Sequences and Time-Series in Medicine

July 25, 2020

Applied Data Science
MMCi Term 4

Matthew Engelhard

Sequential Models in Practice

THE EASY CASE… WE JUST EVALUATE PERFORMANCE

Deidentification of Patient Notes

Table 5. Examples of correctly detected PHI instances (in bold) by the ANN

PHI category	ANN
AGE	Father had a stroke at **80** and died of?another stroke at age Personal data and overall health: Now **63**, despite his FH: Father: Died @ **52** from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to **15**, has not smoked since 15.
CONTACT	History of Present Illness **86F** reports worsening b/l leg pain. by phone, Dr. Ivan Guy. Call w/ questions **86383**. Keith Gilbert, H/O paroxysmal afib VNA **171-311-7974** ======= Medications
DATE	During his **May** hospitalization he had dysphagia Social history: divorced, quit smoking in **08**, sober x 10 yrs, She is to see him on the **29th** of this month at 1:00 p.m. He did have a renal biopsy in teh late **60s** adn thus will look for results, Results**02/20/2087** NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1 Jose Church, M.D. /ray DD: 01/18/20 DT: **01/19/:0** DV: 01/18/20

De-identification of patient notes with recurrent neural networks
Dernoncourt F, Lee JY, Uzuner O, Szolovits P
JAMIA 24(3), 2017, 596–606

- A bidirectional RNN is used to identify PHI (18 HIPAA fields)
- *i2b2*: 889 discharge summaries, >28k PHI tokens
- *MIMIC*: 1635 discharge summaries, >60k PHI tokens
- State of the art sensitivity and F1 metric on both datasets

Train, Validation, Test

MIMIC:

80% train/validation

20% test

i2b2:

60% train/validation

40% test

"All results were computed using the official evaluation script from the i2b2 2014 de-identification challenge."

Table 3. Overview of the i2b2 and MIMIC datasets

Statistics	i2b2	MIMIC
Vocabulary size	46 803	69 525
Number of notes	1304	1635
Number of tokens	984 723	2 945 228
Number of PHI instances	28 867	60 725
Number of PHI tokens	41 355	78 633

Examples of PHI Identified by the RNN

AGE	Father had a stroke at <u>**80**</u> and died of?another stroke at age Personal data and overall health: Now <u>**63**</u>, despite his FH: Father: Died @ <u>**52**</u> from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to <u>**15**</u>, has not smoked since 15.
CONTACT	History of Present Illness <u>**86F**</u> reports worsening b/l leg pain. by phone, Dr. Ivan Guy. Call w/ questions <u>**86383**</u>. Keith Gilbert, H/O paroxysmal afib VNA <u>**171-311-7974**</u> ======== Medications
DATE	During his <u>**May**</u> hospitalization he had dysphagia Social history: divorced, quit smoking in <u>**08**</u>, sober x 10 yrs, She is to see him on the <u>**29th**</u> of this month at 1:00 p.m. He did have a renal biopsy in teh late <u>**60s**</u> adn thus will look for results, Results<u>**02/20/2087**</u> NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1 Jose Church, M.D. /ray DD: 01/18/20 DT: <u>**01/19/:0**</u> DV: 01/18/20

Evaluation Metrics

Precision, or positive predictive value:

$$\frac{\text{true positives}}{\text{all positive predictions}}$$

Recall, or sensitivity:

$$\frac{\text{true positives}}{\text{all condition positives}}$$

F1-score:

