

Clinical Deployment of ML

Sujay Nagaraj



“I think if you work as a radiologist, you are like the coyote that's already over the edge of the cliff but hasn't yet looked down, **people should stop training radiologists now.**

“I think if you work as a radiologist, you are like the coyote that's already over the edge of the cliff but hasn't yet looked down, **people should stop training radiologists now.**

Geoffrey Hinton, 2016



Implementing AI in healthcare

Vector-SickKids Health AI Deployment Symposium, Toronto, Ontario, Canada

Event Date: October 30, 2019

Published Date: March 24, 2020

Erik Drysdale^{1*}, Elham Dolatabadi^{2,3}, Corey Chivers⁴, Vincent Liu⁵, Such Saria⁶, Mark Sendak⁷, Jenna Wiens⁸, Michael Brudno^{1,2,3}, Amelia Hoyt¹, Mjaye Mazwi¹, Muhammad Mamdani^{2,3,9}, Devin Singh¹, Vanessa Allen¹⁰, Carolyn McGregor¹¹, Heather Ross¹², Antonio Szeto¹³, Amol Anand Verma^{2,8}, Bo Wang^{2,11,13}, P. Alison Paprica^{2,3}, Anna Goldenberg^{1,2,3}

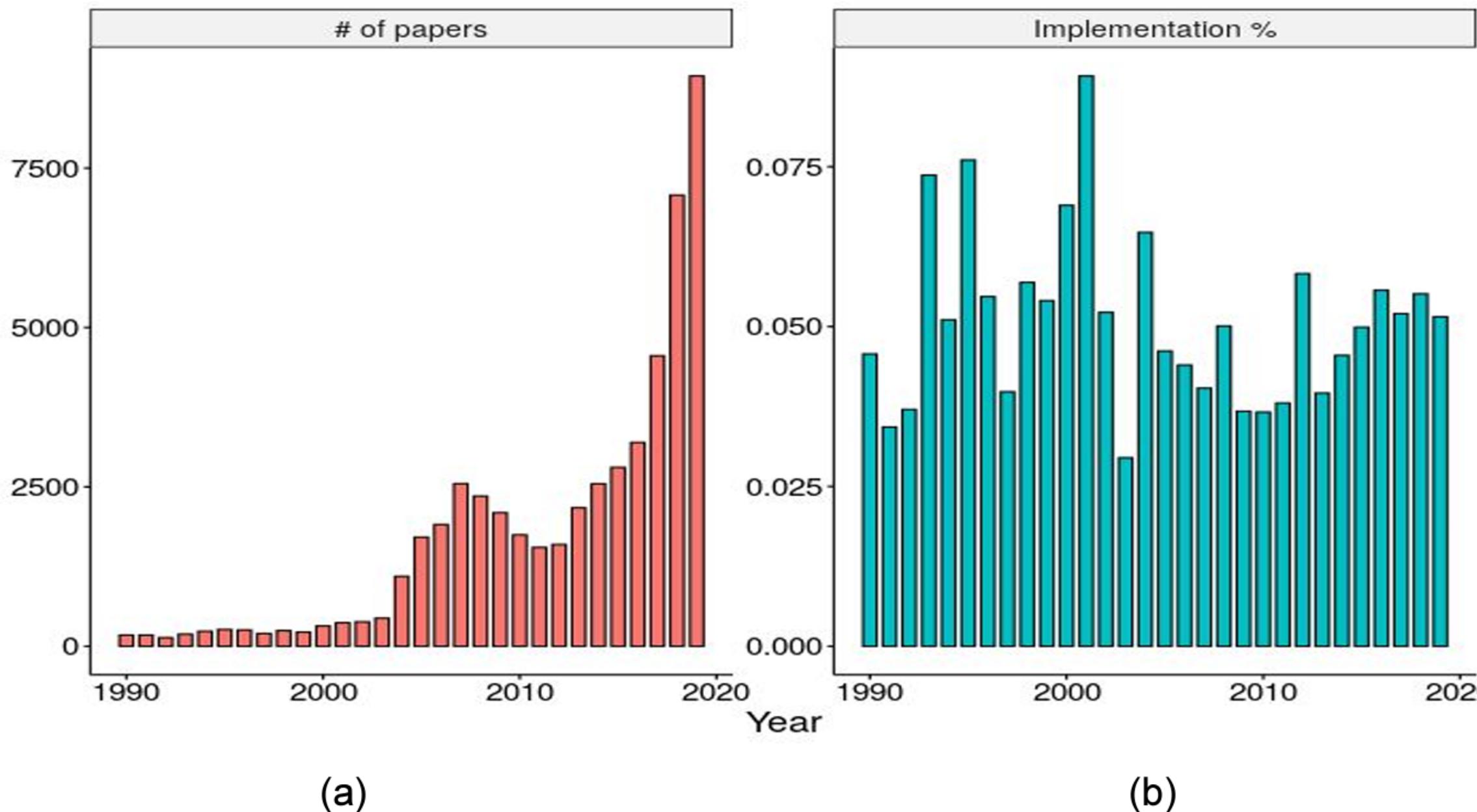


Figure 1: Number of the papers published in PubMed with a reference to “machine learning” or “artificial intelligence” anywhere in the article (a), and the percentage of which have the term “deployment” or “implementation” in them (b).

