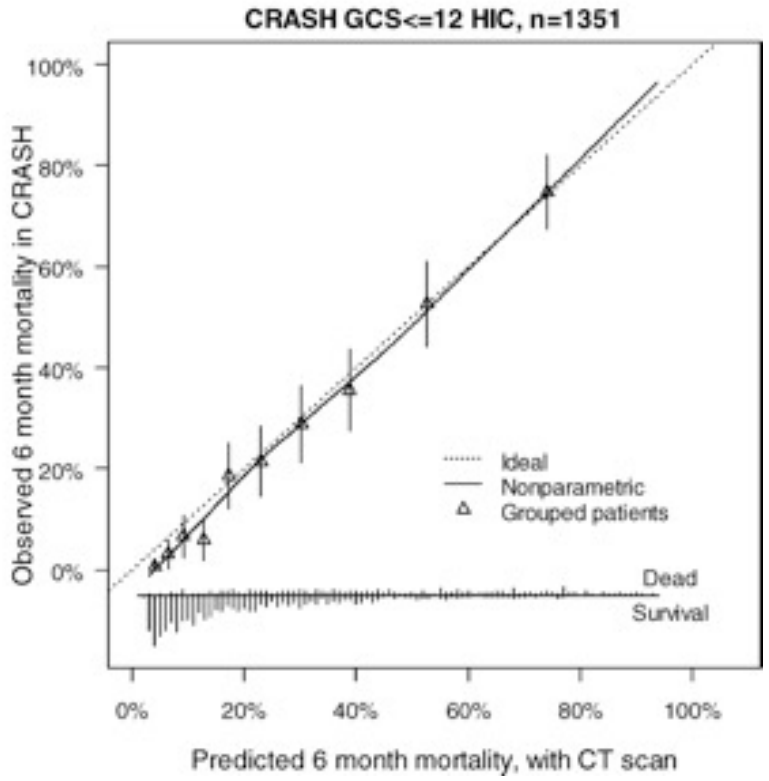
Preferable at external validation, with sufficient N.

Intercept and slope reduce calibration to 2 numbers (“weak calibration”).

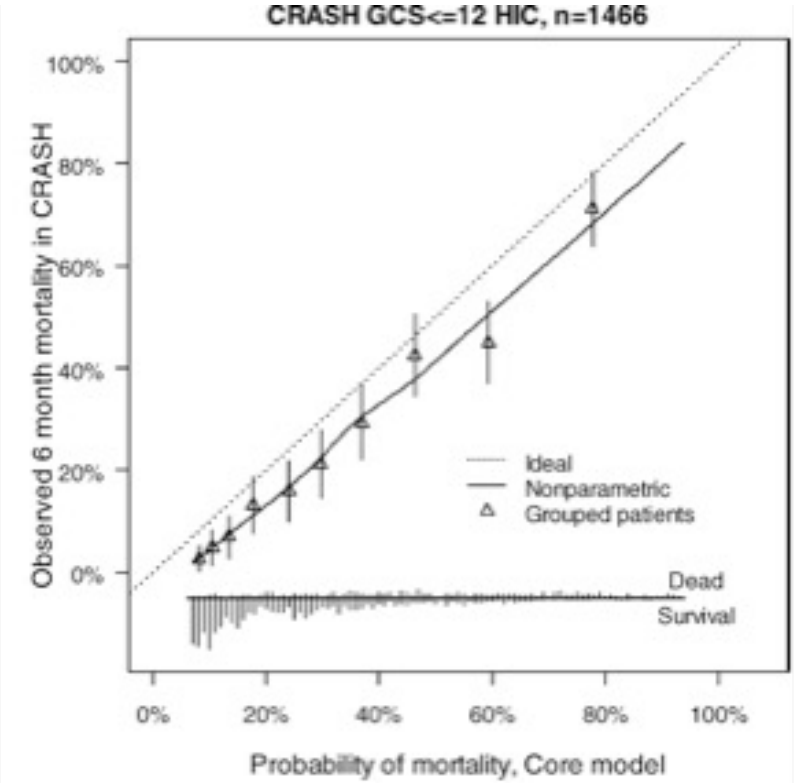
Slope sufficient for internal validation (using bootstrapping or cross-validation), intercept and plotting a curve especially important at external validation.

Example in neurotrauma, external validation

well calibrated



over prediction



Modification of `val.prob()` in rms; **`val.prob.ci.2()`**

Steyerberg et al, PLoS Med 2008

Conclusion 2

- Calibration assessment important to assess trustworthiness of predictions
  - Evaluation is at the population-level
  - Individual-level, per covariate profile, usually utopic

ESSAY | 16 December 2024 | Correction 18 December 2024

Why probability probably doesn't exist (but it is useful to act like it does)

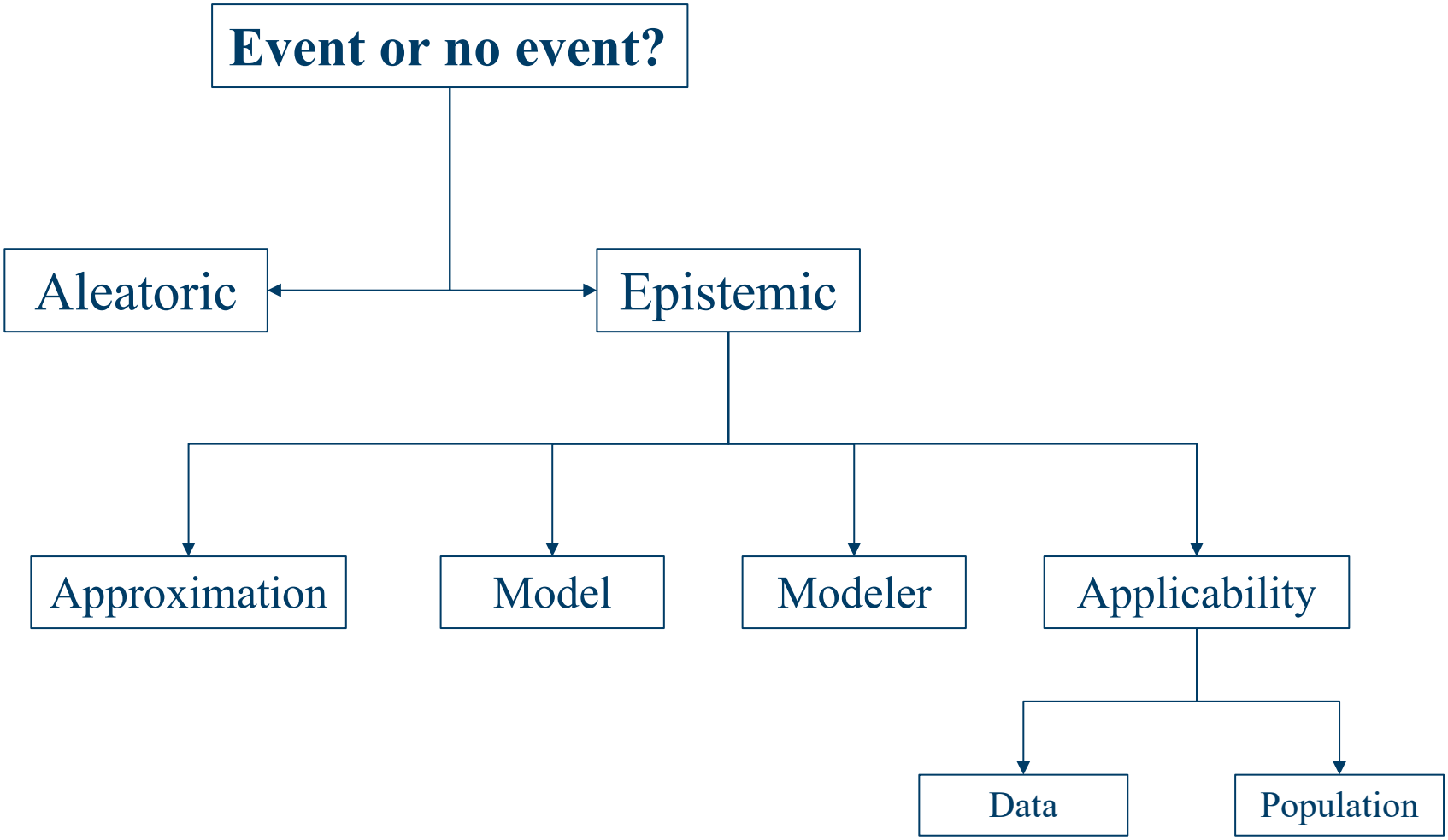
All of statistics and much of science depends on probability — an astonishing achievement, considering no one's really sure what it is.

