THE UNIVERSITY OF MELBOURNE

RMIT UNIVERSITY

IBM

ARC TRAINING CENTRE IN COGNITIVE COMPUTING FOR MEDICAL TECHNOLOGIES

When to trust a classifier for quality assessment of medical evidence?

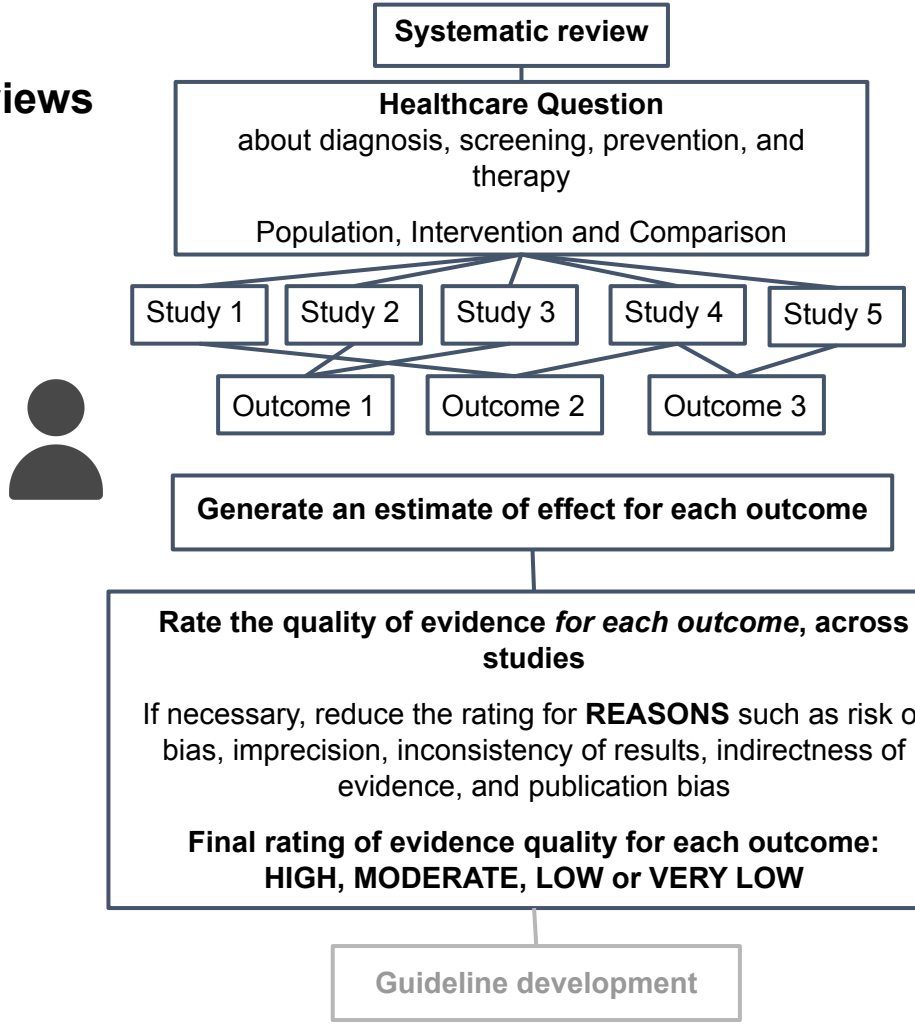


Simon Šuster, Tim Baldwin and Karin Verspoor,

with the help of Jey Han Lau, Antonio Jimeno Yepes, David Martinez Iraola and Yulia Otmakhova

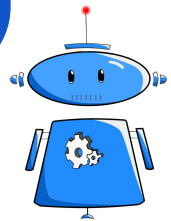
8 June 2022

Constructing systematic reviews and quality assessment



Our goal:

Assume we're given a piece of evidence from a systematic review, predict its quality

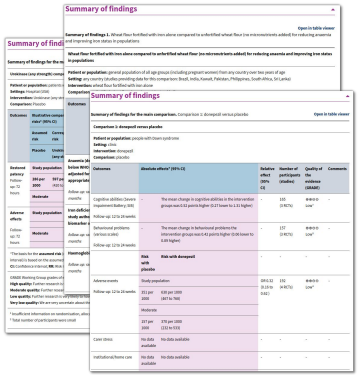


Dataset + Tasks + Models with heterogeneous inputs (structured and non-structured)

