Data 146: Foundations for CPH

Case Study of AI and Radiology

Irene Y. Chen

Announcements

- Final project due Dec 11 5pm
  - Groups of 2-3 people
- Non-graded progress slide due Nov 21 7pm
  - Project goal, method, data, any preliminary results, expected results
  - Short feedback will be given

Example Project Slide

Team Member 1, Team Member 2, …

Research Problem / Project Goal

Idea

- Bullet point
- Bullet point

Impact /vision

What impact your project will introduce

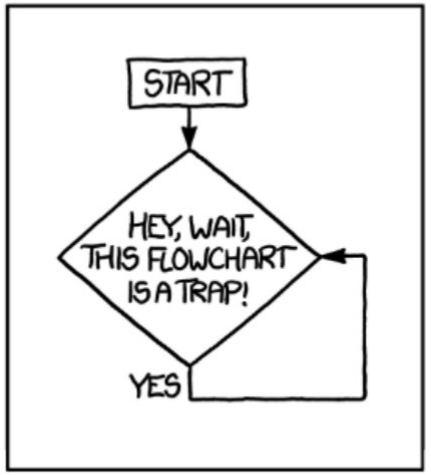
Methods / Data / Model

Idea

- Bullet point
- Bullet point

Idea

- Bullet point
- Bullet point
- Bullet point



Caption: x

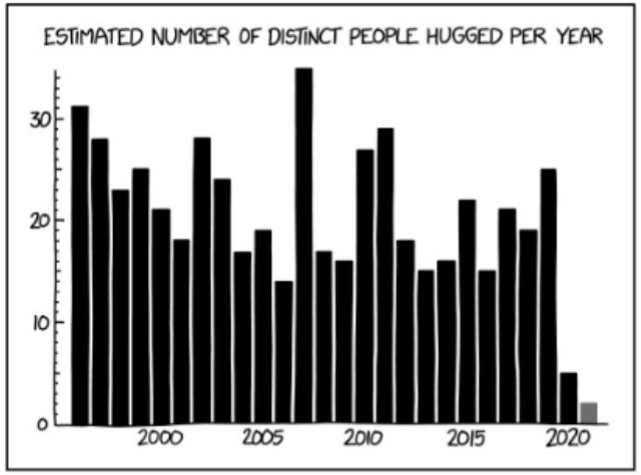
Preliminary Results / Analysis (if any)

Idea

- Bullet point
- Bullet point

Idea

- Bullet point
- Bullet point
- Bullet point



Caption: y

Next Steps / Challenges / Questions / Discussion

Idea

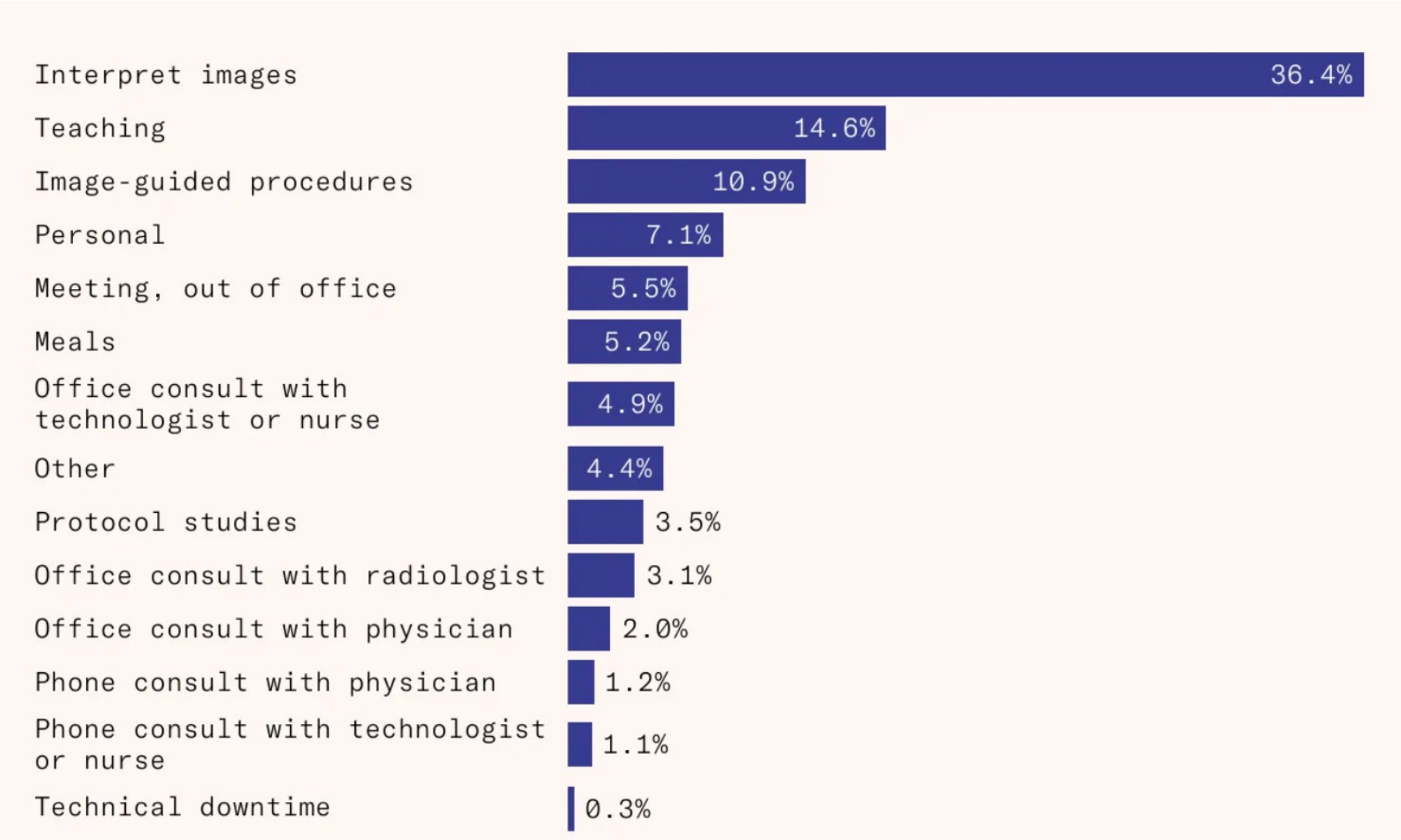
- Bullet point
- Bullet point

Example Project Slide

- Keep these five blocks!
- But feel free to make any changes to the style (e.g., format, font, and visualization)!

“We should stop training radiologists now. It’s just completely obvious that within five years, deep learning is going to do better than radiologists.”

- Geoff Hinton (2016)



Dhanoa et al, “The evolving role of the radiologist: the Vancouver workload utilization evaluation study”, *Journal of American College of Radiology* 2013

Main reasons

1. Misunderstanding of radiology job specifications
2. Benchmarks didn’t show true performance
3. Implementation and regulatory blockers

Main reasons

1. Misunderstanding of radiology job specifications
2. Benchmarks didn’t show true performance
3. Implementation and regulatory blockers

Outline

- **Dataset Shift** (30 mins)
- **Deployment Challenges** (20 mins)

How can we make Data 146 better for you?

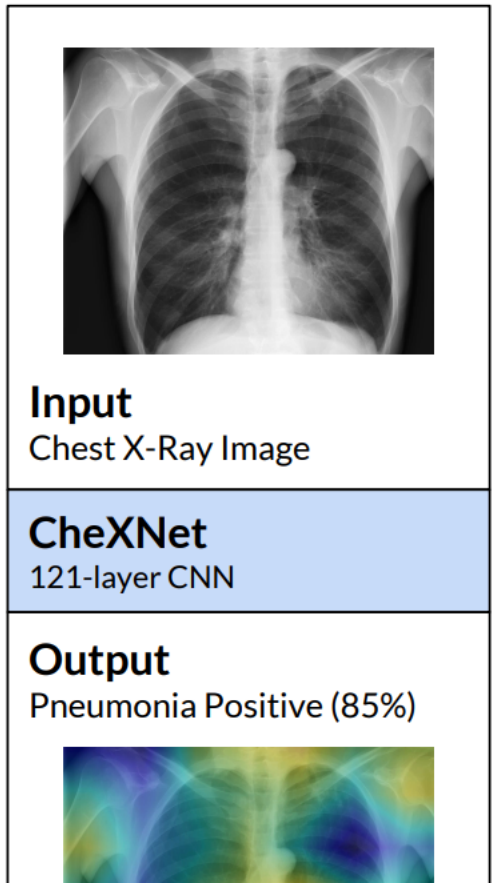
Learning Objective: Understand why high benchmark performance might not translate to impact

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar [*1] Jeremy Irvin [*1] Kaylie Zhu [1] Brandon Yang [1] Hershel Mehta [1]
Tony Duan [1] Daisy Ding [1] Aarti Bagul [1] Robyn L. Ball [2] Curtis Langlotz [3] Katie Shpanskaya [3]
Matthew P. Lungren [3] Andrew Y. Ng [1]

Abstract

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists. Our algorithm, CheXNet, is a 121-layer convolutional neural network trained on ChestX-ray14, currently the largest publicly available chest X-ray dataset, containing over 100,000 frontal-view X-ray images with 14 diseases. Four practicing academic radiologists annotate a test set, on which we compare the performance of CheXNet to that of radiologists. We find that CheXNet exceeds average radiologist performance on the F1 metric. We extend CheXNet to detect all 14 diseases in ChestX-ray14 and achieve state of the art results on all 14 diseases.

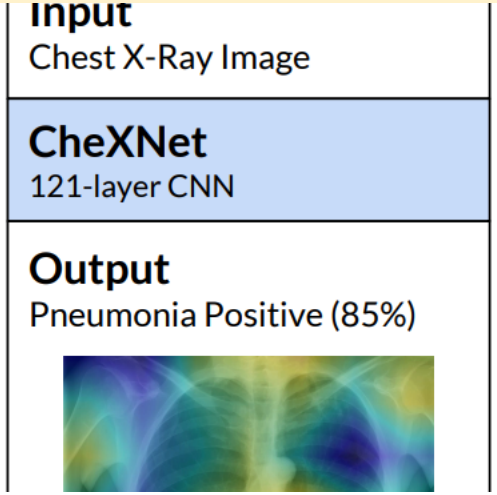


CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar [*1] Jeremy Irvin [*1] Kaylie Zhu [1] Brandon Yang [1] Hershel Mehta [1]
Tony Duan [1] Daisy Ding [1] Aarti Bagul [1] Robyn L. Ball [2] Curtis Langlotz [3] Katie Shpanskaya [3]
Matthew P. Lungren [3] Andrew Y. Ng [1]

Trained and validated on the same dataset!

practicing academic radiologists annotate a test set, on which we compare the performance of CheXNet to that of radiologists. We find that CheXNet exceeds average radiologist performance on the F1 metric. We extend CheXNet to detect all 14 diseases in ChestX-ray14 and achieve state of the art results on all 14 diseases.



Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, Eric Karl Oermann

Published: November 6, 2018 • https://doi.org/10.1371/journal.pmed.1002683

Abstract

Background

There is interest in using convolutional neural networks (CNNs) to analyze medical imaging to provide computer-aided diagnosis (CAD). Recent work has suggested that image classification CNNs may not generalize to new data as well as previously believed. We assessed how well CNNs generalized across three hospital systems for a simulated pneumonia screening task.

Methods and findings

A cross-sectional design with multiple model training cohorts was used to evaluate model

Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, Eric Karl Oermann

Published: November 6, 2018 • https://doi.org/10.1371/journal.pmed.1002683

On 5 different combinations, performance dropped by up to 12.4% (AUC 0.931 to 0.815)

Background

There is interest in using convolutional neural networks (CNNs) to analyze medical imaging to provide computer-aided diagnosis (CAD). Recent work has suggested that image classification CNNs may not generalize to new data as well as previously believed. We assessed how well CNNs generalized across three hospital systems for a simulated pneumonia screening task.

Methods and findings

A cross-sectional design with multiple model training cohorts was used to evaluate model

Lack of children in public medical imaging data points to growing age bias in biomedical AI

Authors:
Stanley Bryan Zamora Hua[1], Nicholas Heller[2], Ping He[1], Alexander J. Towbin[5,6], Irene Y. Chen[3,4], Alex X. Lu*[7], Lauren Erdman*[6,8,9]

Affiliations:

1. Center for Computational Medicine, The Hospital for Sick Children, Toronto, Canada
2. Department of Urology, Cleveland Clinic, Cleveland, OH
3. Computational Precision Health, University of California Berkeley and University of California San Francisco, California
4. Electrical Engineering and Computer Science, University of California Berkeley, California
5. Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, USA
6. Department of Radiology, University of Cincinnati College of Medicine, Cincinnati, USA
7. Microsoft Research, Boston, USA
8. James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, USA
9. Division of Gastroenterology, Cincinnati Children's Hospital Medical Center, Cincinnati, USA

* equal contribution

Corresponding Author: Lauren Erdman, lauren.erdman@cchmc.org

Lack of children in public medical imaging data points to growing age bias in biomedical AI

Authors:
Stanley Bryan Zamora Hua[1], Nicholas Heller[2], Ping He[1], Alexander J. Towbin[5,6], Irene Y. Chen[3,4], Alex X. Lu*[7], Lauren Erdman*[6,8,9]

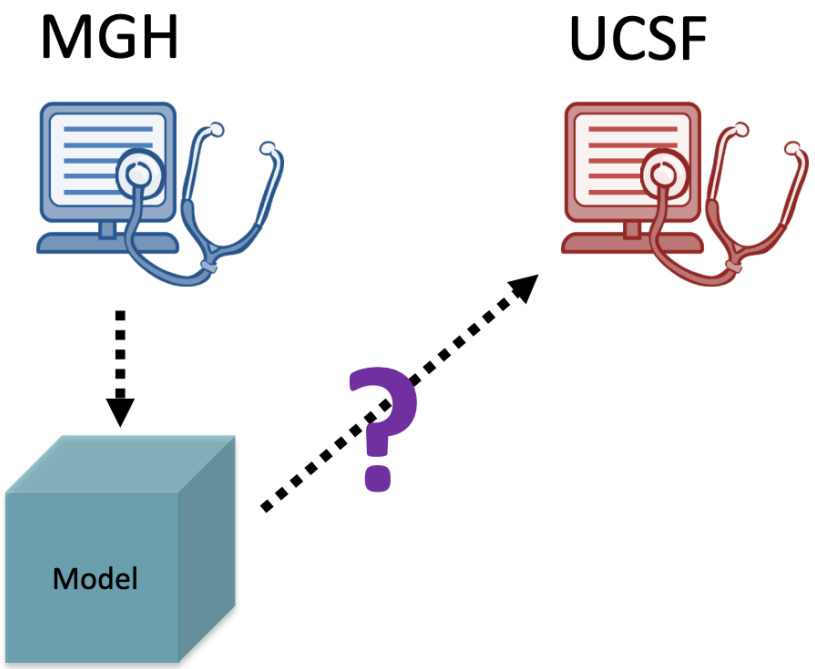
Our systematic review of 181 public medical imaging datasets reveals that children represent just under 1% of available data

