Validation and Trustworthiness of AI based Predictions

March 2024, Aachen

Ewout W. Steyerberg, PhD

Professor of Clinical Biostatistics and Medical Decision Making

Dept of Biomedical Data Sciences

Leiden University Medical Center

Thanks to many for assistance and inspiration

Prediction

“10% risk”

Clinical prediction models

www.cancercalculators.org

evidencio.com/specialties

rasmus MC YouTube Maps Gmail

EVIDENCIO MEDICAL DECISION SUPPORT Models Validations About Pricing

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)

Validated → Trustworthy?

3-year survival after resection in patients with pancreatic cancer: updated amsterdam model

This prediction model for 3-year survival after resection in patients with pancreatic cancer was developed in 2015 in Amsterdam UMC (location AMC) and externally validated in 2019 in 3081 patients from 8 countries USA, UK, Germany, Italy, Sweden, the Netherlands, Korea, Australia).PubMed ID 3192443...

Filtered by: × CE Certified

There were no models found matching your search criteria

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

Trustworthiness of predictions

Mathematical models: Questions of trustworthiness

A Morton - The British journal for the philosophy of science, 1993 - journals.uchicago.edu

… It gives us many of the **predictions** we based our lives upon, and is essential to the marriage between experiment and theory in many parts of science. We rely on mathematical models …

☆ Opslaan 99 Citeren Geciteerd door 181 Verwante artikelen Alle 10 versies

Source: *The British Journal for the Philosophy of Science*, Dec., 1993, Vol. 44, No. 4 (Dec., 1993), pp. 659-674

Mathematical models needed:

- Complex processes
- No simple prediction via a deterministic theory

Modelling assumptions:

- Generally false
- Intelligent guesswork

Medical prediction: y ~ X

Topics: trustworthy predictions

1. What do we need for individual patients?
   - Internal vs external validity
   - Calibration vs discrimination
2. Trustworthy processes to build a prediction model?
   - AI
   - Humans
   - Requirements
3. Types of uncertainty
   - Statistical aspects
   - Model uncertainty
   - Heterogeneity between contexts of practical application

Validated = trustworthy?

- Classic:
  - No validation, only internal → low ranking journal
  - 1 convincing validation → top journal
- Modern
  - Substantial heterogeneity in performance

→ There is no such thing as a validated model

Opinion | Open access | Published: 24 February 2023

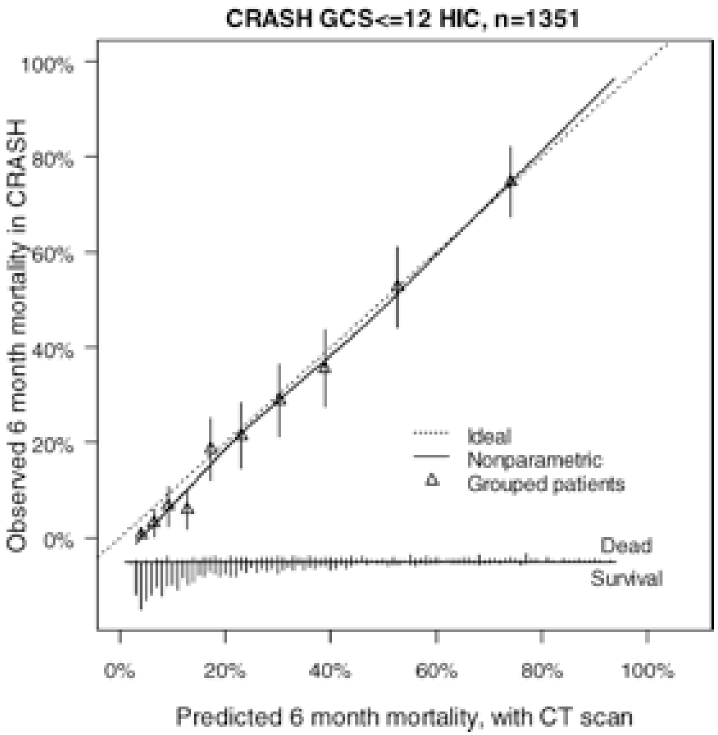
There is no such thing as a validated prediction model

Ben Van Calster, Ewout W. Steyerberg, Laure Wynants & Maarten van Smeden

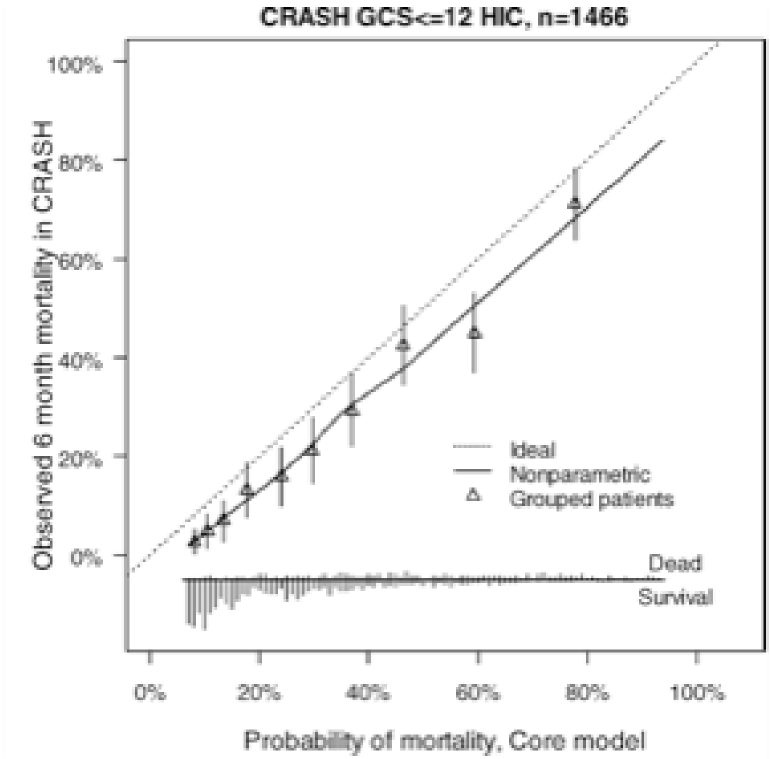
BMC Medicine **21**, Article number: 70 (2023) | Cite this article

Example in neurotrauma, external validation

well calibrated



over prediction

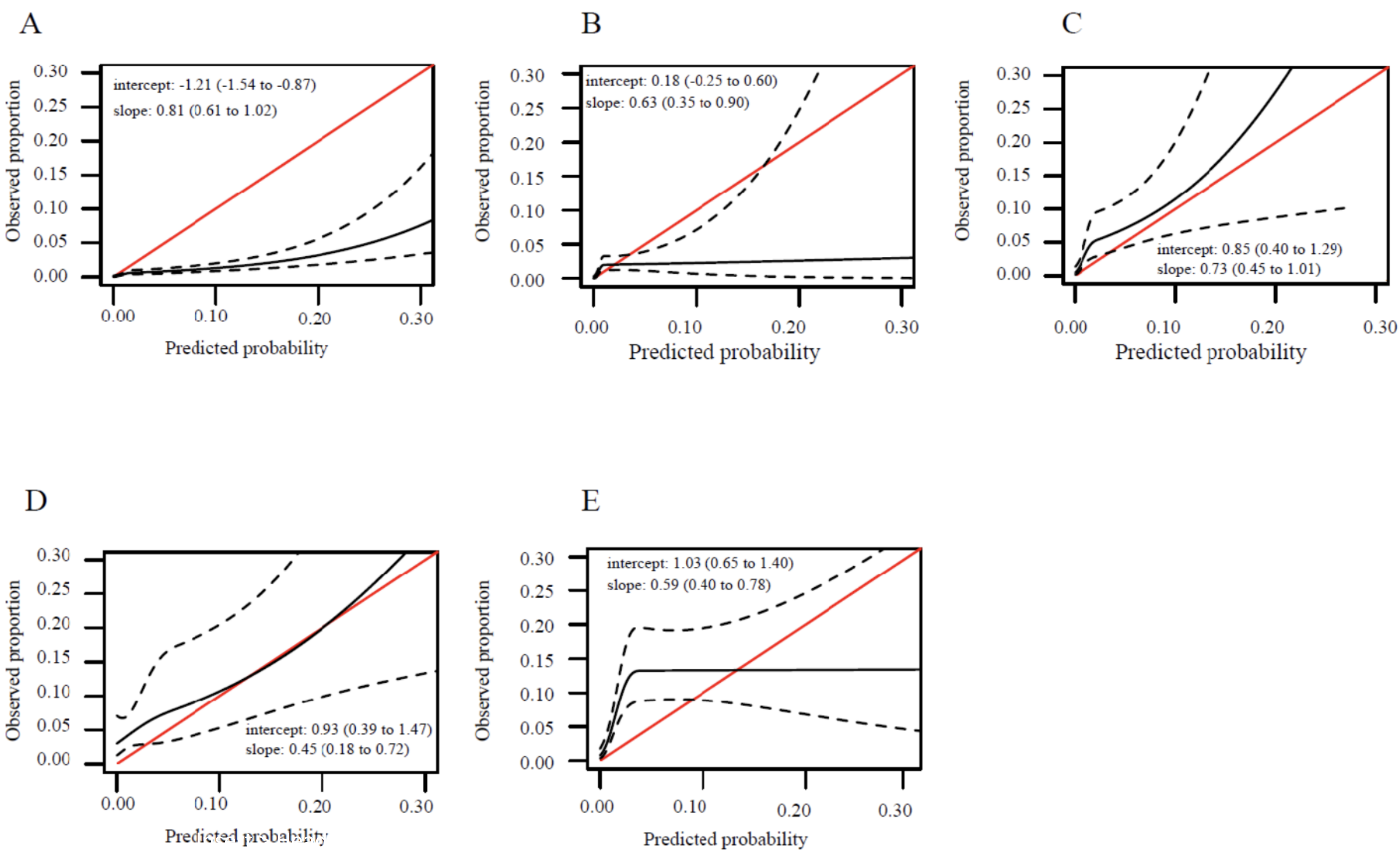


Modification of `val.prob()` in rms; **`val.prob.ci.2()`**

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.