$$\frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

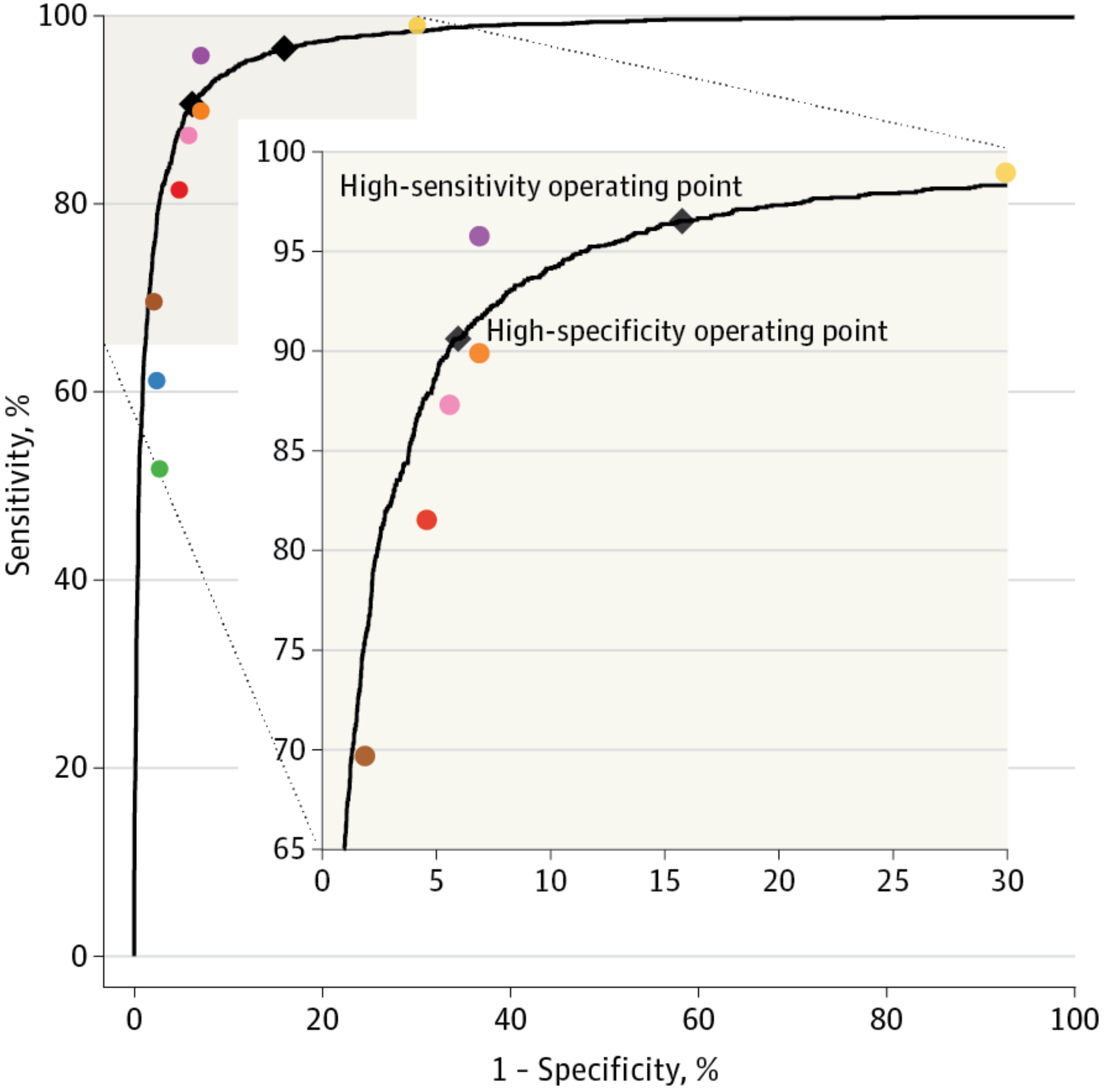
	Condition Positive	**Condition Negative**
Prediction Positive	True Positive	False Positive
Prediction Negative	False Negative	True Negative

RNN Model Outperforms Previous Benchmarks

Table 4. Performance (%) on the PHI as defined in HIPAA

	i2b2			MIMIC		
Model	Precision	Recall	F1	Precision	Recall	F1
Nottingham	**99.000**	96.400	97.680	–	–	–
MIST	91.445	92.745	92.090	95.867	98.346	97.091
CRF	98.560	96.528	97.533	99.060	98.987	99.023
ANN	98.320	97.380	97.848	**99.208**	99.251	**99.229**
CRF + ANN	97.920	**97.835**	**97.877**	98.820	**99.398**	99.108

Deep Learning for Diabetic Retinopathy Classification



$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{total number of positives in the dataset}}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{total number of negatives in the dataset}}$$

<u>Choose an operating point:</u>

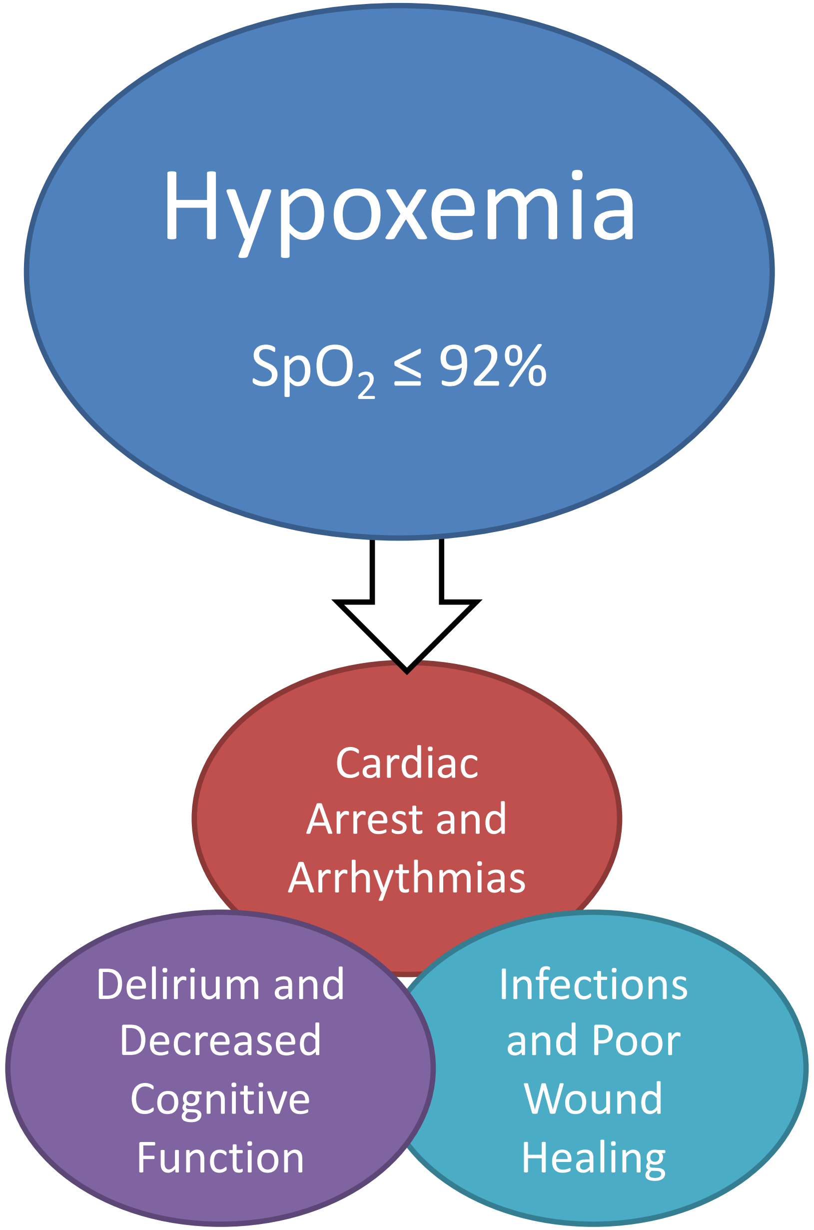
Are we more concerned about false positives, or false negatives?