By David Spiegelhalter

A framework for epistemic uncertainty in predictions

ALEATORIC VS. EPISTEMIC

A framework for epistemic uncertainty in predictions



Aleatoric uncertainty

Randomness can be surprising?

“Do 100 coin flips”

Probability of 6 heads out of 100 coin flips

- **Exactly 6 heads in 100 flips:**
  This is calculated using the binomial probability formula:
  {100}{6} * (0.5)^6 * (0.5)^94, resulting in a very value low (<<1%)
- **6 Heads from a specific flip:**
  0.5^6 = 1/64
- **A streak of 6+ heads in 100 flips:**
  95 opportunities for a 6-flip streak to begin in 100 flips
  a streak of 6 heads (or tails) is likely (>50%)

Unfair comparison: coins <-> patients

Homogeneous

heterogeneous

Reference class problem

Definition

- The dilemma of deciding which specific group (*reference class*) to use when assigning a probability to a unique, individual event.

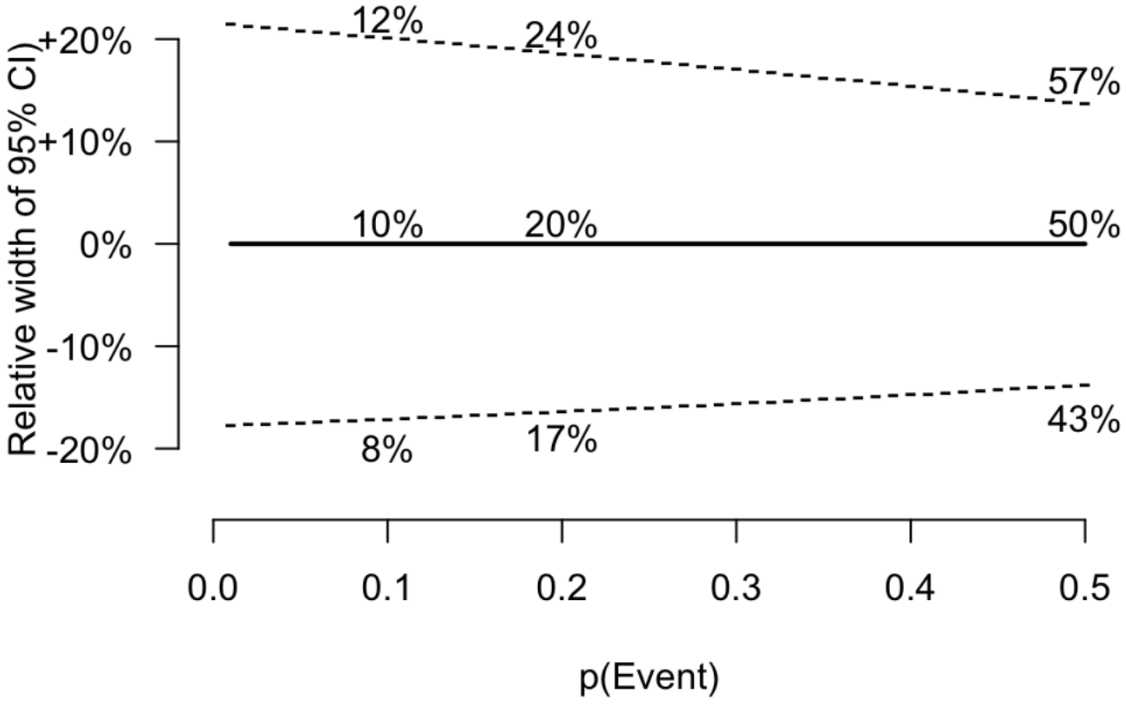
Implications

- Because an individual belongs to multiple, distinct classes, each with different statistical frequencies, choosing the "correct" class is ambiguous
  - Prediction
  - Subgroup effect in randomized controlled trials
- Critical to frequentist interpretation of probability, which relies on long-run frequencies

Approximation uncertainty

- Larger N, more precision
  - Binary events:
    smallest category drives effective sample size (if relatively rare, <20%)

- Any modeling: >100 events for reliable estimation of the average



The 95% confidence interval was from 8% to 12% for p(Event)=10%, or a relative width of –17% to +20%; and from 17% to 24% for p(Event)=20%, or a relative width of –16% to +19%.

Approximation uncertainty

- More complex modeling: reliable estimation of parameters
  - Classic rule: EPP > 10 (Events Per Parameter)
  - Modern: shrinkage > 0.9, with $s=(\chi 2 - p)/\chi 2$

    $\chi 2$: difference in 2LL; p: #parameters in the model

    e.g.: $\chi 2 = 100$; $p=5 \rightarrow s=0.95$
  - EPP>10 equivalent to $s>0.9$ for AUC=0.77 (if p(Event)<20%)

- More complex modeling: reliable estimation of predictions
  - Fisher information matrix

Shrinkage: Copas 1983, Van Houwelingen 1990

BMJ. 2020 Mar 18:368:m441. doi: 10.1136/bmj.m441.

Calculating the sample size required for developing a clinical prediction model

Richard D Riley [1], Joie Ensor [2], Kym I E Snell [2], Frank E Harrell Jr [3], Glen P Martin [4], Johannes B Reitsma [5], Karel G M Moons [5], Gary Collins [6], Maarten van Smeden [5 6 7]

arXiv:2407.09293 [pdf] stat.ME

A decomposition of Fisher's information to inform sample size for developing fair and precise clinical prediction models -- part 1: binary outcomes

Authors: Richard D Riley, Gary S Collins, Rebecca Whittle, Lucinda Archer, Kym IE Snell, Paula Dhiman, Laura Kirton, Amardeep Legha, Xiaoxuan Liu, Alastair Denniston, Frank E Harrell Jr, Laure Wynants, Glen P Martin, Joie Ensor

How to express individual uncertainty?

- Statistical:

  prediction / confidence / credible / uncertainty interval

- Intuitive:

  number of patients like you (translation of `se.pred`)

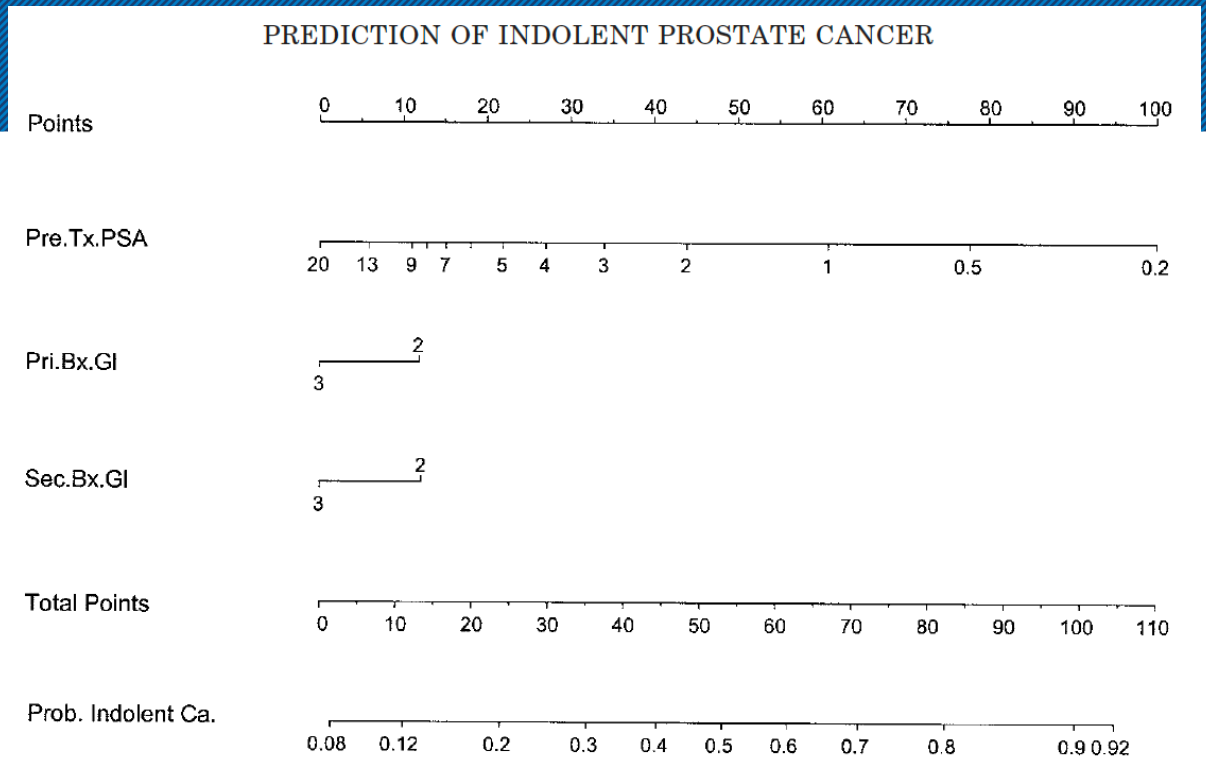
Example on presentation by 'the king of nomograms'

