

BMI 702: Biomedical Artificial Intelligence

Foundations of Biomedical Informatics II, Spring 2024


Lecture 1: Course overview and introduction to
biomedical AI



HARVARD
MEDICAL SCHOOL

Marinka Zitnik
marinka@hms.harvard.edu

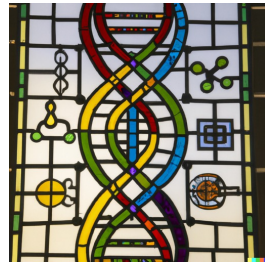
Outline for today's class

-  1. Overview of syllabus
2. What makes biomedical data unique
3. Motivation for machine learning
4. Roadmap for responsible AI

What will you learn in this course?

- **Key data modalities**

- Clinical data
- Networks, graphs, and multimodal datasets
- Language and text
- Images



- **Cutting-edge algorithmic principles underlying AI**

- Self-supervised learning and transfer learning
- Large-scale pre-training and efficient fine-tuning
- Multimodal learning
- Generative AI



- **Broader impacts:**

- Model evaluation, benchmarking, and deployment
- Privacy, safety, and copyright issues of AI



Course staff

- **Marinka Zitnik (Instructor)**
 - Assistant Professor of Biomedical Informatics
 - Associate Faculty at Kempner Institute
 - Associate Member at the Broad Institute
 - Faculty at Harvard Data Science
 - <https://zitniklab.hms.harvard.edu>



Course staff

- **Yasha Ektefaie**
 - 4th year PhD student in BIG program
 - yasha_ektefaie@g.harvard.edu

- **Varun Ullanat**
 - 2023 BMI graduate
 - vullanat@hms.harvard.edu



Dates, times and format

<https://zitniklab.hms.harvard.edu/BMI702>

- **Thursday, 2:00 PM – 4:00 PM ET**
 - No class or assignments due: Week of March 11
- **Location:** Countway Library, Classroom 403
- **Office hours:**
 - Tue, 3-4pm, Countway 423/424
 - Thu, 4-5 pm, Countway 309
 - Fri, 11-12pm, Countway 423/424

Course syllabus

- 14 lectures:
 - Introduction to biomedical AI
 - Lectures are divided into six modules
 - The first lecture in each module introduces ML concepts in the area
 - The following lecture introduces advanced topics in the same area
 - Final lecture on broader considerations of biomedical AI
- **Modules:**
 - **Module 1:** Clinical AI
 - **Module 2:** Trustworthy and Efficient AI
 - **Module 3:** Graph Learning
 - **Module 4:** Language Modeling
 - **Module 5:** Biomedical imaging
 - **Module 6:** Generative AI

Assignments

- **Problem sets:**
 - 3 problem sets
 - Primary form of support are office hours we will host
 - Problem sets must be completed individually
- **Pre-class quizzes:**
 - Open at 9am on Friday, due at 2pm on Thursday
 - Based on the **Required Reading** section of each lecture
- **Quick checks:**
 - Short questions embedded into lectures
 - Check your understanding of the concepts just introduced
 - Your score on them doesn't matter, you must complete them

Grading

Grade Components

Component	Percent of grade (%)
Problem Set 1	20
Problem Set 2	20
Problem Set 3	20
Class Participation	14 (1 point for Lecture 1-14)
Pre-Class Quizzes	26 (2 points per quiz; there is no quiz for Lecture 1)

We Want You to Succeed!

You are more than welcome to visit our office hours and talk with us. We know graduate school can be stressful and we want to help you succeed

Course culture and attendance

■ Course culture:

- Students taking this course come from a wide range of backgrounds
- We hope to foster an inclusive and safe learning environment based on curiosity and research inquiry
- All members of the course community are expected to treat each other with courtesy and respect

■ Attendance:

- The course will be run in a in-person format
- Students must attend all classes unless they have explicit permission from the course instructor

Policies

■ Collaboration policy

- Unless otherwise specified, all work submitted must reflect student's own effort and understanding
- Clearly distinguish your own ideas and knowledge from information derived from other sources:
 - Students must properly cite all submitted work
 - Unless noted otherwise, students are expected to complete assignments, quizzes, and projects individually, not as teams
 - Discussion about course content and materials is acceptable, but sharing solutions is not acceptable

■ Late policy

- Extensions provided in the case of exceptional circumstances
- Email the course instructor to request an extension

Policies

- **We support using LLMs and generative AI:**
 - **Responsibility for content:** Students who use LLMs and generative AI tools in their assignments take full responsibility for the content they submit
 - **Acknowledgment of AI use:** Clearly acknowledge any use of LLMs, specifying the nature and extent of assistance received from AI. Make sure to perform critical thinking, analysis, and synthesis of information
 - **Ethical use and originality:** Use these tools ethically, following the principles of academic honesty. Using AI to plagiarize, misrepresent original work, or fabricate data is prohibited
 - **Instructor discretion:** We may specify assignments where LLMs and generative AI use is encouraged or prohibited

Outline for today's class



1. Overview of syllabus

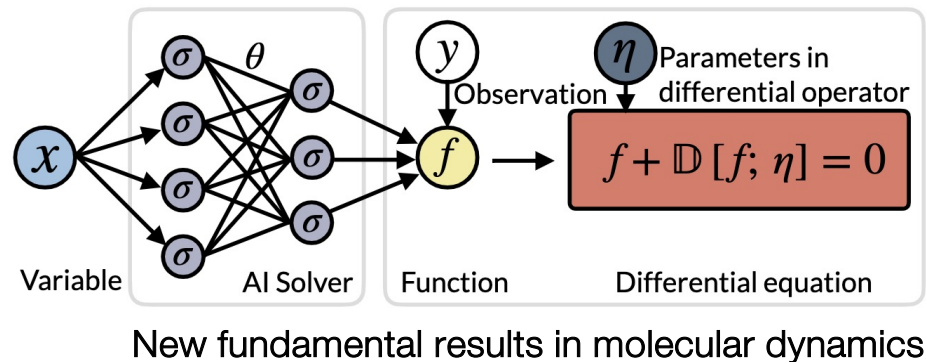
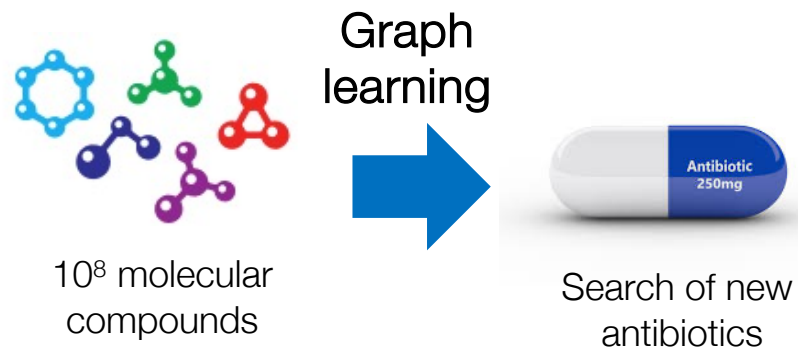
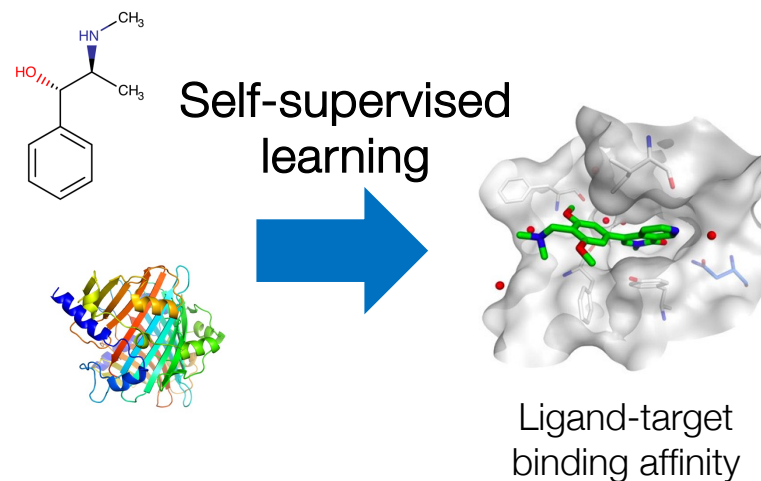
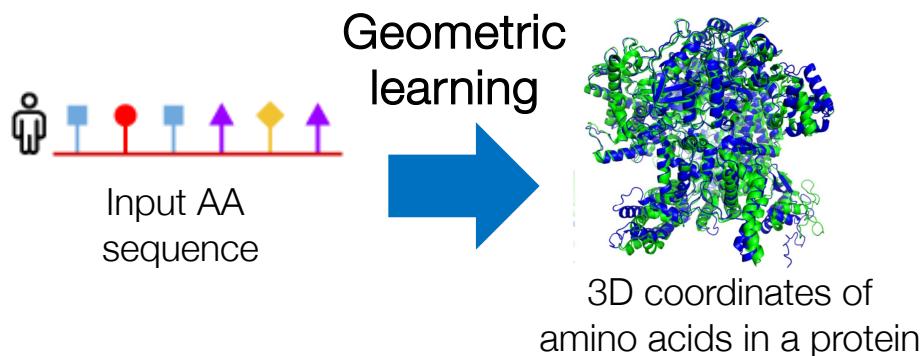