ML4H Research



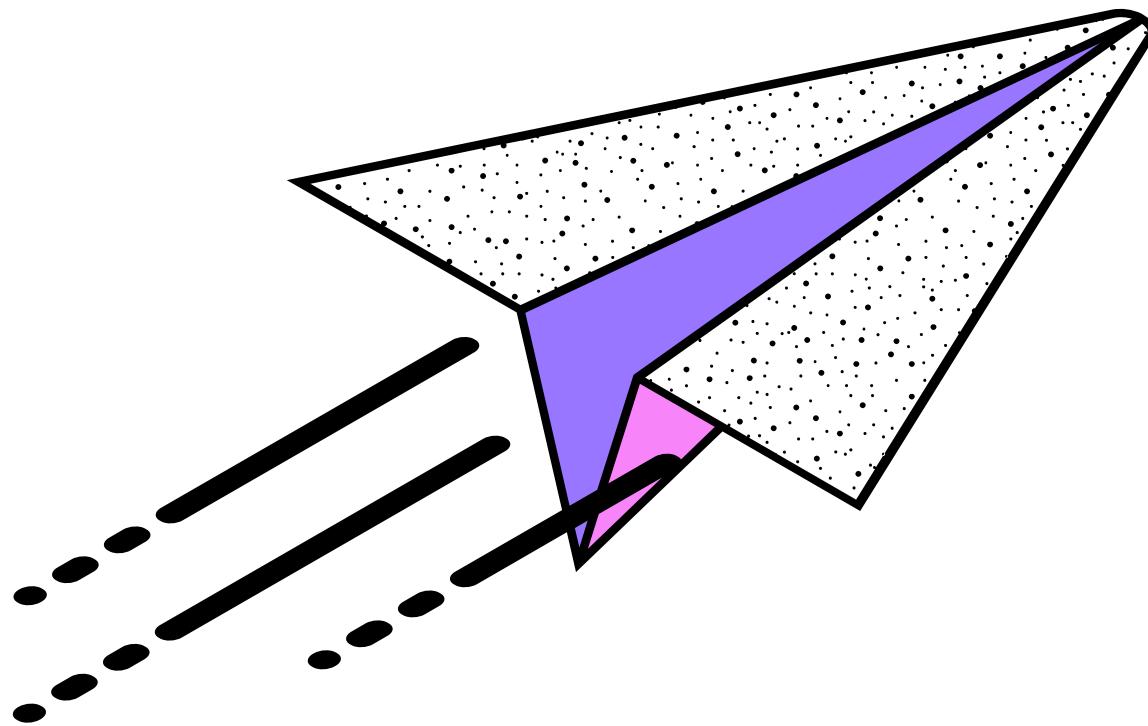
Deployed Projects

ML4H Research



Deployed Projects

Contents



“Engineering” problem

“Infrastructure” problem

“Regulatory” problem

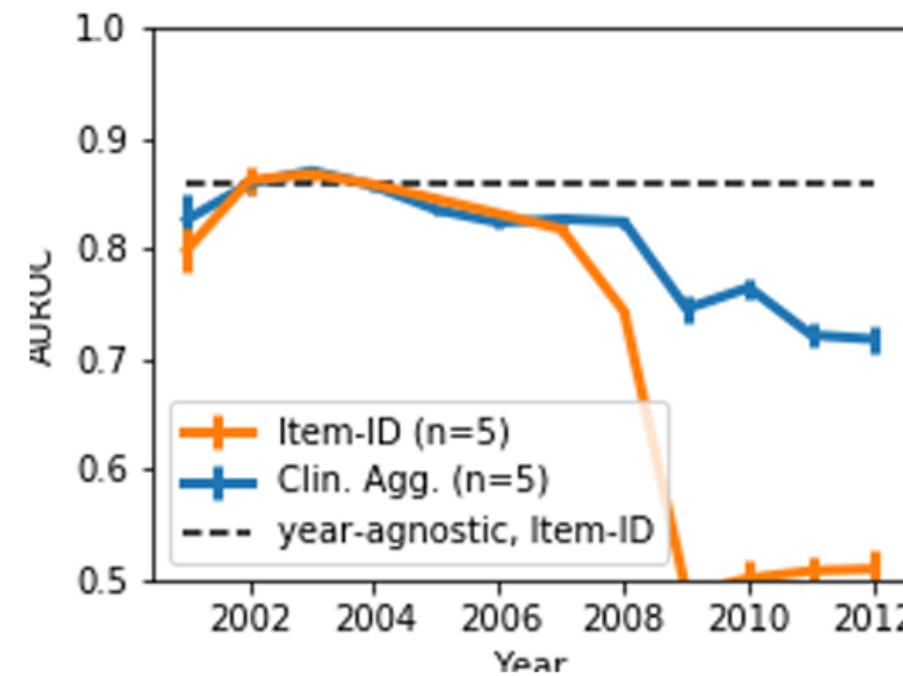
“Design” problem

Case Study at SickKids

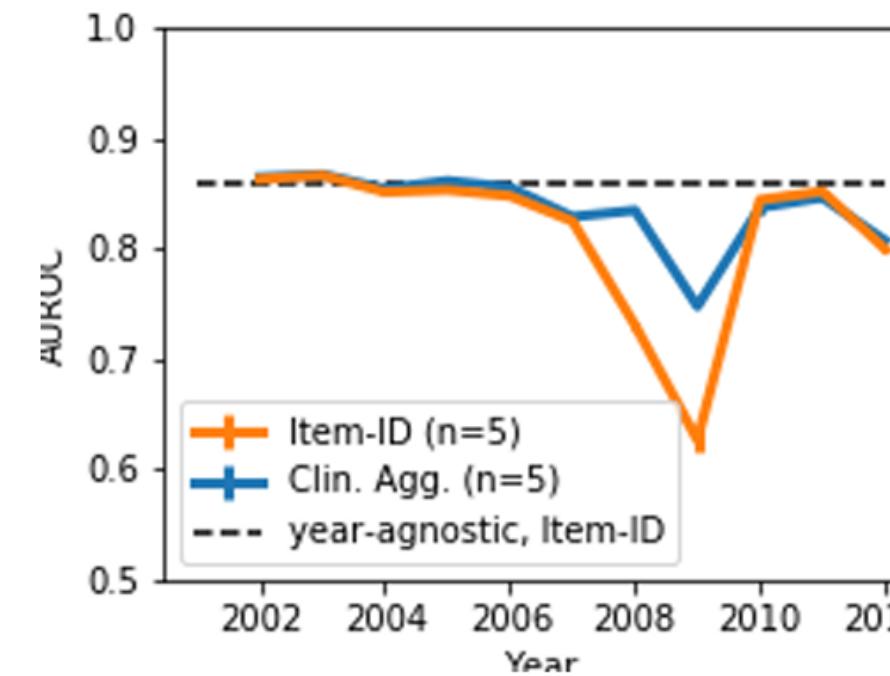
“Engineering” Problem

- **Distribution Shifts**
 - Dataset Shift - varying distributions over time/site
 - Generalizability - scalability
- **Algorithmic Bias**
 - Are algorithms reproducing or exacerbating existing systematic issues in healthcare?
- **Reliability of Metrics**
 - How does AUROC translate to patient outcomes?
 - Are the metrics we use to evaluate models even the right ones to consider?

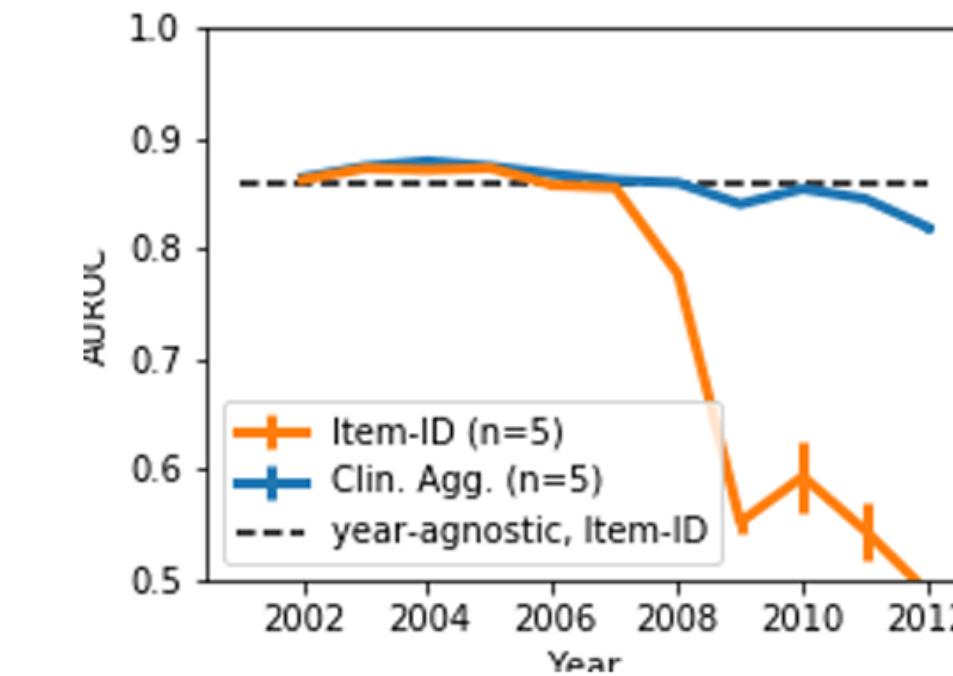
Dataset Shift



(a) Mortality AUC, models trained on 2001-2002 data.

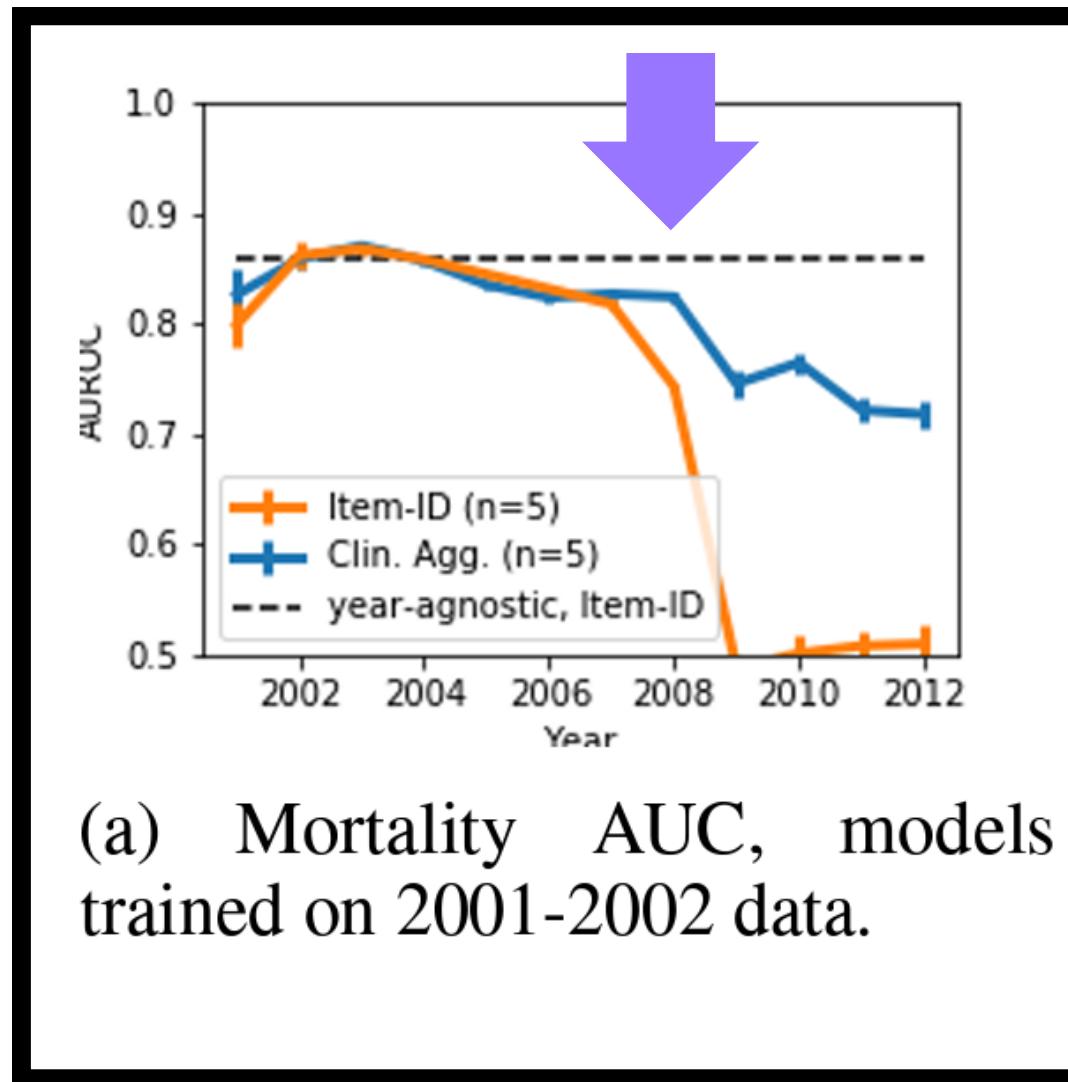


(b) Mortality AUC, models trained yearly on prior year only.