Medical Center, Cincinnati, USA
9. Division of Gastroenterology, Cincinnati Children's Hospital Medical Center, Cincinnati, USA

* equal contribution

Corresponding Author: Lauren Erdman, lauren.erdman@cchmc.org

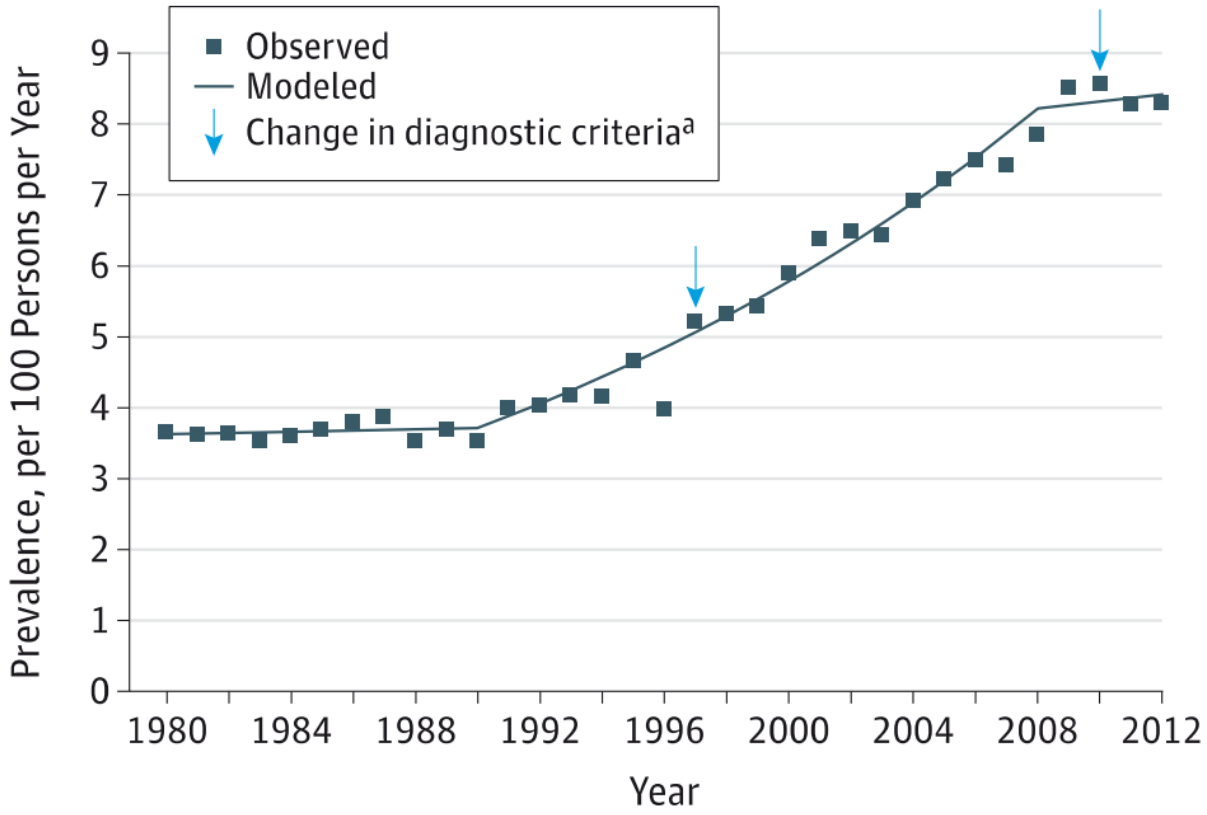
Dataset shift / non-stationarity: *Models often do not generalize*



What kinds of dataset shift might this cause, and why?

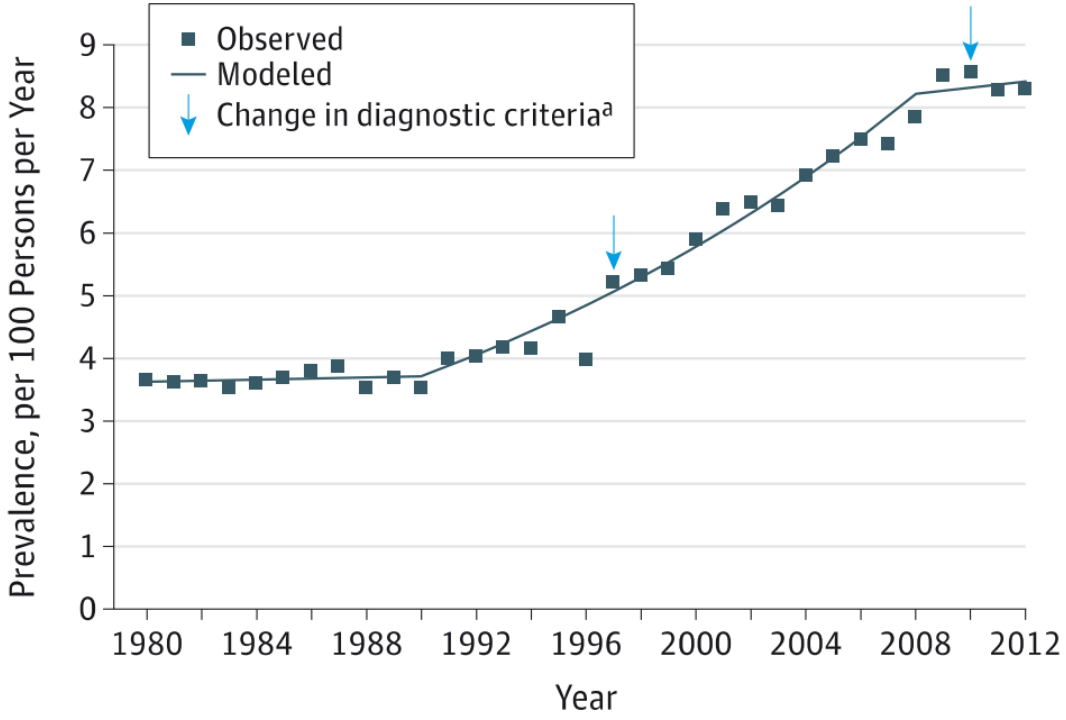
- Machine biases
- Population age/health
- Treatment patterns
- Treatment selection
- Past history of treatments
- Environmental factors
- Socioeconomic factors

Dataset shift / non-stationarity: *Diabetes Onset After 2009*



Why might diabetes go up over time?

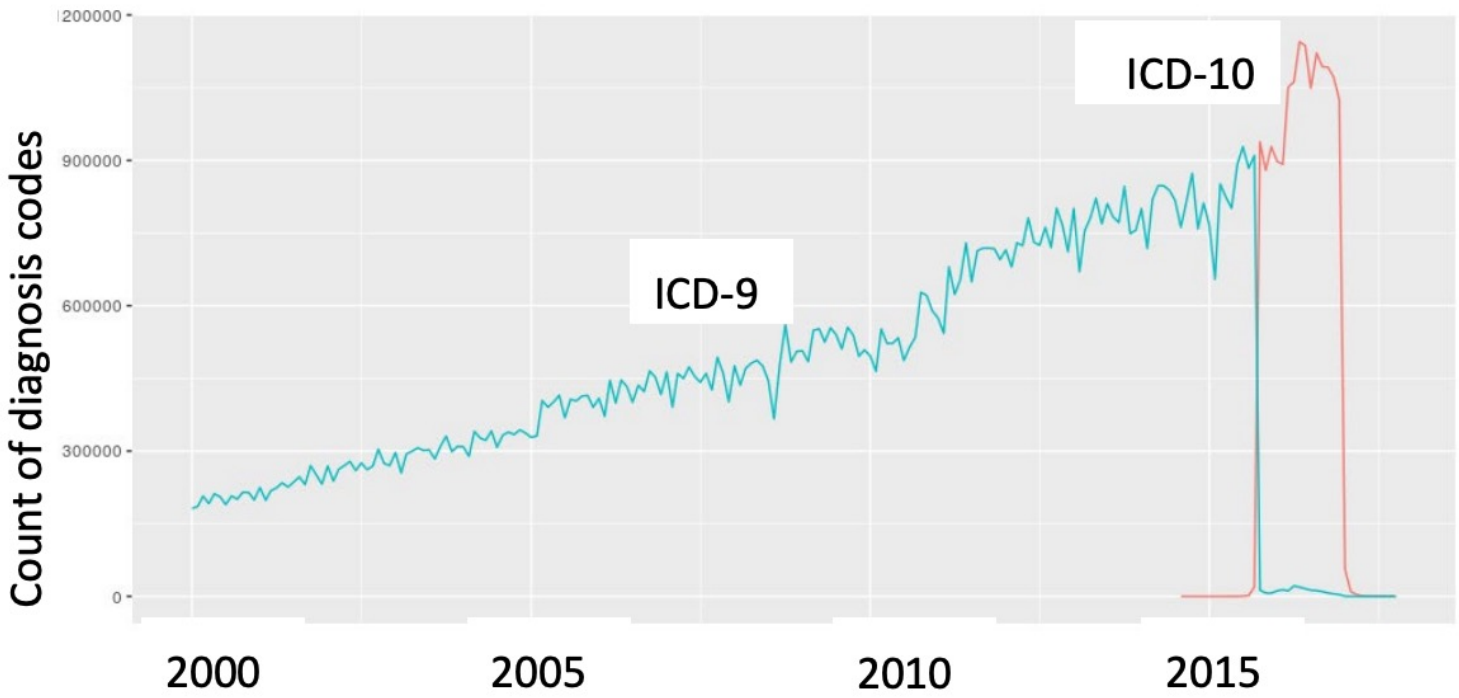
Dataset shift / non-stationarity: *Diabetes Onset After 2009*

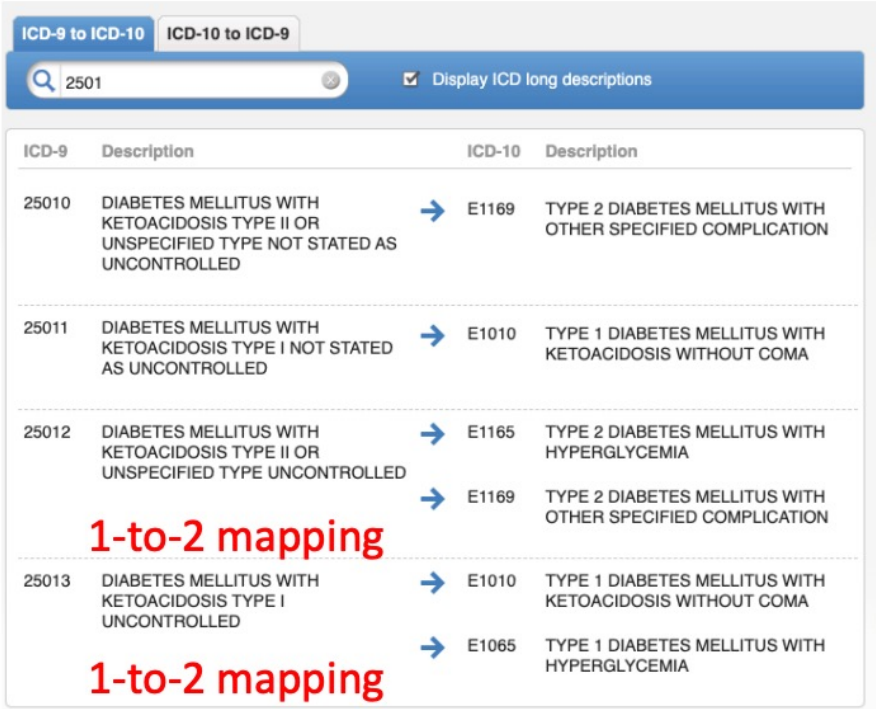


- Better diagnosis criteria
- Insulin / glucose biomarkers
- Obesity rates
- Definition of disease
- Meaning of label
- T1D vs T2D

[Geiss LS, Wang J, Cheng YJ, et al. Prevalence and Incidence Trends for Diagnosed Diabetes Among Adults Aged 20 to 79 Years, United States, 1980-2012. JAMA, 2014.]

Dataset shift / non-stationarity: *Diabetes Onset After 2009*





Significance of features may change over time. Note ICD10 to ICD9 isn't 1-1

[Figure credits: (Left) Mike Oberst, (Right) http://www.icd10codesearch.com/]

Formalizing Dataset Shift

- General Task: Perform well on a “target domain” Q
  - Train: Population **P** (e.g., MGH)
  - Apply: Population **Q** (e.g., UCSF)
- Assumptions: What is changing vs. what is stable?
  - Covariate Shift
  - Label Shift
  - More General Shift

Formalizing Dataset Shift

- General Task: Perform well on a “target domain” Q
- Assumptions: What is changing vs. what is stable?

An Impossible Problem

Given $\{X_i, Y_i\}_{i=1}^n$ from a *source domain* $\boldsymbol{P}(X, Y)$, find a model that performs well on some *target domain* $\boldsymbol{Q}(X, Y)$

$$\min_{f \in \mathcal{F}} \mathbb{E}_{\boldsymbol{Q}}[\ell(Y, f(X)]$$

Minimize the expected loss between truth Y and prediction *f*(X) in domain Q
Find the function *f* that minimizes this

Examples:
- P and Q are two different hospital systems
- P is the past, Q is the future
- …

Not well-posed without further assumptions or information about Q!