2. What makes biomedical data unique

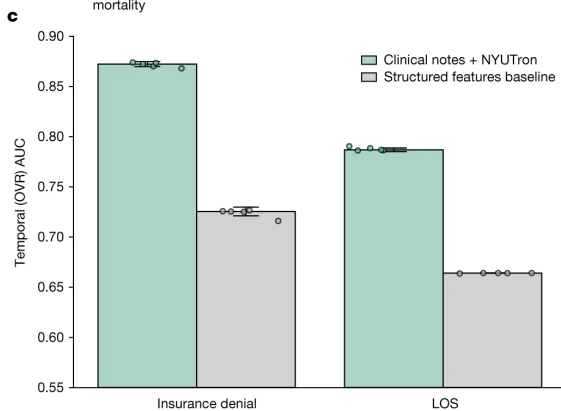
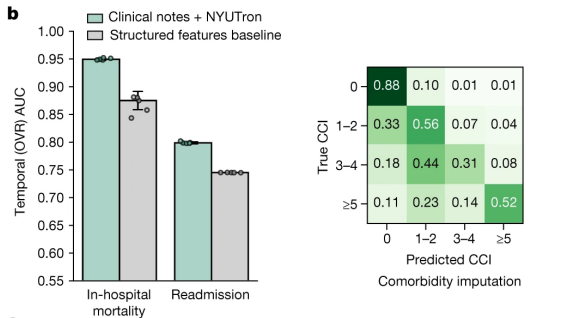
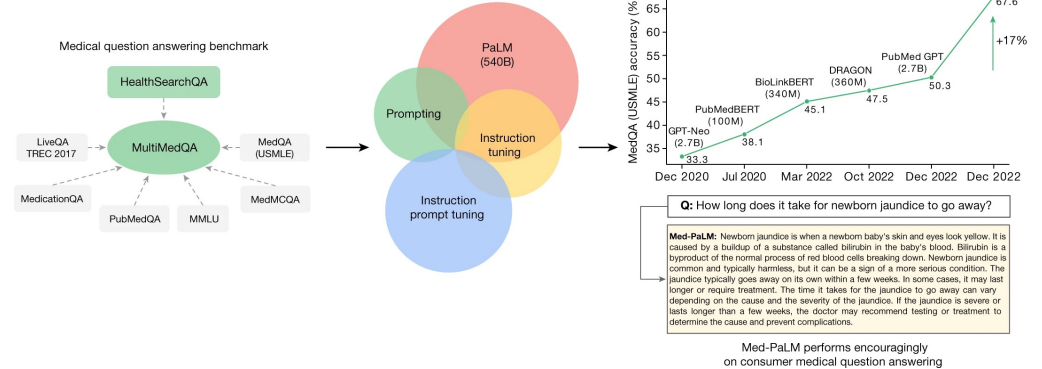
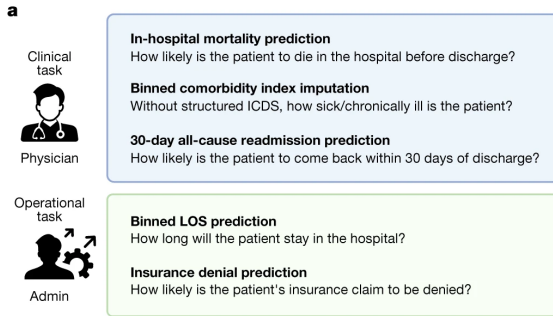
3. Motivation for machine learning

4. Roadmap for responsible AI

AI in medicine



AI in healthcare

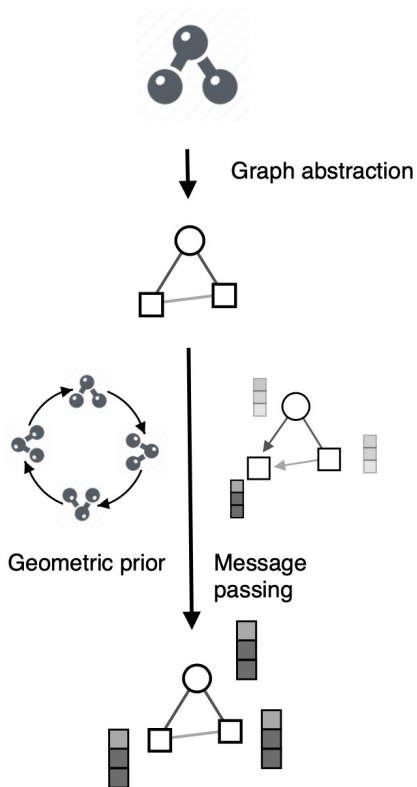


Health system-scale language models are all-purpose prediction engines, *Nature* 2023

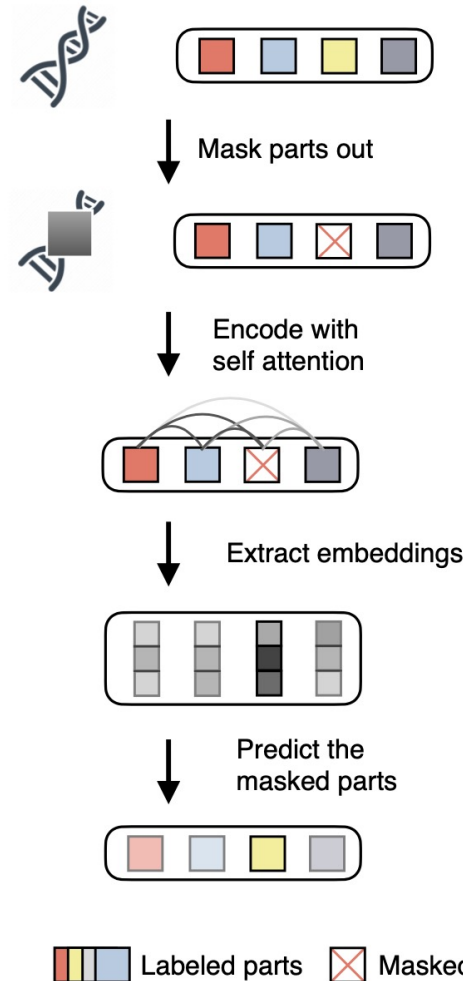
Large language models encode clinical knowledge, *Nature* 2023

Key algorithmic advances

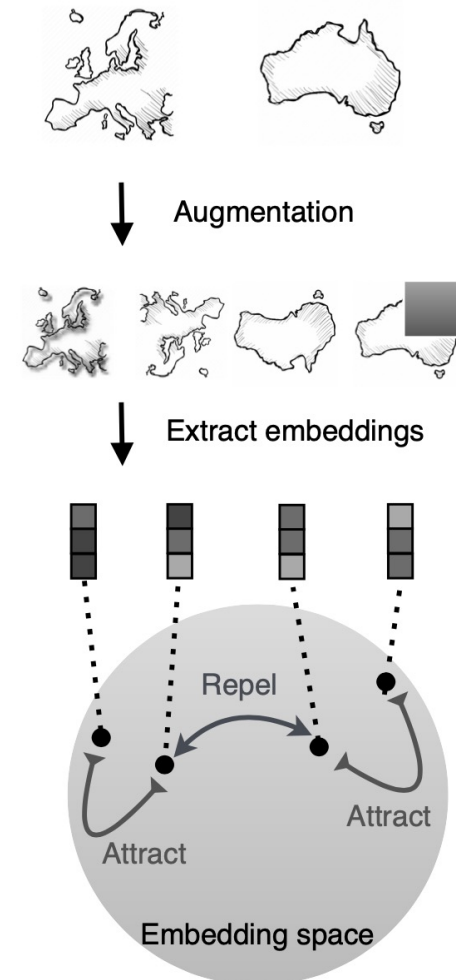
Geometric learning



Self-supervised learning



Generative AI



What makes biomedical data so different?