Gulshan et al. *JAMA* (2016)

Sequential Models in Practice

"GROUND TRUTH" IN MODEL EVALUATION

Predict Hypoxemia during Surgery



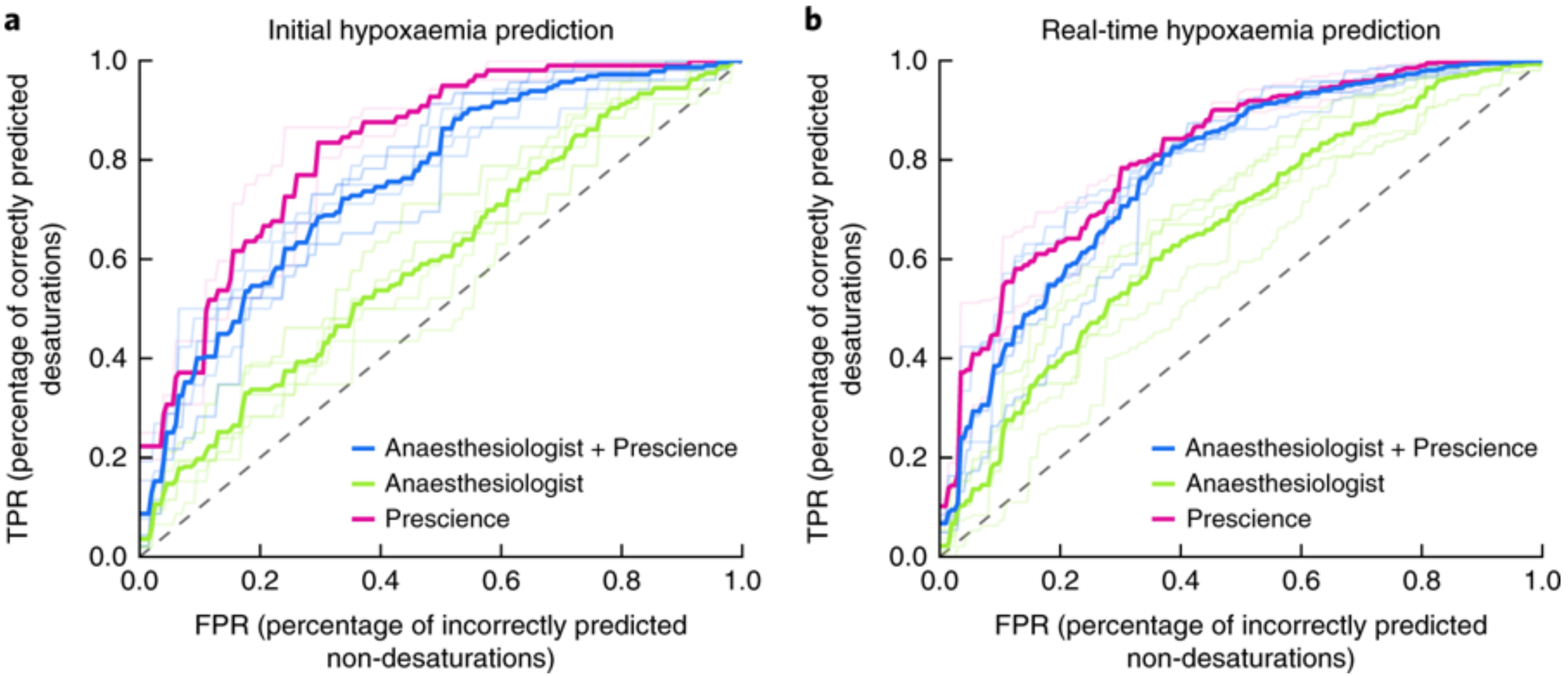
nature
biomedical engineering

Article | Published: 10 October 2018

Explainable machine-learning predictions for the prevention of hypoxaemia during surgery

Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim & Su-In Lee

Nature Biomedical Engineering **2**, 749–760 (2018) | Download Citation



Comparison to Experts

For initial risk prediction:

- Anaesthesiologists performed significantly better with Prescience (AUC = 0.76 versus 0.60; *P* < 0.0001)
- Prescience performed better in a direct comparison with anaesthesiologists (AUC = 0.83; *P* < 0.0001)

For intraoperative real-time (next 5 min) risk prediction:

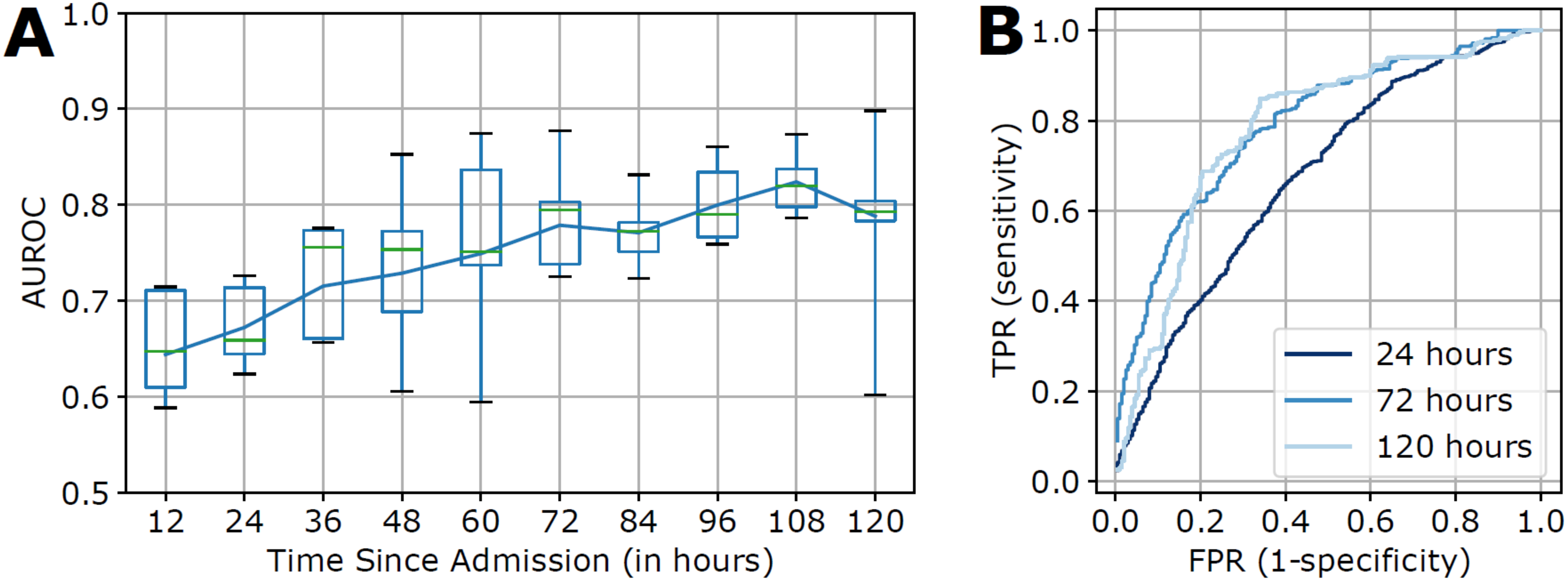
- Anaesthesiologists (AUC = 0.66) again performed better with Prescience (AUC = 0.78; *P* < 0.0001)
- Prescience alone outperformed anaesthesiologists predictions (AUC = 0.81; *P* < 0.0001)

Is this a fair comparison?

- Training examples are episodes of hypoxemia that were not prevented during surgery

- Expert comparison:
  – the expert is predicting likelihood of hypoxemic episodes, some of which were prevented
  – the model has learned to predict hypoxemic episodes that couldn't be avoided

Predict Risk of Requiring Surgery



Turpin et al., *Machine Learning Prediction of Surgical Intervention for Small Bowel Obstruction* (forthcoming)

Was the Decision We're Learning from Correct?

- We learn to replicate surgeons' decisions
- Their risk tolerance depends on many factors, notably the patient's age
- We can't see *counterfactual* outcomes (what would have happened?)
- The same is true in many other problems, *e.g.*, treatment of retinopathy of prematurity

Perinatal and Neonatal Factors (# studies)	Results Across Studies	Summary Effect Estimate (95% CI)
Presentation		
Abnormal presentation (15)	10–, 5↑	1.44 (1.07–1.94)
Breech (4)		1.81 (1.21–2.71)
Other perinatal factors		
Cord complications (14)	13–, 1↑	1.50 (1.00–2.24)
Fetal distress (4)	3–, 1↑	1.52 (1.09–2.12)
Birth injury or trauma (6)	6–	4.90 (1.41–16.94)
Twins or multiple birth (10)	7–, 3↑	1.77 (1.23–2.55)
Maternal hemorrhage (4)	3–, 1↑	2.39 (1.35–4.21)
Birth weight and size		
Total birth weight (decreased) (15)	12–, 2↑, 1↓	
Low birth weight (<2500 g) (15)	8–, 7↑	1.63 (1.19–2.33)
Small for gestational age (10)	7–, 3↑	1.35 (1.14–1.61)
Clinical impression		
Congenital malformation (11)	4–, 7↑	1.80 (1.42–2.82)
Apgar score		
Low 5-minute Apgar score (8)	6–, 2↑	1.67 (1.24–2.26)
Neonatal Status		

Early Autism Risk Prediction

- We'd like to identify at-risk children more promptly
- But, we can only learn from what actually happened…
- See who ends up being diagnosed and try to identify them earlier
- What about children never identified or lost to follow-up?