EvidenceGRADEr in brief

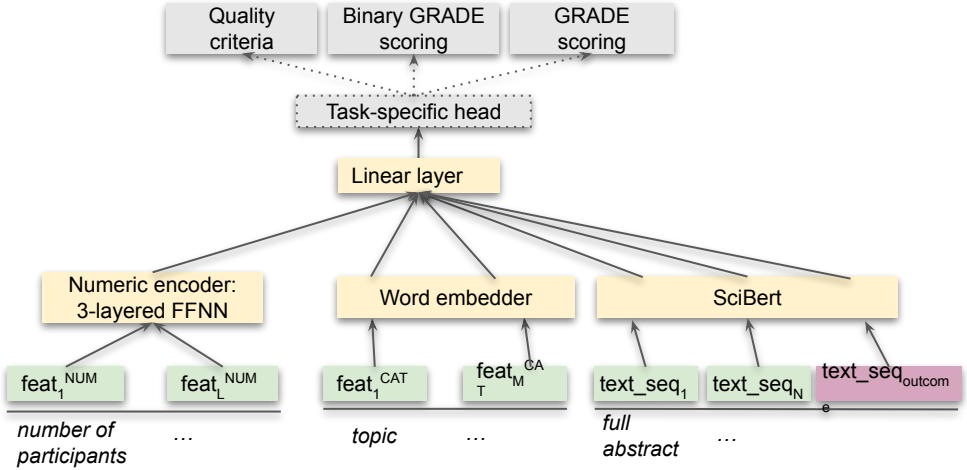
Dataset

~7,000 systematic reviews (majority from 2010-)



Extract data related to quality appraisal from summaries of findings and textual summaries

13,500 outcomes rated for quality using GRADE (with justifications)

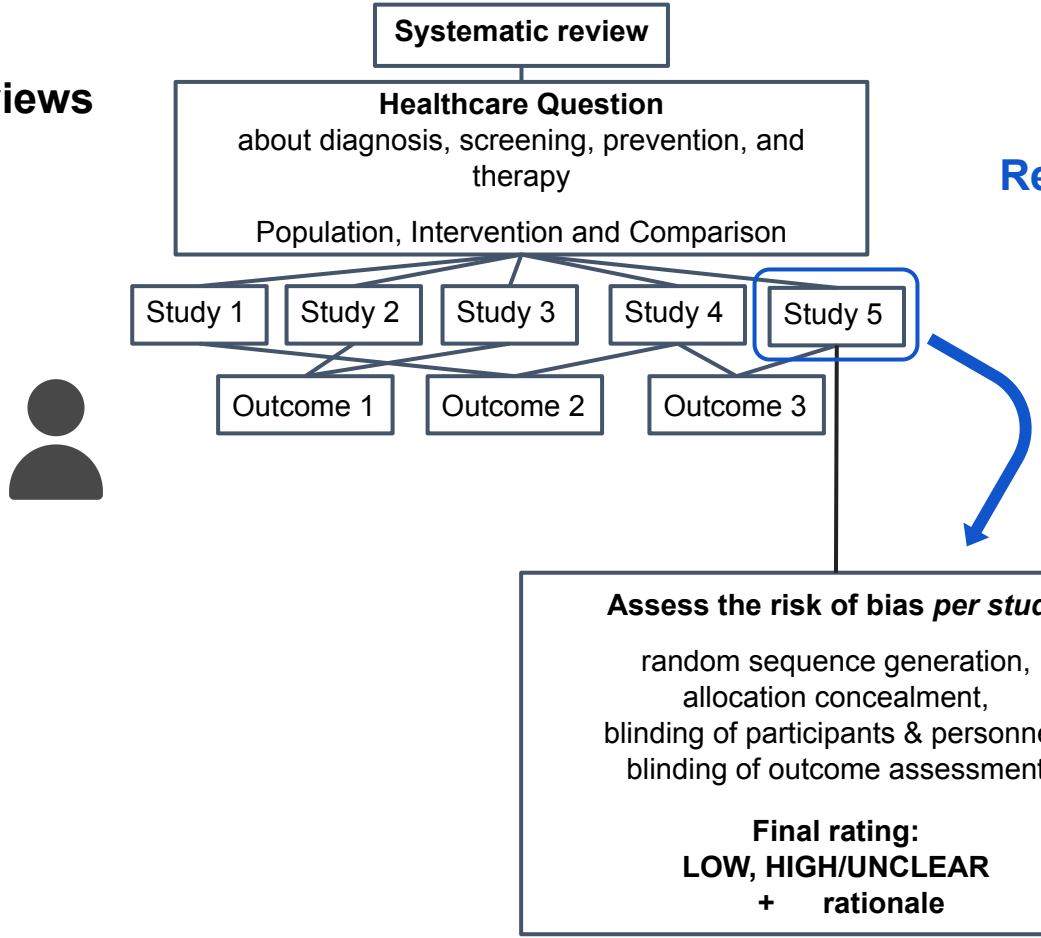


Model

Suster S, Baldwin T, Lau JH, Jimeno Yepes A, Martinez Iraola D, Otmakhova Y, Verspoor K: Automating Quality Assessment of Medical Evidence in Systematic Reviews