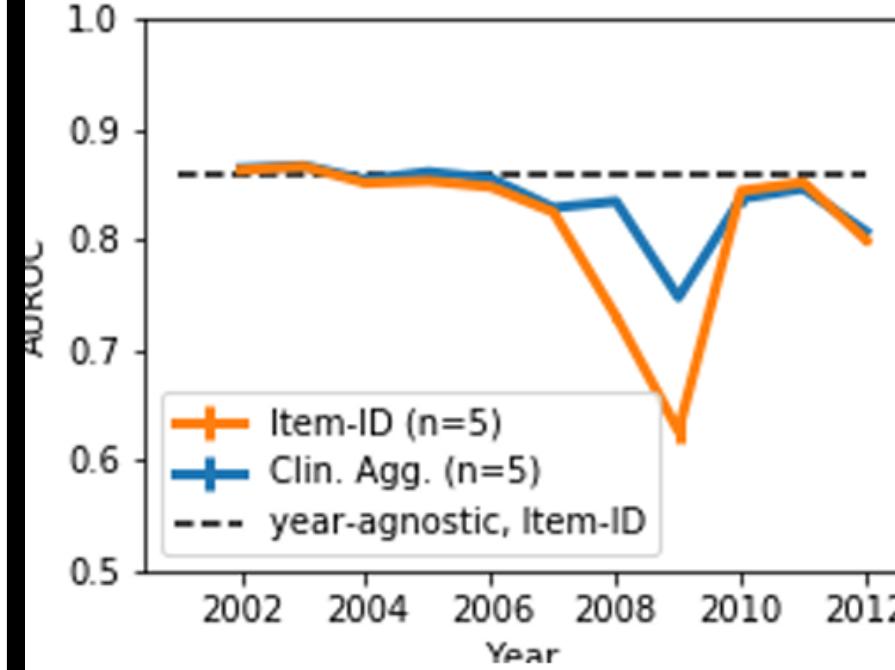


(c) Mortality AUC, models trained yearly on all prior data.

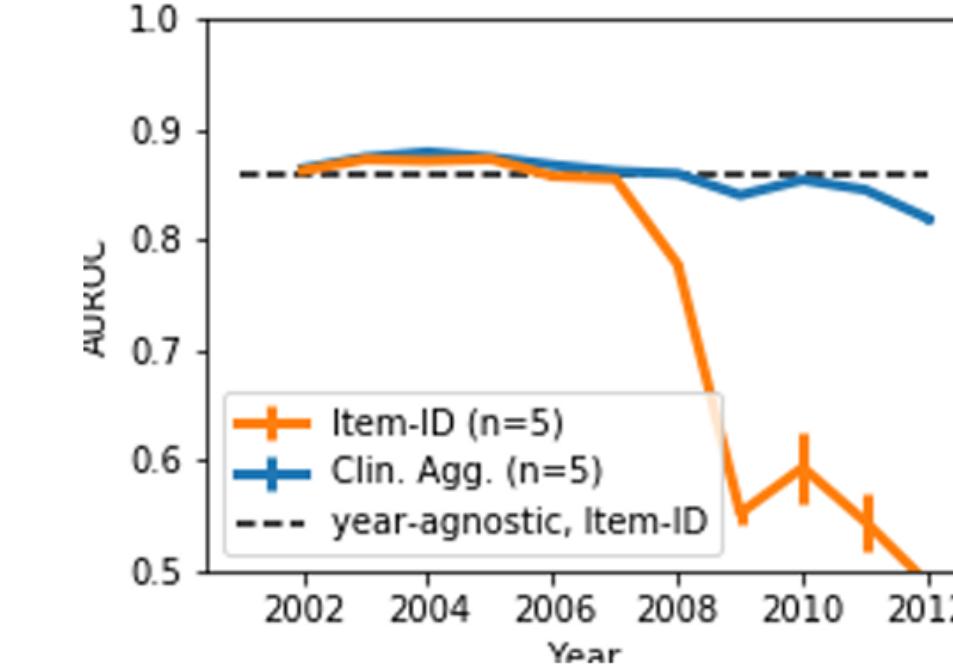
Dataset Shift



(a) Mortality AUC, models trained on 2001-2002 data.

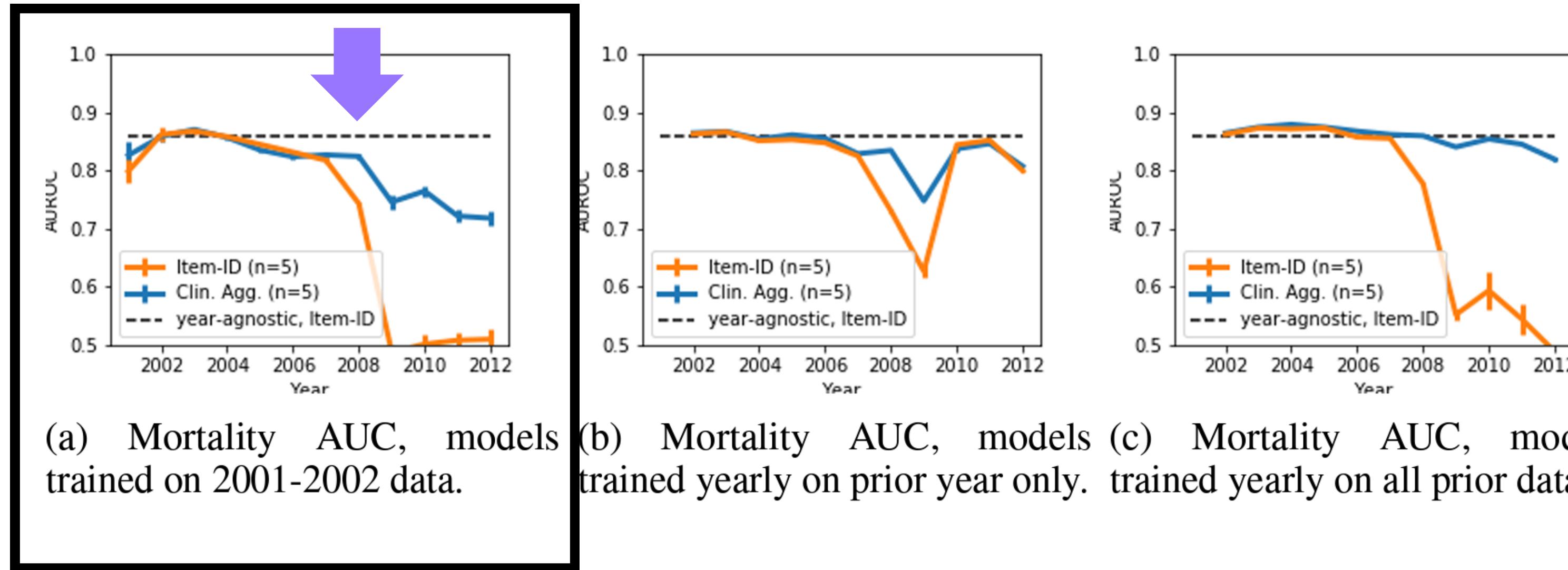


(b) Mortality AUC, models trained yearly on prior year only.



(c) Mortality AUC, models trained yearly on all prior data.