0022-5347/03/1705-1792/0
THE JOURNAL OF UROLOGY®

Vol. 170, 1792–1797, November 2003
Printed in U.S.A.
DOI: 10.1097/01.ju.0000091806.70171.41

COUNSELING MEN WITH PROSTATE CANCER: A NOMOGRAM FOR PREDICTING THE PRESENCE OF SMALL, MODERATELY DIFFERENTIATED, CONFINED TUMORS

MICHAEL W. KATTAN, JAMES A. EASTHAM, THOMAS M. WHEELER, NORIO MARU,



Instructions for Physician: Locate the patient's PSA on the **PreTx PSA** axis. Draw a line straight upwards to the **Points** axis to determine how many points towards having an indolent cancer the patient receives for his PSA. Repeat this process for the remaining axes, each time drawing straight upward to the **Points** axis. Sum the points achieved for each predictor and locate this sum on the **Total Points** axis. Draw a line straight down to find the patient's probability of having indolent cancer.

Exactly --> roughly
100 --> ?

Instruction to Patient: "Mr. X, if we had 100 men exactly like you, we would expect <predicted probability from nomogram * 100 > to have indolent cancer."

Effective sample size: A measure of individual uncertainty in predictions.

Thomassen D, le Cessie S, van Houwelingen HC, Steyerberg EW.

Stat Med. 2024 Mar 30;43(7):1384-1396. doi: 10.1002/sim.10018. Epub 2024 Jan 31.

Effective sample size

Equivalent number of patients with this profile

For binary outcome, depends on

- Covariance (N and exceptionality)
- Predicted risk

$$n_* \approx \frac{\hat{p}_{\text{new}}(1-\hat{p}_{\text{new}})}{x_{\text{new}}^\top \text{Cov}(\hat{\beta}) x_{\text{new}} \cdot (\hat{p}_{\text{new}}(1-\hat{p}_{\text{new}}))^2}$$
$$= \left(x_{\text{new}}^\top \text{Cov}(\hat{\beta}) x_{\text{new}} \cdot (\hat{p}_{\text{new}}(1-\hat{p}_{\text{new}})) \right)^{-1}.$$

THE AMERICAN STATISTICIAN
2026, VOL. 80, NO. 1, 100–108: Statistical Practice
https://doi.org/10.1080/00031305.2025.2542390

Taylor & Francis
Taylor & Francis Group

Effective Sample Size for the Kaplan-Meier Estimator: A Valuable Measure of Uncertainty?

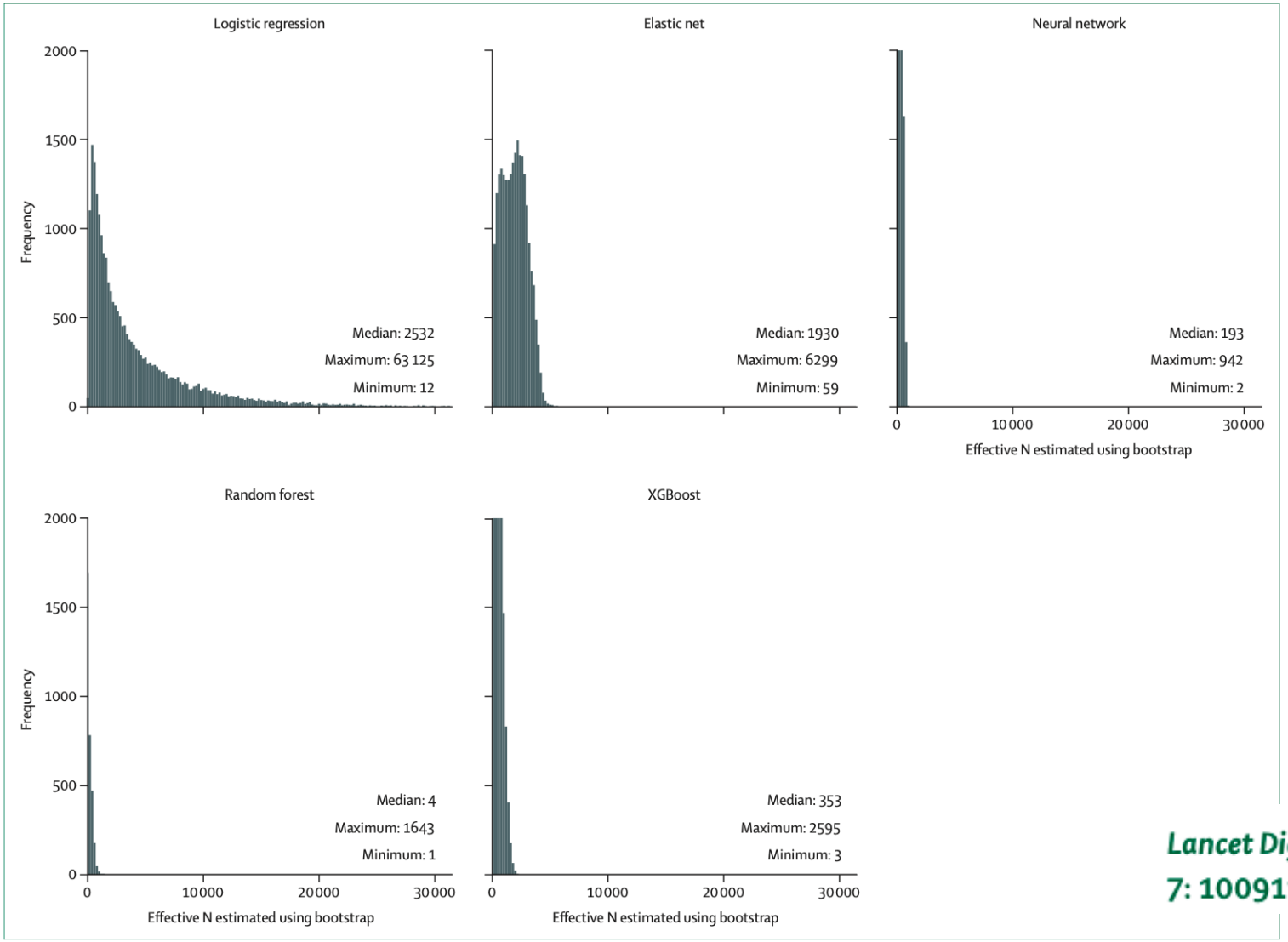
Toby Hackmann[a], Doranne Thomassen[a], Anne M. Stiggelbout[a], Saskia le Cessie[a,b], Hein Putter[a], Liesbeth C. de Wreede[a], Ewout W. Steyerberg[a,c], and on behalf of the 4D PICTURE Consortium[#]

Effective sample size for individual risk predictions: quantifying uncertainty in machine learning models

Doranne Thomassen, Toby Hackmann, Jelle Goeman, Ewout Steyerberg, Saskia le Cessie

Individual prediction uncertainty is a key aspect of clinical prediction model performance; however,

Nnet, RF, XGBoost: low effective N



Lancet Digit Health 2025; 7: 100911

Figure 5: **Distribution of bootstrap-estimated effective sample sizes in the GUSTO dataset collected within the USA (n=23 034) for five fitted prediction models**
All models were used to predict the risk of 30-day mortality using the same set of (candidate) predictors. In all histograms, the x-axis was limited to 30 000 although a few higher effective sample sizes were observed for the logistic regression model (the maximum values are indicated on the plot image). The elastic net model yielded reasonable effective sample sizes with fewer extremes. For the neural network, random forest, and XBoost models, all effective sample sizes were relatively low. Therefore, these models have a high peak in the histogram between 0 and 1000 with bin counts exceeding the limits of this plot. N=sample size.