JMIR Preprints. 12/12/2021:35568

Constructing systematic reviews and quality assessment

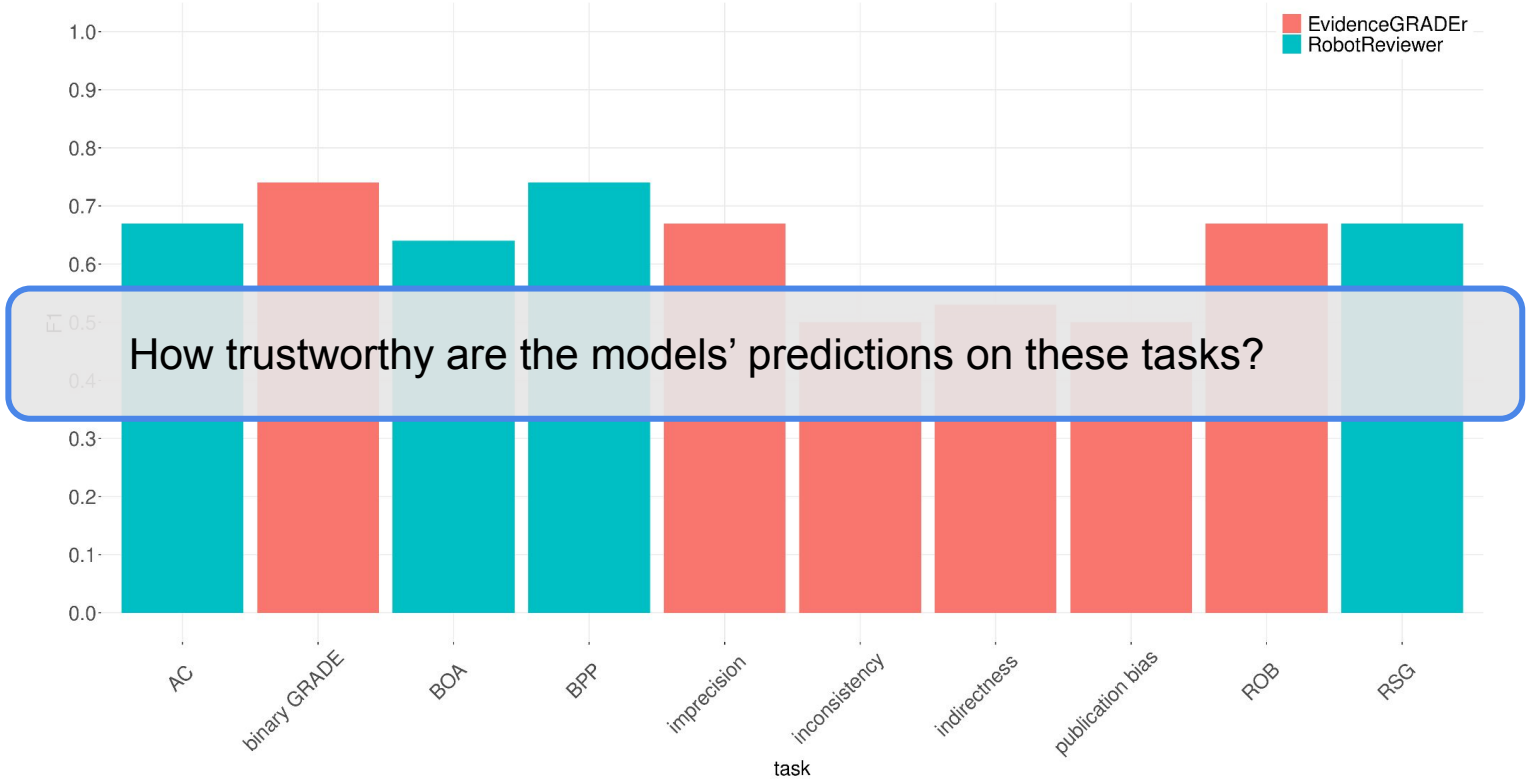


Related & existing work:

RobotReviewer

Given an article abstract/txt, assess its risk of bias

Predictive performance



How trustworthy are the models' predictions on these tasks?