Dataset Shift



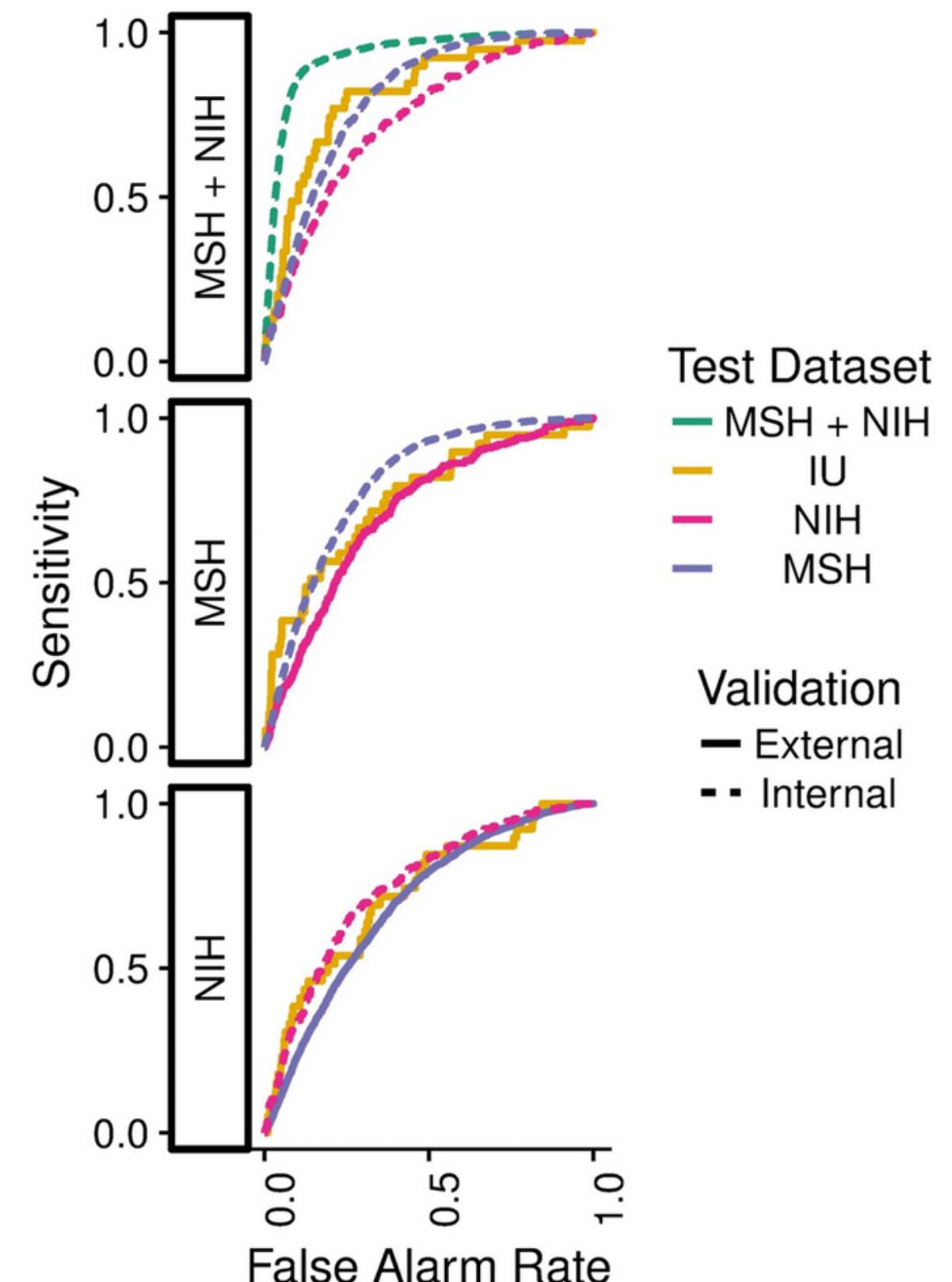
Nestor et al. 2018 ML4H

Other causes of Dataset Shift?

Generalizability

Table 1. Baseline characteristics of datasets by site.

Characteristic	IU	MSH	NIH
Patient demographics			
No. patient radiographs	3,807	42,396	112,120
No. patients	3,683	12,904	30,805
Age, mean (SD), years	49.6 (17.0)	63.2 (16.5)	46.9 (16.6)
No. females (%)	643 (57.3%)	18,993 (44.8%)	48,780 (43.5%)
Image diagnosis frequencies			
Pneumonia, No. (%)	39 (1.0%)	14,515 (34.2%)	1,353 (1.2%)
Emphysema, No. (%)	62 (1.6%)	1,308 (3.1%)	2,516 (2.2%)
Effusion, No. (%)	142 (3.7%)	19,536 (46.1%)	13,307 (11.9%)
Consolidation, No. (%)	26 (0.7%)	25,318 (59.7%)	4,667 (4.2%)
Nodule, No. (%)	104 (2.7%)	569 (1.3%)	6,323 (5.6%)
Atelectasis, No. (%)	307 (8.1%)	16,713 (39.4%)	11,535 (10.3%)
Edema, No. (%)	45 (1.2%)	7,144 (16.9%)	2,303 (2.1%)
Cardiomegaly, No. (%)	328 (8.6%)	14,285 (33.7%)	2,772 (2.5%)
Hernia, No. (%)	46 (1.2%)	228 (0.5%)	227 (0.2%)



Algorithmic Bias

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

Area under the receiver operating characteristics curve		
Race detection in radiology imaging		
Chest x-ray (internal validation)*		
MXR (Resnet34, Densenet121)	0.97, 0.94	
CXP (Resnet 34)	0.98	
EMX (Resnet34, Densenet121, EfficientNet-B0)	0.98, 0.97, 0.99	
Chest x-ray (external validation)*		
MXR to CXP, MXR to EMX	0.97, 0.97	
CXP to EMX, CXP to MXR	0.97, 0.96	
EMX to MXR, EMX to CXP	0.98, 0.98	
Experiments on anatomic and phenotypic confounders		
BMI*		
CXP	0.55, 0.52	
Image-based race detection stratified by BMI†		
EMX, MXR	Multiple results (appendix p 24)	
Breast density*		
EM-Mammo	0.54	
Breast density and age*		
EM-Mammo	0.61	
Disease distribution*		
MXR, CXP	0.61, 0.57	
Image-based race detection for the no finding class*		
MXR	0.94	
Model prediction after training on dataset with equal disease distribution†		
MXR	0.75	
Removal of bone density features*		
MXR, CXP	0.96, 0.94	

“Infrastructure” Problem

- Siloed Data
 - **No single source of all patient data** exists
 - **Digital walls** between (and within) institutions preventing linkage
- **EMR Issues**
 - Not all institutions have the same EMR system, making data linkage even more difficult
 - Feature names can be different across hospitals
- Lack of **computational resources** to store and deploy models
- Clinical interfaces
 - How do you convey algorithmic insights into the EMR system?

Siloed Data

UHN

PMCC

TWH

TGH

**Toronto
Rehab**

Unity Health

St Micheal's

St Joseph's

Sunnybrook

Sinai

Siloed Data

- Family Health Teams / Private Clinics / Hospitals
- Many health systems still use paper records
- Might have competing EMR systems
- Varying **healthcare system models**
 - Hospital system vs insurance system
 - In Ontario, ICES contains billing code data at a provincial level
 - In the USA, private insurers might have these records

Compute Infrastructure

- Storing **vast quantities of data, training deep learning models, hosting streaming inference** is not trivial
- **How do rural sites keep up** with the resources in academic settings?
- Existing rural/urban divide in infrastructure in Canada - can this be exacerbated?
- Do models trained in Toronto even generalize to Kingston or Port Hope?

“Regulatory” Problem

- **521 FDA-approved Medical AI devices**, growing every month
- In Canada, “**Software as a Medical Device**”, is the closest thing we have
 - But it would treat EMRs and Automated Diagnostic Tools in the same category. When they’re clearly not
- Regulatory **oversight needs to be ongoing**
- How do we ensure financial sustainability?
 - Doctors in Canada need to be able to bill for services, how does this work for AI?
 - Needs to be a **financial incentive to adopt tools**

“Design” Problem

- **Lack of well-informed design methodologies**
 - “Move fast and break things” does not work in healthcare
 - Clinicians hate using EMRs (ie: EPIC)
 - Alarm fatigue
- Does a tool **integrate into workflow** efficiently?
- Projects that are **detached from clinical use**
 - Common question in medicine when deciding on performing a Test - does it change practice?
 - Does it provide information to change your prior?
 - Ex: Sepsis prediction horizon - 1 hr vs 24 hr vs 48 hr
- Importance of language...
 - “Deployment” vs “Integration”

ANNALS OF MEDICINE

WHY DOCTORS HATE THEIR COMPUTERS

Digitization promises to make medical care easier and more efficient. But are screens coming between doctors and patients?