Formalizing Dataset Shift

- General Task: Perform well on a “target domain” Q
- Assumptions: What is changing vs. what is stable?

Example: Covariate Shift Assumption

$$P(X) \neq Q(X)$$

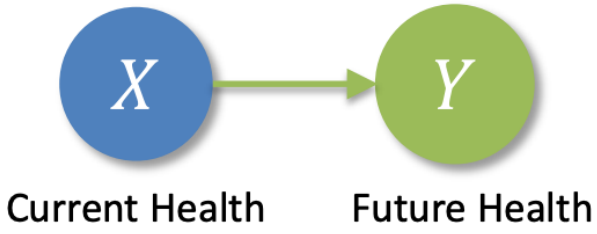
$$P(Y \mid X) = Q(Y \mid X)$$

Why might this be true? One rationale: $P(Y \mid X)$ encodes some "causal" mechanism

Example: Covariate Shift Assumption

$$P(X) \neq Q(X)$$
$$P(Y \mid X) = Q(Y \mid X)$$

Why might this be true? One rationale: $P(Y \mid X)$ encodes some "causal" mechanism



Example: Risk stratification for different patient populations

Common assumption: distribution of future health will be the same. But doesn't usually hold

Example: Label Shift Assumption

$$P(Y) \neq Q(\mathrm{Y})$$

$$P(X \mid Y) = Q(X \mid Y)$$

(flip directionality from previous slide)

Why might this be true? One rationale: $P(X \mid Y)$ encodes some "causal" mechanism

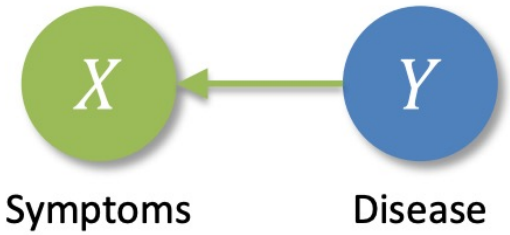
Example: Label Shift Assumption

$$P(Y) \neq Q(Y)$$

$$P(X \mid Y) = Q(X \mid Y)$$

(flip directionality from previous slide)

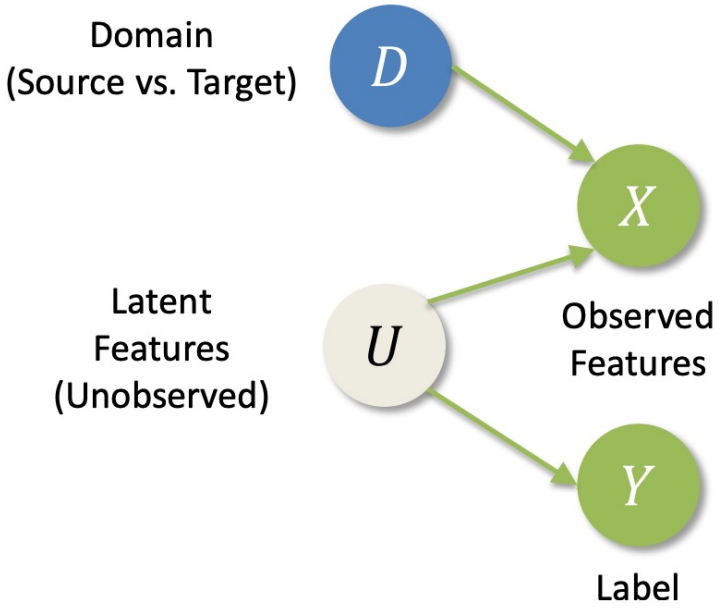
Why might this be true? One rationale: $P(X \mid Y)$ encodes some "causal" mechanism



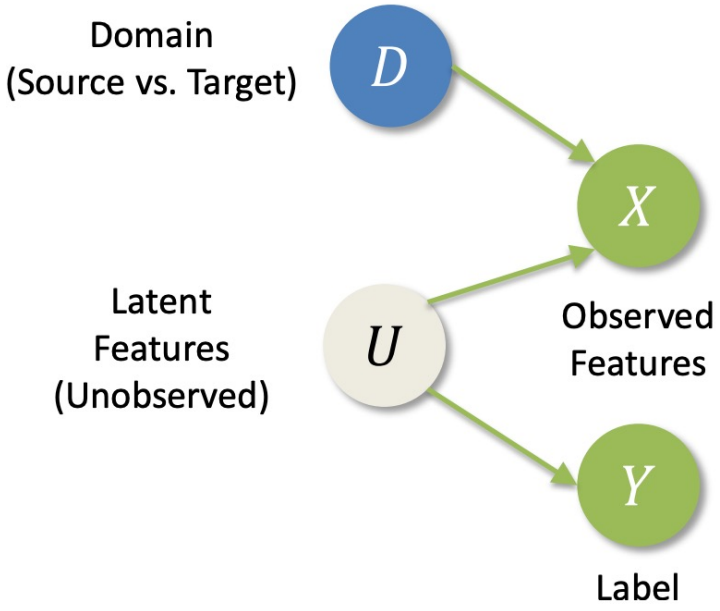
Example: Diagnostic testing under changes in disease prevalence.

Disease informative of symptoms. Prior vs. posterior probability.
Generative model vs. data-conditional posterior-probability inference

Example: “Domain Shift”



Example: “Domain Shift”



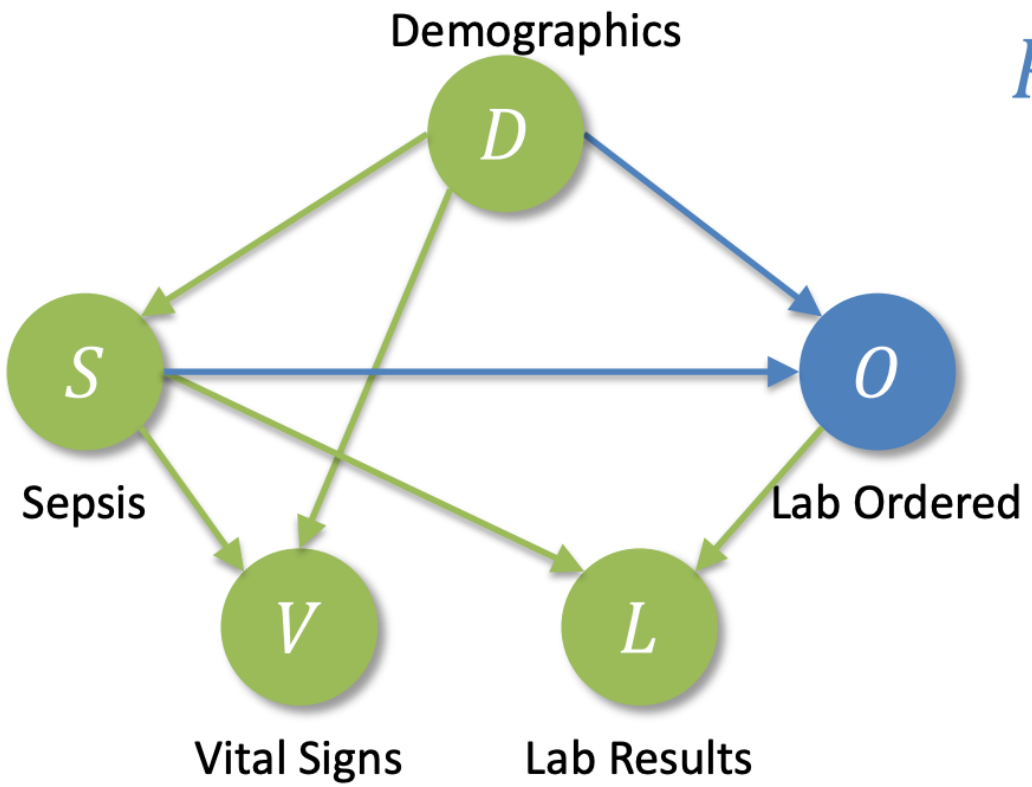
Example: Changes in how features are derived (e.g., ICD-9 versus ICD-10)

We can also view the domain itself as a variable that influences others

Note: So far, we have not discussed how to mitigate these shifts. In this example, more information is required!

Quinonero-Candela et al., (2008). *Dataset Shift in Machine Learning*, MIT Press.

Example: Using causal graphs to reason about shifts



$$P(O \mid D, S) \neq Q(\mathrm{O} \mid \mathrm{D}, \mathrm{S})$$

More fine-grained shifts can be reasoned about as changes in marginal/conditional distributions

Example: Changes in lab ordering patterns across hospitals

$$P(D, S, O, V, L) = P(D)P(S|D)P(V \mid D, S)P(O|D, S)P(L|O, S)$$

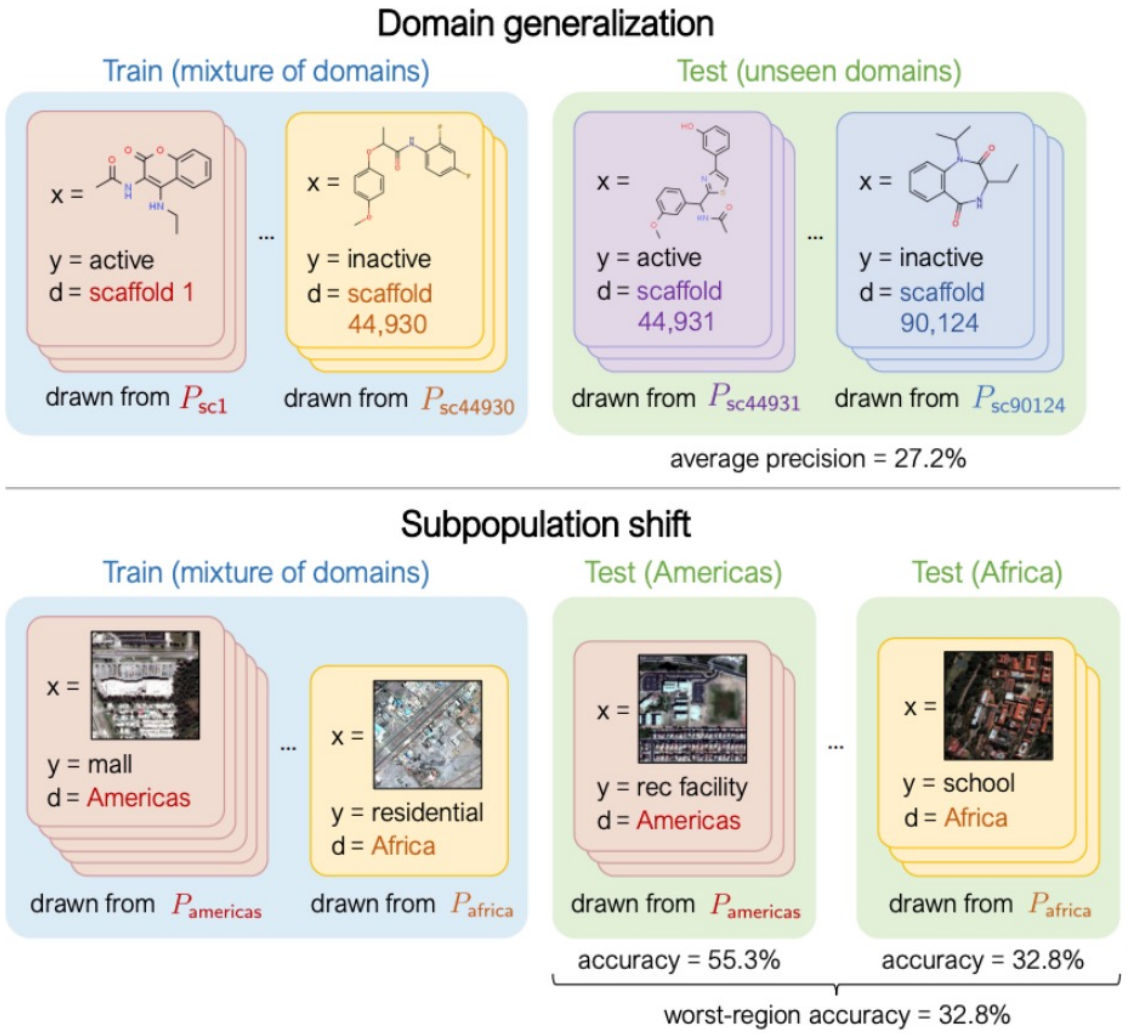
Example from Subbaswany et al. (2021). Evaluating Model Robustness and Stability to Dataset Shift. AISTATS

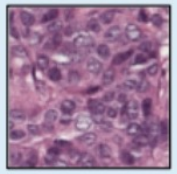
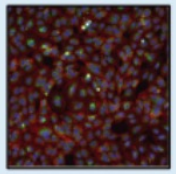
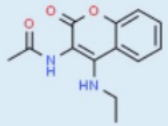
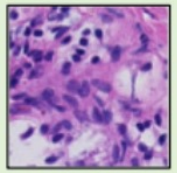
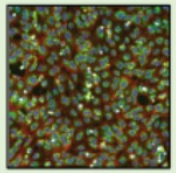
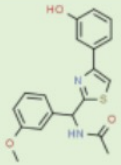
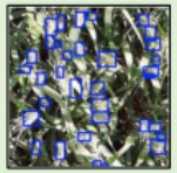
Distribution Shift Benchmarking

WILDS: A Benchmark of in-the-Wild Distribution Shifts

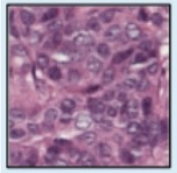
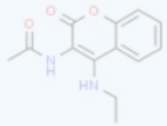
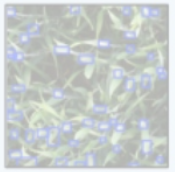
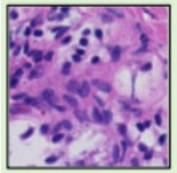
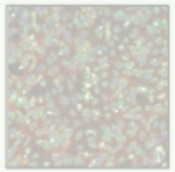
Pang Wei Koh* and Shiori Sagawa*	{pangwei, ssagawa}@cs.stanford.edu
Henrik Marklund	marklund@stanford.edu
Sang Michael Xie	xie@cs.stanford.edu
Marvin Zhang	marvin@eecs.berkeley.edu
Akshay Balsubramani	abalsubr@stanford.edu
Weihua Hu	weihuahu@stanford.edu
Michihiro Yasunaga	myasu@stanford.edu
Richard Lanas Phillips	richard@cs.cornell.edu
Irena Gao	igao@stanford.edu
Tony Lee	tonyhlee@stanford.edu
Etienne David	etienne.david@inrae.fr
Ian Stavness	stavness@usask.ca
Wei Guo	guowei@g.ecc.u-tokyo.ac.jp
Berton A. Earnshaw	berton.earnshaw@recursionpharma.com
Imran S. Haque	imran.haque@recursionpharma.com
Sara Beery	sbeery@caltech.edu
Jure Leskovec	jure@cs.stanford.edu
Anshul Kundaje	akundaje@stanford.edu
Emma Pierson	epierson@microsoft.com
Sergey Levine	svlevine@eecs.berkeley.edu
Chelsea Finn	cbfinn@cs.stanford.edu
Percy Liang	pliang@cs.stanford.edu

Correspondence to: wilds@cs.stanford.edu



	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	**iWildCam**	**Camelyon17**	**RxRx1**	**OGB-MolPCBA**	**GlobalWheat**	**CivilComments**	**FMoW**	**PovertyMap**	**Amazon**	**Py150**
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000
Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	`import numpy as np … norm=np.___`
Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp p=sp.Popen() stdout=p.___`
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

[Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421, 2021.]