- Life or death decisions
 - Need **robust** algorithms
 - Checks and balances built into ML deployment
 - (Also arises in other applications of AI such as autonomous driving)
 - Need **fair** and **accountable** algorithms
- Many questions are about **unsupervised learning**
 - Discovering disease subtypes, or answering question such as “characterize the types of people that are highly likely to be readmitted to the hospital”?
- Many of the questions we want to answer are **causal**
 - Naïve use of supervised machine learning is insufficient

What makes biomedical data so different?

- ML models are increasingly deployed in real-world applications and implemented in clinical settings:
 - It is critical to ensure that these models are behaving responsibly and are trustworthy
- Accuracy alone is no longer enough
- Auxiliary criteria are important:
 - **Explainable predictions and interpretable models**
 - **Fair and non-discriminatory predictions**
 - **Privacy-preserving, causal, and robust predictions**
- This broad area is known as **trustworthy ML**



High-stakes decisions

What makes biomedical data so different?

- Very little labeled data
- Recent breakthroughs in AI depended on lots of labeled data!

Large, diverse data
(+ large models)



Broad generalization



Russakovsky et al. '14

GPT-2

Radford et al. '19

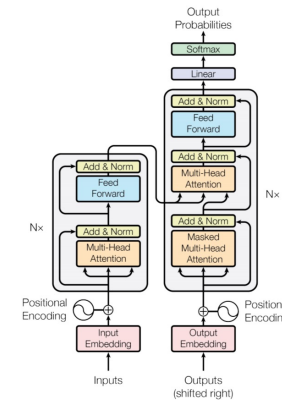


Figure 1: The Transformer - model architecture.

Vaswani et al. '18

What if you don't have a large dataset?

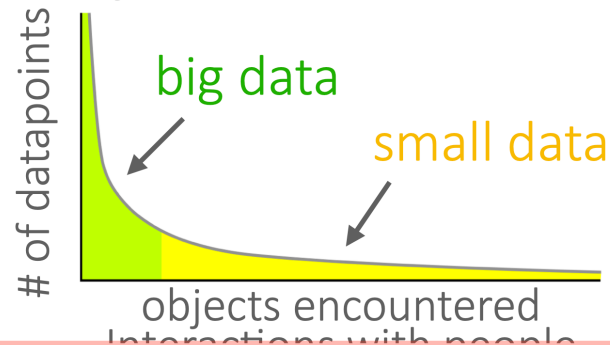
medical imaging robotics personalized education,
translation for rare languages recommendations

What if you want a general-purpose AI system in the real world?

Need to continuously adapt and learn on the job.

Learning each thing from scratch won't cut it.

What if your data has a long tail?



These settings break the supervised learning paradigm.

driving scenarios

What makes biomedical data so different?

- Very little labeled data
 - Motivates **semi-supervised and self-supervised learning**
- Sometimes small numbers of samples (e.g., a rare disease)
 - **Learn as much as possible from other data** (e.g., from healthy patients)
 - Model the problem **carefully**
- Lots of **missing data, varying time intervals, censored labels**


What makes biomedical data so different?

- Difficulty of **de-identifying** data:
 - Need for **data sharing agreements** and **sensitivity**
- Difficulty of **deploying ML**:
 - Commercial electronic health record software is **difficult to modify**
 - Data are often in siloes; everyone recognizes need for **interoperability**, but slow progress
 - **Rigorous testing and iteration** are needed
- Difficulty of **correcting for biases and inequities**:
 - Consideration of ethical and legal issues
 - Health data on which algorithms are trained are likely to be influenced by **many facets of social inequality**