npj | digital medicine

www.nature.com/npjdigitalme

COMMENT OPEN

Check for updates

Perspectives on validation of clinical predictive algorithms

Anne A. H. de Hond[1,2,3]✉, Vaibhavi B. Shah[2], Ilse M. J. Kant[4], Ben Van Calster[3,5], Ewout W. Steyerberg[1,3] and Tina Hernandez-Boussard[2,6,7]

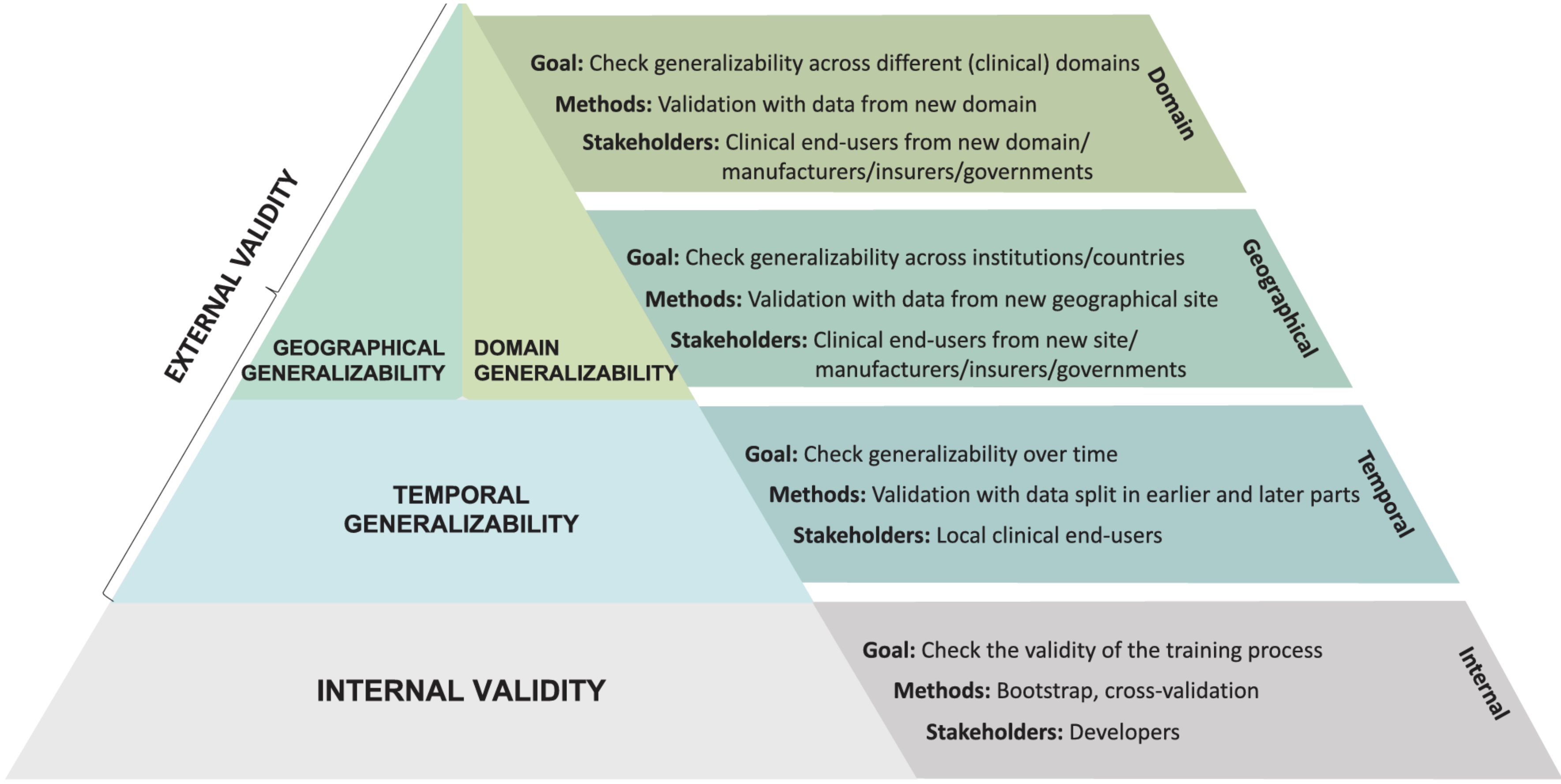


Fig. 1 Generalizability types. Schematic overview of the different types of generalizability with the validation's goals, methods, and stakeholders.

Summary validation

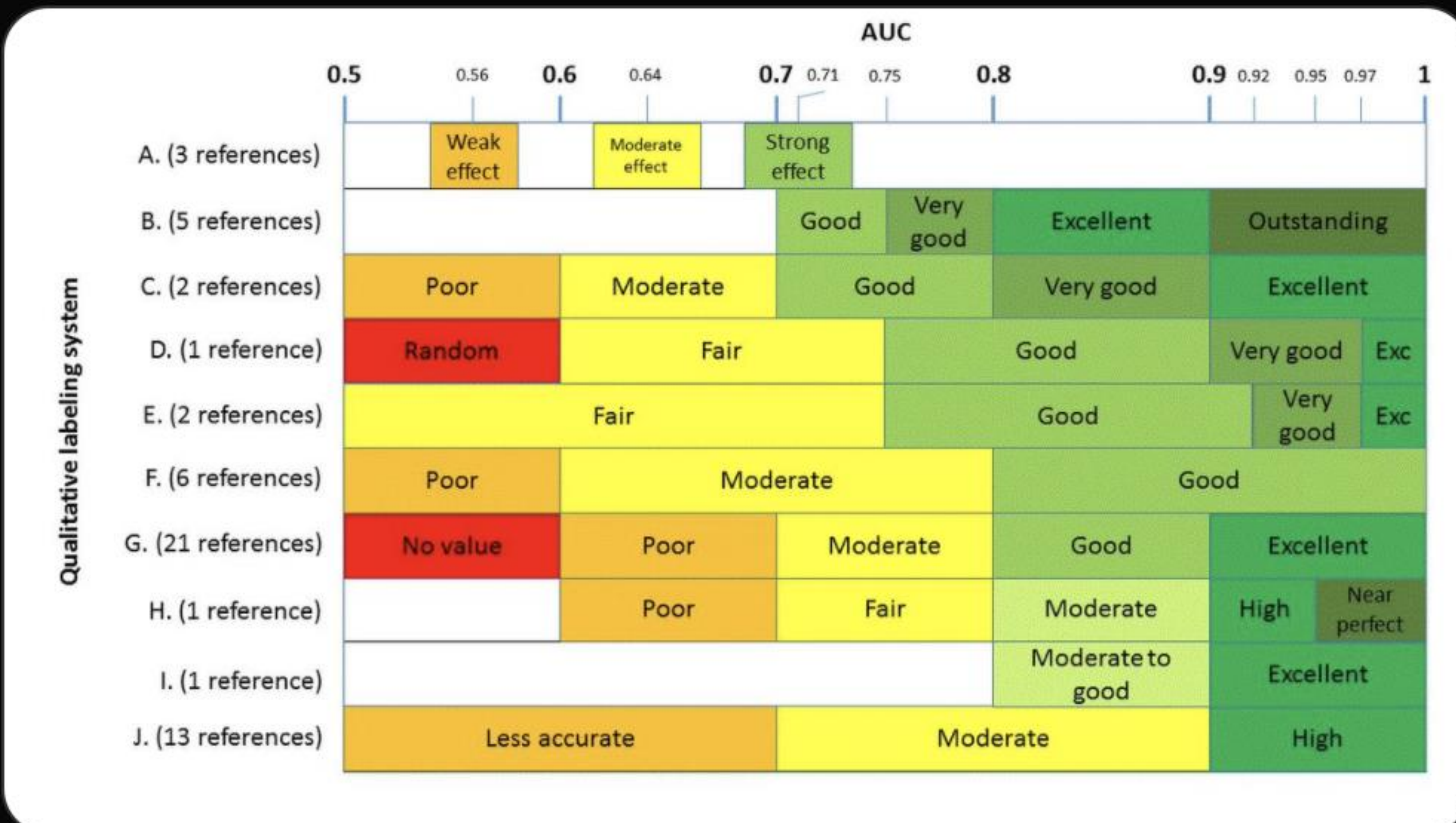
- Internal: minimum, same underlying population
- External:
  - Temporal
  - Geographic
  - Domain
- Efficient: internal-external cross-validation

Performance assessment

- What is the most commonly reported measure for performance of prediction models?
  - Area under the Receiver Operating Characteristic curve (AUC), or concordance (c) statistic
  - Discrimination = spread of predictions between individuals
    - Higher if better predictors in model
    - Higher --> more trustworthy??

