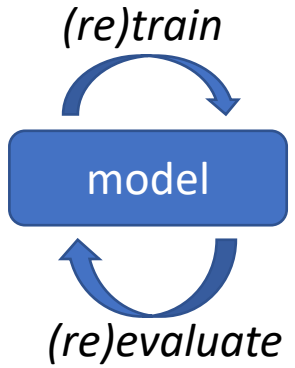
Preferred Quality Metrics for Clinical Prediction Models

What should stakeholders look for to "approve" an algorithm for deployment?



Mike Hughes

Assistant Professor of Computer Science, Tufts University

Inspiration: Model Report Card

M. Mitchell et al. (FAT 2019)*

Model Card - Smiling Detection in Images

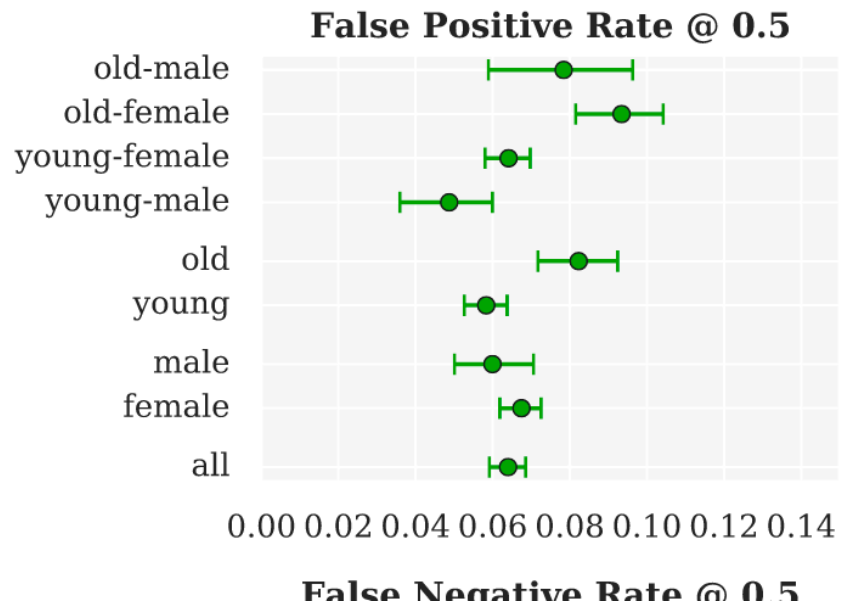
Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

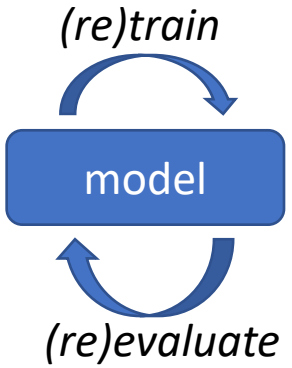
Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Quantitative Analyses



Goal: Model Report Card for Clinical Deployment



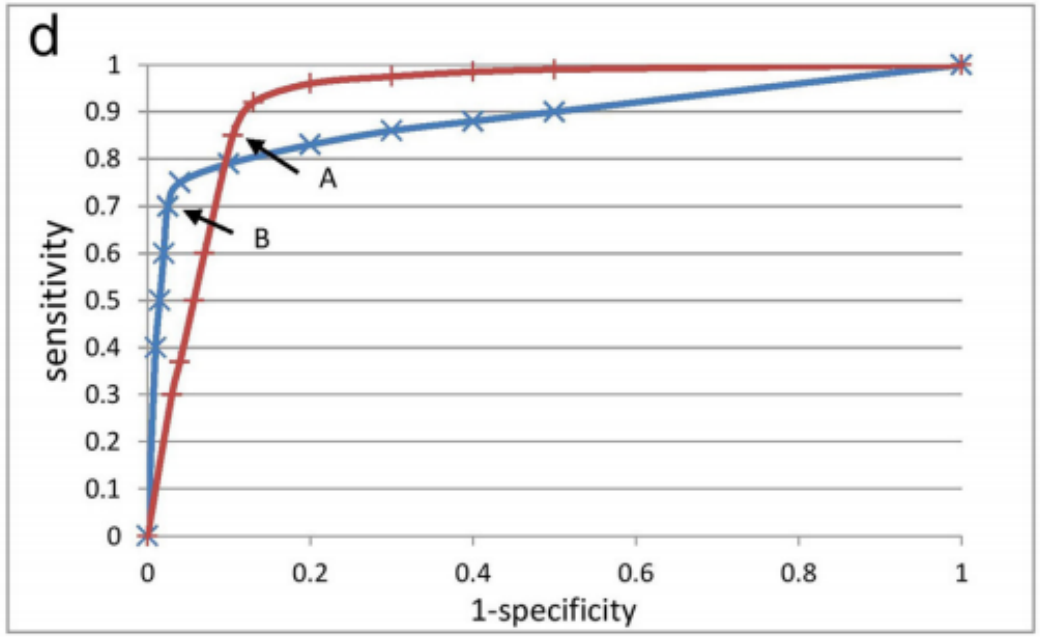
- Choose **task-relevant evaluation**
  - Think about operating conditions and costs of different mistakes
  - Compare to baselines (treat all, treat none) and internal variations
- Show **uncertainty** in all estimates
- Show **external evaluation** (new site? new time window?)
- Study **fairness** via **subgroup analysis** (and intersections of subgroups)