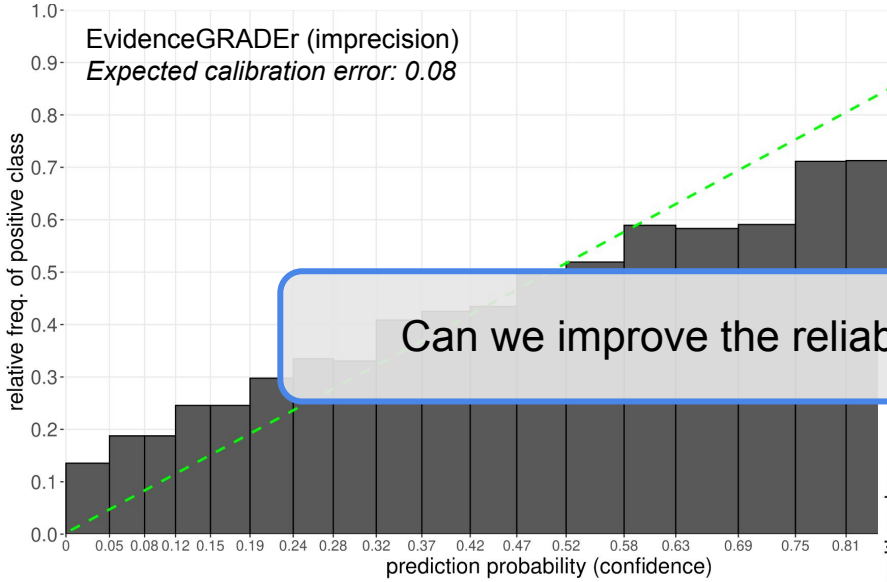
Uncertainty-calibrated classifiers

… are reliable because they know what they don't know

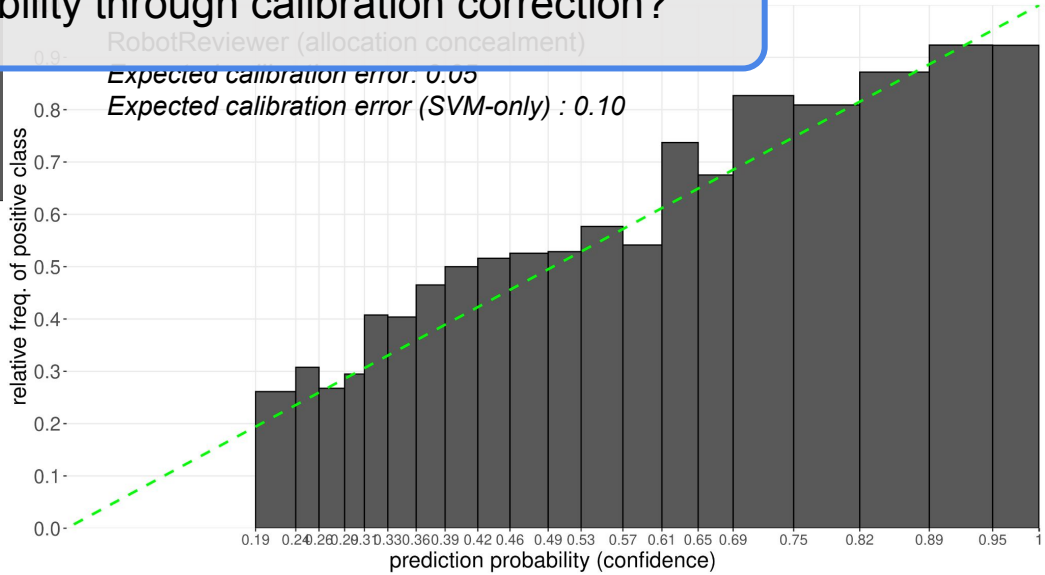
> if a system classifies 100 instances as *y* with probability 0.7, approximately 70 of them should indeed be *y*

But modern neural networks are notorious for over-confidence

Reliability analysis of quality assessment models



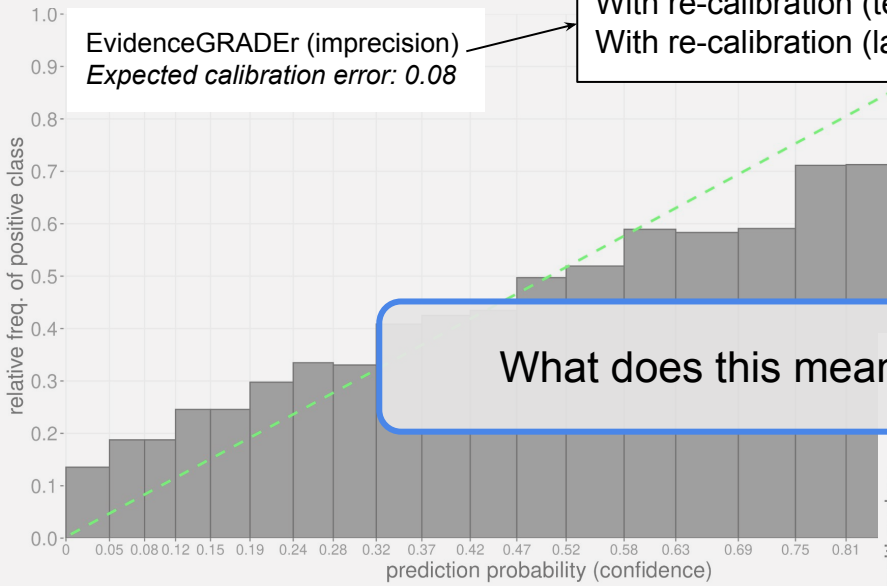
Can we improve the reliability through calibration correction?