PEDIATRICS
OFFICIAL JOURNAL OF THE AMERICAN ACADEMY OF PEDIATRICS

Review Article

Perinatal and Neonatal Risk Factors for Autism: A Comprehensive Meta-analysis

Hannah Gardener, Donna Spiegelman and Stephen L. Buka

Pediatrics August 2011, 128 (2) 344-355; DOI: https://doi.org/10.1542/peds.2010-1036

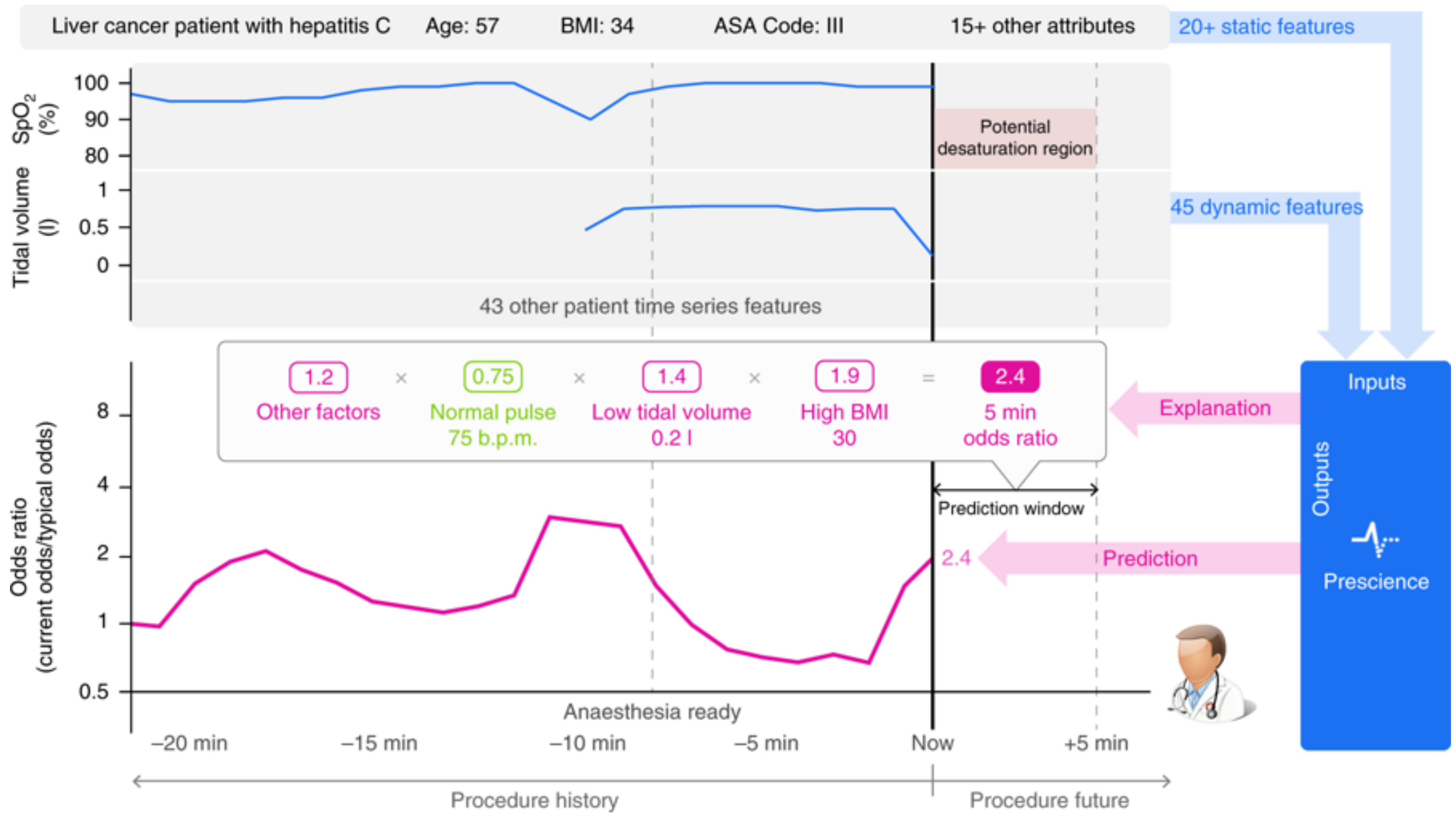
Sequential Models in Practice

WE'RE MAKING GOOD PREDICTIONS… NOW WHAT DO WE DO?