Outline for today's class

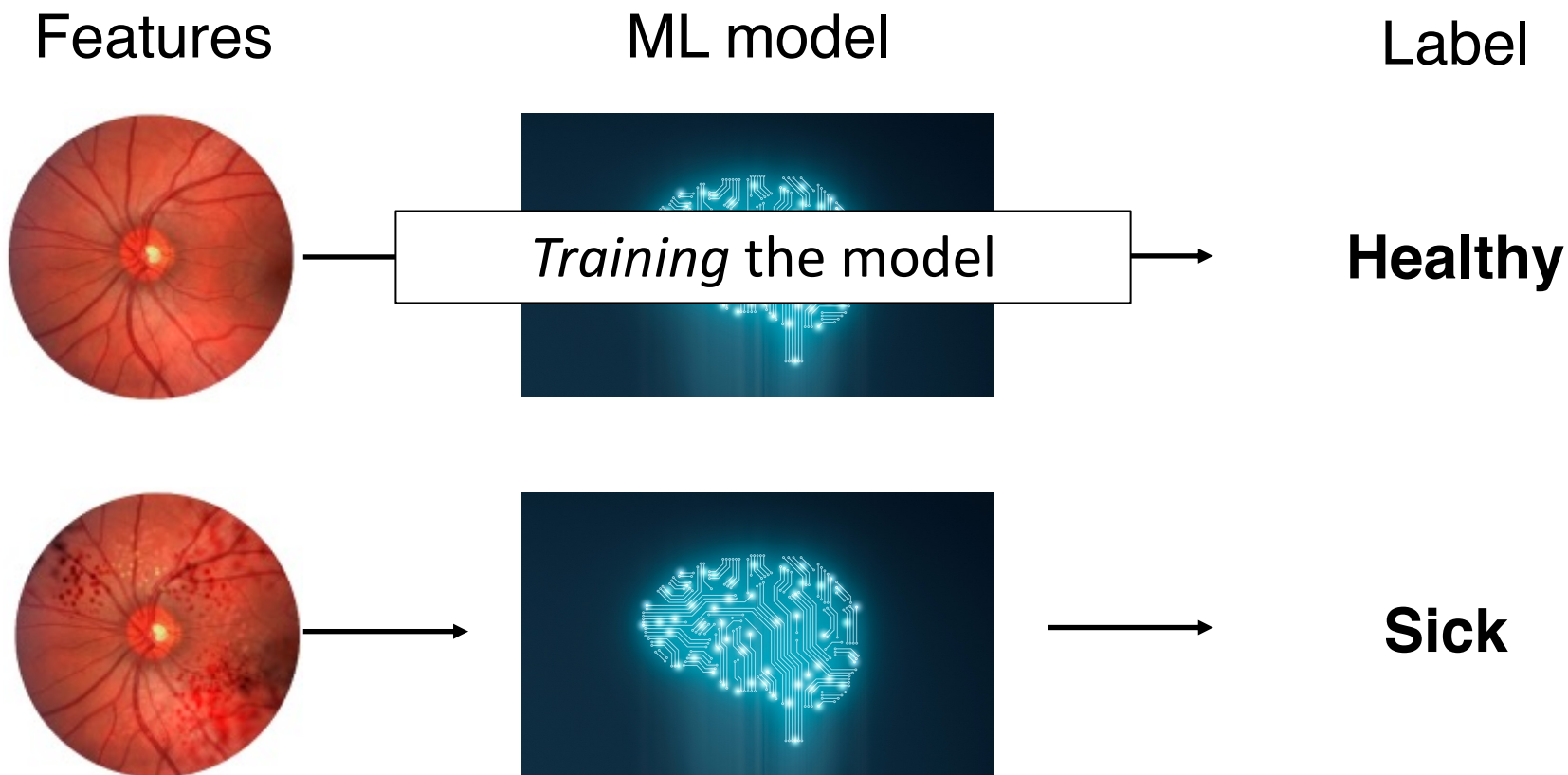
 1. Overview of syllabus

 2. What makes biomedical data unique

 3. Motivation for machine learning

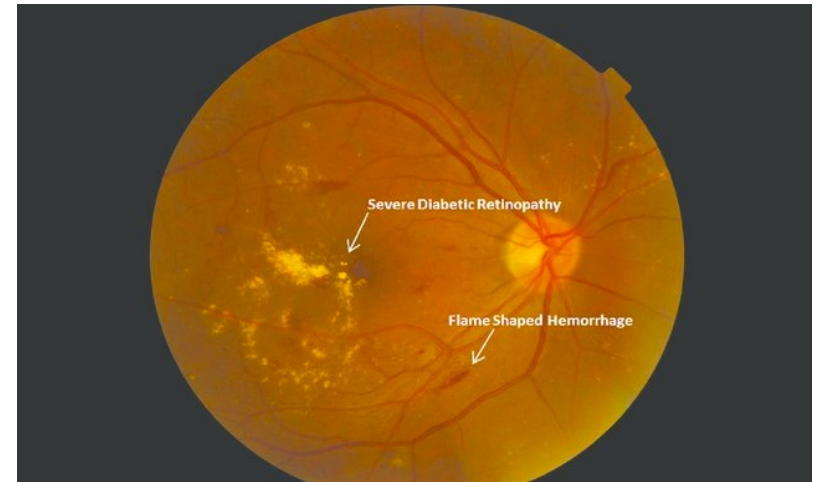
4. Roadmap for responsible AI

Machine learning

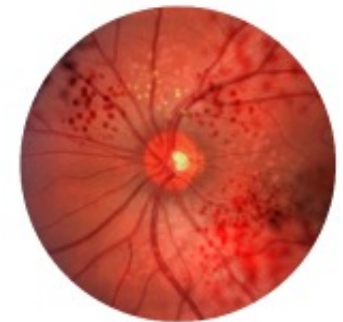


Diagnosing diabetic retinopathy

- Diabetic retinopathy affects blood vessels in the retina that lines the back of the eye
- The most common cause of vision loss among people with diabetes
- Leading cause of vision impairment and blindness among adults



Normal
Retina



Diabetic
Retina

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, 2016

Diagnosing diabetic retinopathy

- 128,000 retinal fundus photographs
- Each image was rated by 3-7 ophthalmologists
- “Off the shelf” deep neural network

This Issue

Views **60,378**

Citations **2**

Altmetric

633

Original Investigation | Innovations in Health Care Delivery

December 13, 2016

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD¹; Lily Peng, MD, PhD¹; Marc Coram, PhD¹; et al

» Author Affiliations

JAMA. 2016;316(22):2402-2410. doi:10.1001/jama.2016.17216



Normal
Retina



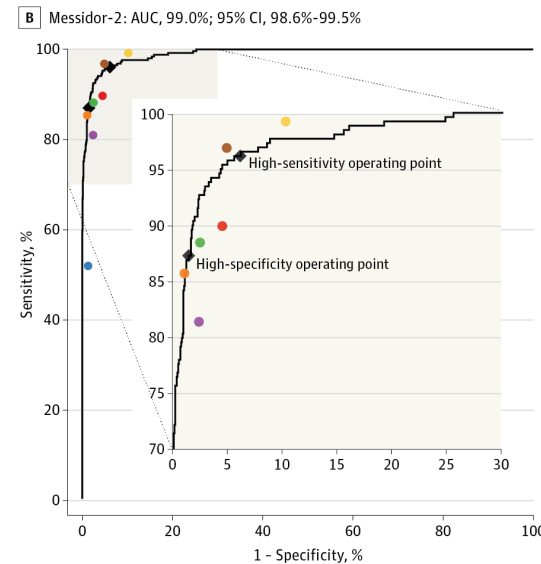
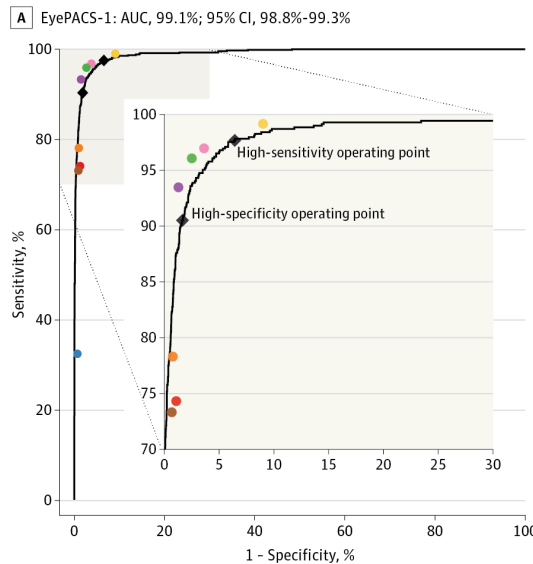
Diabetic
Retina

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, 2016

Diagnosing diabetic retinopathy

Algorithm did better than most individual ophthalmologists in the study

Large data + machine learning ~ human-level performance in diagnostic medical imaging



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Sensitivity (TPR): probability of positive test result, conditioned on individual truly being positive
Specificity (TNR): probability of a negative test result, conditioned on individual truly being negative

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, 2016

Pivotal trial of an autonomous AI-based diagnostic system

Abstract

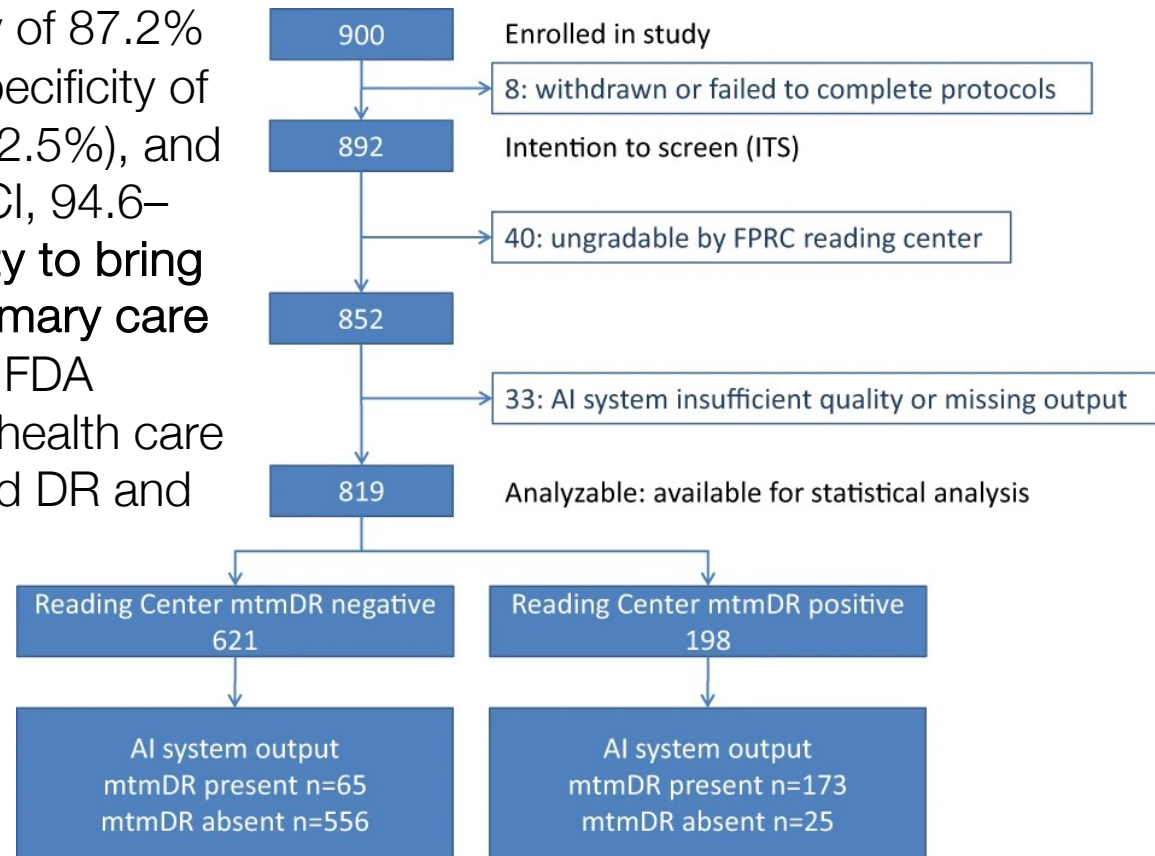
Artificial Intelligence (AI) has long promised to increase healthcare affordability, quality and accessibility but FDA, until recently, had never authorized an autonomous AI diagnostic system. This pivotal trial of an AI system to detect diabetic retinopathy (DR) in people with diabetes enrolled 900 subjects, with no history of DR at primary care clinics, by comparing to Wisconsin Fundus Photograph Reading Center (FPRC) widefield stereoscopic photography and macular Optical Coherence Tomography (OCT), by FPRC certified photographers, and FPRC grading of Early Treatment Diabetic Retinopathy Study Severity Scale (ETDRS) and Diabetic Macular Edema (DME). More than mild DR (mtmDR) was defined as ETDRS level 35 or higher, and/or DME, in at least one eye. AI system operators underwent a standardized training protocol before study start. Median age was 59 years (range, 22–84 years); among participants, 47.5% of participants were male; 16.1% were Hispanic, 83.3% not Hispanic; 28.6% African American and 63.4% were not; 198 (23.8%) had mtmDR.

Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices, *NPJ Digital Medicine*, 2018

Pivotal trial of an autonomous AI-based diagnostic system

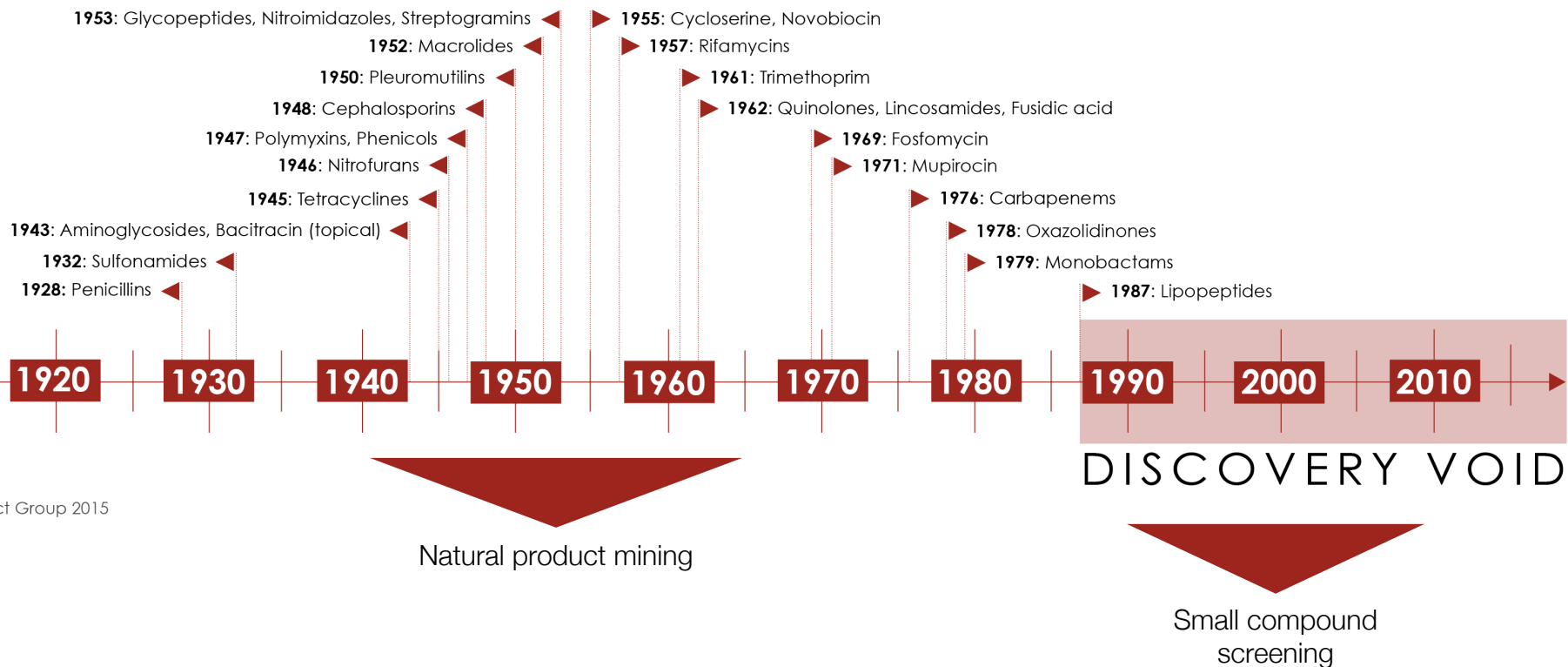
The AI system exceeded all pre-specified superiority endpoints at sensitivity of 87.2% (95% CI, 81.8–91.2%) (>85%), specificity of 90.7% (95% CI, 88.3–92.7%) (>82.5%), and imageability rate of 96.1% (95% CI, 94.6–97.3%), **demonstrating AI's ability to bring specialty-level diagnostics to primary care settings.** Based on these results, FDA authorized the system for use by health care providers to detect more than mild DR and diabetic macular edema.

First FDA authorized autonomous AI diagnostic system in any field of medicine



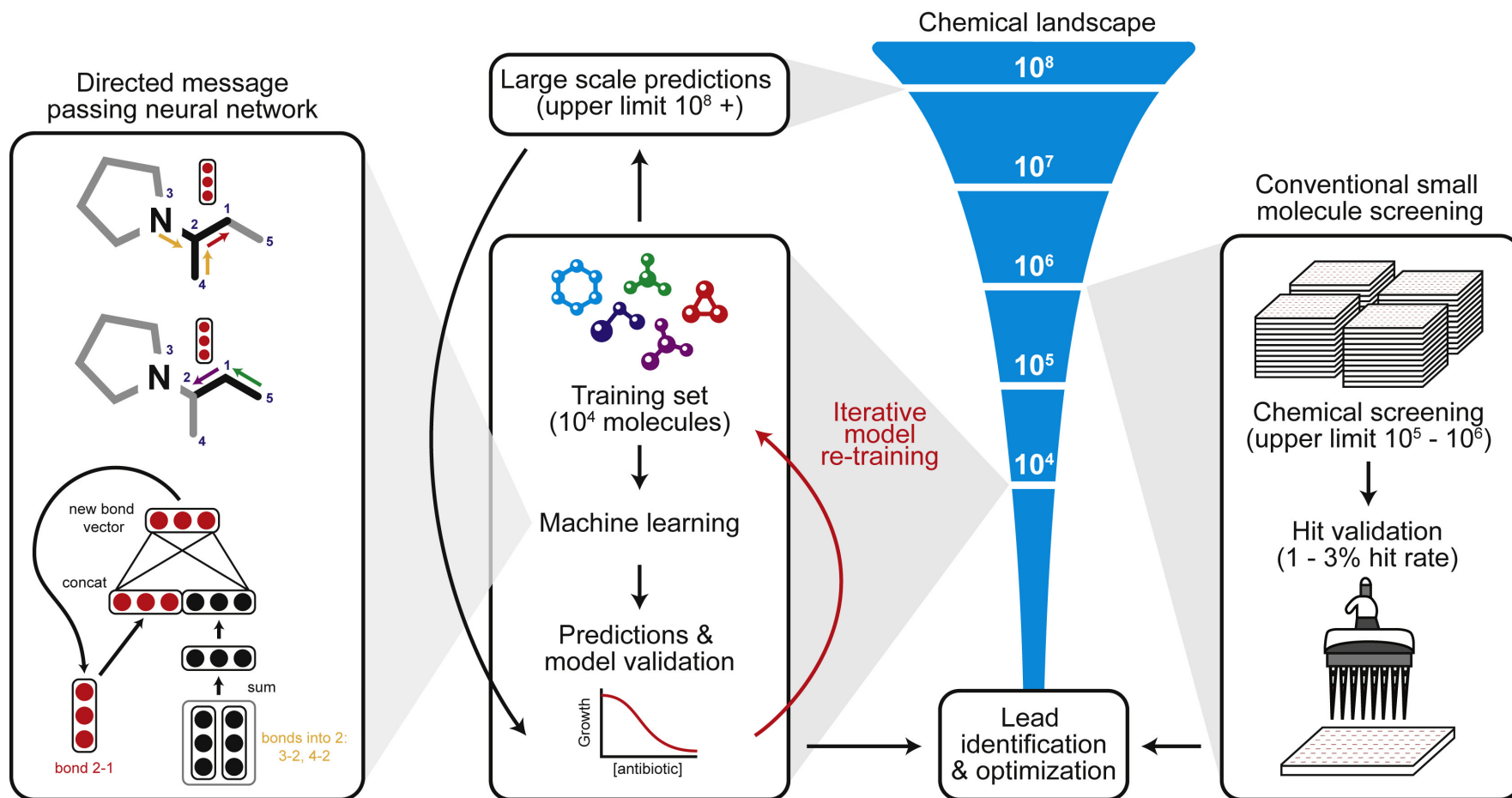
Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices, *NPJ Digital Medicine*, 2018