The currency is sample size

The more complicated (or ‘fancy’) the modeling strategy, the more you have to **pay with sample size**.

Between model variability is large

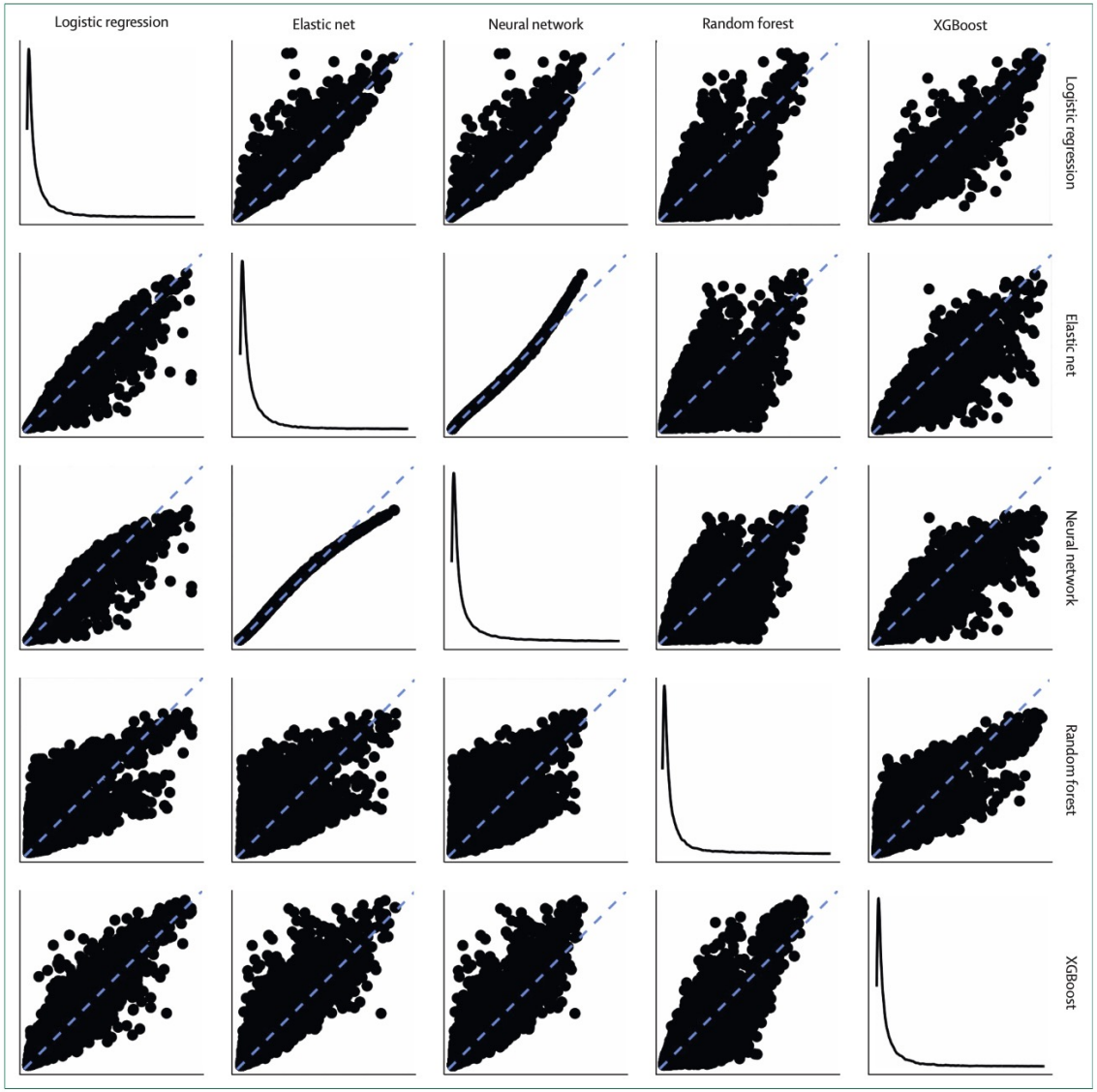


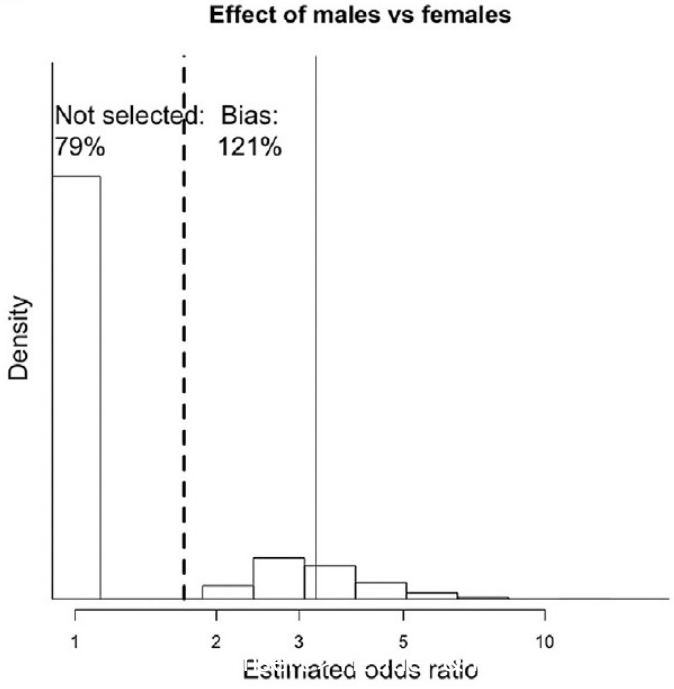
Figure 4: **Predicted 30-day mortality risks for patients in the GUSTO dataset collected within the USA (n=23 034)**
Pairwise comparisons of predictions from logistic regression, elastic net, neural network, random forest, and XGBoost tree-based models. Axes for predicted risks run from 0 to 100%. In the nondiagonal panels, each point represents an individual patient in the data, whose predicted risk from one model (x-axis) is compared to their predicted risk from another model (y-axis). Diagonal panels show density plots of predicted risks in the GUSTO US dataset (n=23 034) for each model. N=sample size.

Lancet Digit Health 2025; 7: 100911

Model uncertainty

- Model structure is determined by the data
  - Regression: stepwise selection
  - Tree methods, such as RF
    - Optimal cut-offs for continuous predictors
    - Interactions
  - Penalty in LASSO, Ridge, Elastic Net

A



Poor performance of clinical prediction models: the harm of commonly applied methods

Ewout W. Steyerberg[a,b,*], Hajime Uno[c], John P.A. Ioannidis[d,e,f,g], Ben van Calster[a,h],

Trustworthiness: poor for human modelers

Red cards and dark skin soccer players

https://psyarxiv.com/qkwst/

Empirical Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(3) 337–356


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245917747646
www.psychologicalscience.org/AMPPS
SAGE