		Domain generalization				Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000
Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	import numpy as np … norm=np.___
Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	import subprocess as sp p=sp.Popen() stdout=p.___
Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

[Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421, 2021.]

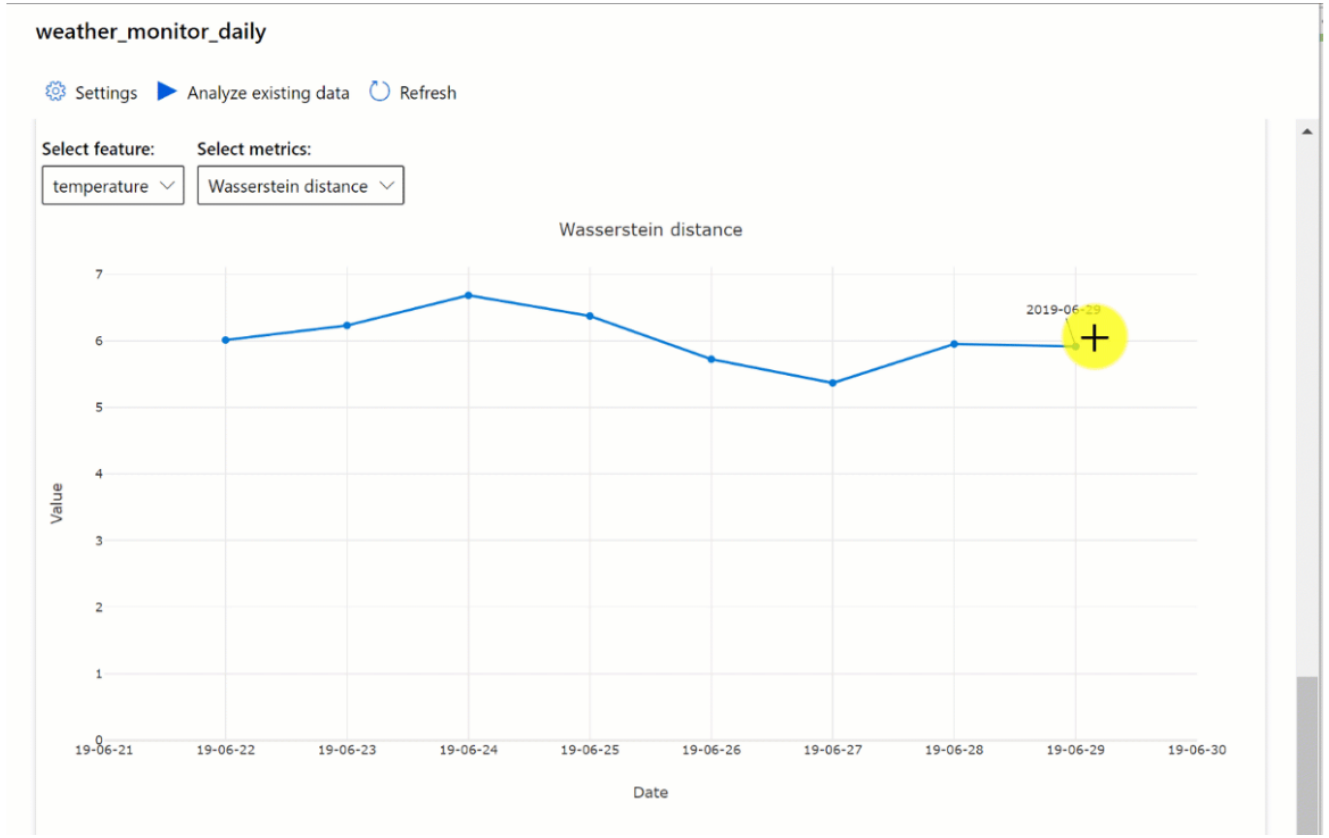
TL;DR: Existing algorithms don't substantially improve over Empirical Risk Minimization (ERM)

where ERM = estimate risk empirically on training data, bc we don't know all possible distributions of datasets

Table 2: The out-of-distribution test performance of models trained with different baseline algorithms: CORAL, originally designed for unsupervised domain adaptation; IRM, for domain generalization; and Group DRO, for subpopulation shifts. Evaluation metrics for each dataset are the same as in Table 1; higher is better. Overall, these algorithms did not improve over empirical risk minimization (ERM), and sometimes made performance significantly worse, except on CivilComments-wilds where they perform better but still do not close the in-distribution gap in Table 1. For GlobalWheat-wilds, we omit CORAL and IRM as those methods do not port straightforwardly to detection settings; its ERM number also differs from Table 1 as its ID comparison required a slight change to the OOD test set. Parentheses show standard deviation across 3+ replicates.

Dataset	Setting	ERM	CORAL	IRM	Group DRO
iWildCam2020-wilds	Domain gen.	31.0 (1.3)	**32.8 (0.1)**	15.1 (4.9)	23.9 (2.1)
Camelyon17-wilds	Domain gen.	**70.3 (6.4)**	59.5 (7.7)	64.2 (8.1)	68.4 (7.3)
RxRx1-wilds	Domain gen.	**29.9 (0.4)**	28.4 (0.3)	8.2 (1.1)	23.0 (0.3)
OGB-MolPCBA	Domain gen.	**27.2 (0.3)**	17.9 (0.5)	15.6 (0.3)	22.4 (0.6)
GlobalWheat-wilds	Domain gen.	**51.2 (1.8)**	—	—	47.9 (2.0)
CivilComments-wilds	Subpop. shift	56.0 (3.6)	65.6 (1.3)	66.3 (2.1)	**70.0 (2.0)**
FMoW-wilds	Hybrid	**32.3 (1.3)**	31.7 (1.2)	30.0 (1.4)	30.8 (0.8)
PovertyMap-wilds	Hybrid	**0.45 (0.06)**	0.44 (0.06)	0.43 (0.07)	0.39 (0.06)
Amazon-wilds	Hybrid	**53.8 (0.8)**	52.9 (0.8)	52.4 (0.8)	53.3 (0.0)
Py150-wilds	Hybrid	**67.9 (0.1)**	65.9 (0.1)	64.3 (0.2)	65.9 (0.1)

Current state of industry on dataset shift



Source: https://docs.microsoft.com/en-us/azure/machine-learning/how-to-monitor-datasets
See also: https://cloud.google.com/solutions/machine-learning/ml-modeling-monitoring-identifyingtraining-server-skew-with-novelty-detection & https://docs.seldon.io/projects/alibi-detect/en/latest/

OpenMIBOOD: Open Medical Imaging Benchmarks for Out-Of-Distribution Detection

Max Gutbrod[1,2], David Rauber[1], Danilo Weber Nunes[1,2], Christoph Palm[1,2]
[1]Regensburg Medical Image Computing (ReMIC), OTH Regensburg, Regensburg, 93053, Germany
[2]Regensburg Center of Health Sciences and Technology (RCHST), OTH Regensburg, Regensburg, 93053, Germany
{max.gutbrod,christoph.palm}@oth-regensburg.de

Abstract

The growing reliance on Artificial Intelligence (AI) in critical domains such as healthcare demands robust mechanisms to ensure the trustworthiness of these systems, especially when faced with unexpected or anomalous inputs. This paper introduces the Open Medical Imaging Benchmarks for Out-Of-Distribution Detection (OpenMIBOOD), a comprehensive framework for evaluating out-of-distribution (OOD) detection methods specifically in medical imaging contexts. OpenMIBOOD includes three benchmarks from diverse medical domains, encompassing 14 datasets divided into covariate-shifted in-distribution, near-

deployment. However, this assumption overlooks the possibility of encountering data from unknown and unseen distributions, known as out-of-distribution (OOD) data. When confronted with such data, AI models often exhibit high confidence in their predictions, even when these predictions are entirely incorrect [18]. Such behavior can result in silent and potentially catastrophic failures, particularly in high-stakes domains like healthcare, where erroneous predictions could directly impact patient safety. To address this, OOD detection methods help distinguish ID from OOD inputs, allowing models to flag or discard unreliable predictions or refer them for human review. Since 2016, numerous OOD detection methods have emerged [72], but a unified, com-

OpenMIBOOD: Open Medical Imaging Benchmarks for Out-Of-Distribution Detection

Max Gutbrod[1,2], David Rauber[1], Danilo Weber Nunes[1,2], Christoph Palm[1,2]
[1]Regensburg Medical Image Computing (ReMIC), OTH Regensburg, Regensburg, 93053, Germany

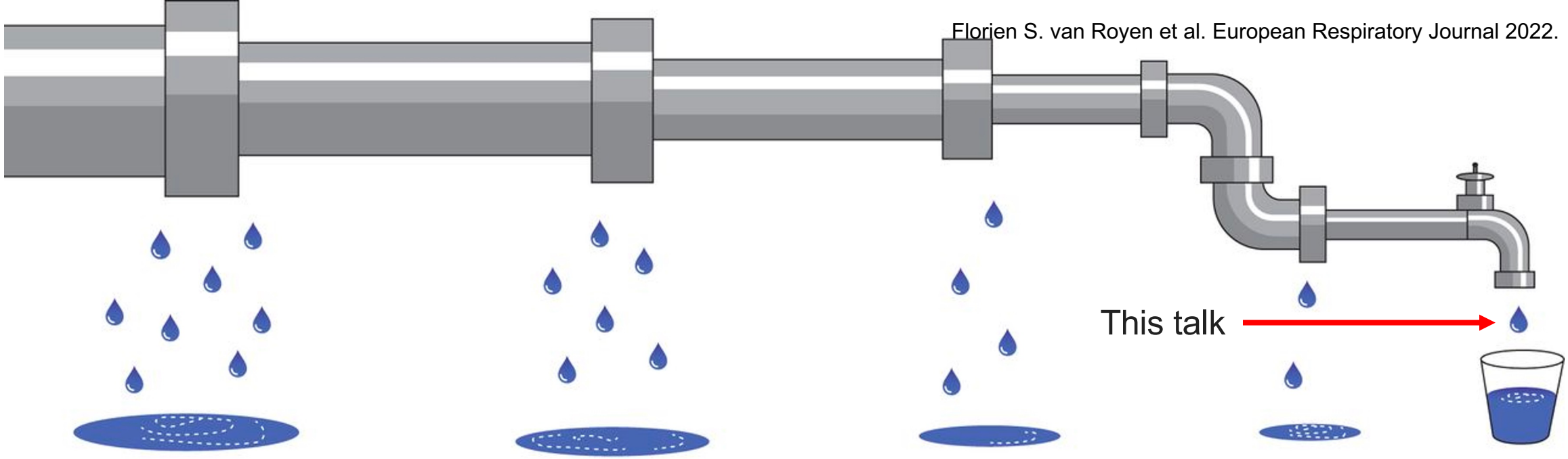
Encompassing 14 datasets divided into covariate-shifted in-distribution, near OOD, and far-OOD categories.

nisms to ensure the trustworthiness of these systems, especially when faced with unexpected or anomalous inputs. This paper introduces the Open Medical Imaging Benchmarks for Out-Of-Distribution Detection (OpenMIBOOD), a comprehensive framework for evaluating out-of-distribution (OOD) detection methods specifically in medical imaging contexts. OpenMIBOOD includes three benchmarks from diverse medical domains, encompassing 14 datasets divided into covariate-shifted in-distribution, near-

confidence in their predictions, even when these predictions are entirely incorrect [18]. Such behavior can result in silent and potentially catastrophic failures, particularly in high-stakes domains like healthcare, where erroneous predictions could directly impact patient safety. To address this, OOD detection methods help distinguish ID from OOD inputs, allowing models to flag or discard unreliable predictions or refer them for human review. Since 2016, numerous OOD detection methods have emerged [72], but a unified, com-

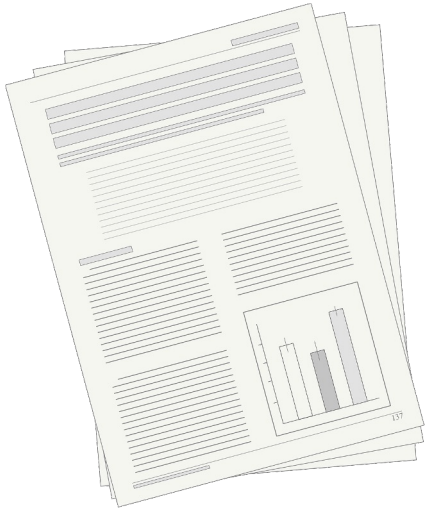
Even if model performance generalizes, the implementation can face challenges

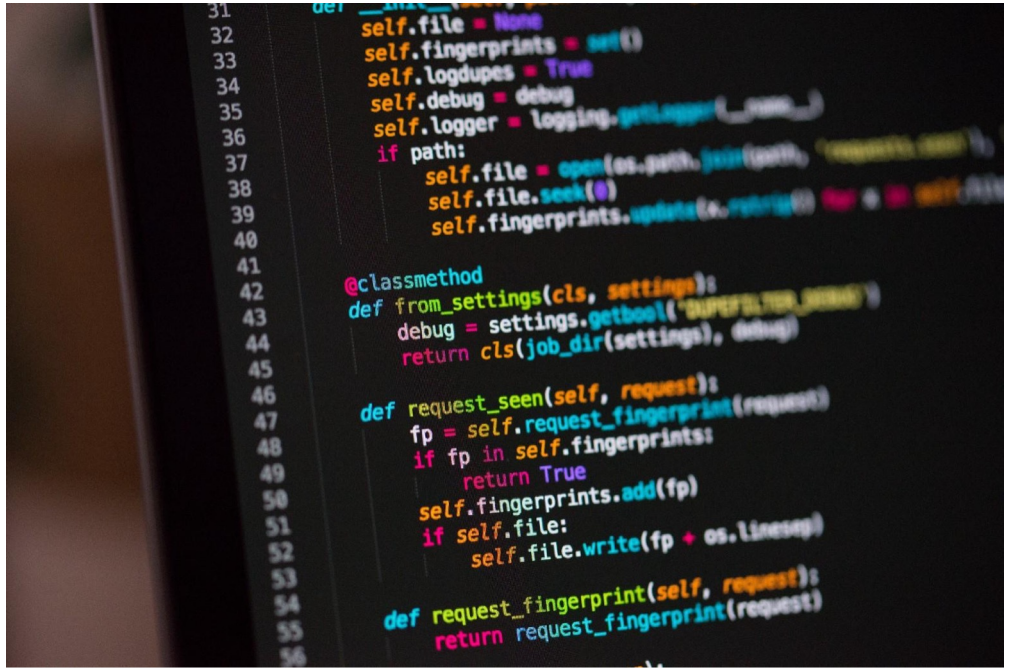
Florien S. van Royen et al. European Respiratory Journal 2022.