Ben Van Calster @BenVanCalster · 25 mei

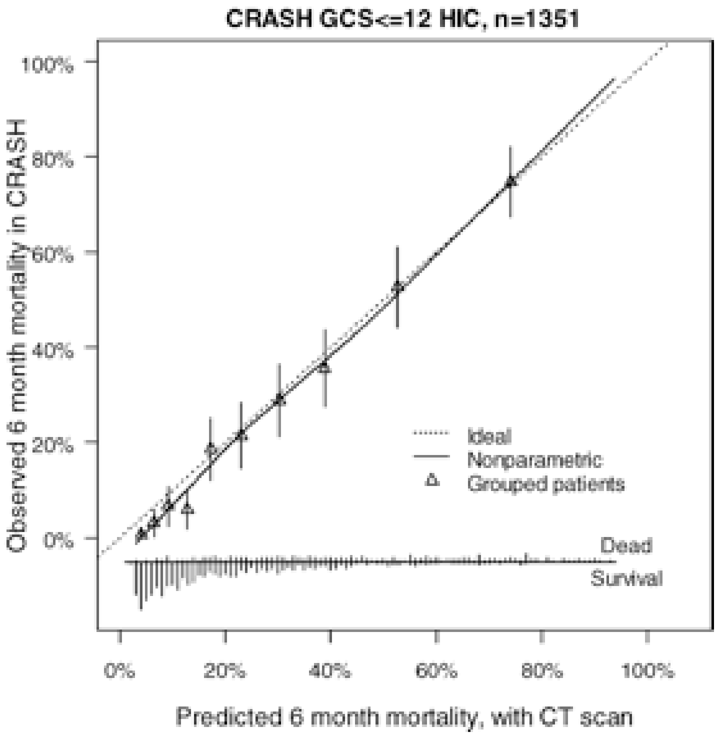
I searched the literature some day about how people think about different AUC values. Interesting...



Trustworthiness for individuals

- Calibration = reliability of predictions per individual

well calibrated



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

Steyerberg et al, PLoS Med 2008

Trustworthiness for individuals

- Calibration = reliability of predictions per individual
  - True risk estimates UTOPIAN

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]

  - Calibration underreported

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]

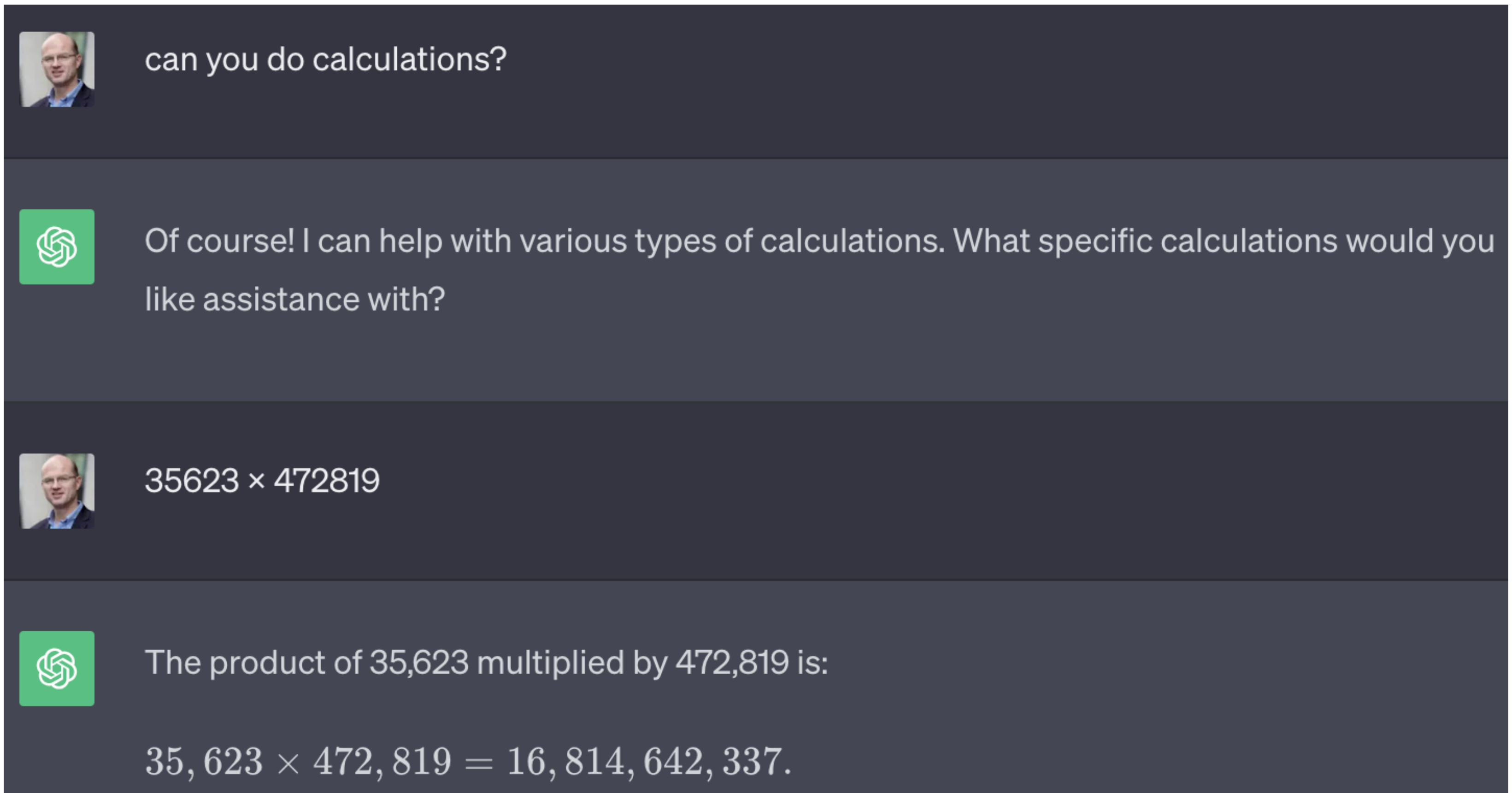
Topics: trustworthy predictions

1. What do we need for individual patients?
    - Internal vs external validity
    - Calibration vs discrimination
2. Trustworthy processes to build a prediction model?
    - AI
    - Humans
3. Types of uncertainty
    - Statistical aspects
    - Model uncertainty
    - Heterogeneity between contexts of practical application

Trustworthiness of ChatGPT

ChatGPT: may be hallucinating

- Simple calculations:

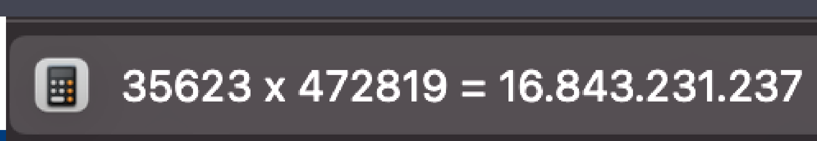

can you do calculations?

Of course! I can help with various types of calculations. What specific calculations would you like assistance with?

35623 × 472819

The product of 35,623 multiplied by 472,819 is:

$$35,623 \times 472,819 = 16,814,642,337.$$



are you sure about this calculation?

Apologies for the mistake. Let me correct that for you.

$$35,623 \times 472,819 = 16,814,506,737.$$

My apologies for the earlier response. Let me recalculate that for you:

35623 × 472819 = 16,827,062,037

So, the correct product of 35,623 multiplied by 472,819 is 16,827,062,037.

35623 x 472819 = 16.843.231.237

Trustworthiness and AI

- Relation to evidence base
- Other popular terms related to AI
  - Fairness
  - Equity

→ Let's ask ChatGPT

what are important ethical concerns regarding AI for predictive purposes?

Ethical concerns regarding AI for predictive purposes are numerous and evolving as AI technologies advance. Here are some important ethical considerations:

1. **Bias and Fairness:**
   AI models can perpetuate or even amplify biases present in the data they are trained on. Ensuring fairness and addressing bias is a critical concern, especially in predictive models related to sensitive aspects like race, gender, or socioeconomic status.
2. **Privacy:**
   Predictive AI often relies on vast amounts of personal data. Safeguarding privacy, obtaining informed consent, and ensuring data is used appropriately and securely are essential ethical considerations.
3. **Transparency and Explainability:**
   Understanding how an AI model arrives at its predictions is crucial for trust and accountability. Models should be transparent and explainable to users, making it easier to identify potential biases and errors.
4. **Accountability and Responsibility:**
   Defining responsibility and accountability for the actions and decisions made by AI systems is challenging. Determining who is responsible when things go wrong, or when the AI system produces harmful predictions, is a key ethical concern.
5. **Autonomy and Decision-making:**
   Predictive AI can influence decision-making processes in various domains, from lending to criminal justice. Striking the right balance between automation and human intervention, preserving human autonomy, and avoiding undue influence is vital.