Antibiotic discovery timeline



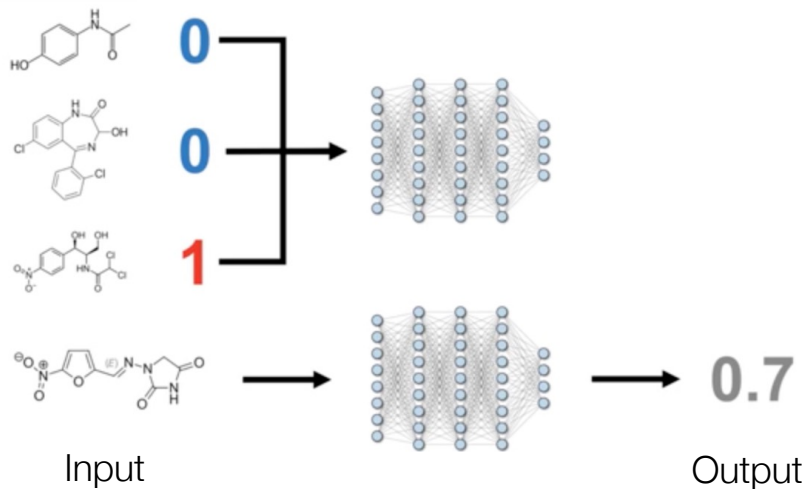
© ReAct Group 2015

GNN to learn molecular structure



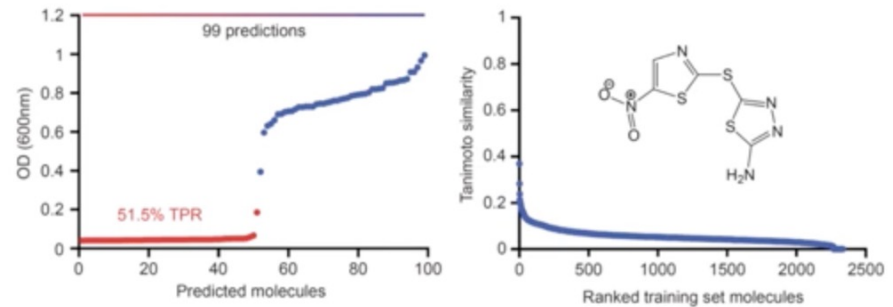
Experimental setup

Training Dataset
(Human Medicines and Natural Products)



Data: 2,335 molecules (human medicines and natural products) screened for growth inhibition

Empirical Validation
(Broad Repurposing Hub)



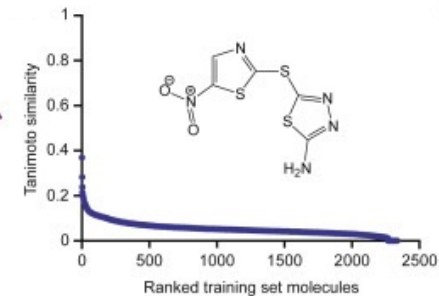
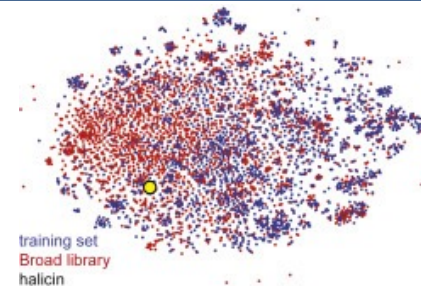
Data: 6,111 molecules (at various stages of investigation for human diseases) in the Repurposing Hub

Task: Test top 99 predictions & prioritize based on similarity to known antibiotics or predicted toxicity

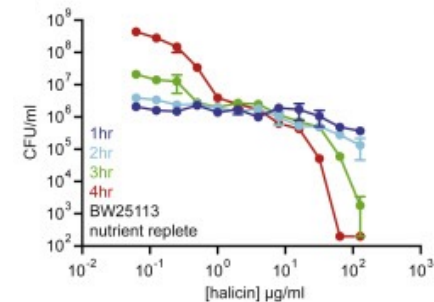
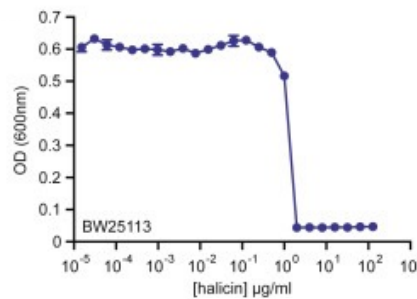
Chemical screening results

Halicin, initially developed as anti-diabetic drug (but discontinued due to poor results in testing), is identified and verified through experiments as a promising antibiotic

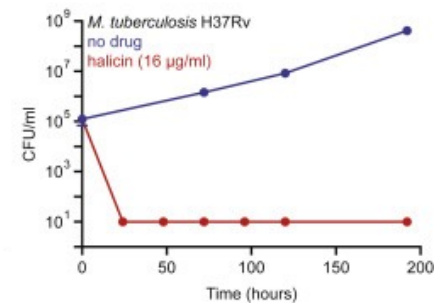
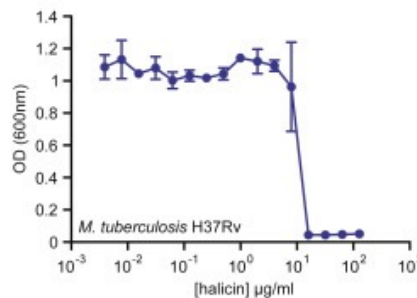
Halicin predicted to be antibacterial