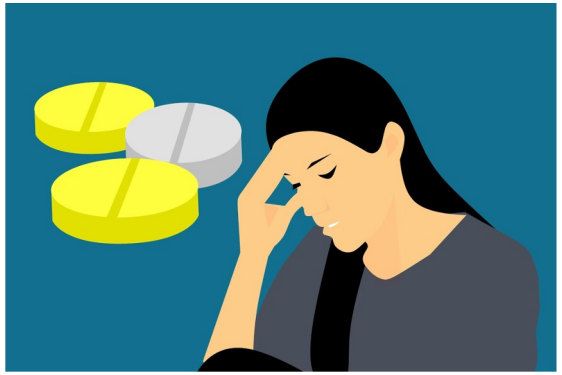


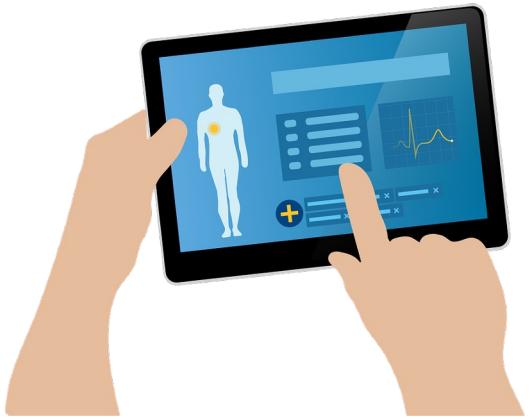
Not fit for purpose	No validation	No implementation	Not adopted
Developed on wrong patient population	Lack of data or incentive to pursue validation studies	No impact on decision making or patient (health) outcomes	Prediction (perceived as) not useful
Expensive or non-available predictors	Incompletely reported prediction model	No software developed to implement and use the model	Predictions not trusted
Time intensive to use model	Poorly developed or overfitted model	Requirements for adherence to (medical device) regulations	Model not transparent enough, or no tools available to enhance its use in practice
Outcome measured unreliably	Proprietary model code	Cost(-effectiveness) of using proprietary model	Model (perceived as) outdated

How deployments and impact are depicted in the literature

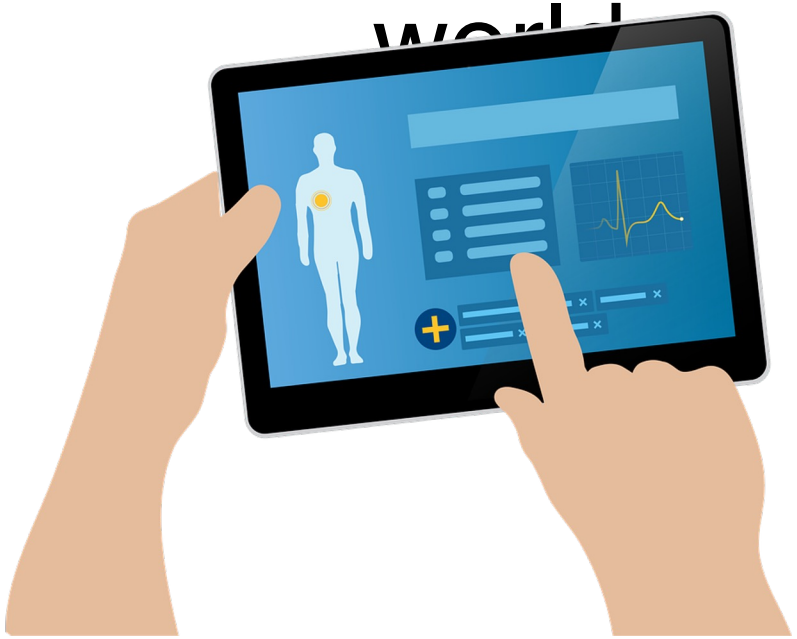


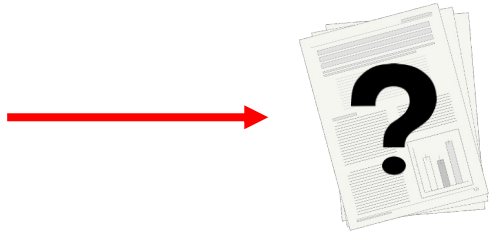


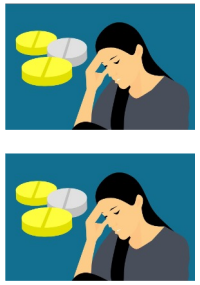


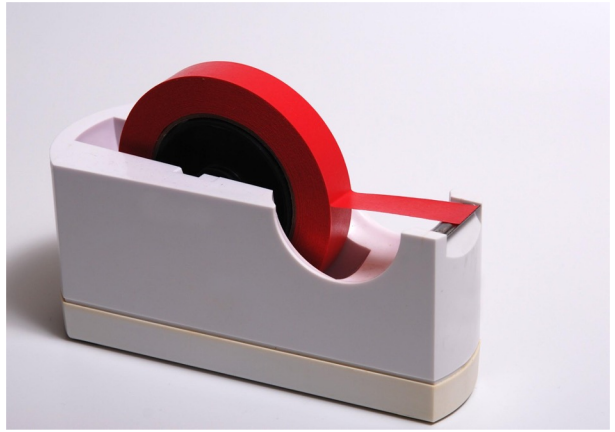


How deployments and impact happen in the real world









Who deploys AI in health?

Operations

Quality improvement initiatives

- May be randomized
- Usually no IRB approval

Led by CMO, CQO, or CMIO

Success? People stop complaining or quality measures improve.

Research

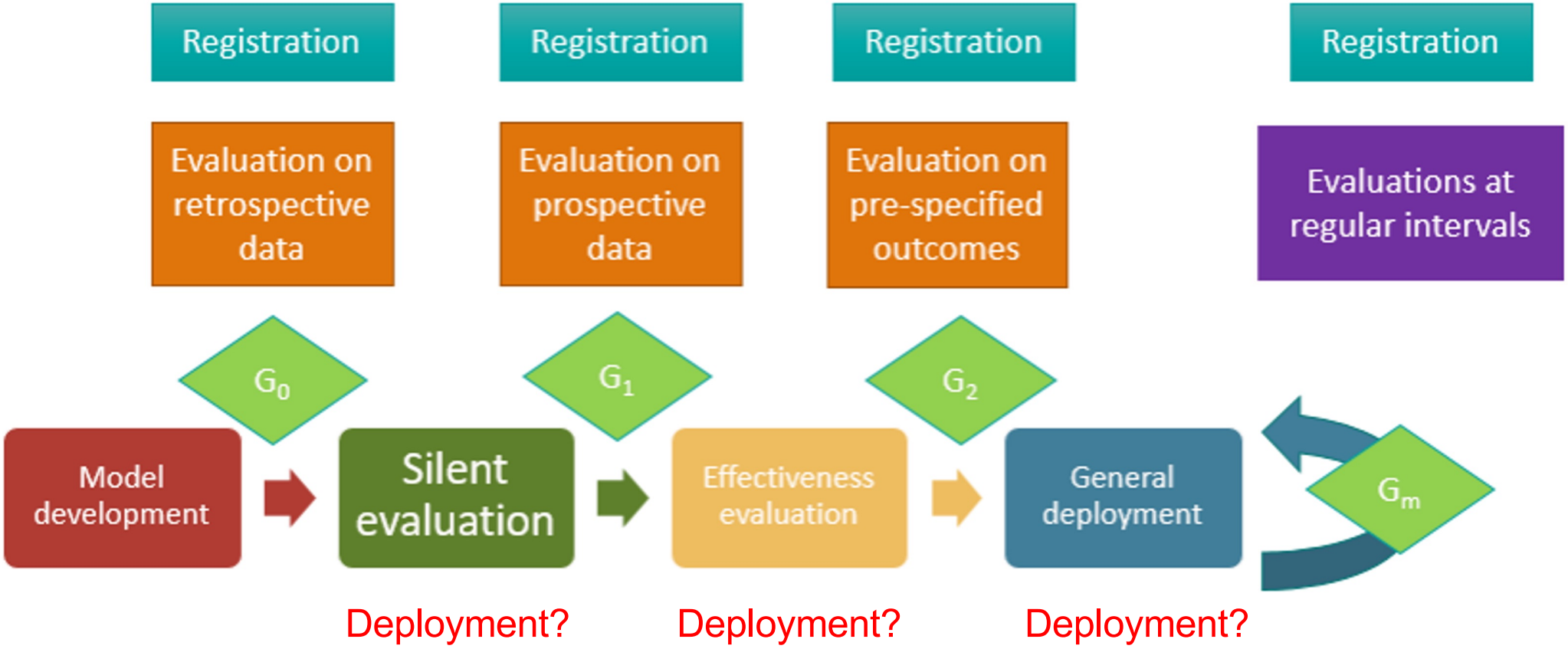
Clinical trials

- May be randomized
- IRB approval required

Led by individual researchers

Success? Generalizable knowledge gets published.

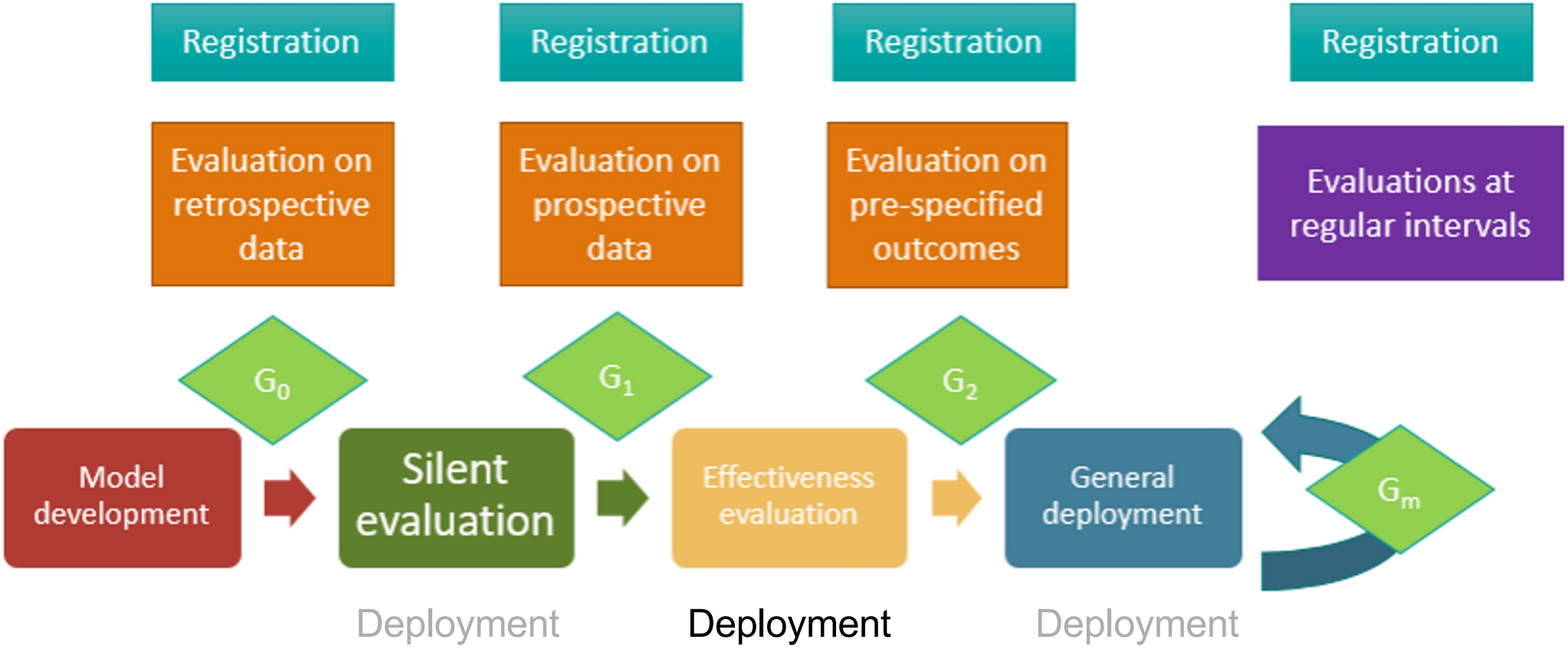
What is deployment?



Slide courtesy of Michael Pencina, PhD

Bedoya et al., JAMIA. 2022; 1-6, https://doi.org/10.1093/jamia/ocac078

What is deployment?