6. **Security:**
   Ensuring the security of AI systems is essential to prevent malicious use, data breaches, and other potential risks associated with predictive AI.
7. **Long-term Effects and Unintended Consequences:**
   Predictive AI can have unforeseen and far-reaching consequences, including social, economic, and cultural impacts. Anticipating and mitigating these consequences is a significant ethical consideration.
8. **Informed Consent and User Rights:**
   Obtaining meaningful informed consent from individuals whose data is used in predictive models is important. Individuals should have rights over their data and should be informed about how their data is being used.
9. **Data Quality and Representativeness:**
   Ensuring that the data used to train predictive models is of high quality and representative of the population it aims to predict for is essential to avoid skewed or inaccurate predictions.
10. **Human Augmentation vs. Replacement:**
    Balancing the use of AI for predictive purposes to enhance human capabilities (augmentation) versus replacing human decision-making entirely raises important ethical questions regarding the role of AI in society.

Addressing these ethical concerns is crucial to ensure the responsible and beneficial development and deployment of AI for predictive purposes.

Can you group these concerns further to fewer groups?

Certainly! We can group these ethical concerns into broader categories view:

1. **Data and Model Integrity:**
   a. **Bias and Fairness**
   b. **Data Quality and Representativeness**
   c. **Transparency and Explainability**
2. **Privacy and Consent:**
   a. **Privacy**
   b. **Informed Consent and User Rights**
3. **Accountability and Responsibility:**
   a. **Accountability for AI Actions**
   b. **Security**
   c. **Long-term Effects and Unintended Consequences**
4. **Human-Centric Concerns:**
   a. **Autonomy and Decision-making**
   b. **Human Augmentation vs. Replacement**

Trustworthy models?

- Modeling flexibility: friend or foe?
  Human oversight on:
  - Classical modeling: selection of predictors; nonlinearity; interactions
  - AI: hyperparameters; technique CART / RF / XGBoost / nnet / ..

NEWS | 12 October 2023

Reproducibility trial: 246 biologists get different results from same data sets

Wide distribution of findings shows how analytical choices drive conclusions.

Trustworthiness: poor for human modelers

Red cards and dark skin soccer players

https://psyarxiv.com/qkwst/

Empirical Article

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

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

DOI: 10.1177/2515245917747646
www.psychologicalscience.org/AMPPS
SAGE

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 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

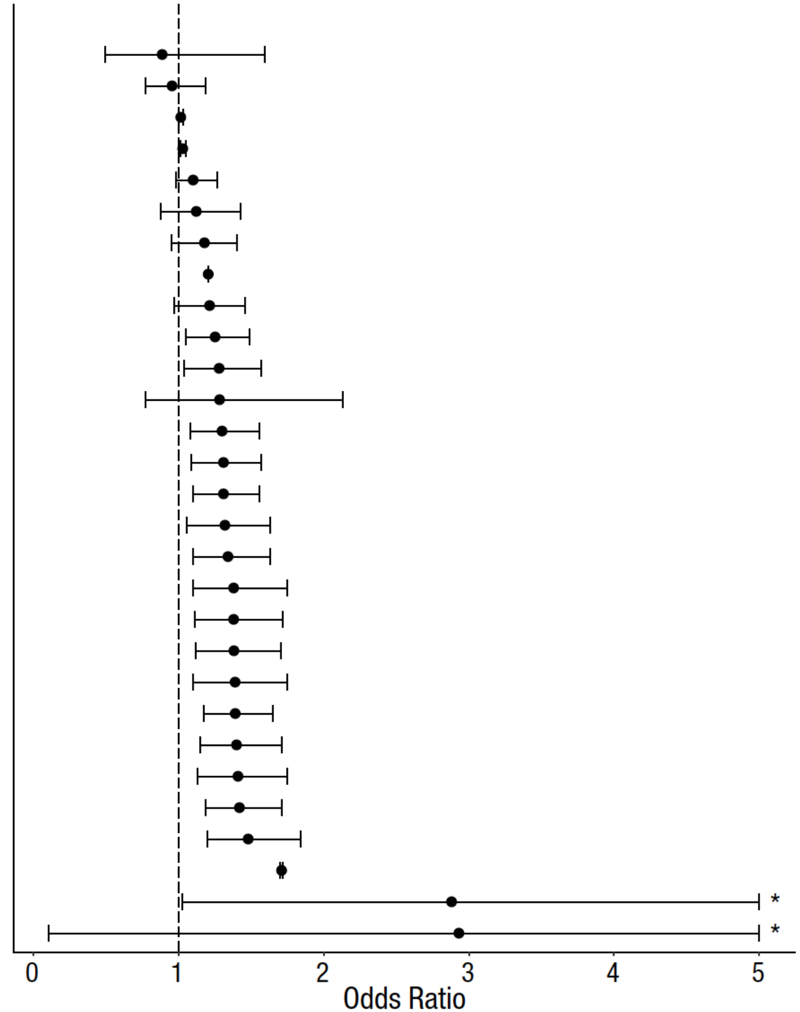


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that 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 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

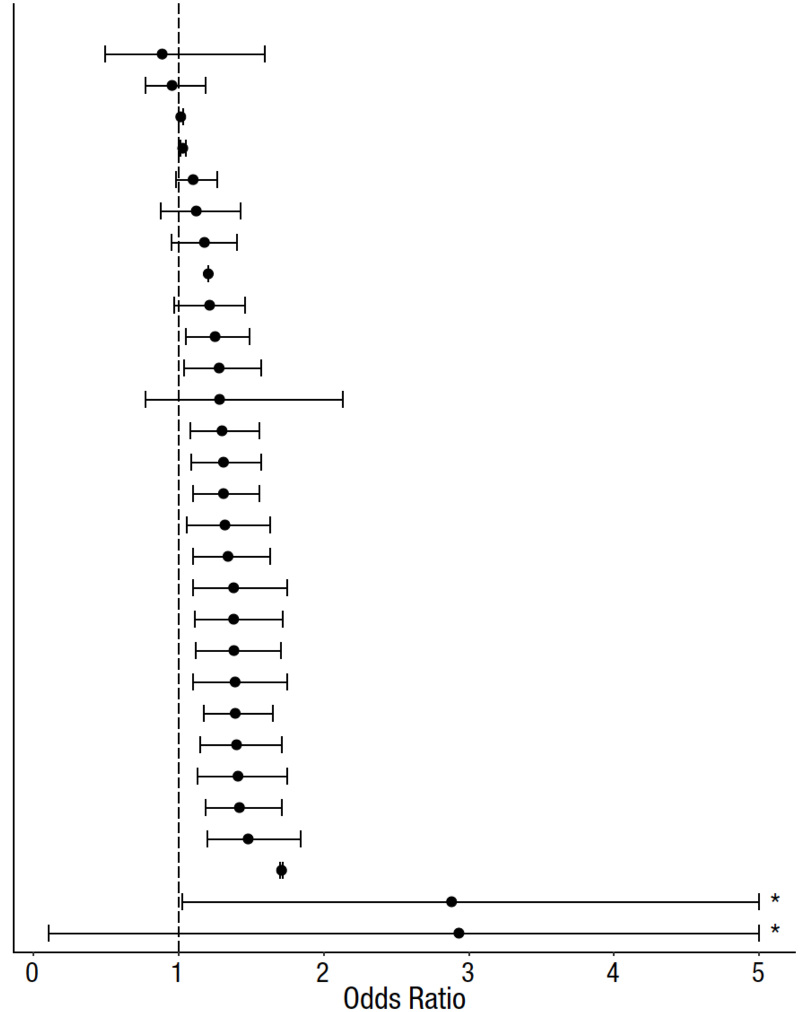


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players’ skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that 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.