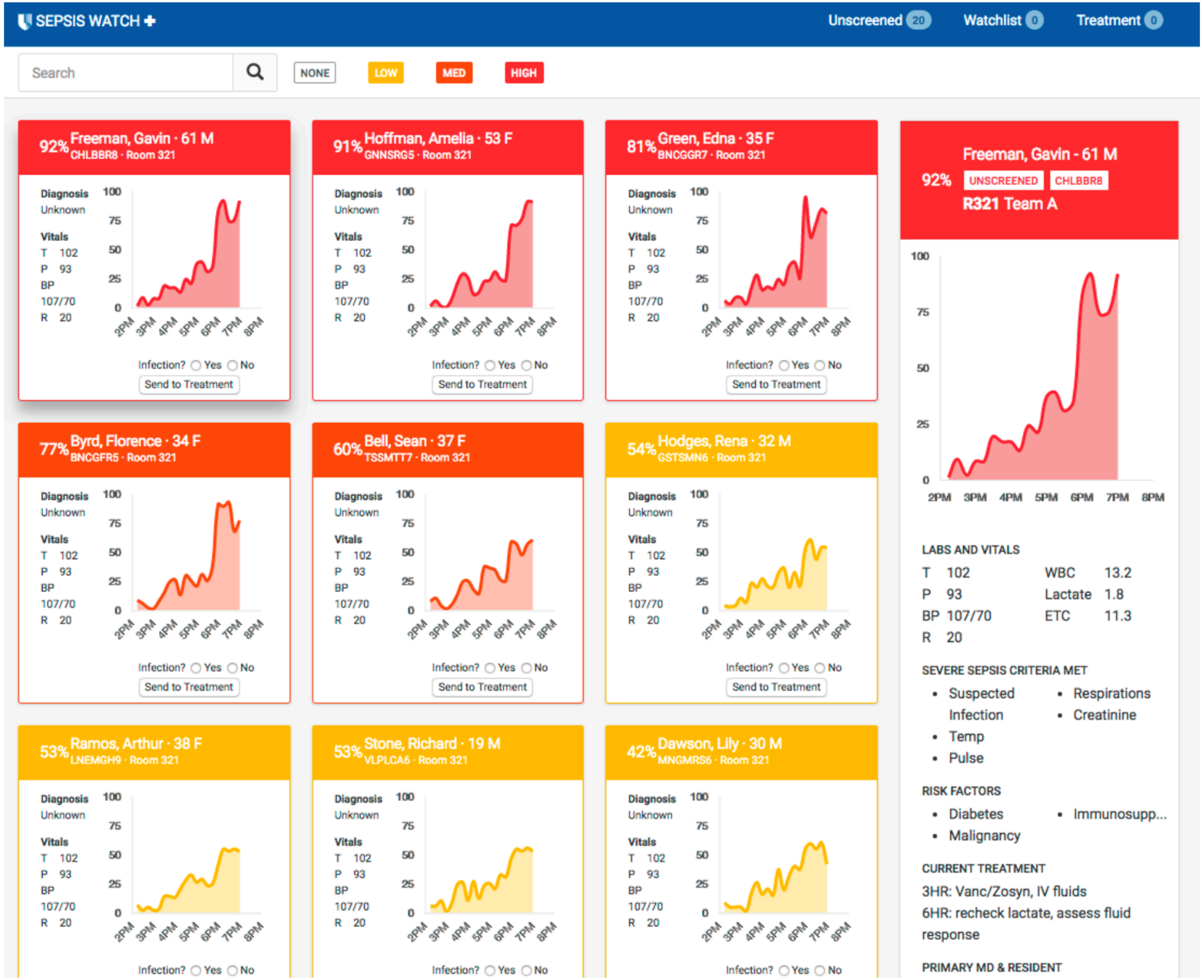
Silent Deployment

- Implement the model
  - Real-time data acquisition and processing (e.g. work from DIHI pipeline rather than curated dataset)
  - Determine criteria (available e.g. at admission) to trigger its application
  - Additional data used to refine the model
- Secondary / ongoing prospective evaluation
  - Ensure model generalizes beyond initial data collection
  - Collection of secondary outcomes / metrics

Passive Intervention (Dashboard)



Sepsis Watch Dashboard



Active Intervention

- Enroll patients at routine eye exam
- Upload fundoscopic images, receive system recommendation (refer or not refer)
- Weekly review by ophthalmologist to ensure no referrals missed
    - Refer if needed and missed by the system
    - Otherwise, take no action