Slide courtesy of Michael Pencina, PhD

Bedoya et al., JAMIA. 2022; 1-6, https://doi.org/10.1093/jamia/ocac078

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-022-01894-0

Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis

Roy Adams[1,2], Katharine E. Henry[2,3], Anirudh Sridharan[4], Hossein Soleimani[5], Andong Zhan[2,3], Nishi Rawat[6], Lauren Johnson[7], David N. Hager[8], Sara E. Cosgrove[8], Andrew Markowski[9], Eili Y. Klein[10], Edward S. Chen[8], Mustapha O. Saheed[10], Maureen Henley[7], Sheila Miranda[11], Katrina Houston[7], Robert C. Linton[4], Anushree R. Ahluwalia[7], Albert W. Wu[6,8,12,13,14] ✉ and Suchi Saria[1,3,8,12,15] ✉

nature medicine

ARTICLES

https://doi.org/10.1038/s41591-022-01895-z

Check for updates

Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing

Katharine E. Henry[1,2,16], Roy Adams[2,3,16], Cassandra Parent[4,16], Hossein Soleimani[5], Anirudh Sridharan[6], Lauren Johnson[7], David N. Hager[8], Sara E. Cosgrove[8], Andrew Markowski[9], Eili Y. Klein[10], Edward S. Chen[8], Mustapha O. Saheed[10], Maureen Henley[7], Sheila Miranda[11], Katrina Houston[7], Robert C. Linton II[6], Anushree R. Ahluwalia[7], Albert W. Wu[8,12,13,14] ✉ and Suchi Saria[1,2,8,12,15] ✉

TREWS ≠ TREWScore

Science Translational Medicine

Current Issue First release papers

HOME > SCIENCE TRANSLATIONAL MEDICINE > VOL. 7, NO. 299 > A TARGETED REAL-TIME EARLY WARNING SCORE (TREWSCORE) FOR SEPTIC SHOCK

RESEARCH ARTICLE | SEPSIS

A targeted real-time early warning score (TREWScore) for septic shock

KATHARINE E. HENRY, DAVID N. HAGER, PETER J. PRONOVOST, AND SUCHI SARIA Authors Info & Affiliations

SCIENCE TRANSLATIONAL MEDICINE • 5 Aug 2015 • Vol 7, Issue 299 • p. 299ra122 • DOI: 10.1126/scitranslmed.aab3719

Adams *et al.* and Henry *et al.*





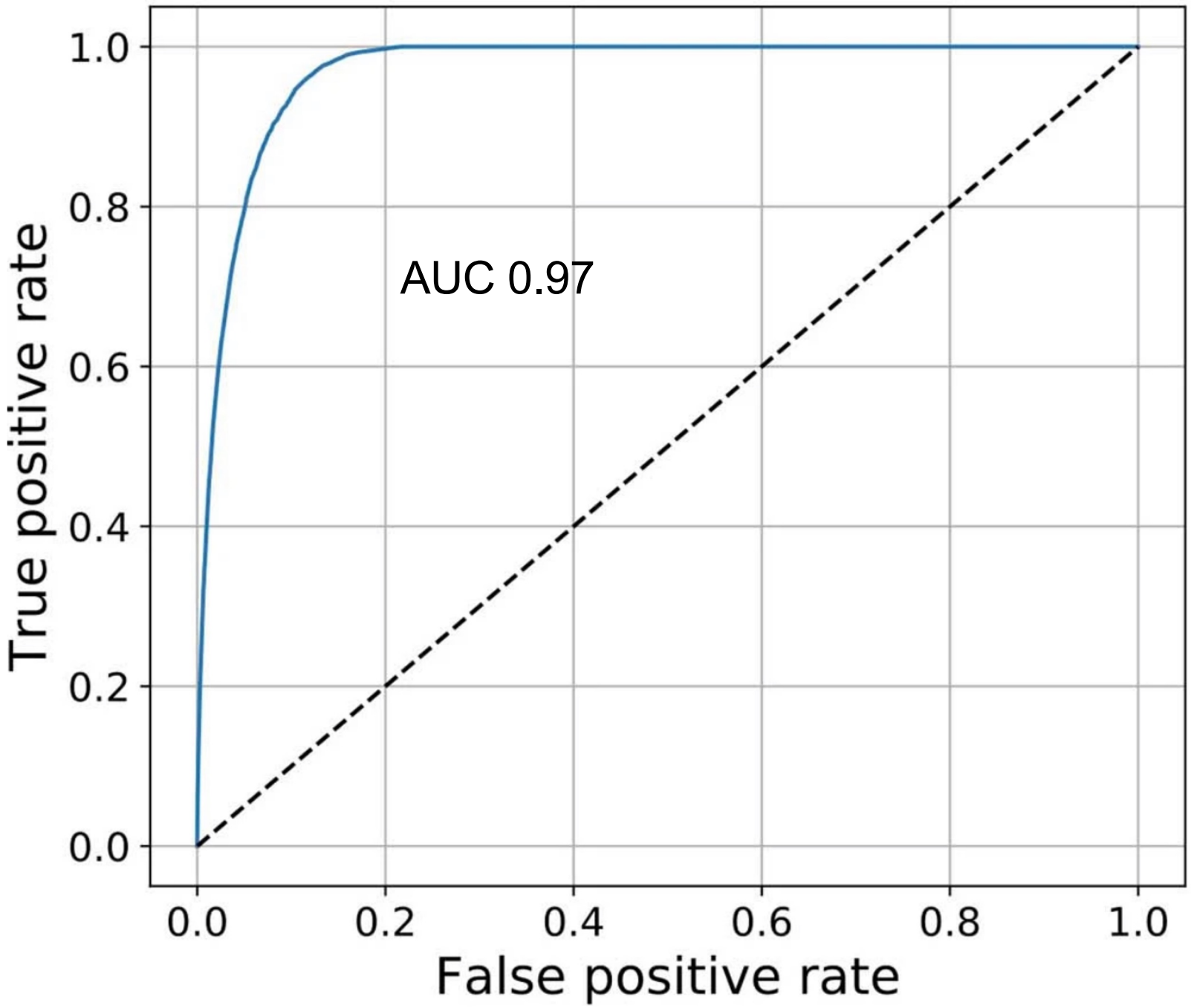
Five hospitals

Trained a model on patients who presented to the ED or were admitted to an inpatient unit to predict

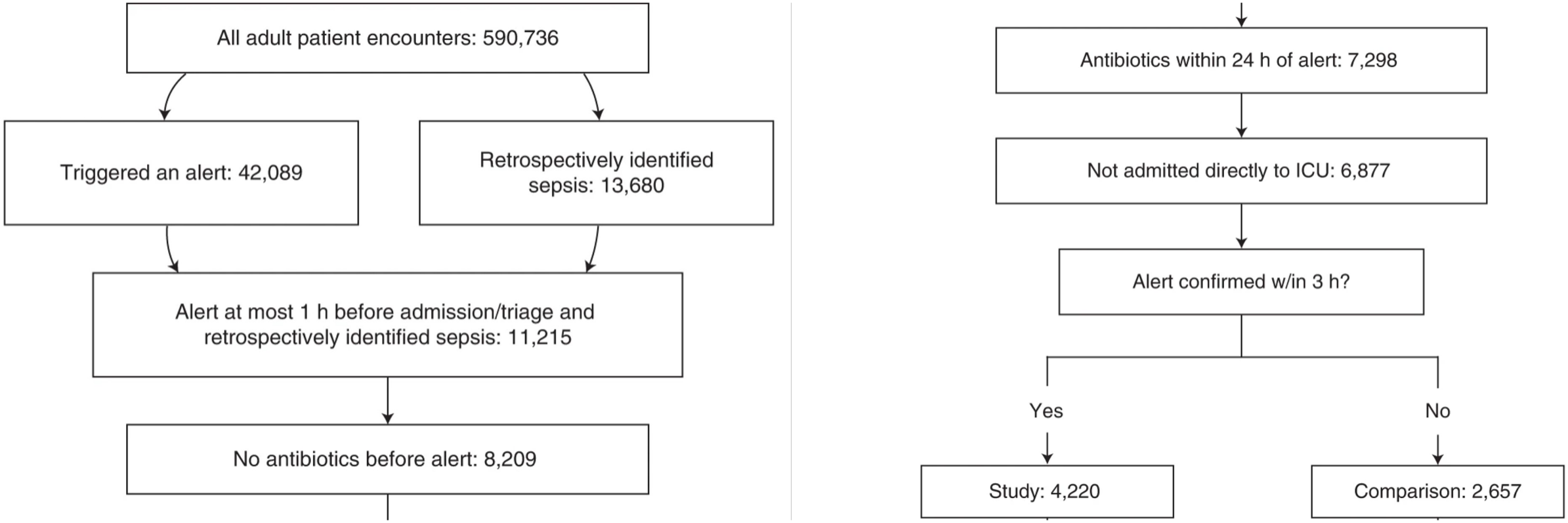
- Sepsis onset (TREWS model)

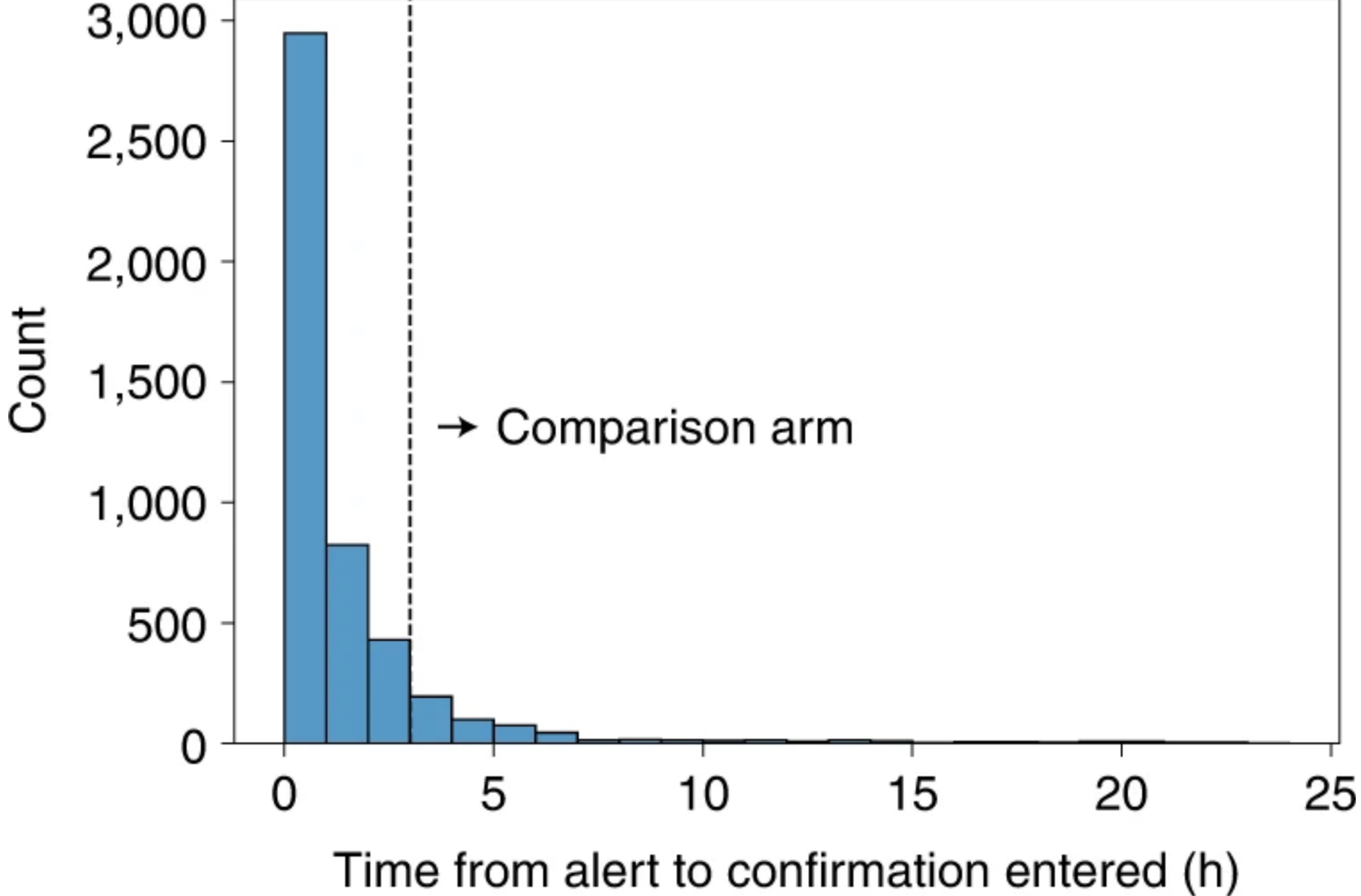
Evaluated the model

- Retrospectively
- Prospectively
  - Primary outcome: All-cause in-hospital mortality
  - Factors driving clinician adoption


AUC 0.97
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Adams *et al.*


All adult patient encounters: 590,736
Triggered an alert: 42,089
Retrospectively identified sepsis: 13,680
Alert at most 1 h before admission/triage and retrospectively identified sepsis: 11,215
No antibiotics before alert: 8,209
Antibiotics within 24 h of alert: 7,298
Not admitted directly to ICU: 6,877
Alert confirmed w/in 3 h?
Yes
No
Study: 4,220
Comparison: 2,657


3,000
2,500
2,000
1,500
1,000
500
0
Count
→ Comparison arm
0
5
10
15
20
25
Time from alert to confirmation entered (h)

Summary

TREWS Severe Sepsis met at 15:44 3/16/2018. More Detail

Please order missing bundle items under Step 3.