- Recommended Metrics:
  - **Precision-recall curves** plus ROC curves, not just AUROC aka C-statistic
  - **Calibration curves**
  - **Net benefit** (inspired by decision curves)
  - **C-for-benefit statistic** for clinical trials (AUROC when can't know counterfactual)

Idea: **Precision-Recall curve** (not just ROC)

Use Precision-Recall when:

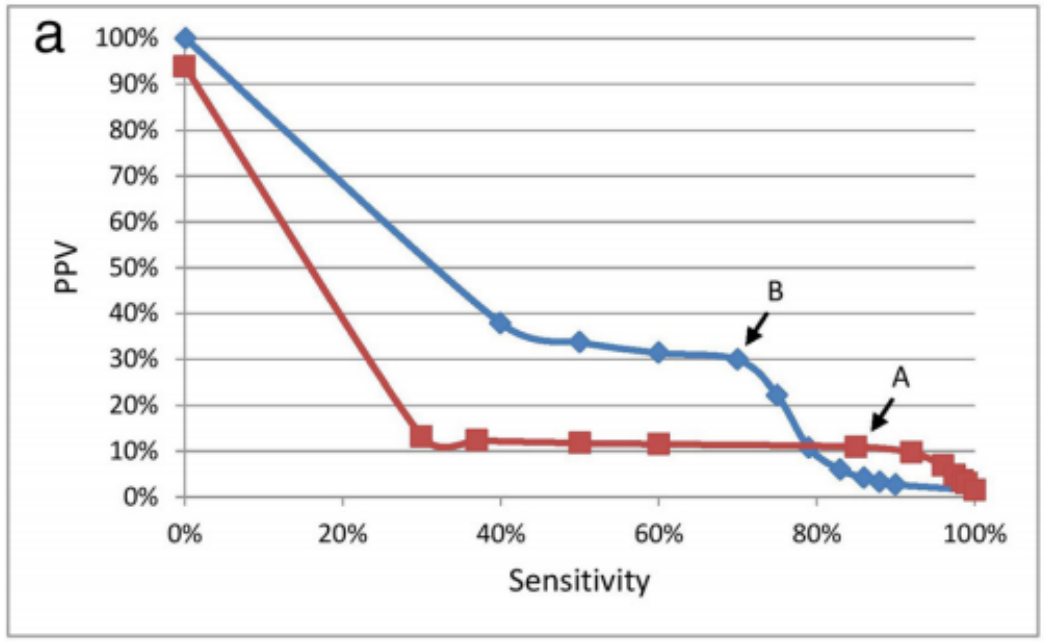
- Data has **significant class imbalance**
- False alarm rates matter



ROC: red is better (area)

Why the *C*-statistic is not informative to evaluate early warning scores and what metrics to use

Santiago Romero-Brufau[1,2*], Jeanne M. Huddleston[1,2,3], Gabriel J. Escobar[4] and Mark Liebow[5]



But, blue is better for alarm fatigue

Idea: Assess **calibration** (not just discrimination)

Does a prediction of 10% chance mortality mean 10% of those subjects will die?