By Atul Gawande

November 5, 2018

Why Are Digital Health Care Systems Still Poorly Designed, and Why Is Health Care Practice Not Asking for More? Three Paths Toward a Sustainable Digital Work Environment

Monitoring Editor: Rita Kukafka

Reviewed by Alessandro Jatoba and Berglind Smaradottir

[Johanna Persson](#), MSc, PhD^{✉#1} and [Christofer Rydenfält](#), MSc, PhD^{#1}

Case Study: Sepsis Watch



Repairing Innovation

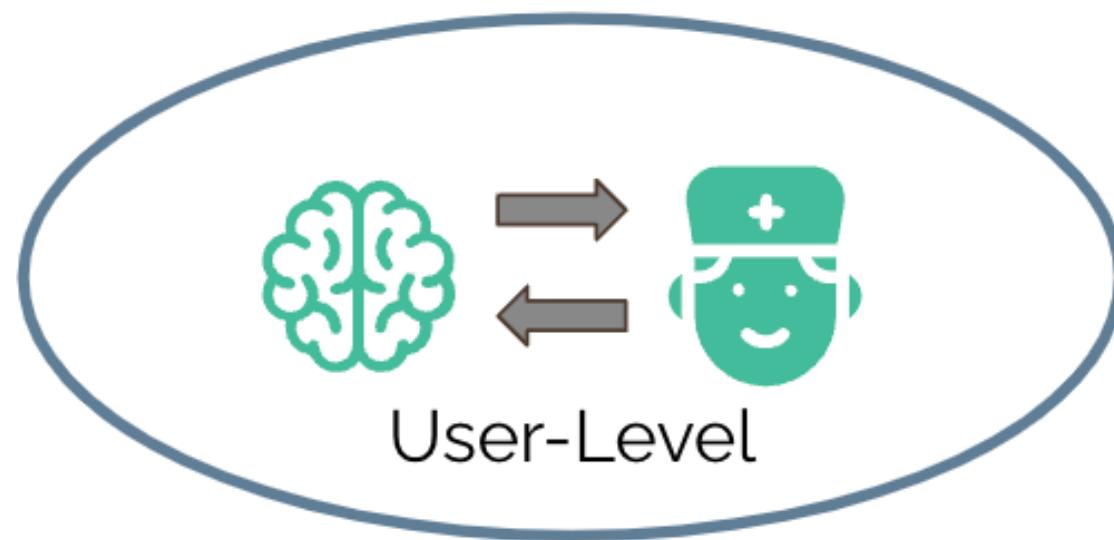
A Study of Integrating
AI in Clinical Care

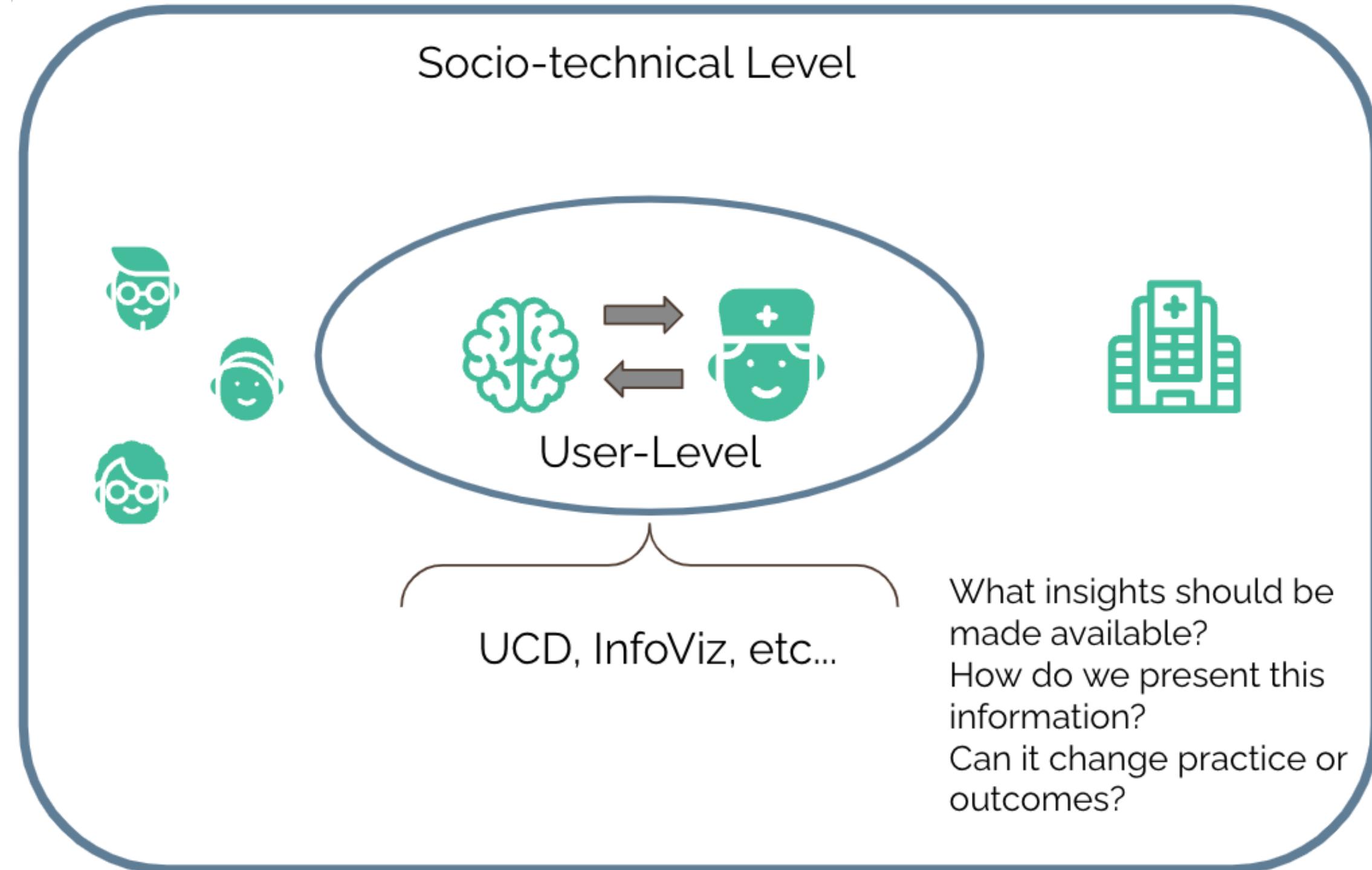
Madeleine Clare Elish
Elizabeth Anne Watkins

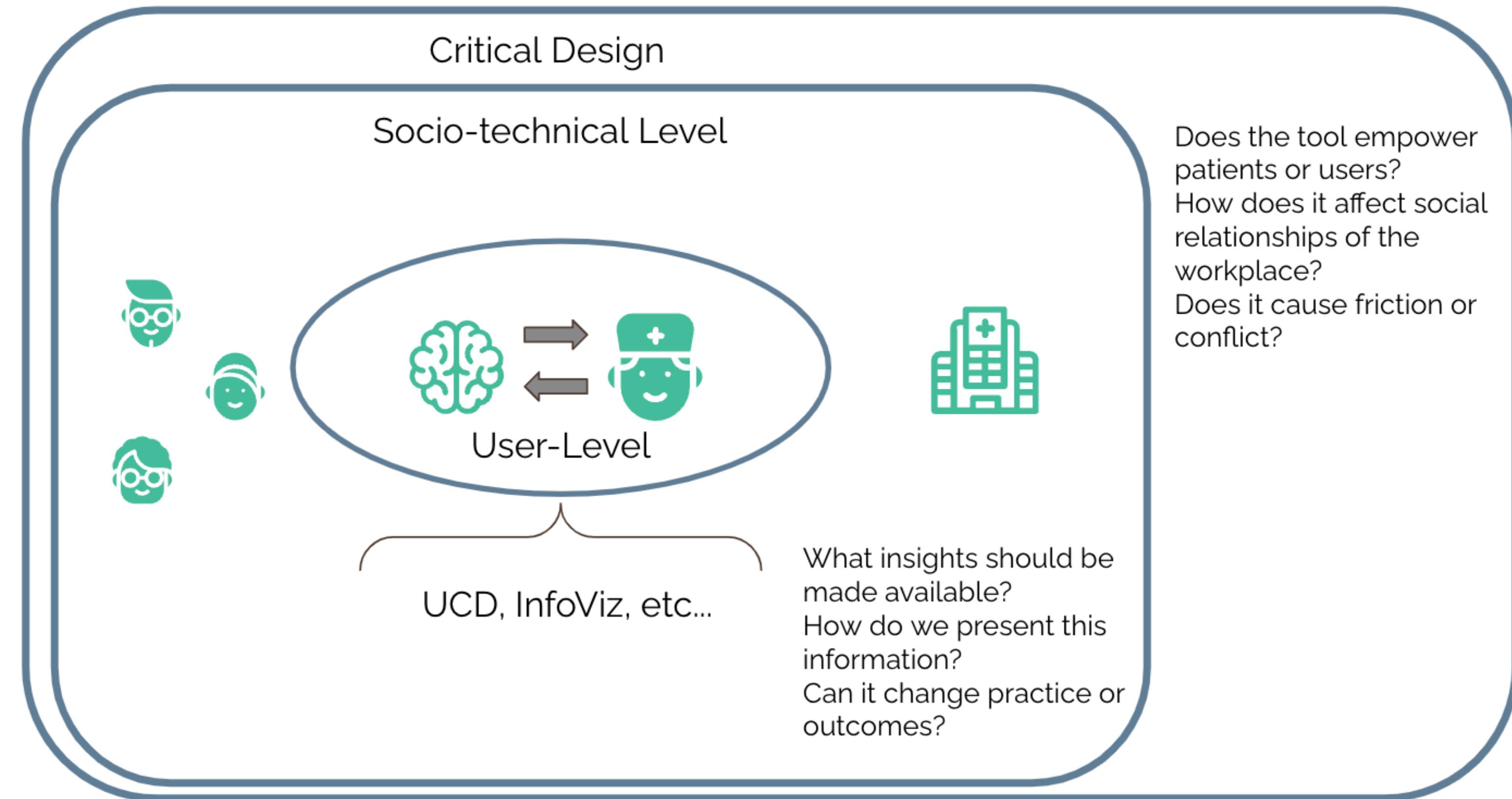
Case Study: Sepsis Watch

*“Elish and Watkins chronicle the integration of Sepsis Watch through a **sociotechnical lens**: one that acknowledges the **human labor required to harmonize a technical system with existing organizational and social structures**”*