Nursing Assessment Expand

Severe Sepsis Evaluation Re-evaluate in 1 hr OFF Skip to Sepsis Bundle OFF

1 Please indicate whether infection is suspected No Infection Suspected Enter or Edit Infection Source

Unknown Source
by UNKNOWN at 21 secs ago

2 Below, we list likely sources of organ dysfunction that tri
believe are not due to infection.

Creatinine > 1.5 mg/dL
Criteria met on 11/22/2017 at 10:35:00 AM with a value of 3.5

Lactate > 2 mmol/L
Criteria met on 11/22/2017 at 10:34:00 AM with a value of 2.5

Bilirubin measurements not due to infection
Customized by UNKNOWN 1 min ago Re-enable

Nursing assessment questions (automatically expands in nurse view)

"More Detail" expands alert explanation to show factors behind the alert

Provider indicates whether the patient has a suspected source of infection

Provider confirms if there is evidence of organ dysfunction

Organ dysfunctions that are not attributed to sepsis are grayed out and remembered to prevent future false alerts based on the same criteria

Adams *et al.*

	Treatment	Comparison	*P* value[a]
All included	$n = 4{,}220$	$n = 2{,}657$	
In-hospital mortality, no. (rate)	617 (14.6%)	509 (19.2%)	<0.001
			<0.001
SOFA progression at 72 h[b]	-0.8 ± 2.7	-0.4 ± 2.9	0.001
Median length of stay (h)[c]	156 (99–260)	190 (118–323)	0.001

Environmental factors		
High alert level	True if the total number of TREWS alerts in the past 24 h in that unit exceeded the median for that unit and was greater than two alerts in the past 24 h	Providers may have alert fatigue if there have been a lot of alerts in the past day and be less likely to respond to new alerts
High admit volume	True if the total number of admissions in the past 3 h in that unit exceeded the median for that unit and the number of new admissions was greater than two	Providers are busier when there are many new admissions to the unit and may be less likely to respond to alerts in a timely way
Alert occurred 7:00–15:00	True if alert occured between 7:00 and 15:00	This corresponds to the morning/early afternoon hospital shift, which tends to have fewer new admissions in most units
Alert occurred 15:00–23:00	True if alert occured between 15:00 and 23:00	This corresponds to the late afternoon/ evening shift, which tends to have increased rates of new admissions and buildup of volume in the ED
Alert occurred 23:00–7:00	True if alert occured between 23:00 and 7:00	This corresponds to the overnight shift, which tends to have higher total patient volume in the ED from buildup through the day, sparser provider coverage and fewer new admissions
Provider factors		
ED provider	True if provider caring for the patient at the time of the alert was an ED provider	ED providers interact with patients earlier in their stay when there is more uncertainty and have a higher patient load per hour
Provider experience with alert	True if provider evaluated a previous alert within the past 30 d	Providers who are more familiar with the alert, may be more aware of the alert and be more likely to respond again

CCI, Charlson comorbidity index; CHF, congestive heart failure; CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disorder; SAPS, Simplified Acute Physiology Score; WBC, white blood cell.

Table 4 | Associations between patient, environmental and provider factors and provider evaluation of TREWS alerts

Factor (number of patients with that factor present out of 3,775 patients in the study population)	Unadjusted risk ratio (95% CI)	Adjusted risk ratio (95% CI)
Patient presentation factors		
Absence of key sepsis symptoms ($n=968$)	1.01 (0.98–1.04)	0.99 (0.96–1.03)
Alternative diagnosis ($n=2,114$)	0.99 (0.96–1.02)	1.00 (0.97–1.03)
Condition at risk for fluid overload ($n=1,926$)	1.02 (1.00–1.04)	1.01 (0.98–1.04)
Acute general severity ($n=1,887$)	0.98 (0.96–1.01)	0.97 (0.94–1.01)
Chronic complexity ($n=2,733$)	1.04 (1.00–1.08)	1.02 (0.97–1.08)
Advanced age ($n=1,810$)	**1.05 (1.02–1.10)**	**1.06 (1.03–1.10)**
Environmental factors		
High alert level ($n=1,749$)	**0.96 (0.93–0.99)**	**0.94 (0.91–0.96)**
High admit volume ($n=1,557$)	1.01 (0.98–1.05)	0.99 (0.96–1.03)
Alert occurred 7:00–15:00 ($n=1,310$)	**1.06 (1.04–1.09)**	**1.03 (1.01–1.06)**
Alert occurred 15:00–23:00 ($n=1,686$)	**0.94 (0.92–0.97)**	0.98 (0.95–1.00)
Alert occurred 23:00–7:00 ($n=779$)	1.00 (0.95–1.03)	1.01 (0.97–1.04)
Provider factors		
ED provider ($n=3,455$)	**1.35 (1.24–1.49)**	**1.22 (1.14–1.32)**
Provider experience with alert ($n=1,574$)	**1.25 (1.21–1.29)**	**1.22 (1.19–1.26)**

Associations in bold indicated confidence intervals that exclude zero.

Table 5 | Associations between patient, environmental and provider factors and provider dismissal of TREWS alerts

Factor (number of patients with that factor present out of 2,463 patients in the study population)	Unadjusted risk ratio (95% CI)	Adjusted risk ratio (95% CI)
Patient presentation factors		
Absence of key sepsis symptoms ($n=576$)	1.01 (0.86–1.19)	**1.28 (1.06–1.45)**
Alternative diagnosis ($n=1,409$)	**1.27 (1.14–1.42)**	1.11 (0.97–1.32)
Condition at risk for fluid overload ($n=1,286$)	1.10 (0.97–1.21)	1.08 (0.97–1.22)
Acute general severity ($n=1,271$)	**1.39 (1.23–1.56)**	**1.46 (1.28–1.66)**
Chronic complexity ($n=1,823$)	**0.87 (0.76–0.98)**	0.90 (0.75–1.05)
Advanced age ($n=1,232$)	**0.74 (0.65–0.81)**	**0.69 (0.60–0.75)**
Environmental factors		
High alert level ($n=1,113$)	0.91 (0.80–1.01)	1.01 (0.90–1.13)
High admit volume ($n=1,031$)	**0.83 (0.73–0.94)**	0.98 (0.86–1.12)
Alert occurred 7:00–15:00 ($n=885$)	**0.87 (0.74–0.99)**	1.12 (0.99–1.28)
Alert occurred 15:00–23:00 ($n=1,079$)	1.04 (0.92–1.16)	**1.20 (1.09–1.33)**
Alert occurred 23:00–7:00 ($n=499$)	**1.15 (1.03–1.29)**	**1.19 (1.07–1.36)**
Provider factors		
ED provider ($n=2,297$)	**0.39 (0.34–0.43)**	**0.47 (0.40–0.54)**
Provider experience with alert ($n=1,167$)	**0.58 (0.48–0.64)**	**0.66 (0.56–0.73)**

Associations in bold indicated confidence intervals that exclude zero.

Hinson *et al.*





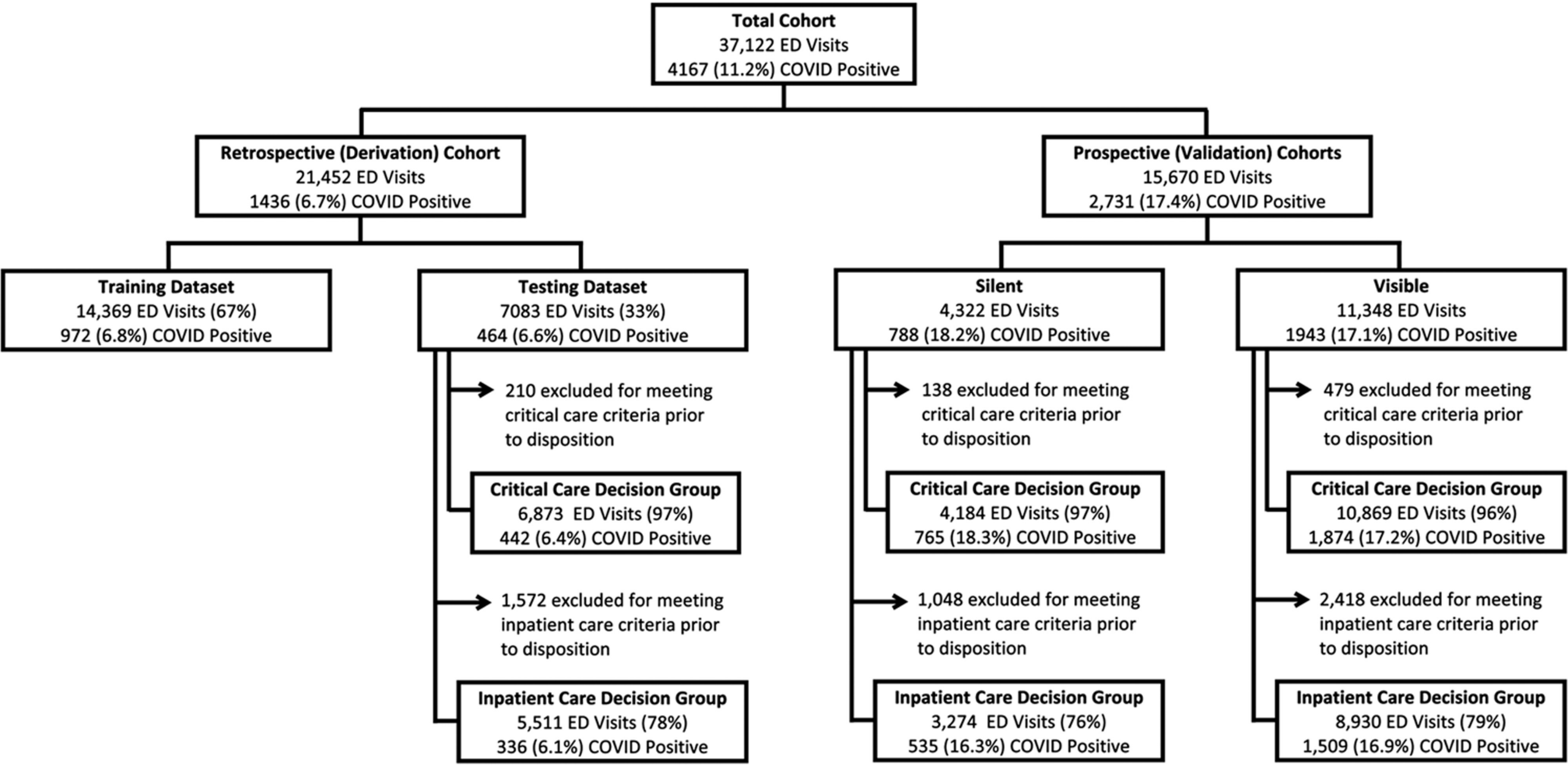
Five hospitals

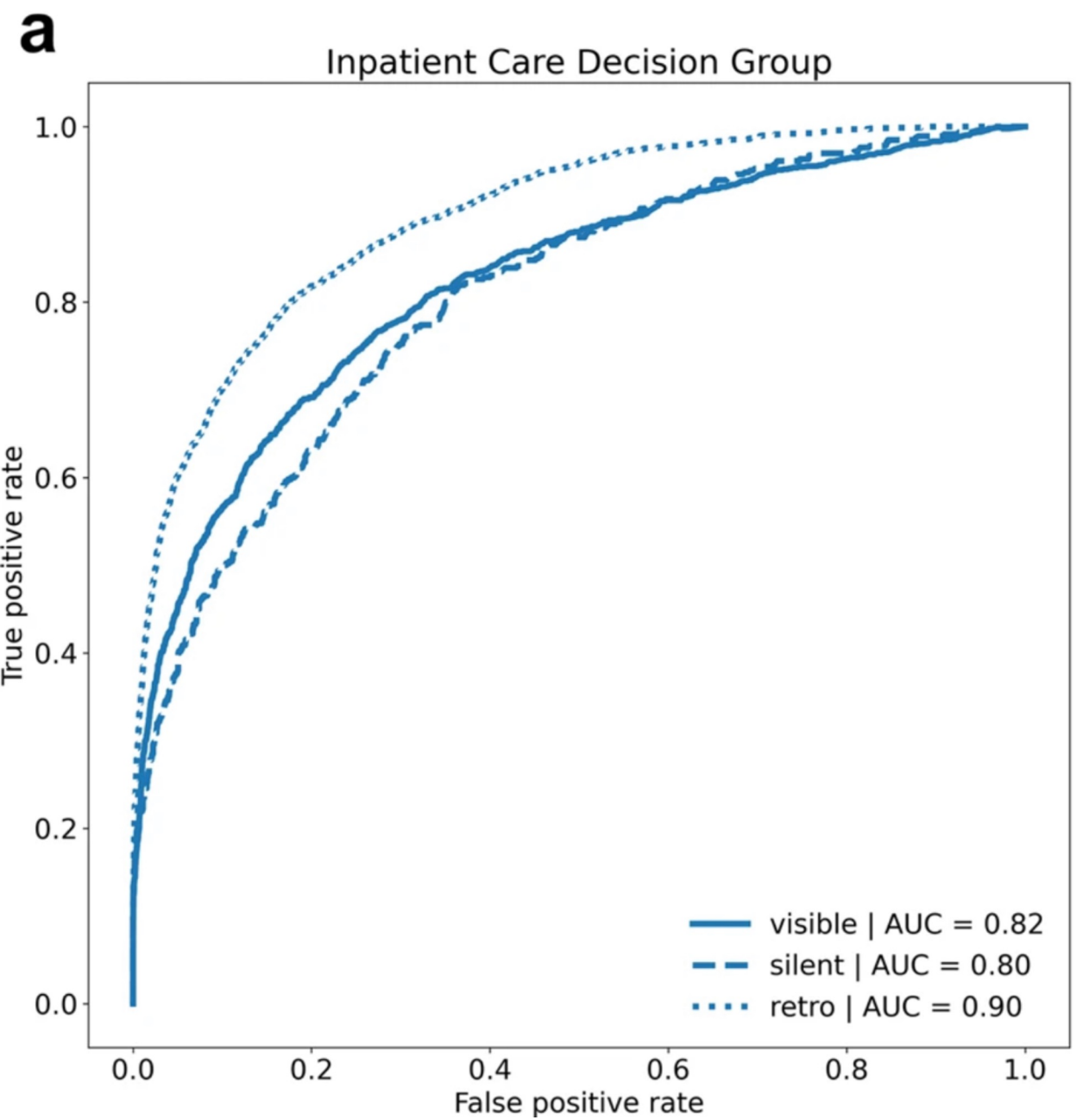
Trained a model on patients with suspected Covid-19 to support Emergency Dept triage by predicting