Models with high AUC and high accuracy can have terrible calibration (cause harm if mis-used)

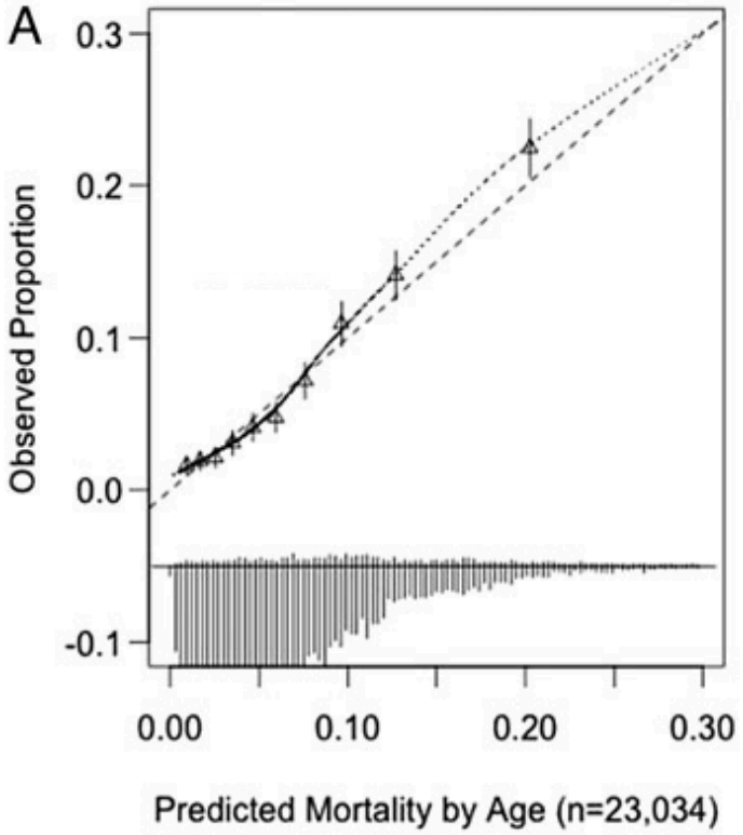
Use when:

- Probabilities produced by model will be used in decision making

Caveats:

- Non-linear/deep models are often criticized as poorly calibrated, but…
- Models can (and should) be post-hoc calibrated
  - e.g. isotonic regression

E. Steyerberg and Y. Vergouwe (Euro. Heart J. 2014)



Idea: Assess **net benefit** of binary predictions

Requires selecting specific operating point (threshold).

Critical to assess in terms of real costs (e.g. hours of human lifetime) of each possible mistake (false positive / false negative)

If hard to select costs, average over a plausible *distribution* over costs

Compare to simple baselines (treat all, treat none)

"Decision Curve Analysis"
A. Vickers and E. Elkin
(Med. Decision Making 2006)