- Sepsis prediction tool at Duke
- Integration of **AI breaks social structures in units**, and **requires repair work** (in this case from the nurses)
- They “mediated professional hierarchies and performed emotional labor to strategically communicate patients’ risk scores to doctors”
- This **work is often hidden and undervalued**
- Can lead to excess strain/fatigue for a provider and for inter-provider relationships







Participatory Design

- Origins in Scandinavian movements seeking workplace democratization - inherently political
- Involvement of **stakeholders and users and designers, together designing and co-creating**
- Requires participation (rooted in power and agency)
 - Questions of who participates and who gets to participate
- Difficult in settings of **asymmetrical power**
 - **Power imbalances in healthcare**
- Not new in healthcare:
 - Conflicts may tend to arise because the well-established scientific rationality, culture, and biomedical approach in healthcare may clash with epistemological cultural assumptions of PD

On Participation

Qualitative research

Original research

Conditionally positive: a qualitative study of public perceptions about using health data for artificial intelligence research 

Melissa D McCradden¹, Tasmie Sarker²,  P Alison Paprica^{2, 3}

Research

Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study

Melissa D. McCradden, Ami Baba, Ashirbani Saha, Sidra Ahmad, Kanwar Boparai, Pantea Fadaiefard and Michael D. Cusimano

February 18, 2020 8 (1) E90-E95; DOI: <https://doi.org/10.9778/cmajo.20190151>

A novel approach to machine learning-based automated vascular catheter access detection in a paediatric critical care setting

² ² Sujay Nagaraj¹, Andrew J. Goodwin^{2,4}, Sebastian Goodfellow, Robert W. Greer, Danny Eytan^{2,3} Anna Goldenberg^{5,6,7}, and Mjaye L. Mazwi²

1. Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
2. Department of Critical Care Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada
3. Department of Medicine, Technion, Haifa, Israel
4. School of Electrical and Information Engineering, University of Sydney, Sydney, Australia
5. SickKids Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada
6. Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
7. Vector Institute, Toronto, Ontario, Canada

Introduction

- **High-frequency physiological waveform data** sampled at up to 500 Hz (i.e.: EKG, ABP, etc.)
- Data can be difficult to interpret due to artefacts
- Traditional analysis requires artefact removal. However, **certain artefacts are relevant and important**
- Catheter accesses generate a **unique** characteristic artefact
- Detection of such artefacts in real-time provides valuable clinical insight