Reliability analysis of quality assessment models with calibration correction

With re-calibration (temperature scaling): 0.03
With re-calibration (label smoothing) : 0.03

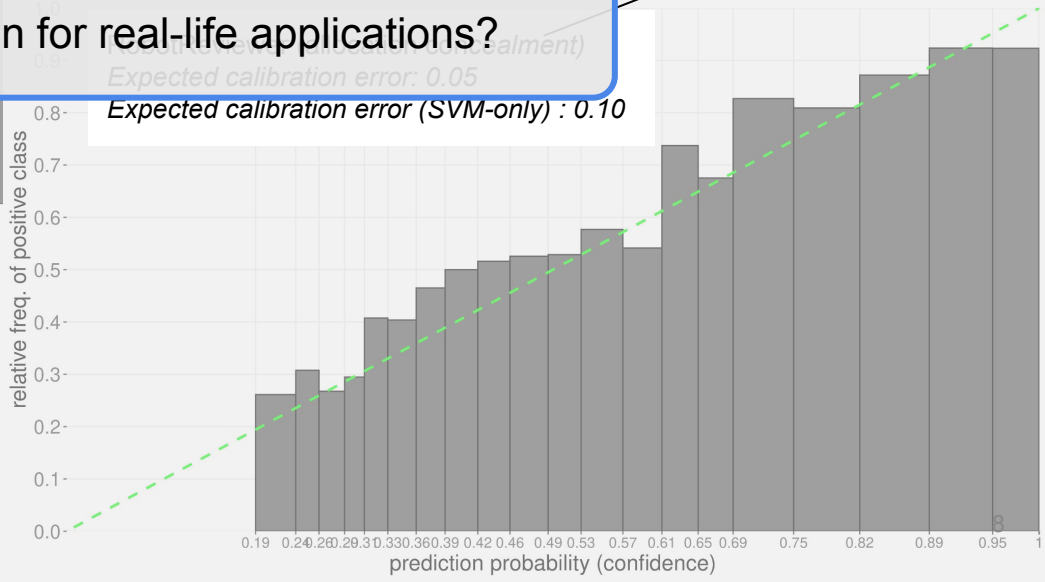
EvidenceGRADEr (imprecision)
Expected calibration error: 0.08



What does this mean for real-life applications?

With re-calibration (temperature scaling): 0.05, 0.07 (SVM-only)

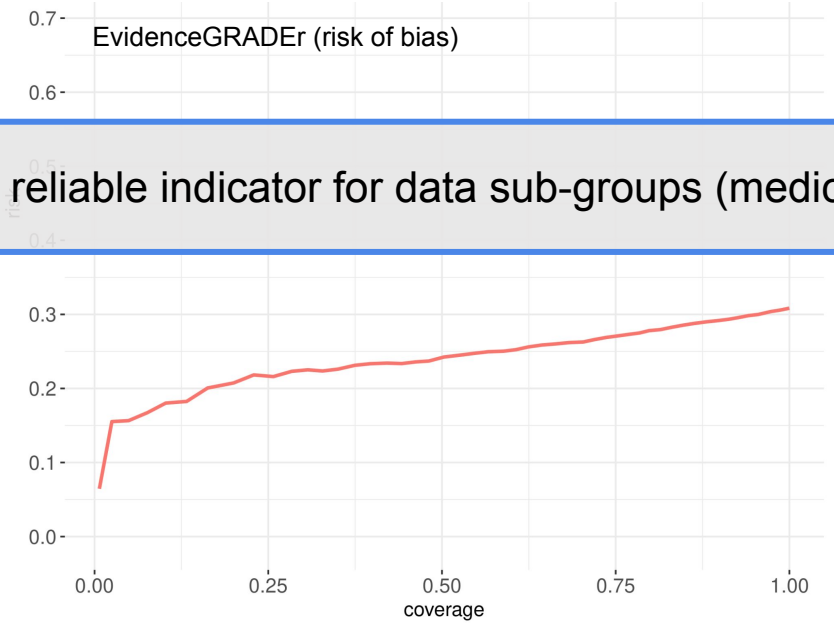
Expected calibration error: 0.05
Expected calibration error (SVM-only) : 0.10