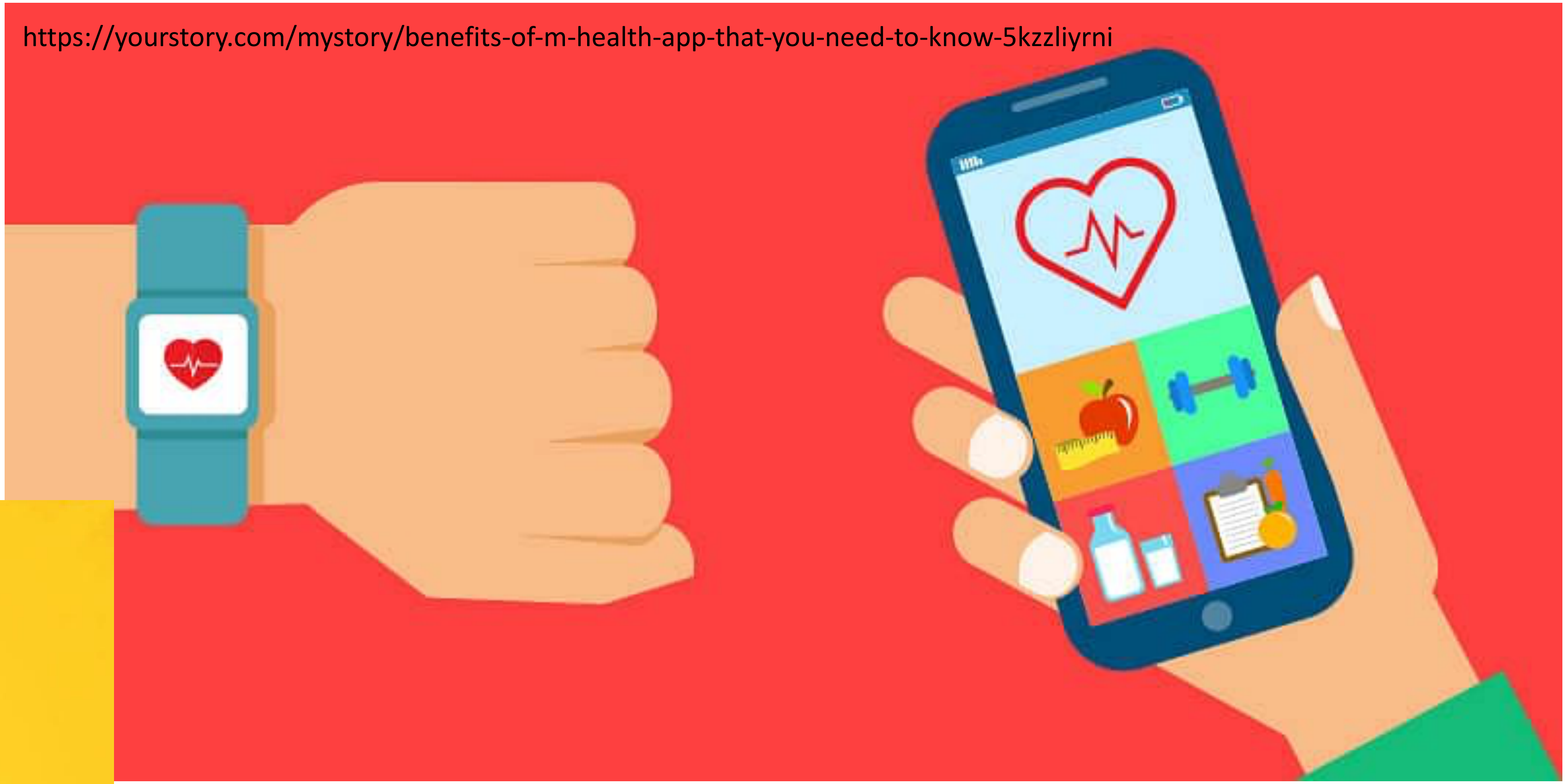
Active Intervention

- In this case, the intervention is implemented at the health system level (routine procedures were changed)
- Alternatively, could consider a provider-centered intervention
- Or a patient-centered intervention

Let the market decide… ??

In short, FDA won't regulate if it:

- Doesn't provide *specific* treatment recommendations
- Automates *routine* provider tasks



https://www.fda.gov/medical-devices/device-software-functions-including-mobile-medical-applications/examples-software-functions-which-fda-will-exercise-enforcement-discretion

Sequential Models in Practice

WE KNOW WHAT TO DO, BUT WHEN DO WE DO IT?

Clinically Meaningful Performance Measures

	Model-based referral	No model-based referral
Child has autism	Earlier Diagnosis (true positive)	Referral via Current Mechanisms (false negative)
Child does not have autism	Unnecessary Specialist Visit (false positive)	No Action Taken or Needed (true negative)