Arterial Blood Pressure Waveform

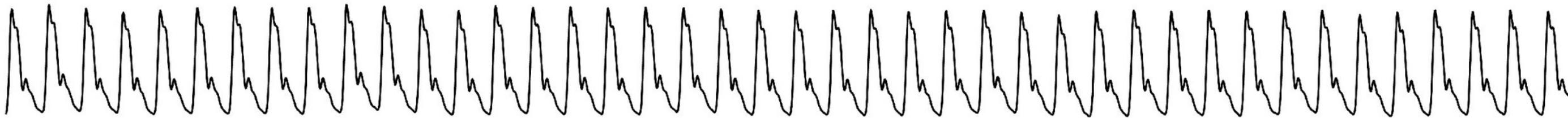
1 Hour



Arterial Blood Pressure



1 Minute



1 Second

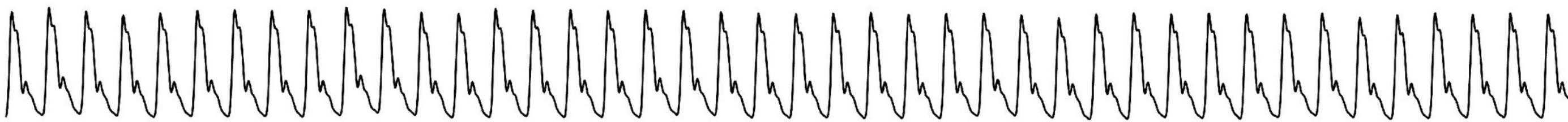


Arterial Blood Pressure Waveform

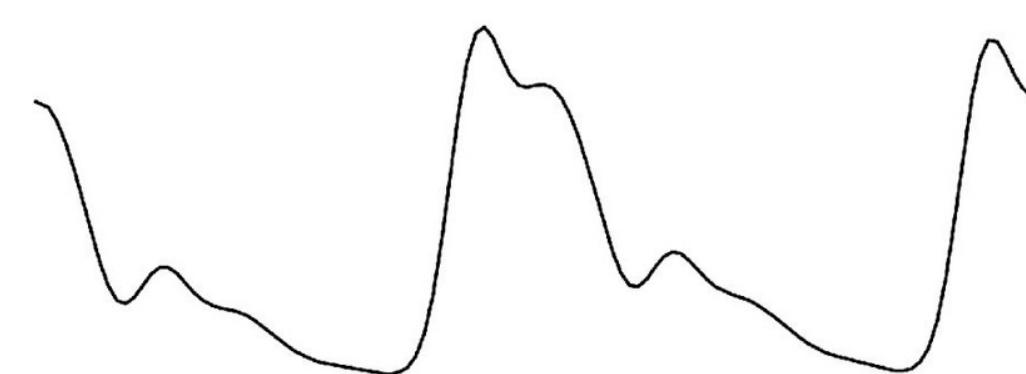
1 Hour



1 Minute



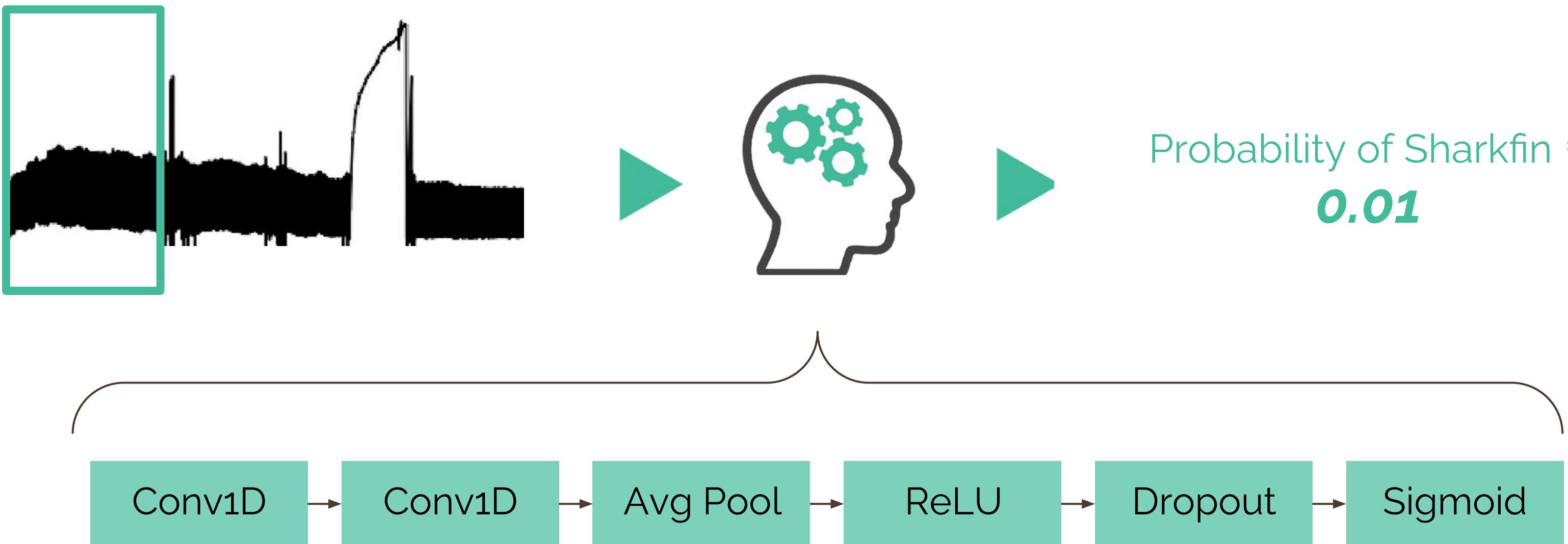
1 Second



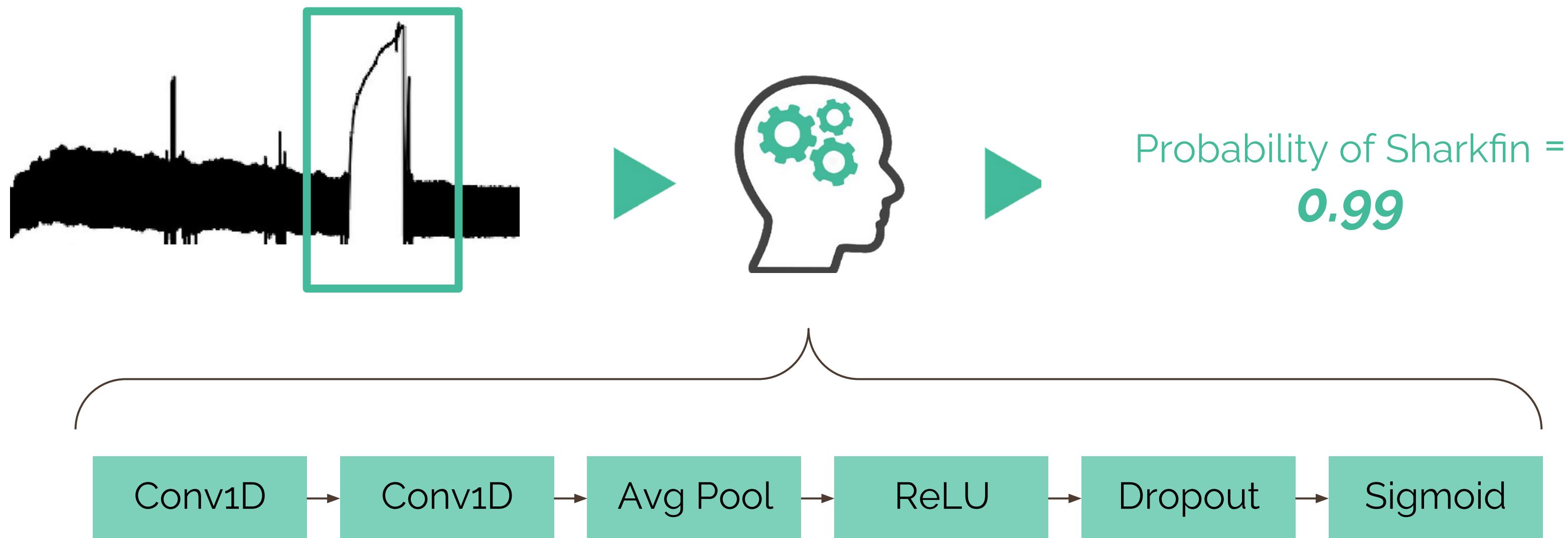
Objectives

Develop and deploy a machine-learning tool capable of accurately detecting catheter access events (sharkfins) in real-time

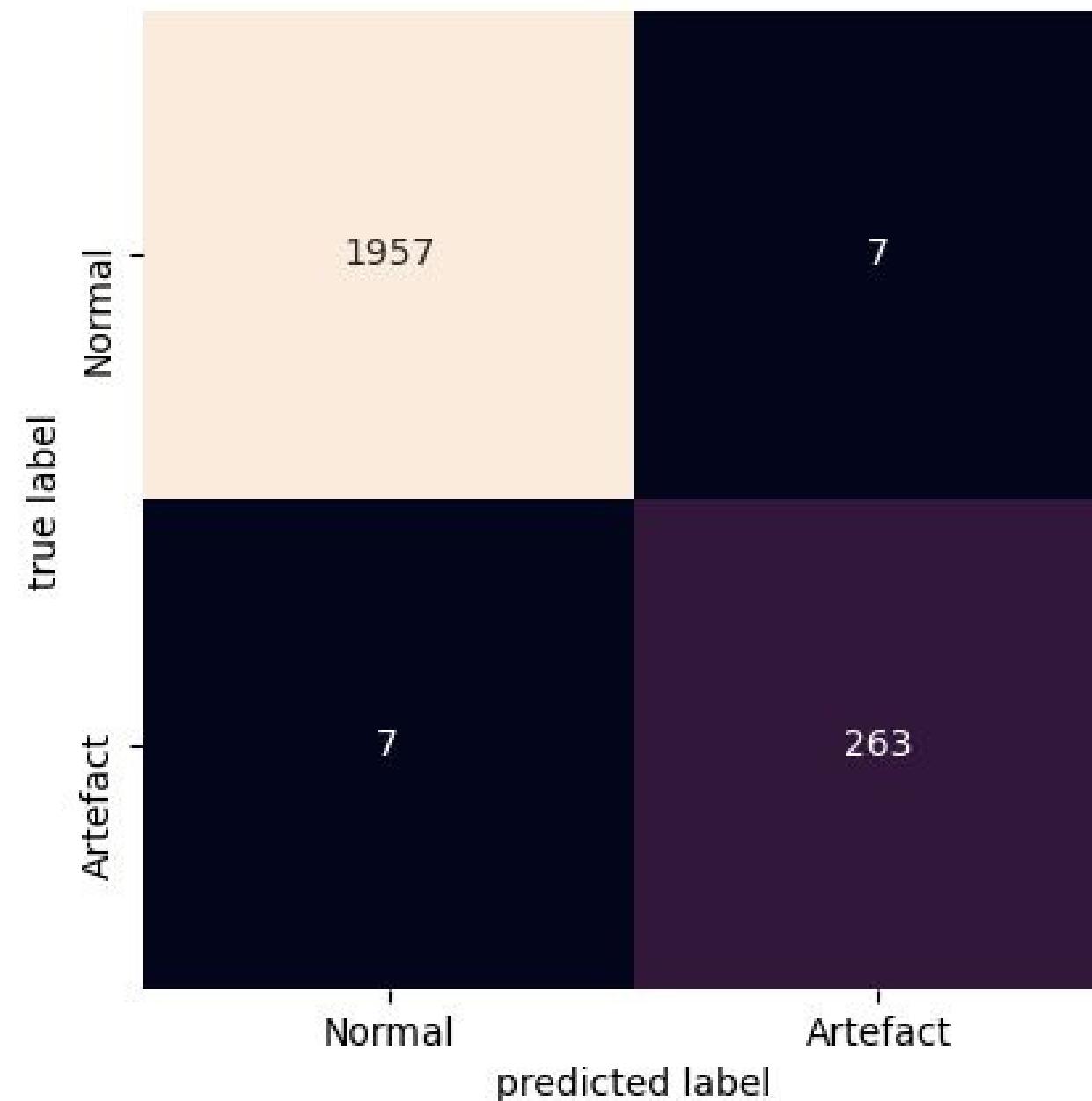
Model Training Objective



Model Training Objective



Model Performance



Accuracy: **0.99**
Precision: **0.97**
Recall: **0.97**
F1 score: **0.97**
ROC AUC score: **0.99**

Model Testing



1-minute windows of n=270 sharkfins

Goal: learn to detect the characteristic shape of this artefact

Real-world data is much noisier, greater class-imbalance, etc.