Selective classification

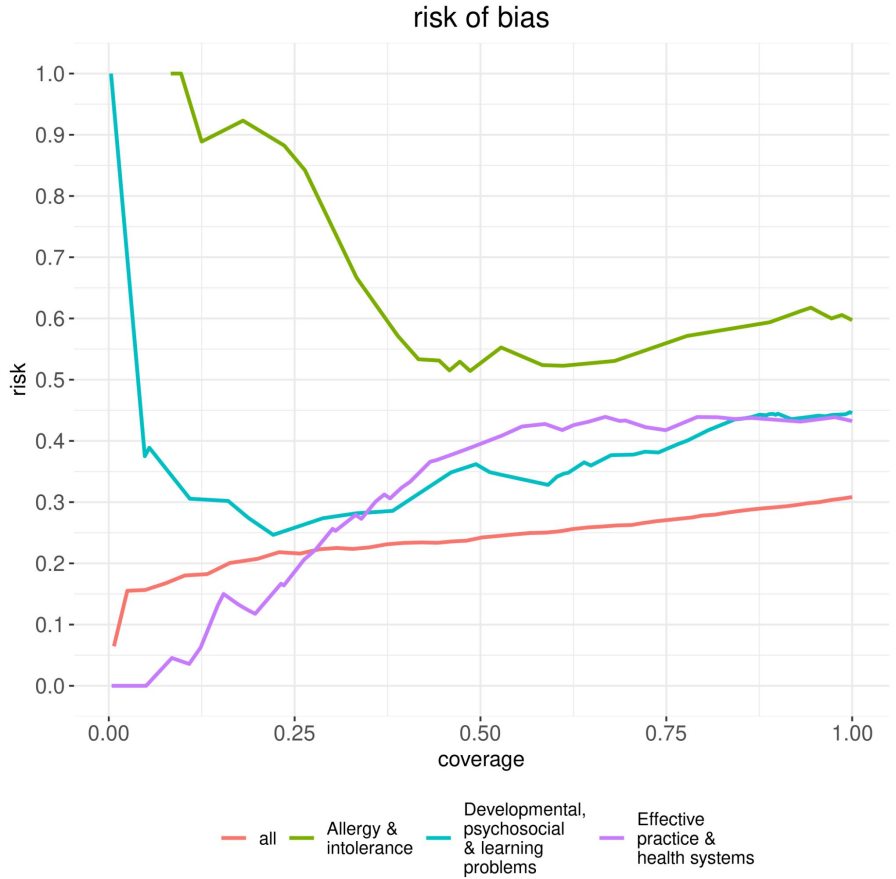
Assume the ability to decide which predictions should be trusted (kept) and which not