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

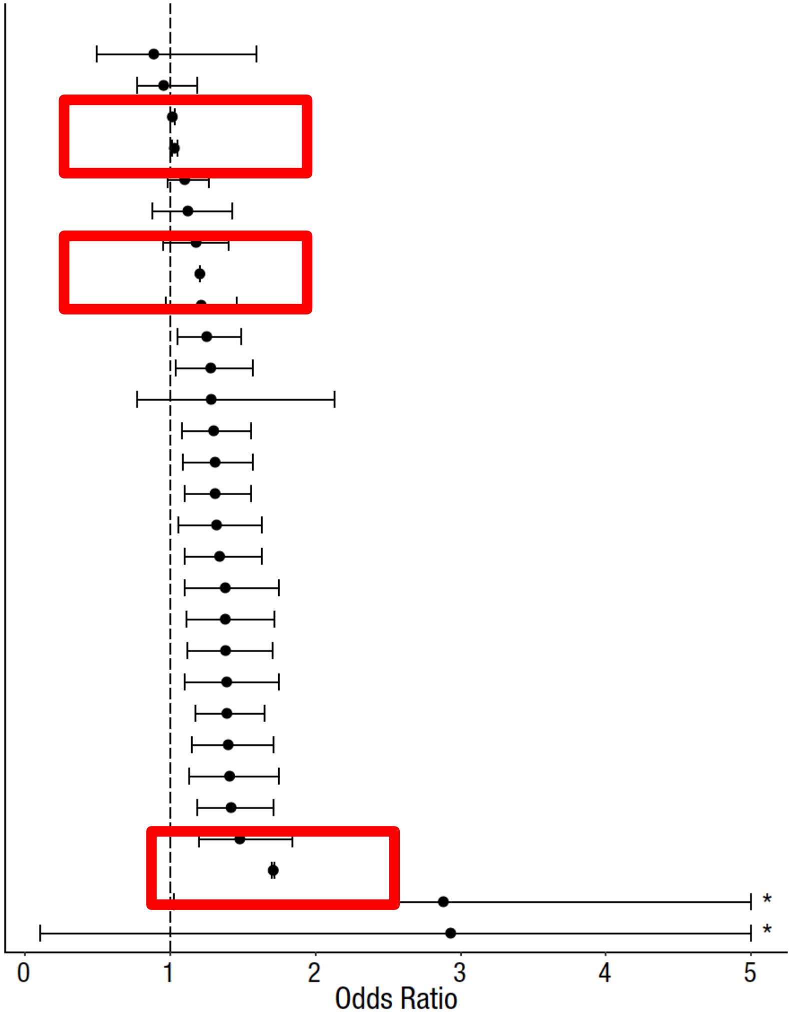


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that 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”**

Corrigendum

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

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

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(4) 580

DOI: 10.1177/2515245918810511
www.psychologicalscience.org/AMPP
SAGE

“ .. the authors forgot to add a citation of the *Nature* commentary to the final published version of the *AMPPS* article or to note that the main findings had been previously publicized via the commentary, the online preprint, research presentations at conferences and universities, and media reports by other people. The authors regret the oversight.”

Explainable to humans = trustworthy?

- Explainable AI
  - Algorithm trustworthy: if predictions are based on factors that are acceptable to domain experts instead of on 'spurious correlations'
  - SHAP (SHapley Additive exPlanations) values
    "By using SHAP values, researchers and practitioners can gain a deeper understanding of how different features influence model predictions, leading to improved model interpretability and trust."
    ChatGPT3.5

Topics: trustworthy predictions

1. What do we need for individual patients?
   - Internal vs external validity
   - Calibration vs discrimination
2. Trustworthy processes to build a prediction model?
   - AI
   - Humans
3. Types of uncertainty
   - Statistical aspects
   - Model uncertainty
   - Heterogeneity between contexts of practical application

Approaches to uncertainty quantification

- Sample size
  - specifically **#events** for binary outcome prediction
- 'patients like you' and exceptionality
  - For risk communication: aleatory uncertainty
  - For uncertainty communication: epistemic

Example on presentation by 'the king of nomograms'

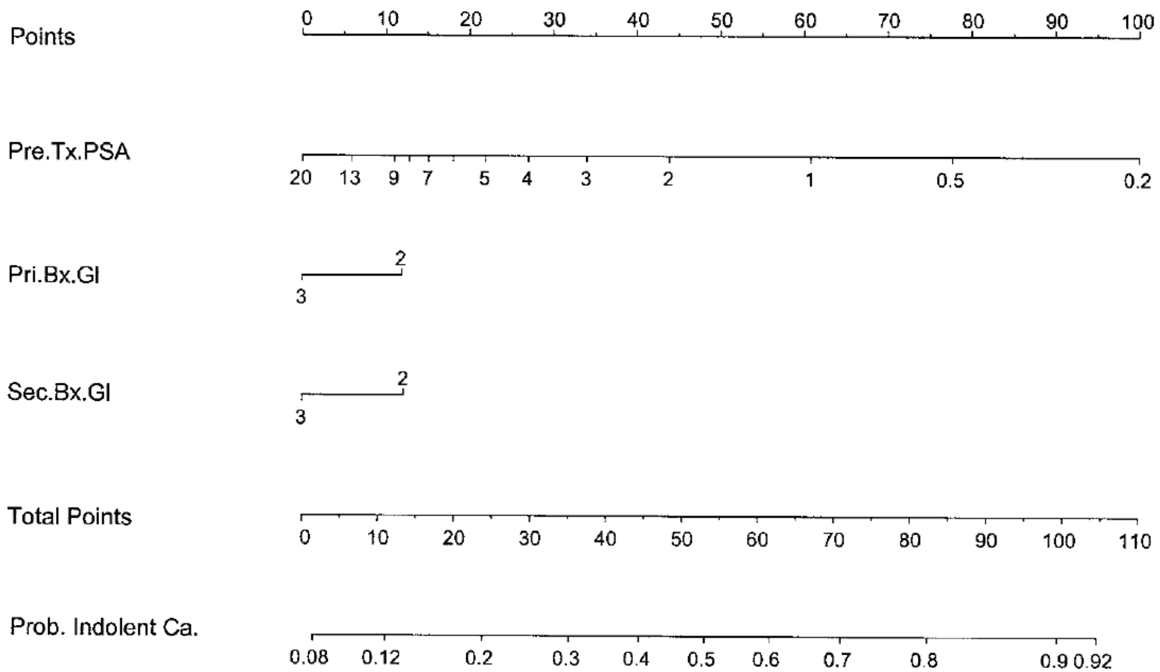
0022-5347/03/1705-1792/0
The Journal of Urology®

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,

PREDICTION OF INDOLENT PROSTATE CANCER



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.

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."

N = 100?