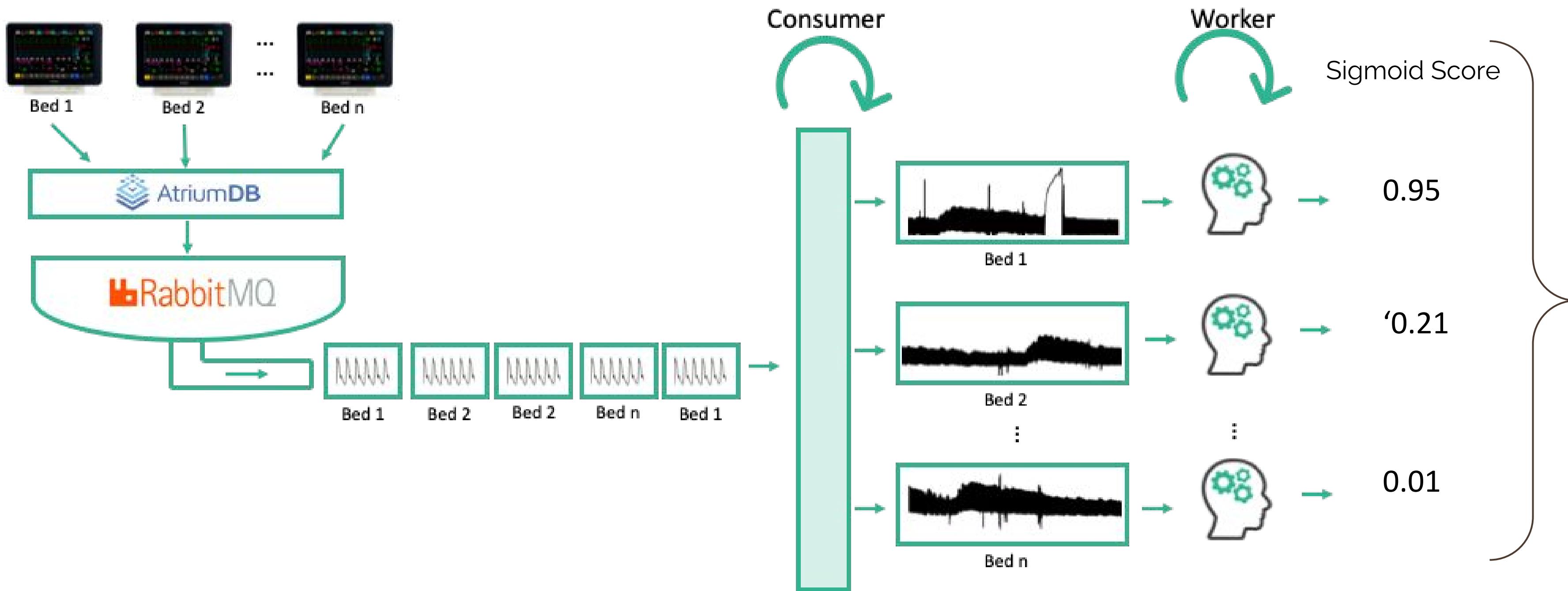
Simulated Deployment

Several months of continuous waveform from 1 device (n=548 sharkfins)

Goal: simulate prospective deployment (“online” inference + class imbalance + noise)

Provides us with a means to tune hyperparameters (window length, slide, smoothing)

Real-time Implementation



Path to Translation - where we are now



- 1) Correlate observation to artefact
- 2) Verify model's performance on real-time, streaming data
- 3) Test the deployment

- 1) Compare to current documentation practices

Path to Translation - where we are going



- 1) Correlate observation to artefact
- 2) Verify model's performance on real-time, streaming data
- 3) Test the deployment

- 1) Compare to current documentation practices

- 1) What to do with the results?
- 2) How to present the results? What information is important?

Applications

1. Quality Improvement

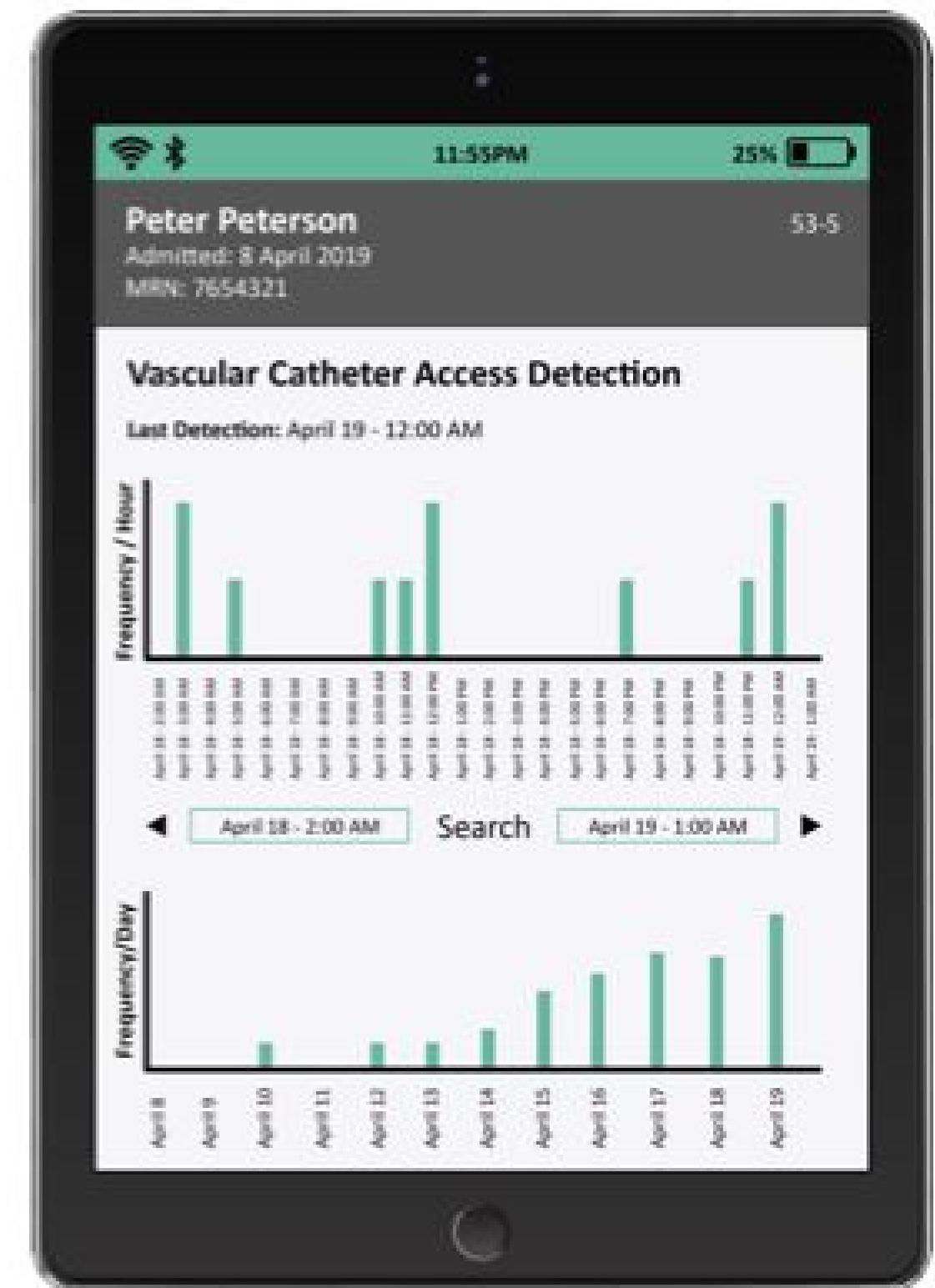
- I. Initiatives to reduce line access using highly accurate information about line utilization patterns

2. Risk Assessment

- I. Ascertaining whether changing patterns of line access can be a proxy to changing patient status

3. Data Science

- I. Accurately time-aligning biomarkers and medication administration
- II. Identifying periods of time where waveform data does not reflect patient physiology



Roadblocks

- **Dataset shift**
 - Models that work well on curated test-sets don't always translate to prospective 24/7 streaming inference
- Pipeline for **streaming inference was not trivial**
 - Have had students do whole masters / PhD projects on this
- **Labour-intensive prospective validation / silent-trial**
- **Pandemic-related closure** of the unit

Driving Factors

- Amazing, **multidisciplinary team**
 - Clinicians, engineers, ML
- Solving a **low-hanging fruit problem**
 - Easily interpretable, strong buy-in from stakeholders
 - Automating an existing documentation task vs reinventing medicine
- Adequate **infrastructure** to build and host ML models
 - Infrastructure has been iterated upon for years
- **Institution-level buy-in** that allow retention of talent

Relevant Reading:

How to validate Machine Learning Models Prior to Deployment: Silent trial protocol for evaluation of real-time models at ICU

Sana Tonekaboni, Gabriela Morgenshtern, Azadeh Assadi, Aslesha Pokhrel, Xi Huang, Anand Jayarajan, Robert Greer, Gennady Pekhimenko, Melissa McCradden, Fanny Chevalier, Mjaye Mazwi, Anna Goldenberg *Proceedings of the Conference on Health, Inference, and Learning, PMLR 174:169-182, 2022.*

> [Front Digit Health](#). 2022 Aug 16:4:929508. doi: 10.3389/fdgh.2022.929508. eCollection 2022.

The silent trial – the bridge between bench-to-bedside clinical AI applications

Jethro C C Kwong ^{1 2}, Lauren Erdman ^{2 3 4}, Adree Khondker ⁵, Marta Skreta ³, Anna Goldenberg ^{2 3 4}, Melissa D McCradden ^{2 6 7 8}, Armando J Lorenzo ^{1 5}, Mandy Rickard ⁵

REVIEW | VOLUME 2, ISSUE 10, E537-E548, OCTOBER 2020

 Download Full Issue

Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension

Xiaoxuan Liu, MBChB • Samantha Cruz Rivera, PhD • David Moher, PhD • Prof Melanie J Calvert, PhD •

Prof Alastair K Denniston, PhD   and the SPIRIT-AI and CONSORT-AI Working Group [†] • Show footnotes

REVIEW | VOLUME 2, ISSUE 10, E549-E560, OCTOBER 2020

 Download Full Issue

Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension

Samantha Cruz Rivera, PhD • Xiaoxuan Liu, MBChB • An-Wen Chan, MD • Prof Alastair K Denniston, PhD  

Prof Melanie J Calvert, PhD • and The SPIRIT-AI and CONSORT-AI Working Group [†] • Show footnotes

Questions?



s.nagaraj@mail.utoronto.ca