reduce the coverage to reduce the risk of error

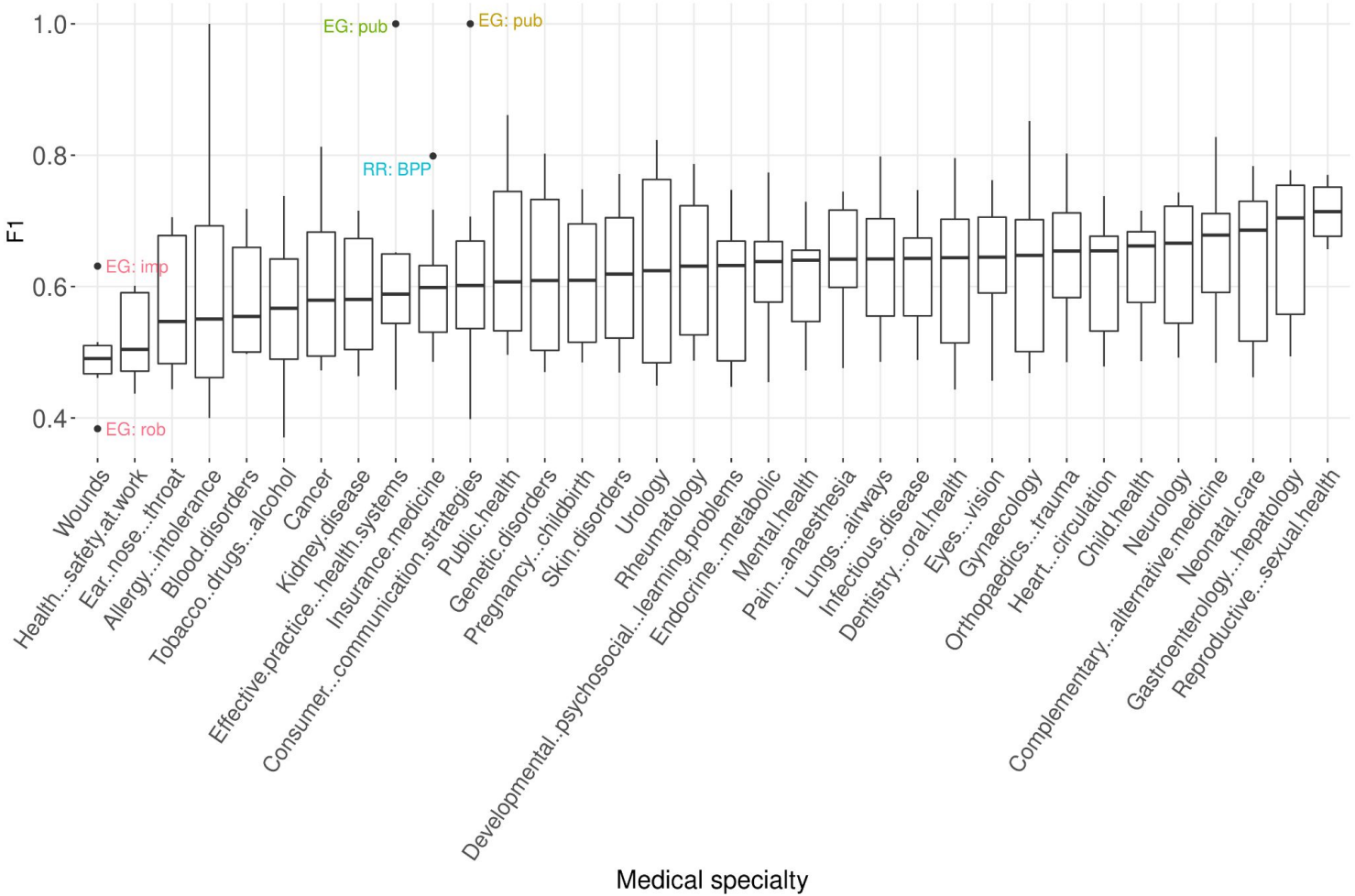


Is this a reliable indicator for data sub-groups (medical specialties)?

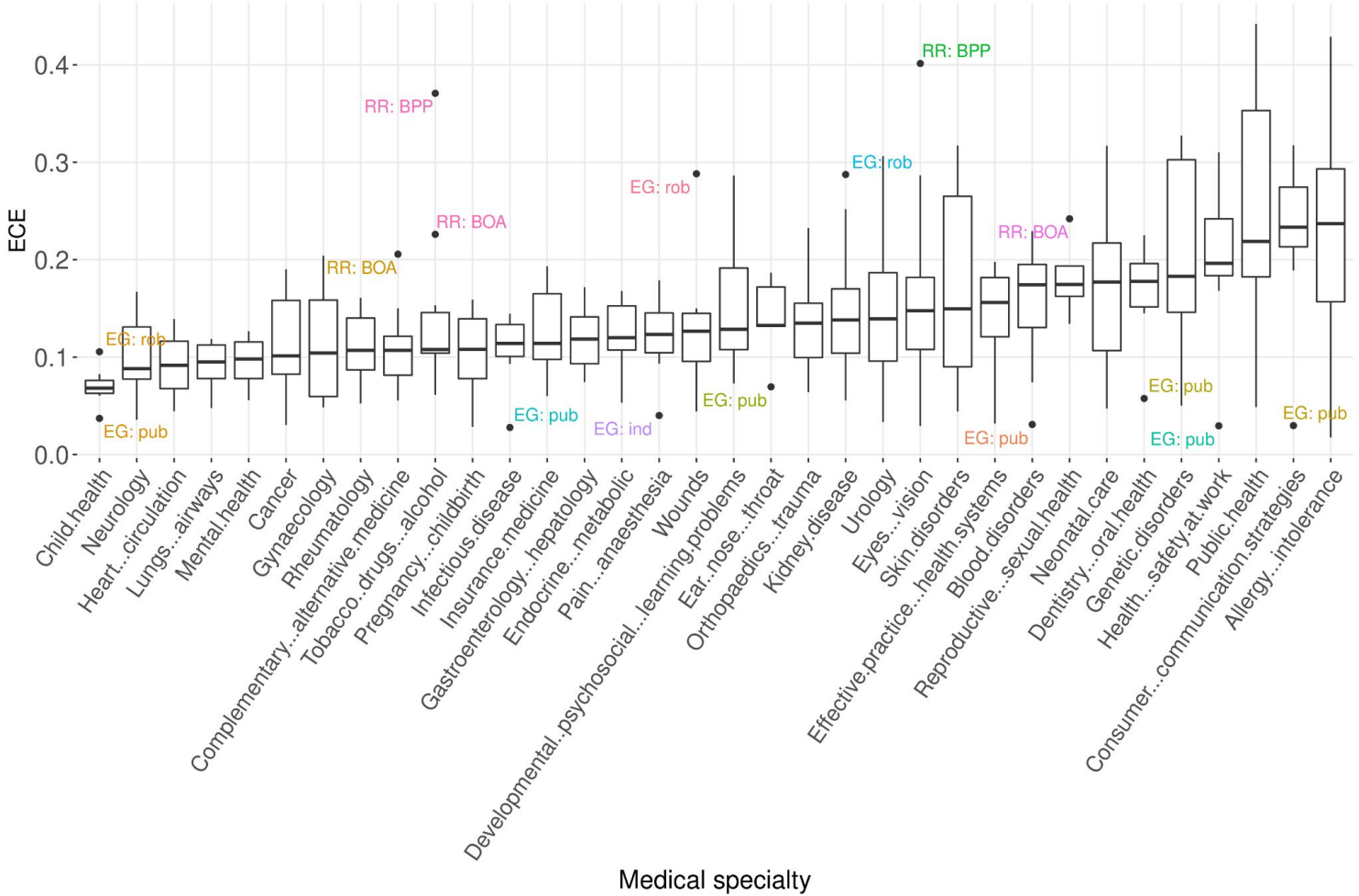
Selective classification
on average vs. three medical specialties with worst performance