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

R. Silberzahn[1], E. L. Uhlmann[2], D. P. Martin[3], P. Anselmi[4], F. Aust[5], E. Awtrey[6], Š. Bahník[7], F. Bai[8], C. Bannard[9], E. Bonnier[10], R. Carlsson[11], F. Cheung[12], G. Christensen[13], R. Clay[14], M. A. Craig[15], A. Dalla Rosa[4], L. Dam[16], M. H. Evans[17], I. Flores Cervantes[18], N. Fong[19], M. Gamez-Djokic[20], A. Glenz[21], S. Gordon-McKeon[22], T. J. Heaton[23], K. Hederos[24], M. Heene[25], A. J. Hofelich Mohr[26], F. Högden[5], K. Hui[27], M. Johannesson[10], J. Kalodimos[28], E. Kaszubowski[29], D. M. Kennedy[30], R. Lei[15], T. A. Lindsay[26], S. Liverani[31], C. R. Madan[32], D. Molden[33], E. Molleman[16], R. D. Morey[34], L. B. Mulder[16], B. R. Nijstad[16], N. G. Pope[35], B. Pope[36], J. M. Prenoveau[37], F. Rink[16], E. Robusto[4], H. Roderique[38], A. Sandberg[24], E. Schlüter[39], F. D. Schönbrodt[25], M. F. Sherman[37], S. A. Sommer[40], K. Sotak[41], S. Spain[42], C. Spörlein[43], T. Stafford[44], L. Stefanutti[4], S. Tauber[16], J. Ullrich[21], M. Vianello[4], E.-J. Wagenmakers[45], M. Witkowiak[46], S. Yoon[19], and B. A. Nosek[3,47]

[1]Organisational Behaviour, University of Sussex Business School; [2]Organisational Behaviour Area, INSEAD Asia Campus;

- 29 teams involving 61 analysts; same dataset; same research question: **whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players**
- Estimated odds ratios 0.89 –2.93 (median 1.3)
- 20 teams: statistically significant positive effect, 9: non-significant relation

Estimated ~~odds ratios~~ relative risks by 29 research teams

Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Models	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93

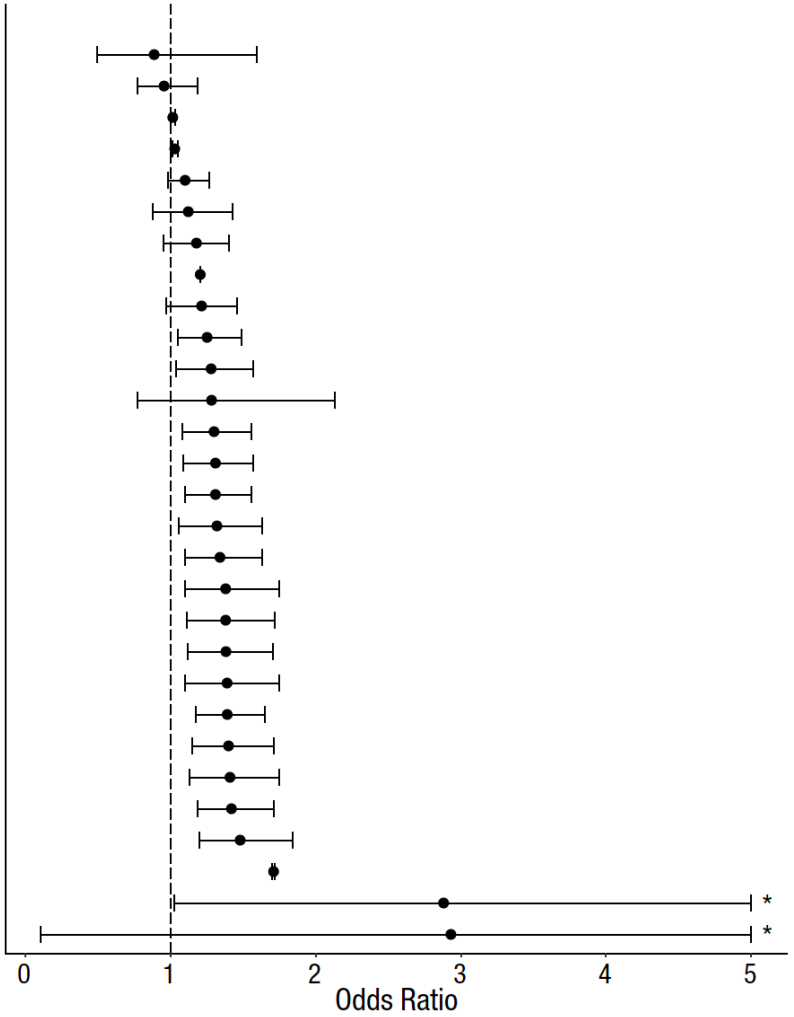


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that the smallest reported effect size is at the top and the largest is at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot; the actual upper bounds of the confidence intervals were 11.47 for Team 21 and 78.66 for Team 27. OLS = ordinary least squares; WLS = weighted least squares.

“Logistic regression”

Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Model	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93

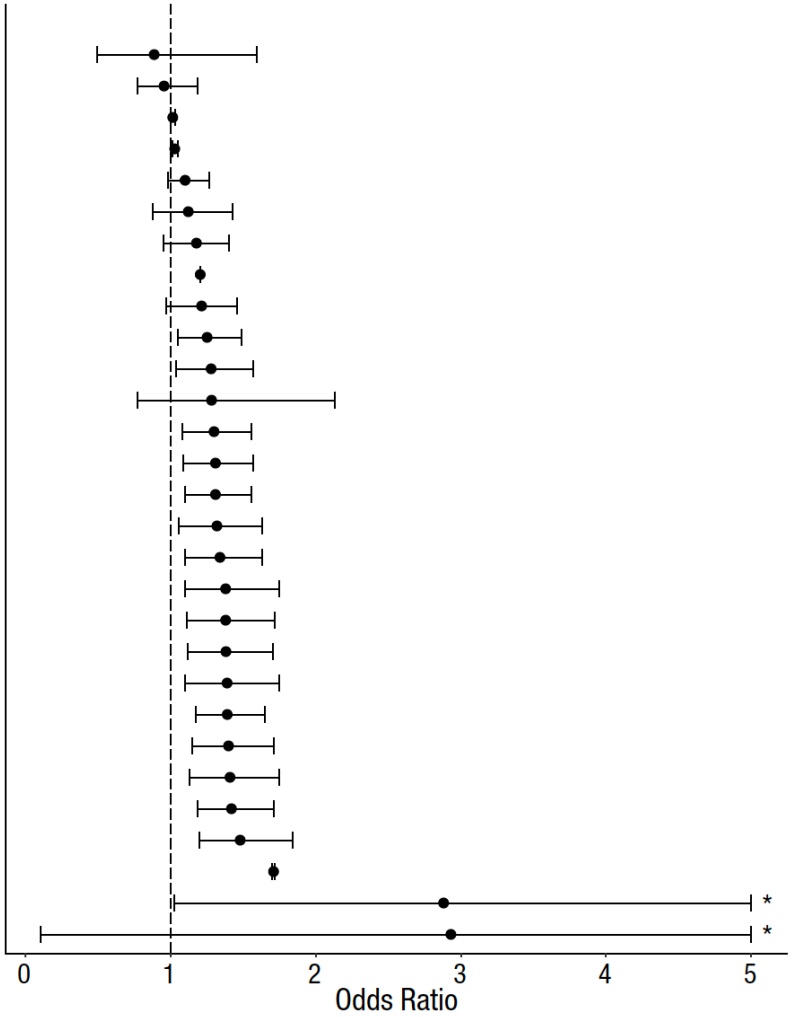


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that the smallest reported effect size is at the top and the largest is at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot; the actual upper bounds of the confidence intervals were 11.47 for Team 21 and 78.66 for Team 27. OLS = ordinary least squares; WLS = weighted least squares.