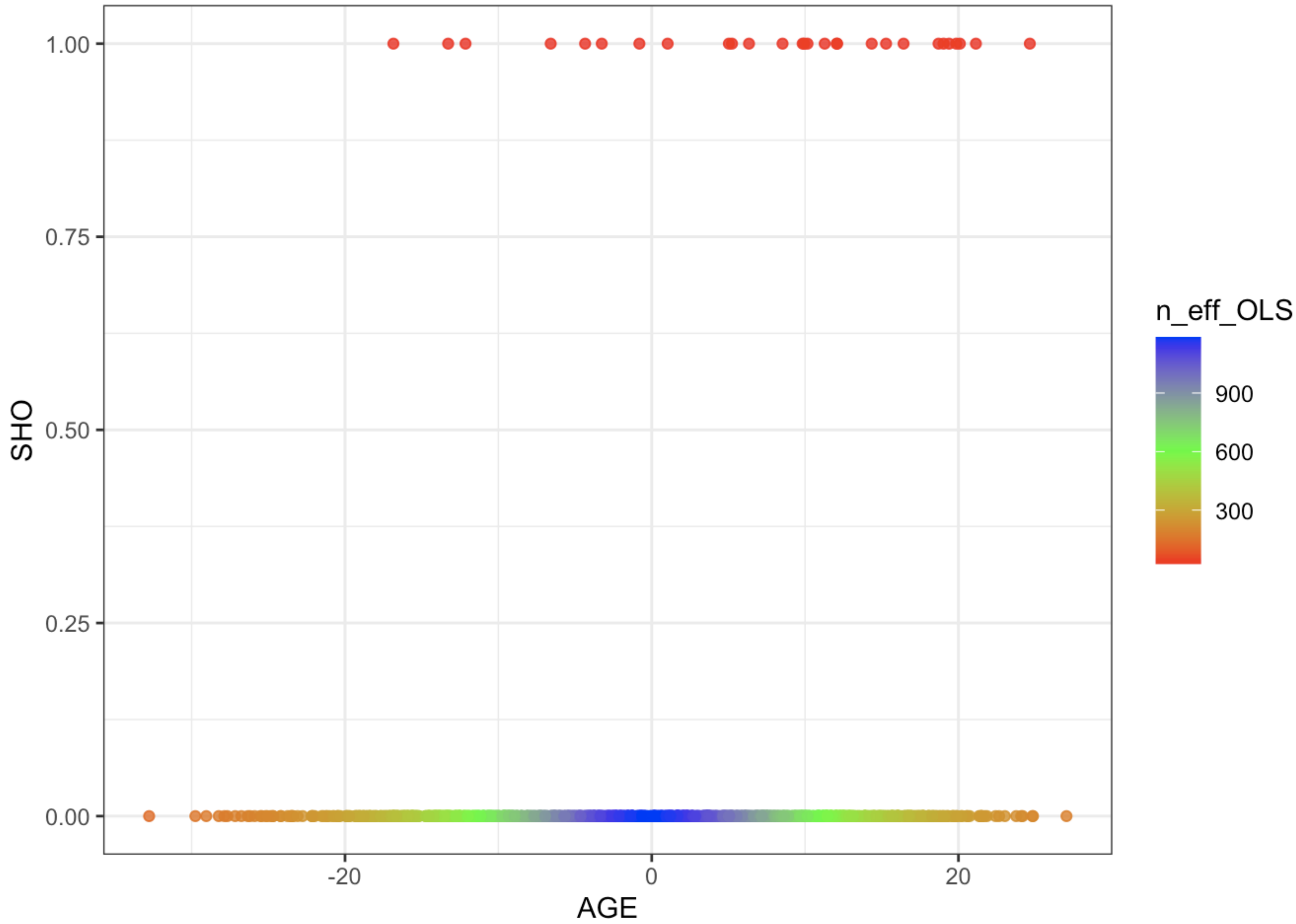
Effective n visualisations

Doranne Thomassen

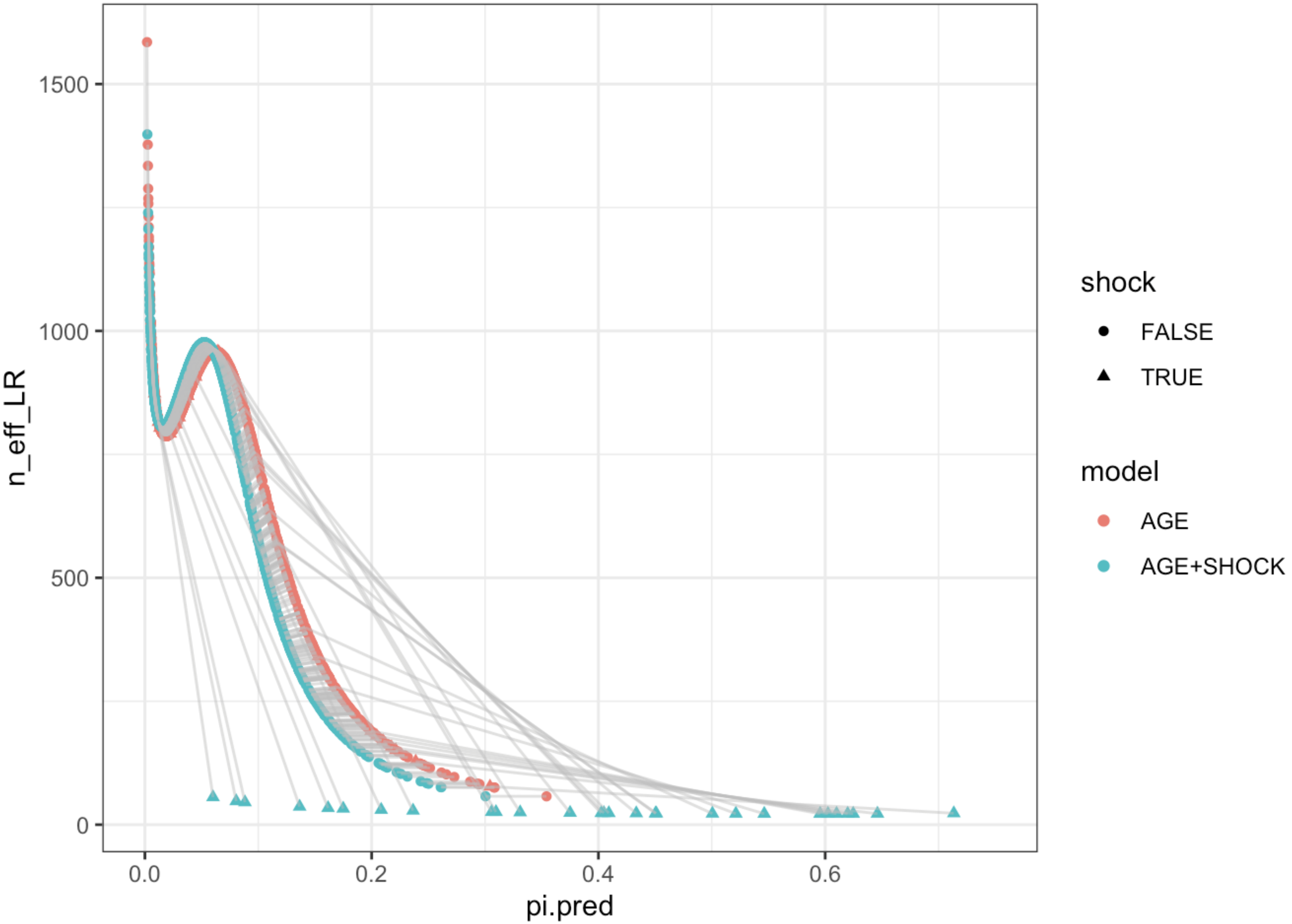
2021-10-27

GUSTO data, n=1200
(out of 40,830)

Uncertainty for rare, strong predictor: SHOCK



Predictions with and without SHOCK in the model



AI: arXiv paper

To Trust or Not to Trust a Regressor: Estimating and Explaining Trustworthiness of Regression Predictions

Kim de Bie[1] Ana Lucic[1] Hinda Haned[1]

"Trustworthiness expresses whether a prediction is aligned with the train data"

- Distance between the new patient and similar patients from the training data estimates the trustworthiness of a prediction; resembles the reference data

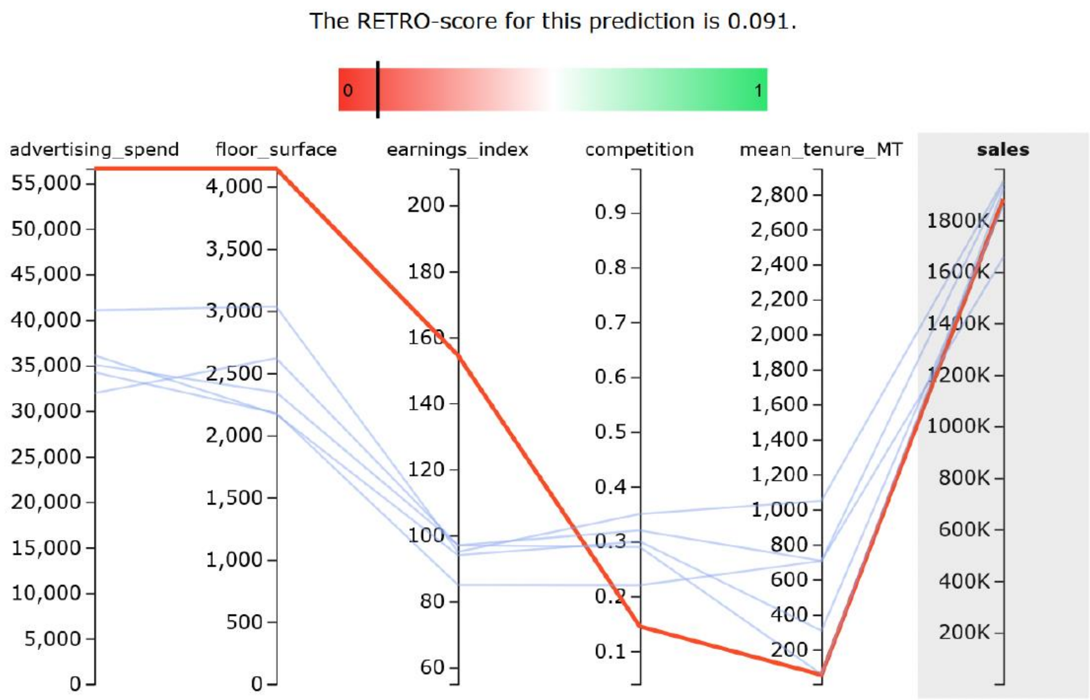


Figure 1. Example RETRO-VIZ output for an untrustworthy prediction. This model predicts sales based on five features. The RETRO-

Claim

”Effective N is an attractive concept to address epistemiologic uncertainty”

- Analytic solutions for regression models
  - Minimum certainty, say, n>10, for model specification: selection / shrinkage?
- Approximate solutions for machine learning models
  - Bootstrap
- Effective N: conditional on the model

Model uncertainty

- 246 biologists modeling
- 61 analysts in 29 teams on the Red Card problem
- …

- Comparisons between classic vs machine learning

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,*]

Arjun (Raj) Manrai
@arjunmanrai

(Thread) The paper by Evangelia et al. in @JClinEpi on 'logistic regression = machine learning' for medicine has generated many reactions. This paper may be misinterpreted by #MachineLearning cynics and enthusiasts alike

Arjun (Raj) Manrai @arjunmanrai · 12 feb.

There are notable absences, such as many of the seminal contributions of deep learning to image analysis in medicine (e.g. Gulshan et al. JAMA 2016 and Esteva et al. Nature 2017). 7/n

Original Investigation | Innovations in

December 13, 2016

Development and Vali
Detection of Diabetic
Photographs

Varun Gulshan, PhD[1]; Lily Peng, MD, PhD[1]; Marc Coram

» Author Affiliations | Article Information

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.201

Machine Learning Website

ure
rnal of science

d: 25 January 2017

tologist-level classification
with deep neural network

Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swett

118 (02 February 2017) | Download Citation

um to this article was published on 28 June 2017

Differences in discrimination

	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

Fig. 4. Differences in discriminative ability between LR and ML models, overall and according to risk of bias (n = 282 comparisons).