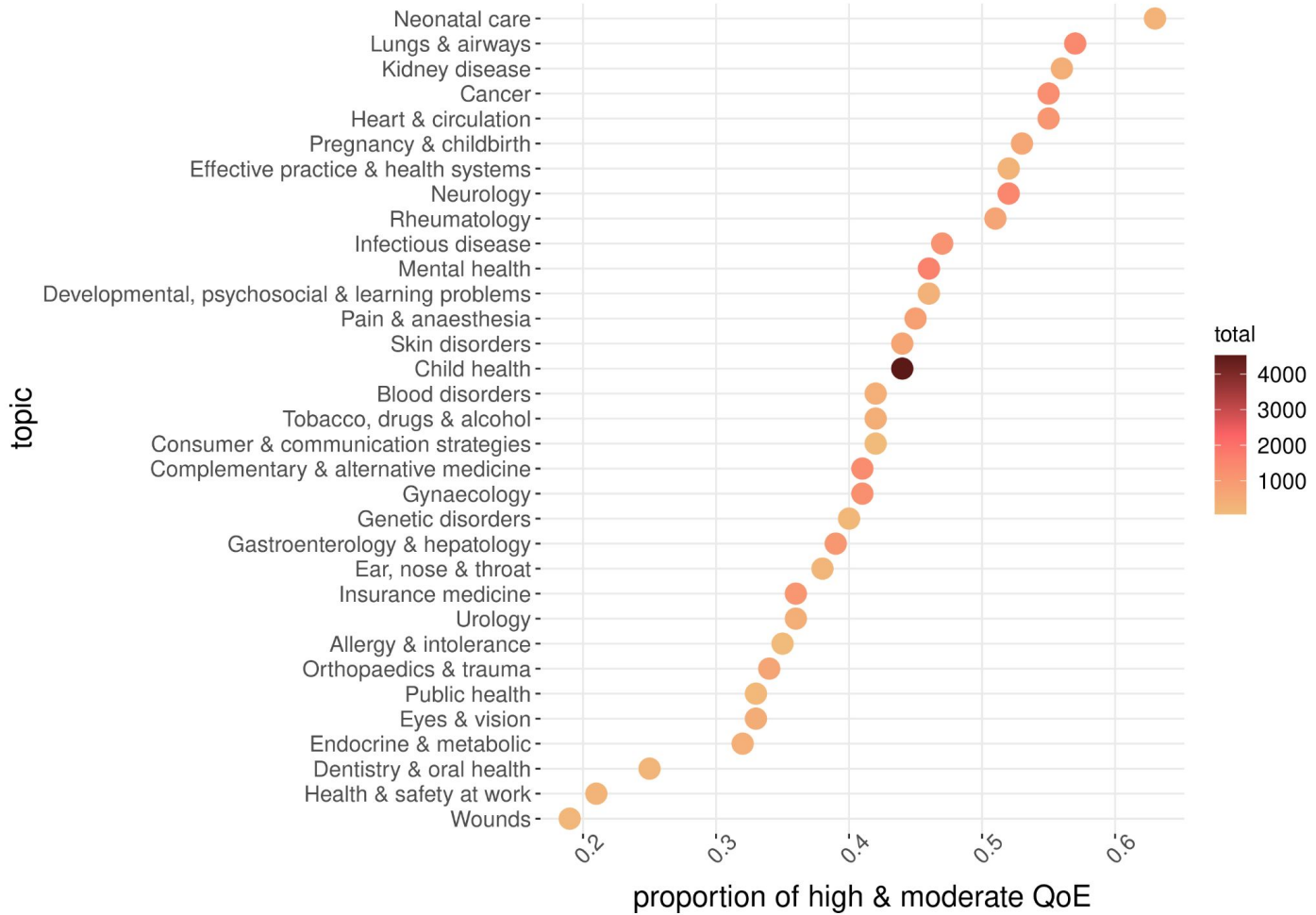
Performance disparity across medical specialties



Disparity in reliability across medical specialties



Disparity in availability of high/moderate-quality evidence



Conclusion

- Reliability of quality assessment models
- Re-calibration
- Selective classification for practical use
- Disparity across medical specialties