Claimed trust in results

Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Models	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93

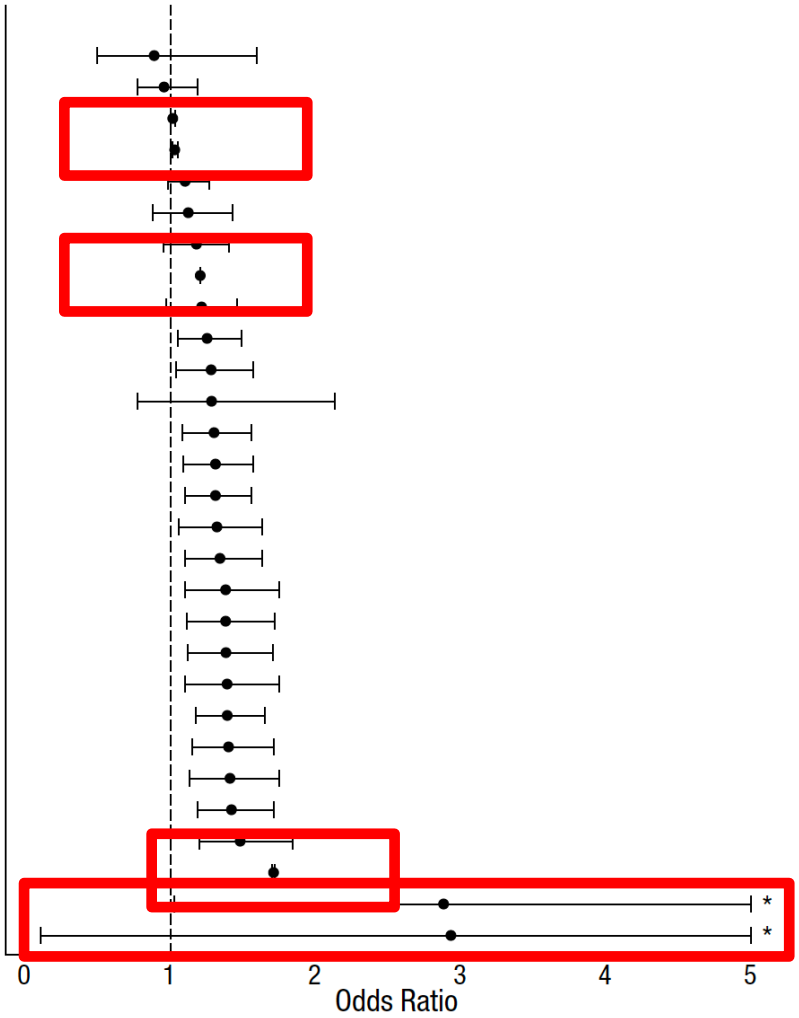


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that the smallest reported effect size is at the top and the largest is at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot; the actual upper bounds of the confidence intervals were 11.47 for Team 21 and 78.66 for Team 27. OLS = ordinary least squares; WLS = weighted least squares.

Trustworthiness: poor for human modelers

- 29 teams involving 61 analysts; same dataset; same research question: whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players
- Estimated odds ratios 0.89 –2.93 (median 1.3).
- 20 teams: statistically significant positive effect, 9: non-significant relation.
- **21 unique combinations of covariates**

- **"Variation in analysis of complex data may be difficult to avoid, even by experts with honest intentions"**

→ Can better training help?

avoid 'mistakes'

encourage 'good practice'

Extra slides

Applicability

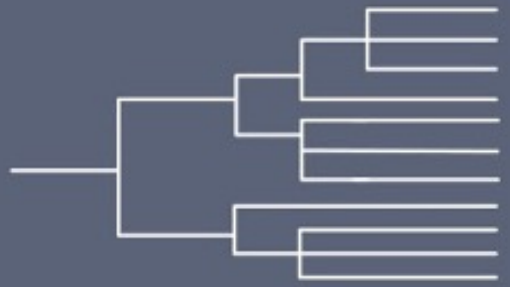
- Data
  - Definitions (X; Y)
  - Coding
  - Missing values and the underlying mechanisms
  - …
- Population
  - Selection
  - Care context
  - Treatment
  - …

Uncertainty

Aleatory

Integration over distribution of expected parameter values

Epistemic



logic tree of technically defensible interpretations

Riley *et al. BMC Medicine* (2023) 21:502
https://doi.org/10.1186/s12916-023-03212-y

BMC Medicine

OPINION

Open Access

Clinical prediction models and the multiverse of madness

Richard D. Riley[1,2*], Alexander Pate[3], Paula Dhiman[4], Lucinda Archer[1,2], Glen P. Martin[3] and Gary S. Collins[4]

Abstract

Background Each year, thousands of clinical prediction models are developed to make predictions (e.g. estimated risk) to inform individual diagnosis and prognosis in healthcare. However, most are not reliable for use in clinical practice.

Conclusions Instability is concerning as an individual's predicted value is used to guide their counselling, resource prioritisation, and clinical decision making. If different samples lead to different models with very different predictions for the same individual, then this should cast doubt into using a particular model for that individual. Therefore, visualising, quantifying and reporting the instability in individual-level predictions is essential when proposing a new model.

Uncertainty illustration

Statistics > Methodology

[Submitted on 20 Jun 2025]

The fundamental problem of risk prediction for individuals: health AI, uncertainty, and personalized medicine

Lasai Barreñada, Ewout W Steyerberg, Dirk Timmerman, Doranne Thomassen, Laure Wynants, Ben Van Calster

Background: Clinical prediction models for a health condition are commonly evaluated regarding performance for a population, although decisions are made for individuals. The classic view relates uncertainty in risk estimates for individuals to sample size (estimation uncertainty) but uncertainty can also be caused by model uncertainty (variability in modeling choices) and applicability uncertainty (variability in measurement procedures and between populations). Methods: We used real and synthetic data for ovarian cancer diagnosis to train 59400 models with variations in estimation, model, and applicability uncertainty. We then used these models to estimate the probability of ovarian cancer in a fixed test set of 100 patients and evaluate the variability in individual estimates. Findings: We show empirically that estimation uncertainty can be strongly dominated by model uncertainty and applicability uncertainty, even for models that perform well at the population level. Estimation uncertainty decreased considerably with increasing training sample size, whereas model and applicability uncertainty remained large. Interpretation: Individual risk estimates are far more uncertain than often assumed. Model uncertainty and applicability uncertainty usually remain invisible when prediction models or algorithms are based on a single study. Predictive algorithms should inform, not dictate, care and support personalization through clinician–patient interaction rather than through inherently uncertain model outputs. Funding: This research is supported by Research Foundation Flanders (FWO) grants G097322N, G049312N, G0B4716N, and

Ovarian cancer diagnosis

- Risk of malignancy of patients with ovarian tumor scheduled for surgery

- Data from 1133 patients from Leuven, 1999-2015

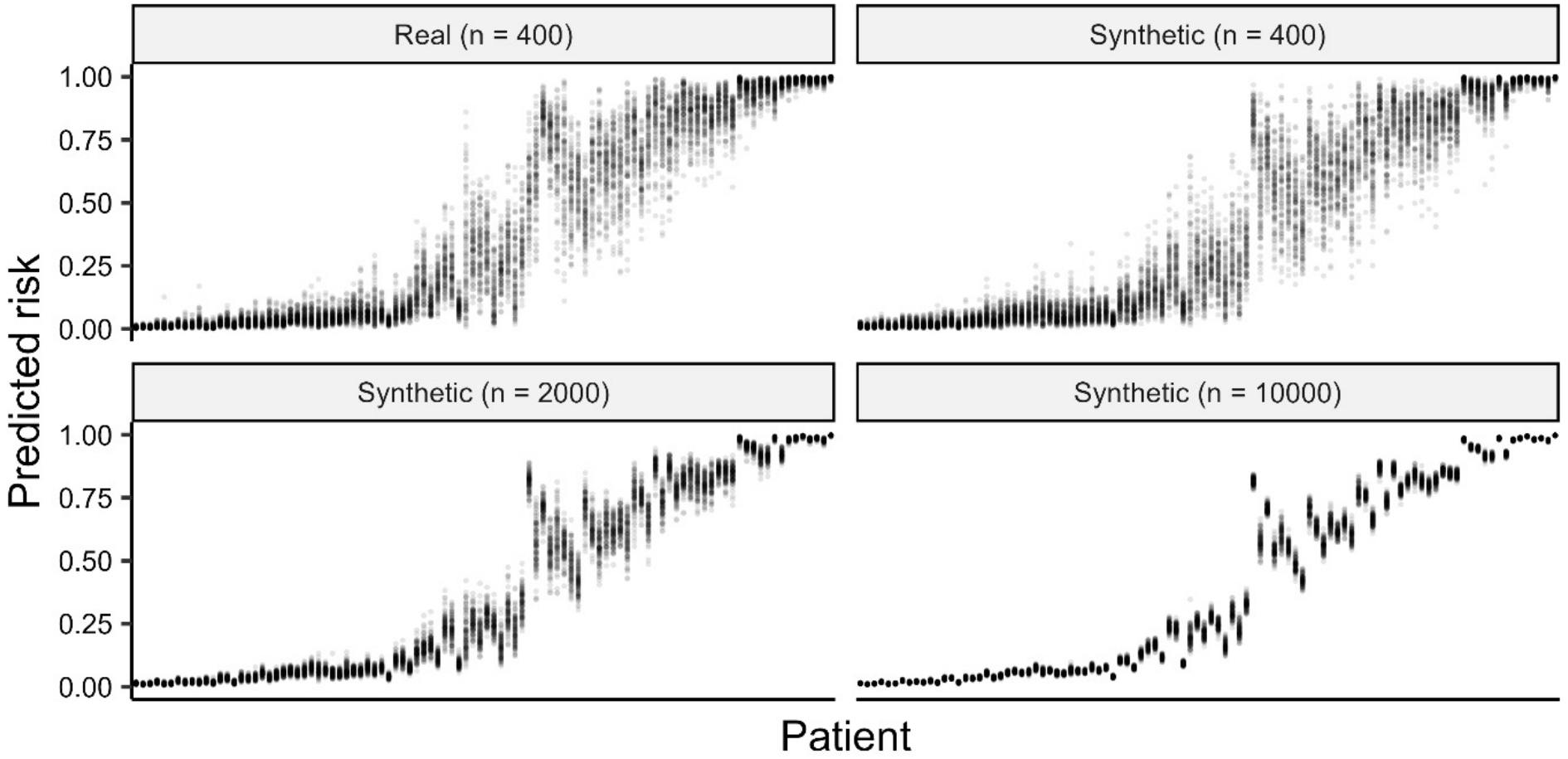
- Main model:
  - Logistic regression
  - age, max lesion diameter, proportion solid tissue, CA125, bilateral tumors (Y/N), papillations with blood flow (Y/N)
  - First 4 predictors: rcs with 3 knots (2 df)