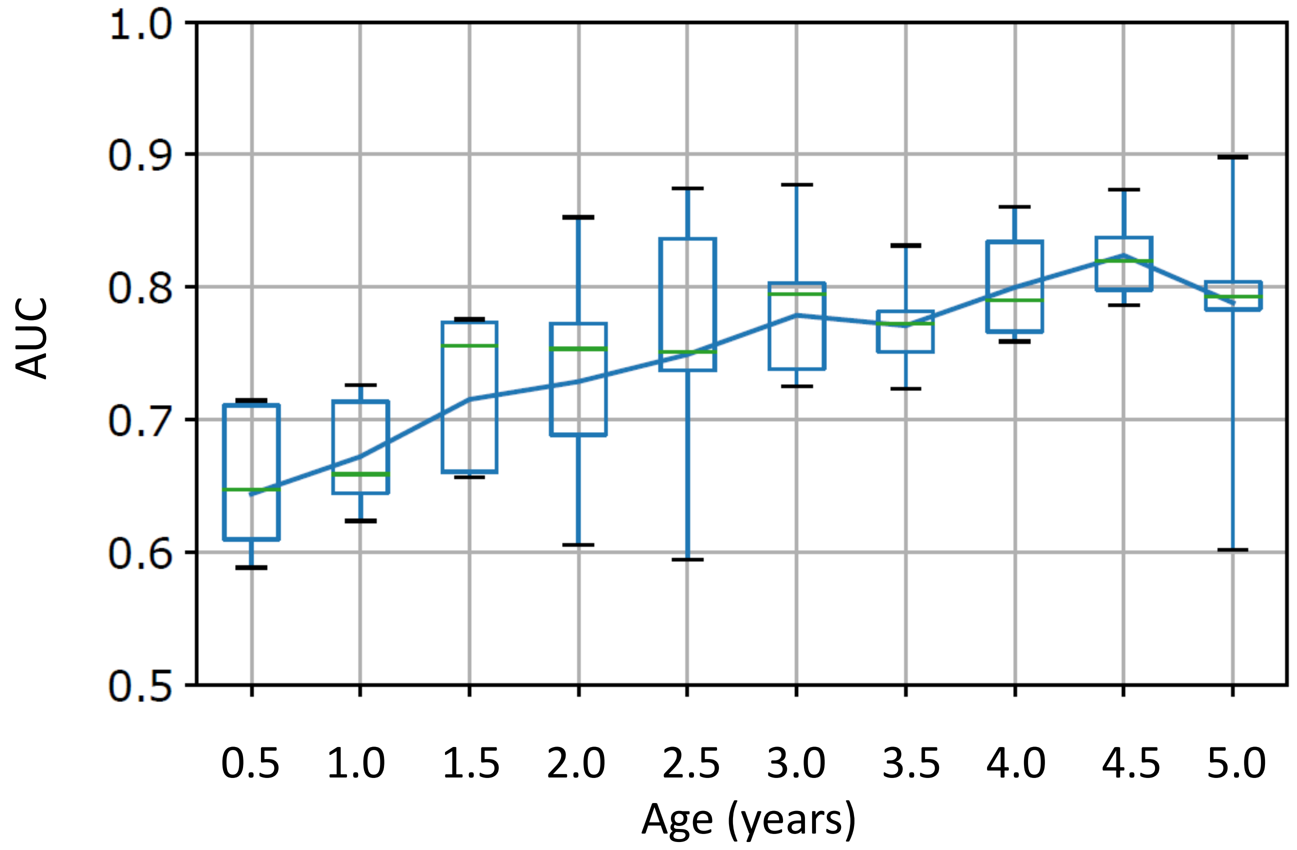
Decision: initiate referral

model is NOT used to rule out diagnoses

Measures of Success:

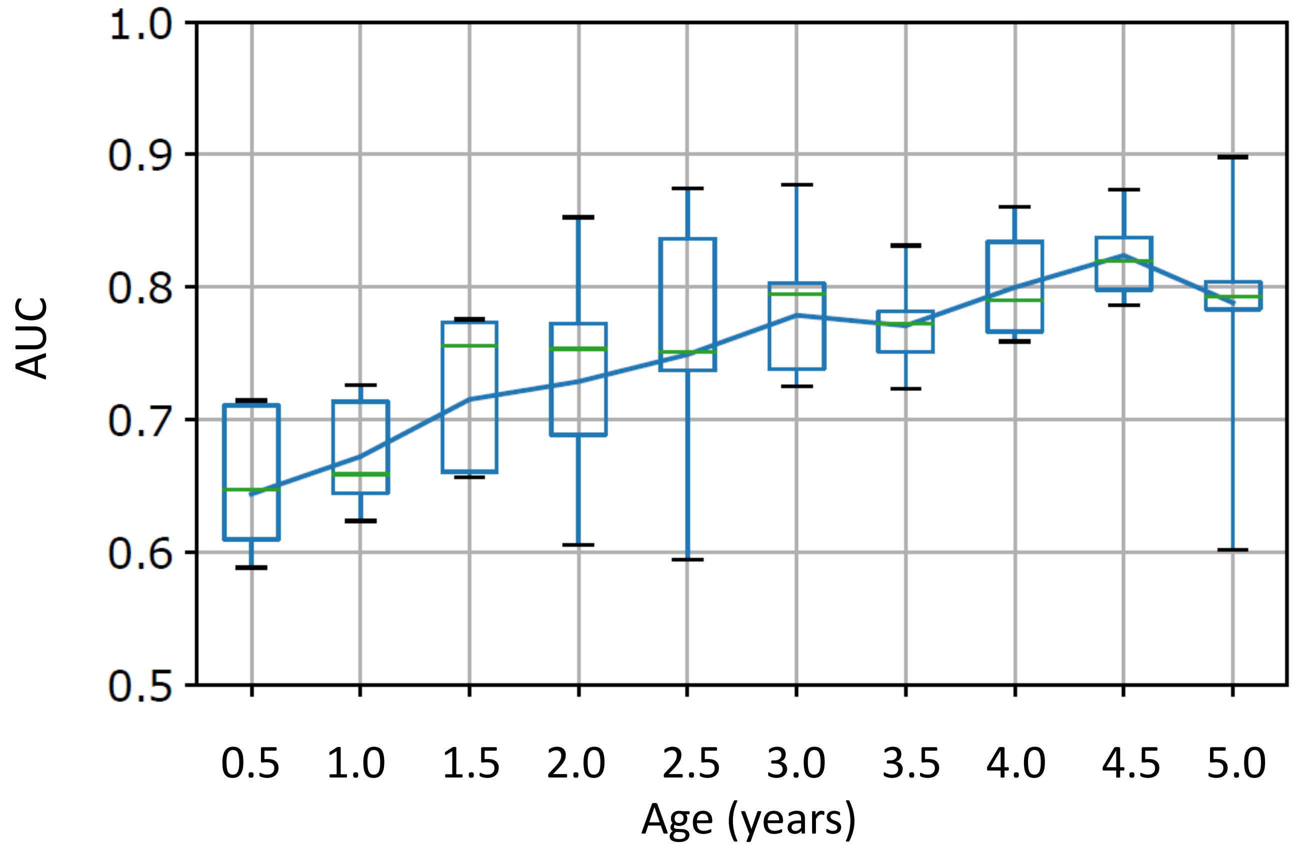
- How much earlier, on average, can we diagnose and intervene?
- How many children are unnecessarily referred?

When do we intervene?



- How good is model performance at time *t*?
- How concerning are false positives and/or false negatives at time *t*?
- How urgent is intervention at time *t*?
- Considering all these factors, what is our threshold for initiating referral, and does it change over time?

When do we intervene?



Some work in this area…
But not enough.

- Train model with a loss that is weighted based on time-varying importance of good predictions
- Many predictions over time, but only *act* once
- Connection to Secretary problem

Discussion Topics

- Implementation in the health system
- Structure of intervention
- Timing of intervention