Topics: trustworthy predictions

1. What do we need for individual patients?
    - Internal vs external validity
    - Calibration vs discrimination
2. Trustworthy processes to build a prediction model?
    - AI
    - Humans
3. Types of uncertainty
    - Statistical aspects
    - Model uncertainty
    - Heterogeneity between contexts of practical application

Heterogeneity

- Study design
- Selection of subjects
- Disease domain
- Measurement of covariates
- Measurement of outcomes
- Associations of covariates with outcome
- Overall outcome rates

Heterogeneity in performance

Performance measure	IMPACT models	
Mortality	Discrimination: 56 validations Calibration: 31 validations	
	Mean[a]	Range
AUC	0.79	0.65-0.90
Calibration slope	1.1	0.42-2.3
Calibration intercept	-0.22	-3.3-0.93

Steyerberg et al, *PLoS Med 2008*

Dijkland S et al; *J Neurotrauma 2019*

Received: 4 June 2017 | Revised: 23 March 2019 | Accepted: 6 June 2019
DOI: 10.1002/sim.8296

RESEARCH ARTICLE

WILEY Statistics in Medicine

Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration

Ewout W. Steyerberg[1,2] | Daan Nieboer[2] | Thomas P.A. Debray[3,4] | Hans C. van Houwelingen[1]

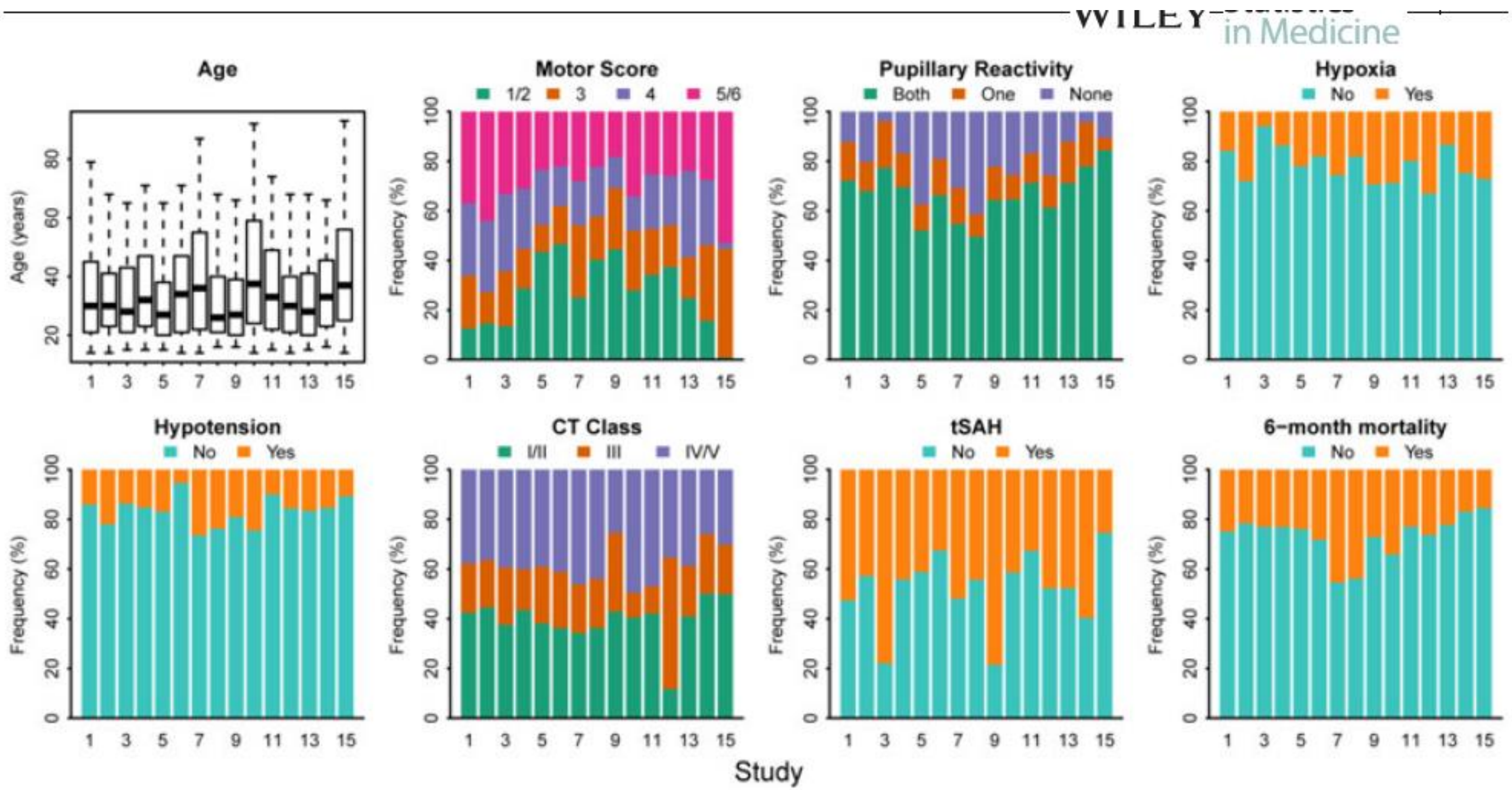
Statistics in Medicine. 2019;**38**:4290–4309.

15 cohorts: 11 RCTs, 4 Observational studies

Nr.	Name	Enrollment period	Type[1]	n
1	TINT	1991–1994	RCT	1118
2	TIUS	1991–1994	RCT	1041
3	SLIN	1994–1996	RCT	409
4	SAP	1995–1997	RCT	919
5	PEG	1993–1995	RCT	1510
6	HIT I	1987–1989	RCT	350
7	UK4	1986–1988	OBS	791
8	TCDB	1984–1987	OBS	603
9	SKB	1996–1996	RCT	126
10	EBIC	1995–1995	OBS	822
11	HIT II	1989–1991	RCT	819
12	NABIS	1994–1998	RCT	385
13	CSTAT	1996–1997	RCT	517
14	PHARMOS	2001–2004	RCT	856
15	APOE	1996–1999	OBS	756

[1] Type of study, RCT: randomized controlled trial, OBS: observational cohort

Heterogeneity in case-mix



Heterogeneity in predictor effects

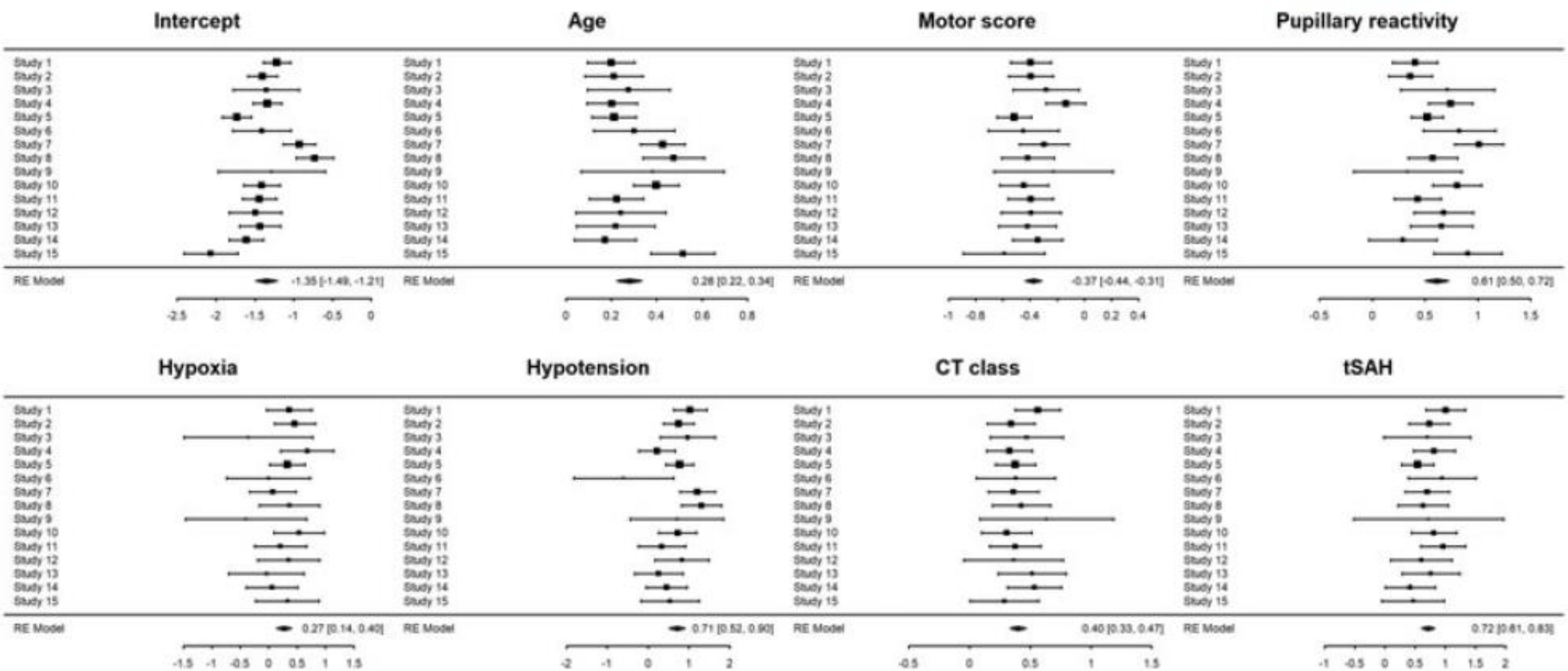


FIGURE 2 Forest plots showing estimated multivariable logistic regression coefficients and associated 95% confidence interval per study. The largest heterogeneity was noted for pupillary reactivity ($\tau = 0.17$) and hypotension ($\tau = 0.27$)