- 10 parameters, 37% prevalence, assumed AUC=0.88 (cf literature): minimum required sample size n=359 (Riley et al 2020)

Illustration: ovarian cancer diagnosis

- Random test set of 100 patients, remaining 1033 is training pool
- Randomly 400 cases from training pool
- Impute missing CA125 values (31%) using regression imputation
- Train model on n=400
- Apply imputation model and prediction model to test set (n=100)
- For variation in sample size, create synthetic data sets

Logistic regression



Variations regarding model/modeler uncertainty

Modeling algorithm	Logistic regression (LR) Random forest (RF) Extreme gradient boosting (XGB)
Continuous variables (LR only)	Linear Dichotomization at median Categorization in 4 groups using quartile split Multivariable fractional polynomials (MFP) Restricted cubic splines with 3 knots
Variable selection (LR only)	None Backward elimination with alpha 0.01 Backward elimination with alpha 0.20
Penalization (LR only)	No Ridge, lambda tuned using AIC[b]
Minimum node size (RF)/ Maximum depth (XGB only)[c]	2 20 Tuned (10 fold CV oOn logloss)

Variations regarding applicability uncertainty

Applicability (Data)	Definition of size of lesion and solid component	Based on maximum diameters Based on volumes
	Handling of missing data for CA125	Regression imputation Median imputation conditional on outcome Missing indicator method
Applicability (Population)	Training data population	Leuven (Belgium) Malmö (Sweden) Rome (Italy)

Key results

59400 models with variations in approximation, model, and applicability uncertainty

Estimation uncertainty decreased considerably with increasing training sample size, whereas model and applicability uncertainty remained large

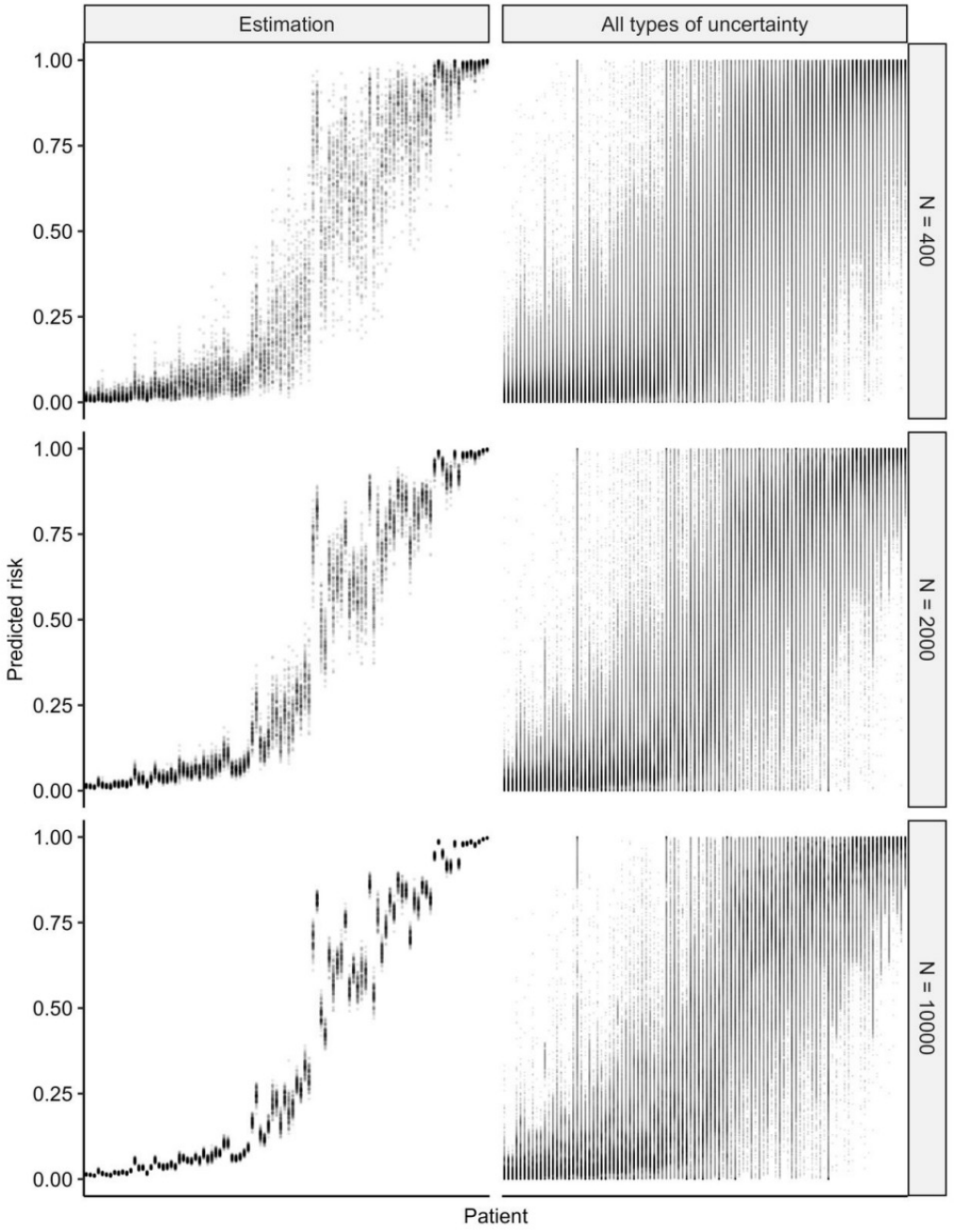
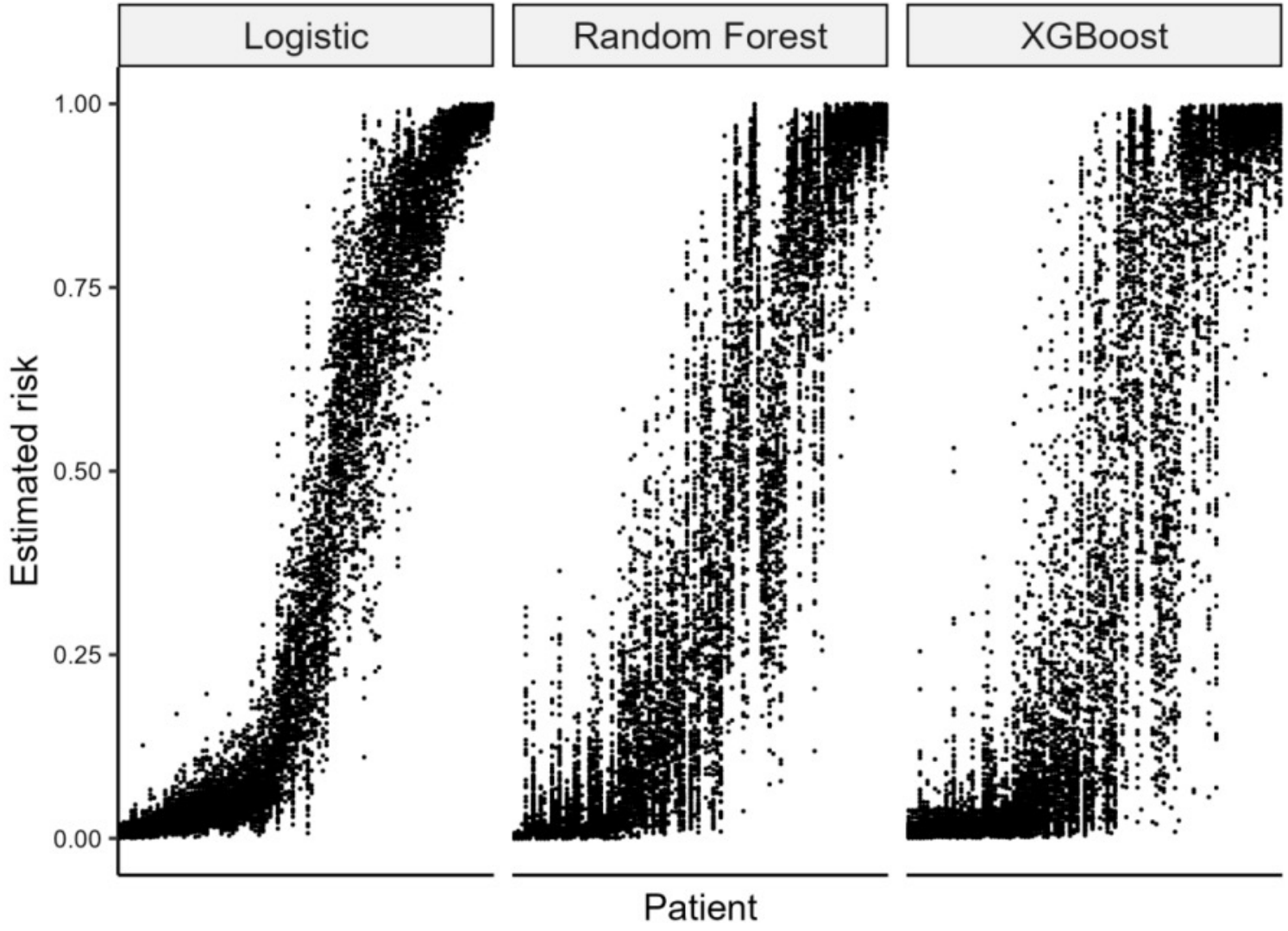