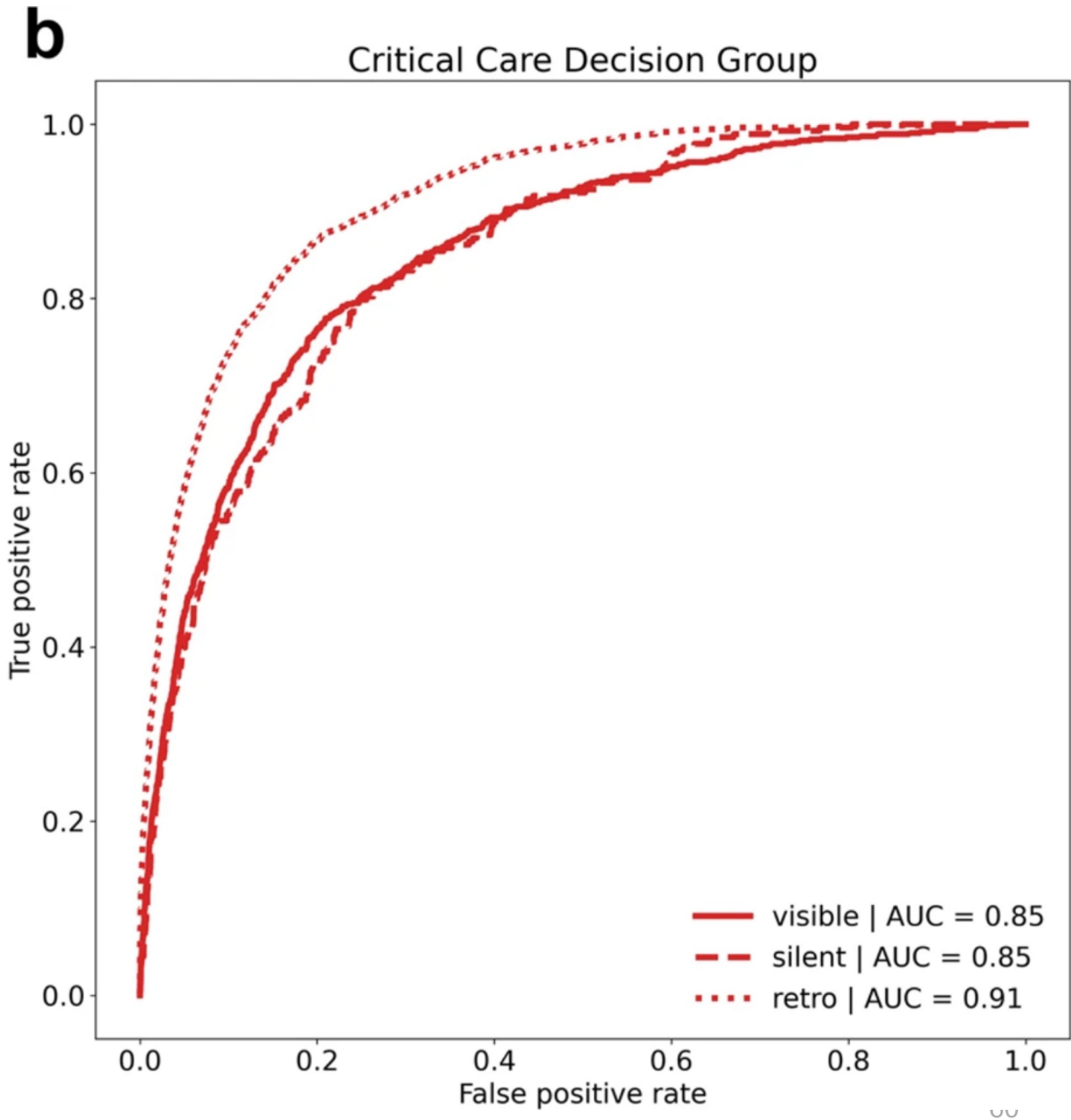
- Hospitalization within 72 hours
- ICU within 24 hours

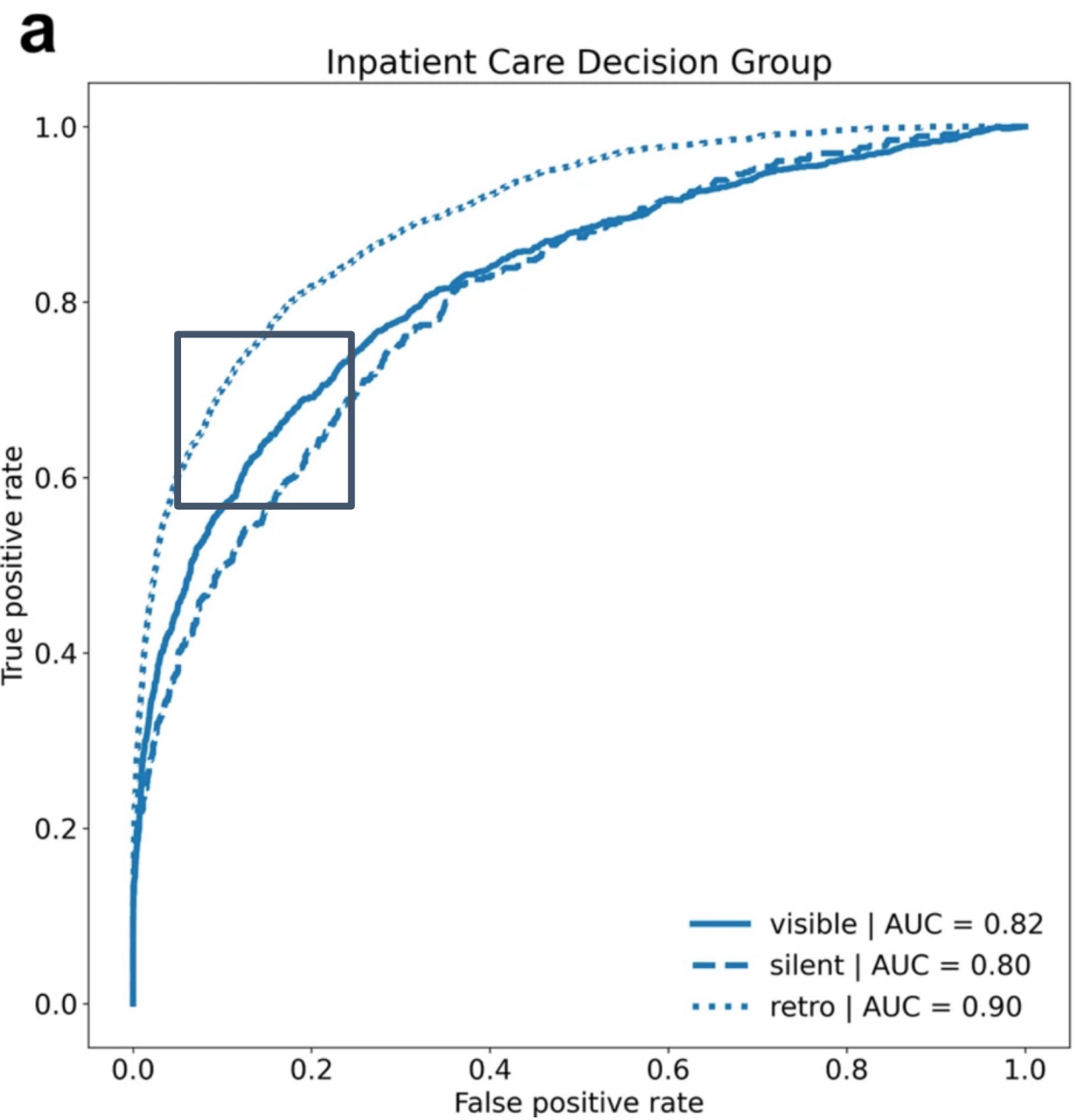
Evaluated the model

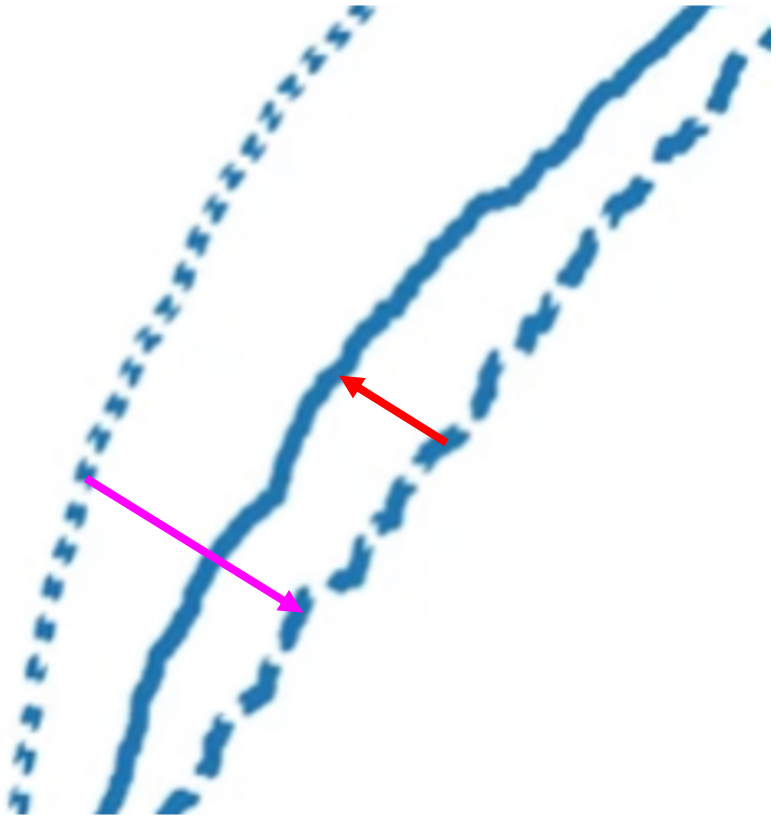
- Retrospectively
- Prospectively (silently)
- Prospectively (visible) linked to non-interruptive guidance


Total Cohort
37,122 ED Visits
4167 (11.2%) COVID Positive
Retrospective (Derivation) Cohort
21,452 ED Visits
1436 (6.7%) COVID Positive
Prospective (Validation) Cohorts
15,670 ED Visits
2,731 (17.4%) COVID Positive
Training Dataset
14,369 ED Visits (67%)
972 (6.8%) COVID Positive
Testing Dataset
7083 ED Visits (33%)
464 (6.6%) COVID Positive
210 excluded for meeting critical care criteria prior to disposition
Critical Care Decision Group
6,873 ED Visits (97%)
442 (6.4%) COVID Positive
1,572 excluded for meeting inpatient care criteria prior to disposition
Inpatient Care Decision Group
5,511 ED Visits (78%)
336 (6.1%) COVID Positive
Silent
4,322 ED Visits
788 (18.2%) COVID Positive
138 excluded for meeting critical care criteria prior to disposition
Critical Care Decision Group
4,184 ED Visits (97%)
765 (18.3%) COVID Positive
1,048 excluded for meeting inpatient care criteria prior to disposition
Inpatient Care Decision Group
3,274 ED Visits (76%)
535 (16.3%) COVID Positive
Visible
11,348 ED Visits
1943 (17.1%) COVID Positive
479 excluded for meeting critical care criteria prior to disposition
Critical Care Decision Group
10,869 ED Visits (96%)
1,874 (17.2%) COVID Positive
2,418 excluded for meeting inpatient care criteria prior to disposition
Inpatient Care Decision Group
8,930 ED Visits (79%)
1,509 (16.9%) COVID Positive


a
Inpatient Care Decision Group
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
visible | AUC = 0.82
silent | AUC = 0.80
retro | AUC = 0.90


b
Critical Care Decision Group
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
visible | AUC = 0.85
silent | AUC = 0.85
retro | AUC = 0.91





Why does the model perform (slightly) better after it becomes visible to clinicians?

Hinson *et al.*

Takeaways

- Multicenter retrospective validation
- Prospective validation and clinical effectiveness evaluation
- Improved mortality for high-risk patients
- Relatively quick turnaround time
  - 9 months from initial patient entering derivation cohort until deployment

Limitations

- Retrospective + prospective validation
- Pre-post design
  - Model deployment started December 2020
  - Covid-19 vaccines arrived December 2020
  - Which improved mortality?
- Weird modeling decisions
  - All continuous variables converted to discrete variables to handle missingness
  - Mostly Python's fault

The sklearn implementation of RandomForest does not handle missing values internally without clear instructions/added code. So while remedies (e.g. missing value imputation, etc.) are readily available within sklearn you DO have to deal with missing values before training the model. Apr 22, 2020

Main reasons

1. Misunderstanding of radiology job specifications
2. Benchmarks didn’t show true performance
3. Implementation blockers

Main reasons

1. Misunderstanding of radiology job specifications
2. Benchmarks didn’t show true performance
3. Implementation blockers

Q: How could we have known this earlier?

“In retrospect, he believes he spoke too broadly in 2016, he said in an email.

He didn’t make clear that he was speaking purely about image analysis, and was **wrong on timing but not the direction**, he added.”

- New York Times (2025)

Themes for the rest of the class

- AI and the Workforce
- Health Datasets
- Measurement and Evaluation
- AI Policy and Regulation
- Interpretability
- Real-world Impact and Ethics

More reading

- Lohr, "Your A.I. Radiologist Will Not Be With You Soon", New York Times, May 2025
- Mousa, "AI Isn't Replacing Radiologists", Works in Progress Blog, Sept 2025.
- Oakden-Rayner, "Medical AI Safety: Doing it Wrong", Personal blog, Jan 2019

Your A.I. Radiologist Will Not Be With You Soon

Experts predicted that artificial intelligence would steal radiology jobs. But at the Mayo Clinic, the technology has been more friend than foe.

AI isn't replacing radiologists

Radiology combines digital images, clear benchmarks, and repeatable tasks. But demand for human radiologists is at an all-time high.

WORKS IN PROGRESS AND DEENA MOUSA
SEP 25, 2025

212 27 37 Share

Medical AI Safety: Doing it wrong.

JANUARY 21, 2019 ~ LAURENOAKDENRAYNER

Last post, I said we have a safety problem in medical AI. I even suggested

“One day maybe we can cure all disease with the help of AI… Maybe within the next decade or so, I don't see why not.”

- Demis Hassabis (2025)

Summary

- ✓ **Course logistics** (5 mins)
- ✓ **Dataset shift** (25 mins)
- ✓ **Deployment challenges** (20 mins)

SCAN ME

How can we make Data 146 better for you?

Next Class: Electronic health records and where they come from