Heterogeneity in predictions

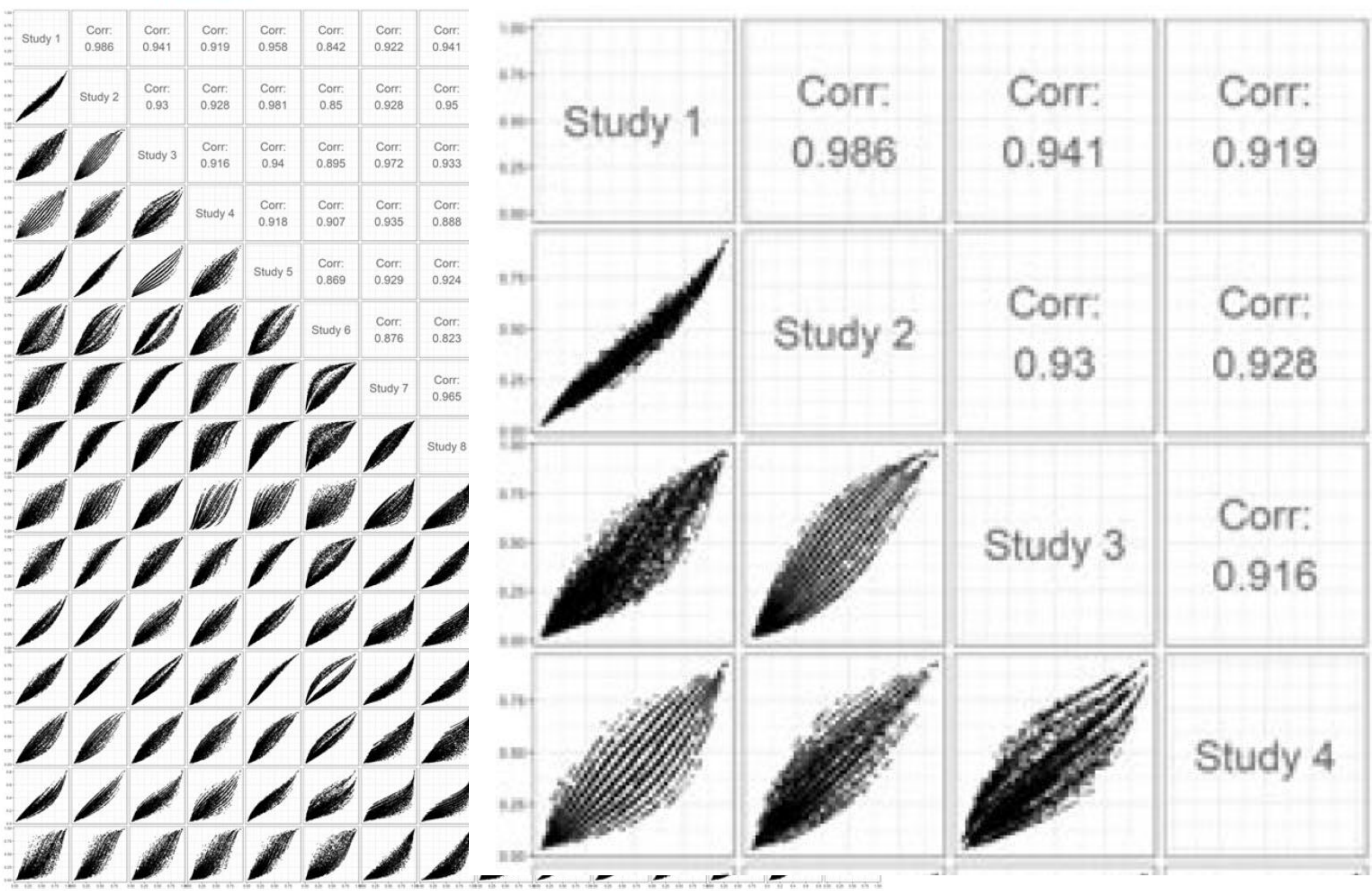
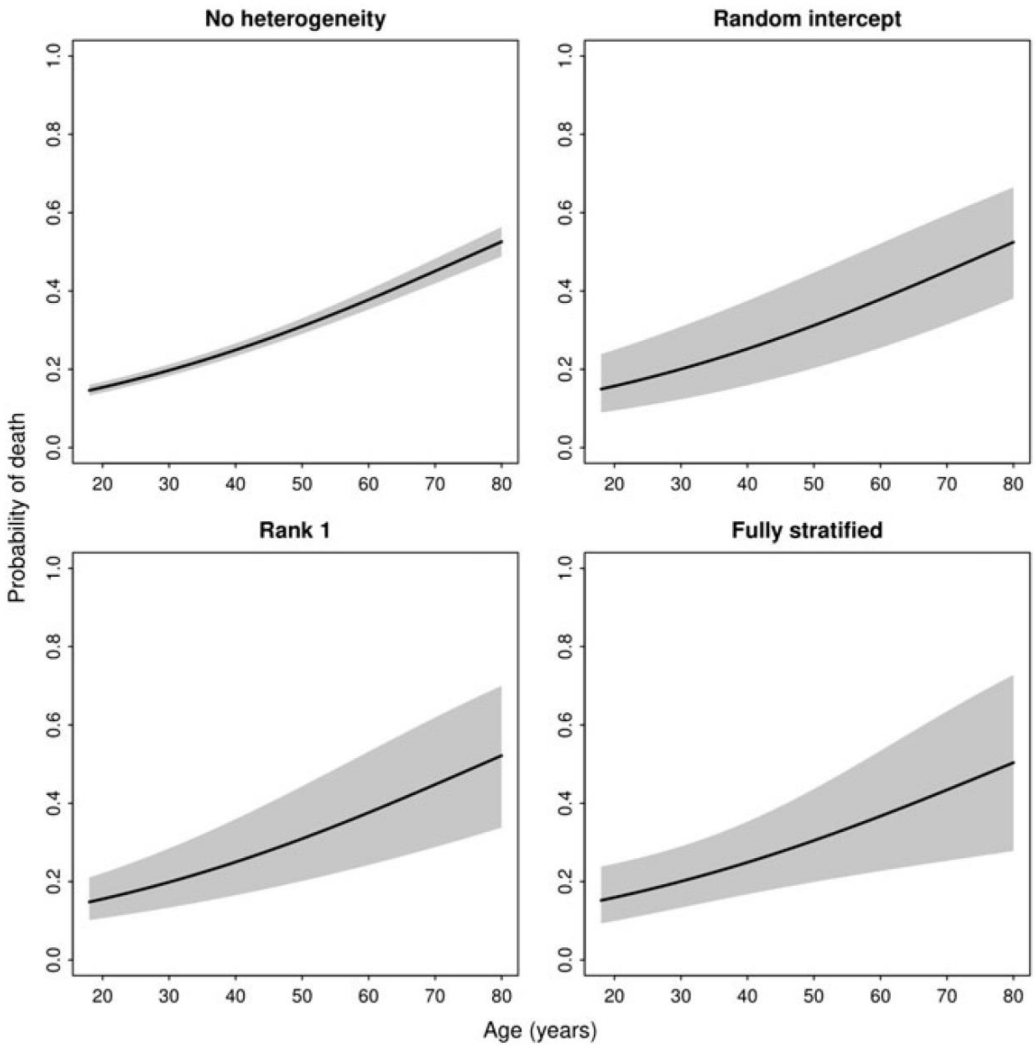


FIGURE 3 Correlation between predictions of study-specific models in a pairwise comparison between studies: 1-to-1 comparisons of predictions for all patients in the individual patient data set ($n = 11\,022$)

Heterogeneity in individual predictions



$$\text{Odds}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \exp(\alpha)\exp(\boldsymbol{X\beta}').$$

$$\text{logit}\left(p_{ij}\right) = \alpha_j + \boldsymbol{X}_{ij}\boldsymbol{\beta}',$$
$$\alpha_j \sim N\left(\alpha, \tau_\alpha^2\right).$$

$$\text{logit}\left(p_{ij}\right) = \alpha_j + \gamma_j \boldsymbol{X}_{ij}\boldsymbol{\beta}',$$
$$\left(\alpha_j, \gamma_j\right) \sim MVN\left((\alpha, 1), \mathbf{T}\right)$$

$$\text{logit}\left(p_{ij}\right) = \alpha_j + \boldsymbol{X}_{ij}\boldsymbol{\beta}_j{}',$$
$$\left(\alpha_j, \boldsymbol{\beta_j}\right) \sim MVN(\boldsymbol{\mu}, \mathbf{T}).$$

Conclusions on trustworthy predictions

- Epistemic uncertainty: under the influence of the modeler
  - Larger sample sizes
  - Modest modeling, limit flexibility

- Heterogeneity: assess differences between settings
  - Study design
  - Distribution and effects of covariates
  - Differences between predictions

- Model predictions suffer from multiple sources of uncertainty
  - Transparency: for policy makers / physicians / patients
  - Context dependency: Local versus global models