Fig. 2. Visualization of the uncertainty in the estimated risks for individual patients. Panels on the left show estimation uncertainty for the main model based on training sample sizes of 400 (top), 2000 (middle) and 10000 (bottom). Each panel contains the individual estimated risks (y-axis) for 100 test patients (x-axis) by 100 models based on randomly selected training datasets. The main model used unpenalized logistic regression with restricted cubic splines, with regression imputation for

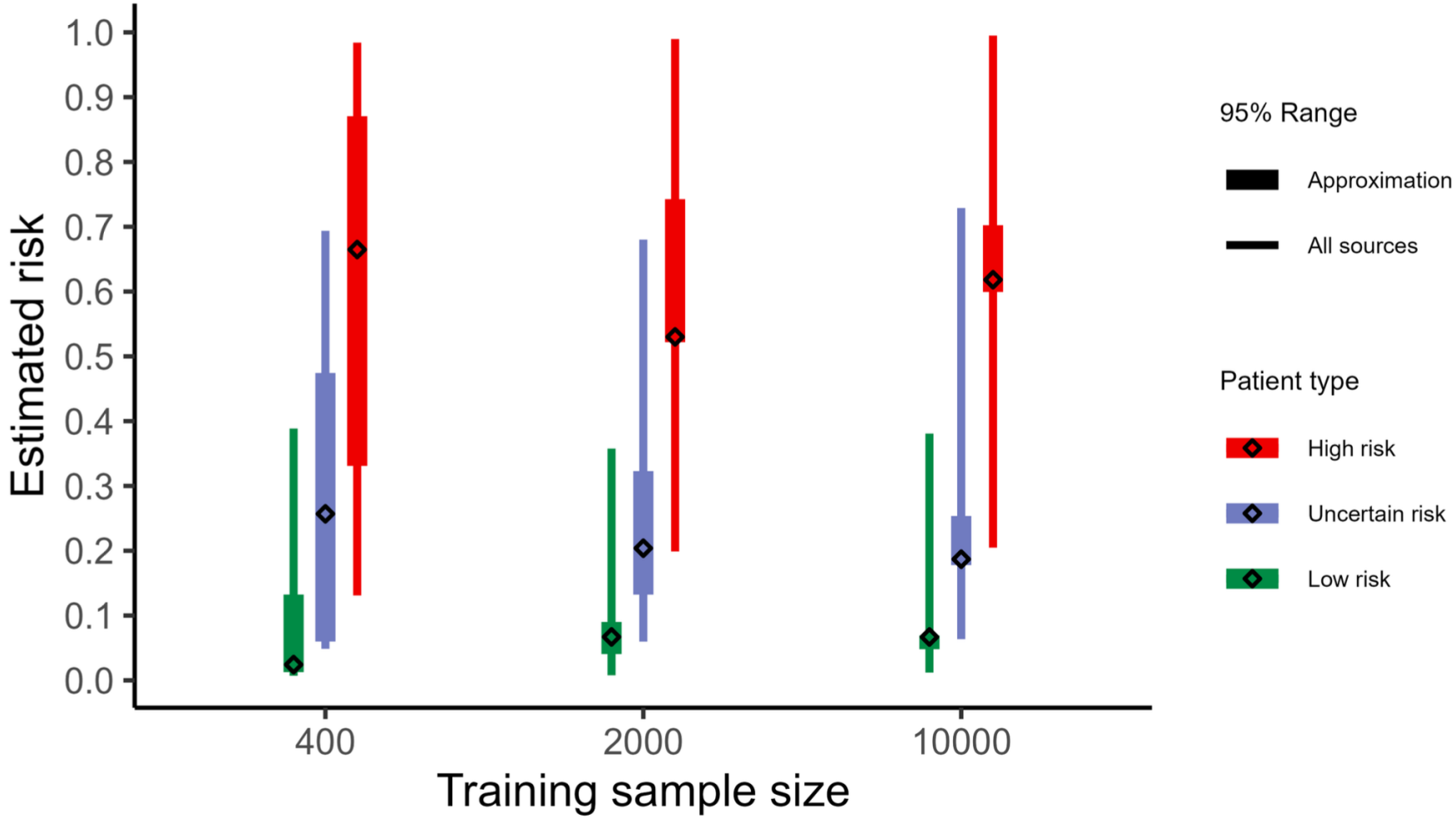
Estimation uncertaity larger in more flexible models

Fig. S1. Scatterplot of individual risk by patient to illustrate estimation uncertainty in real data (N =400). Each patient has 100 individual risks generated by models where the only variation was training sample.



Caution is needed

Illustration for 3 patients



Conclusions

- Sample size / approximation uncertainty is part of epistemic uncertainty
  - Large sample size avoids part of the problem, but only a small part
  - Other sources of uncertainty do not decrease with higher N
  - The influence of the modeler is large
  - More flexibility --> more data hungry
- Quantifying uncertainty completely is impossible (multiverse too large)
- Great caution is needed when interpreting individual risk estimates
  - Risk communication, shared decision making?
    - *Decisions with individualized estimates better than without (Spiegelhalter)*
  - Be careful with model explorations
  - Address heterogeneity between settings during development & validation

Further research

- Illustrate the fundamental problem of epistemic uncertainty
  - More case studies
  - Predictions and treatment benefit
- Guide model development: sacrifice population-level performance for individual-level certainty
  - Focus on key predictors
- Stabalize predictions for rare profiles
  - Effective N>10?
- Does 2nd order uncertainty lead to different decisions being optimal
  - Loss function is asymmetric?
- Risk communication
  - 95% CI *(PREDICT)* / Effective N / ...