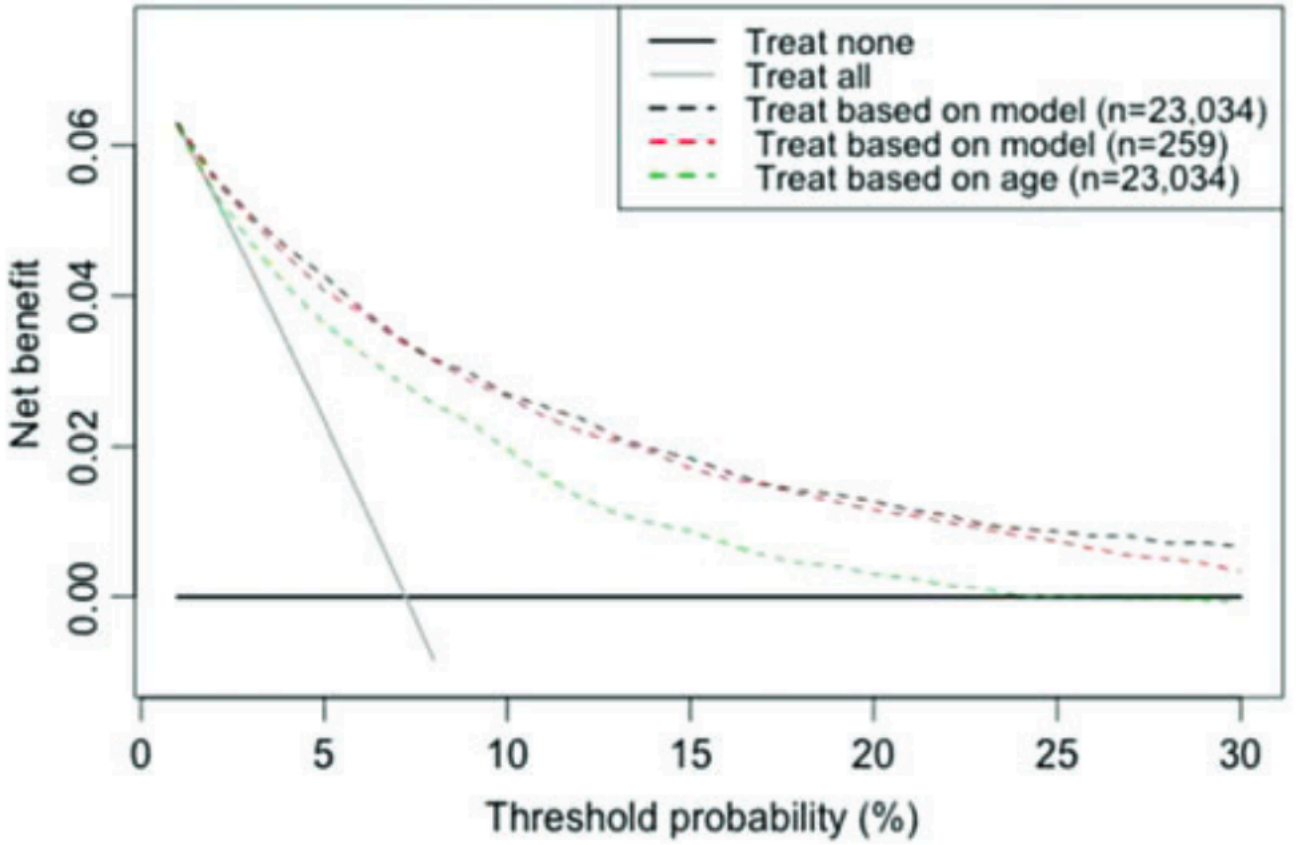


Figure Credit: Steyerberg et al. 2014

Idea: C-for-benefit for 2-arm clinical trials

D. van Klaveren et al.
(J. Clinical Epi. 2018)

Classic C statistic: (aka AUROC)

$$\Pr(\hat{y}(x_i) > \hat{y}(x_j) | y_i = 1, y_j = 0)$$

Given two random examples, one known positive and one negative,
What is probability the model will rank positive one higher?

New C-for-benefit:

$$\Pr\left(\hat{b}(m_i) > \hat{b}(m_j) | b(m_i) > b(m_j)\right)$$

Given two random **matched pairs**, one with known better net benefit,
What is probability the model will rank the better one higher?

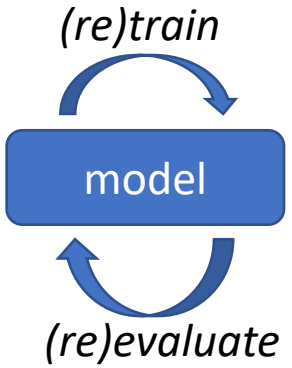
Matched pair: Similar prediction but different treatment arms

Use when: Want to predict benefit, but can only measure each subject under one treatment

Extensions: calibration-for-benefit, ROC-for-benefit, etc.

Caveat: *Active research*

Summary: Model Report Card should….



- Choose **task-relevant evaluation**
  - Think about operating conditions and costs of different mistakes
  - Compare to baselines (treat all, treat none) and internal variations
- Show **uncertainty** in all estimates
- Show **external evaluation** (new site? new time window?)
- Study **fairness** via **subgroup analysis** (and intersections of subgroups)

- Recommended Metrics (use when appropriate):
  - **Precision-recall curves** plus ROC curves, not just AUROC aka C-statistic
  - **Calibration curves**
  - **Net benefit** (inspired by decision curves)
  - **C-for-benefit statistic** for clinical trials (AUROC when can't know counterfactual)