Halicin against *E. coli*

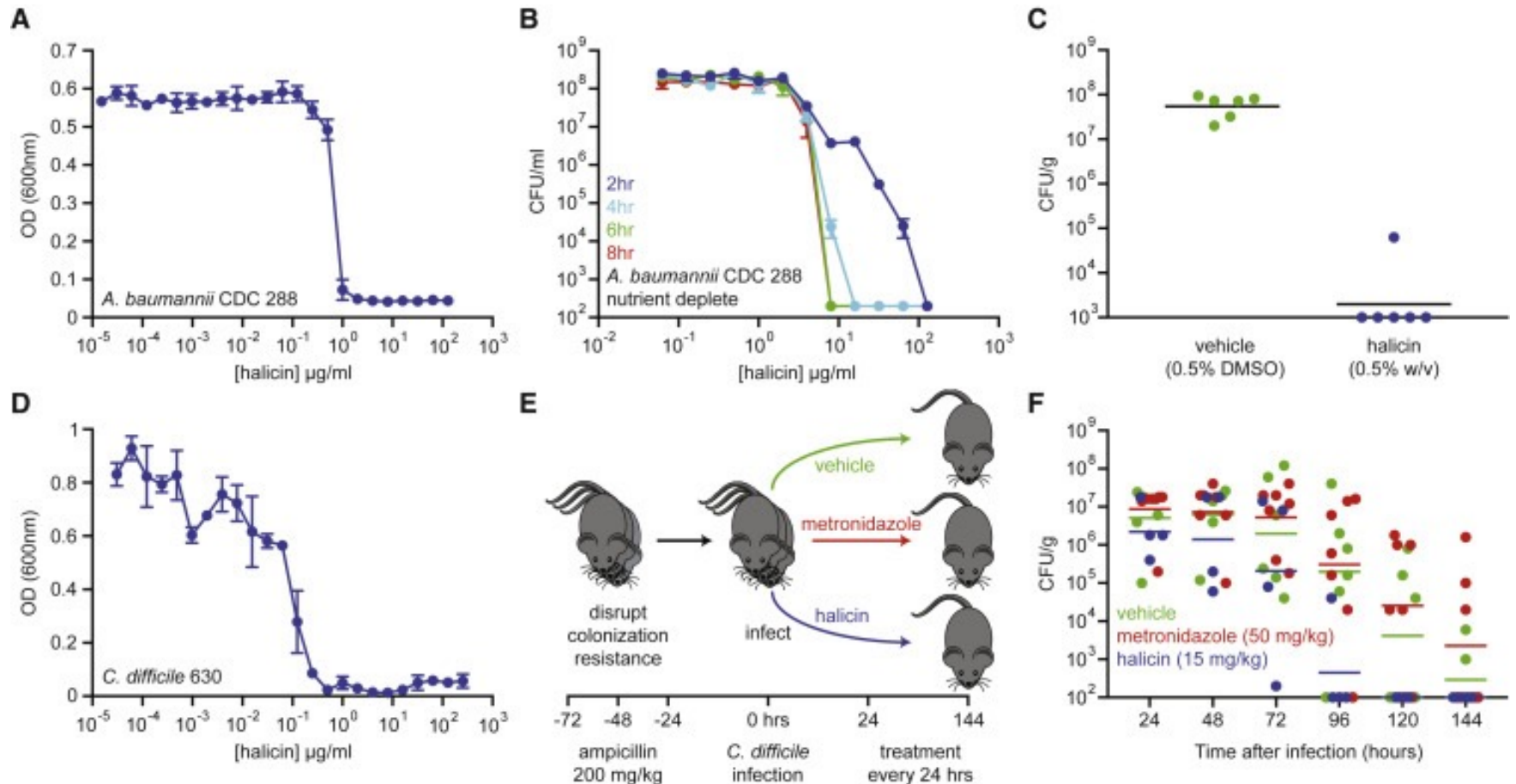


Halicin against *M. tuberculosis*



Chemical screening results

Halicin's efficacy in murine models of infection



Outline for today's class

 1. Overview of syllabus

 2. What makes biomedical data unique

 3. Motivation for machine learning

4. Roadmap for responsible AI



Roadmap to develop, validate and implement ML methods

Choosing the right problems

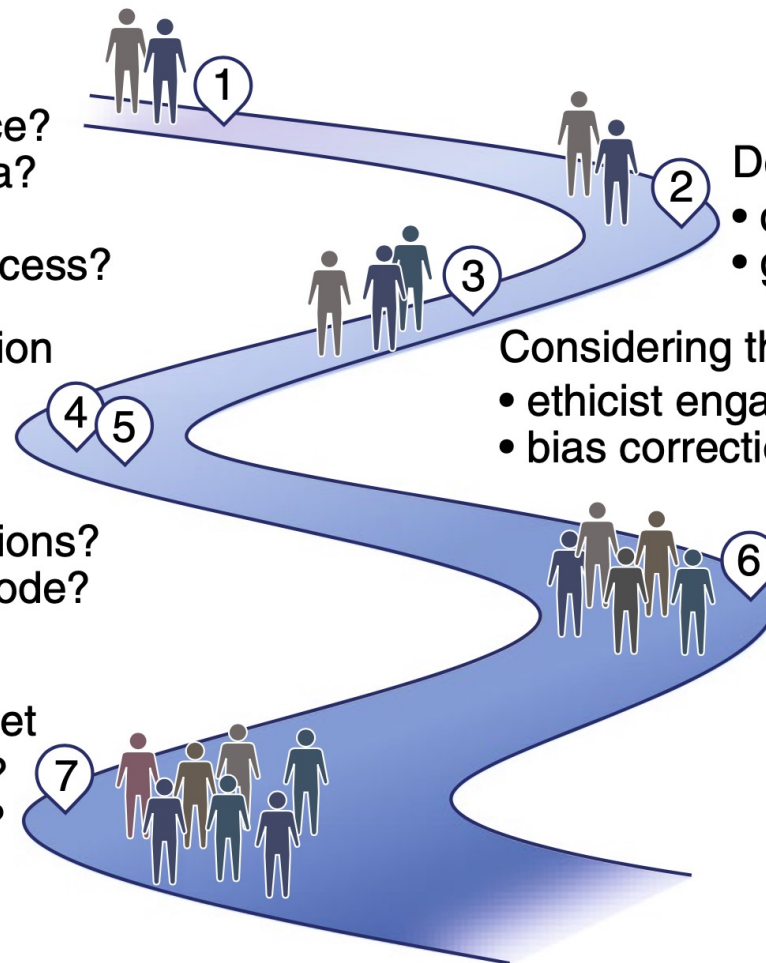
- clinical relevance?
- appropriate data?
- collaborators?
- definition of success?

Rigorous evaluation and thoughtful reporting

- model use?
- sensical predictions?
- shared model/code?
- failure modes?

Making it to market

- medical device?
- model updates?



Developing a useful solution

- data provenance?
- ground truth?

Considering the ethical implications

- ethicist engagement?
- bias correction?

Deploying responsibly

- prospective performance?
- clinical trial?
- safety monitoring?

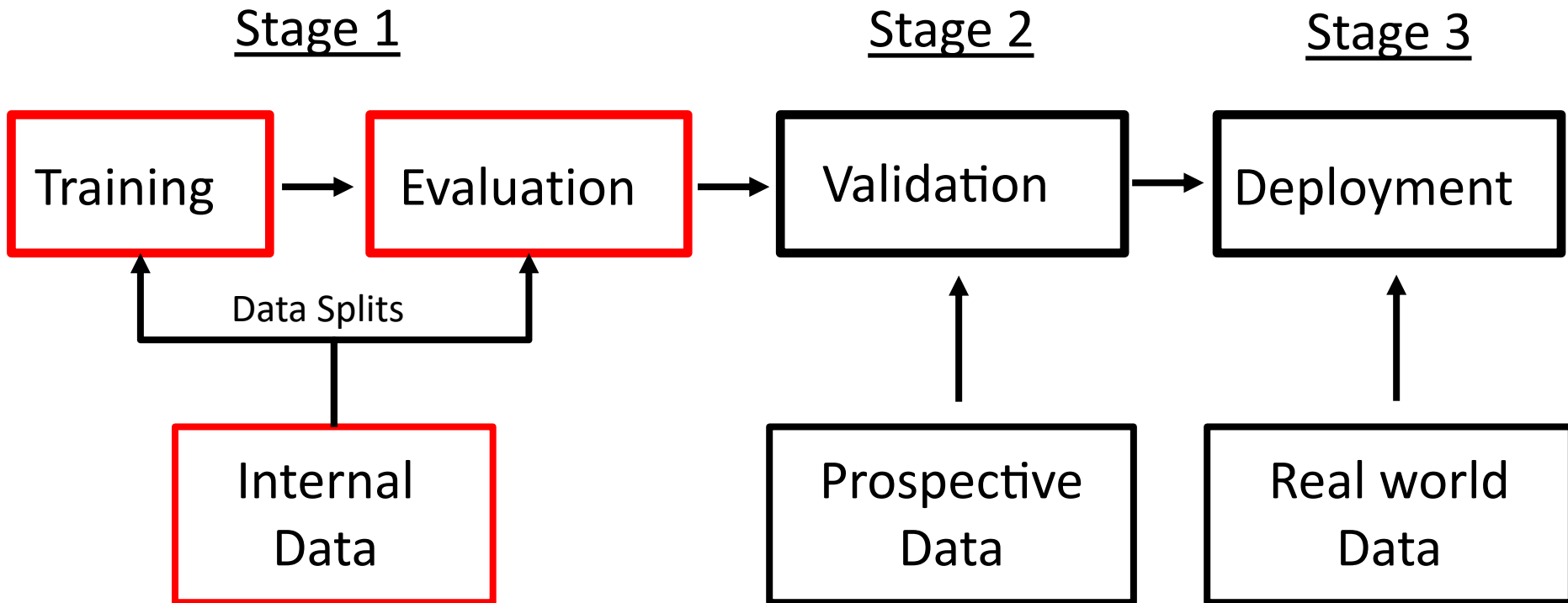
Key ML elements (1/2)

- **Examples:**
 - Also known as ‘samples’ or ‘observations’, basic units of analysis
 - Primary data objects being manipulated by an ML model
- **Features:**
 - Properties of a given example, also known as ‘covariates’
 - For example, the gene expression values associated with a gene or the sequence patterns associated with a genomic window
- **Labels/Outcomes/Target variables:**
 - Outcomes are what we want to predict in supervised learning
 - For example, the functional class assigned to a gene or the binary classification of whether a given genomic window contains a promoter
 - In classification, the **outcome is a category**, known as ‘label’ or ‘class’
 - In regression, the **outcome is a real number**

Key ML elements (2/2)

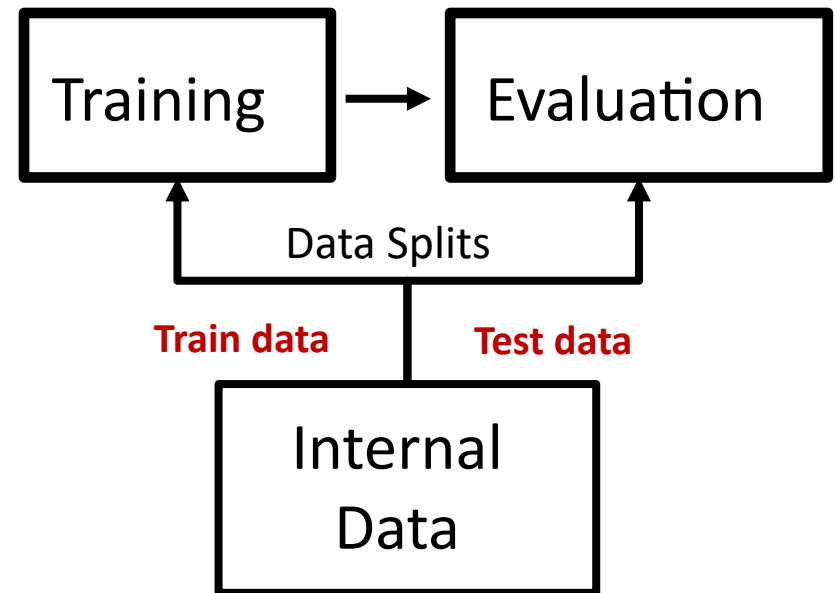
- **Training set:** Examples and associated outcomes used to fit an ML model
 - **Positives:** Examples with the outcome of interest in a binary classifier
 - **Negatives:** Examples with the alternative outcome in a binary classifier
- **Test set:** Examples and associated outcomes that are used to evaluate model performance
 - No examples are shared between training and test sets
- Once we identify the specific ML problem that will be solved, we must train a model and determine how to properly evaluate its performance
- Performance evaluation is often executed using **cross-validation**, whereby examples are iteratively randomized into a **training set** used to train a model and a held-out **test set** used to quantify model performance
- **Prediction set:**
 - Examples whose associated outcomes are truly not known, where a fitted model is applied to make predictions
 - Also known as a **prospective set**

Roadmap for ML development



Stage 1: Algorithm development

- Stage 1 focuses on designing and developing initial models
- Use historical or retrospective data not originally collected for developing ML models
- Most of the data used as **training data** and a small part serve as a **held-out test set**



The importance of labels

- The labels, or ground-truth diagnoses, are often very hard, expensive and time consuming to obtain, but are usually the **most important part of building an ML system**
- In medicine they are **often provided by physicians or other healthcare workers** -> time consuming and noisy: Doctors don't always agree!
- **The quality of your labels will “upper-bound” the performance of the system** – we cannot be more accurate than the labels!

Evaluating performance

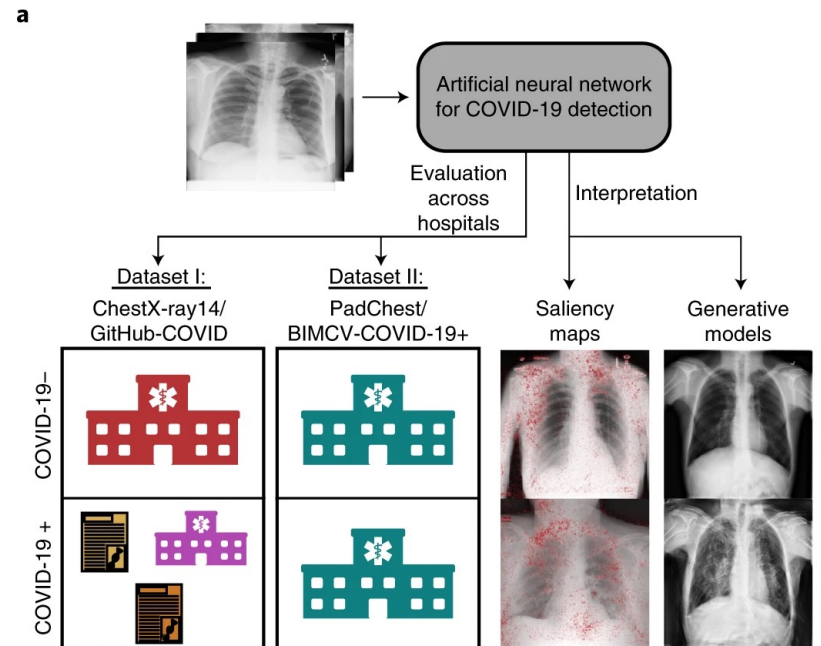
- Ideally, you would like the system to optimize for something you care about, e.g., outcomes, cost, etc. but those are only measurable in Stage 2/3
- Instead, we use **proxy metrics** during Stage 1, e.g.:
 - Classification accuracy
 - Sensitivity/Specificity
 - Precision/Positive predictive value ($PPV = \#TP / \#pos\text{-calls}$)
 - Area under the ROC curve (AUROC)
 - Area under the precision-recall curve (AUPRC)
- There is not usually an objectively good metric
 - The choice of a metric is application-dependent

Stage 1: Challenges

- ML models can fail at this stage for various reasons:
 - Not enough data and insufficient model capability – model is not a meaningful advance over current methods
 - Model is good but is hard to integrate into biomedical and clinical workflows
 - Model looks good but is subtly overfit or confounded in a way that is very hard to detect

Stage 1: Shortcut learning challenge

- Neural network model is trained to detect COVID-19 using radiographs from either of the two datasets
- Model is evaluated on both datasets to learn how performance may drop in deployment
- Interpretability methods are applied to infer what the model learned and which features were important for its decisions
- **Dataset I:** Radiographs from multiple hospital systems as well as cropped images from publication figures
- **Dataset II:** Radiographs from multiple hospitals from a single regional hospital system.

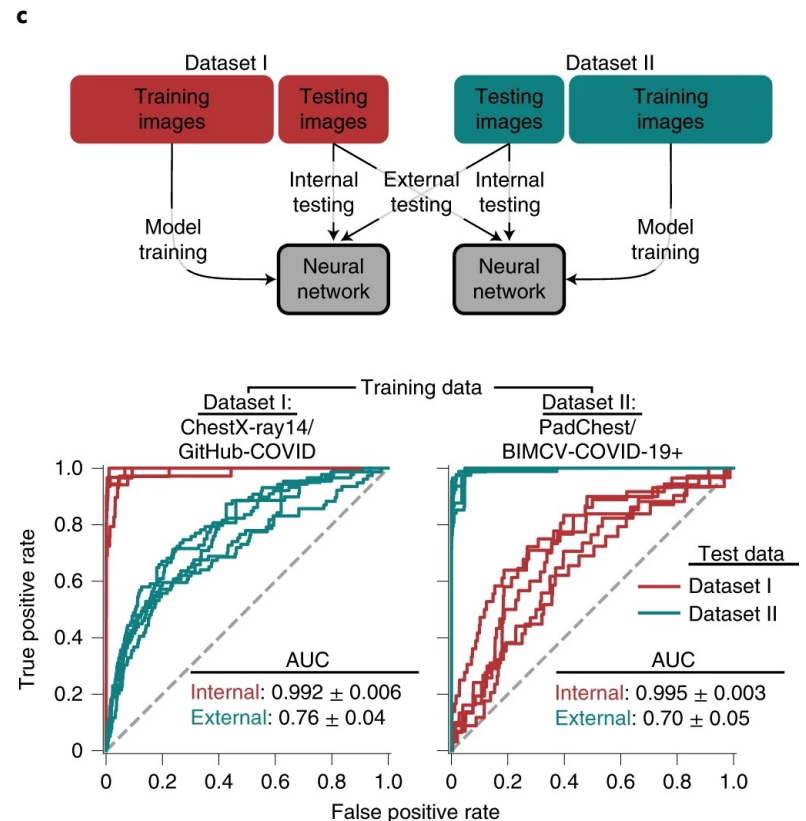


b

	Dataset I			Dataset II		
	Combined	Chest-X-ray14	GitHub-COVID	Combined	PadChest	BIMCV-COVID-19+
No. radiographs	112,528	112,120	408	97,866	96,270	1,596
No. patients	31,067	30,805	262	64,954	63,939	1,105
% COVID-19+	0.2	0	76.5	1.6	0	100
% AP images	39.9	40	26	5.6	4.7	58.1

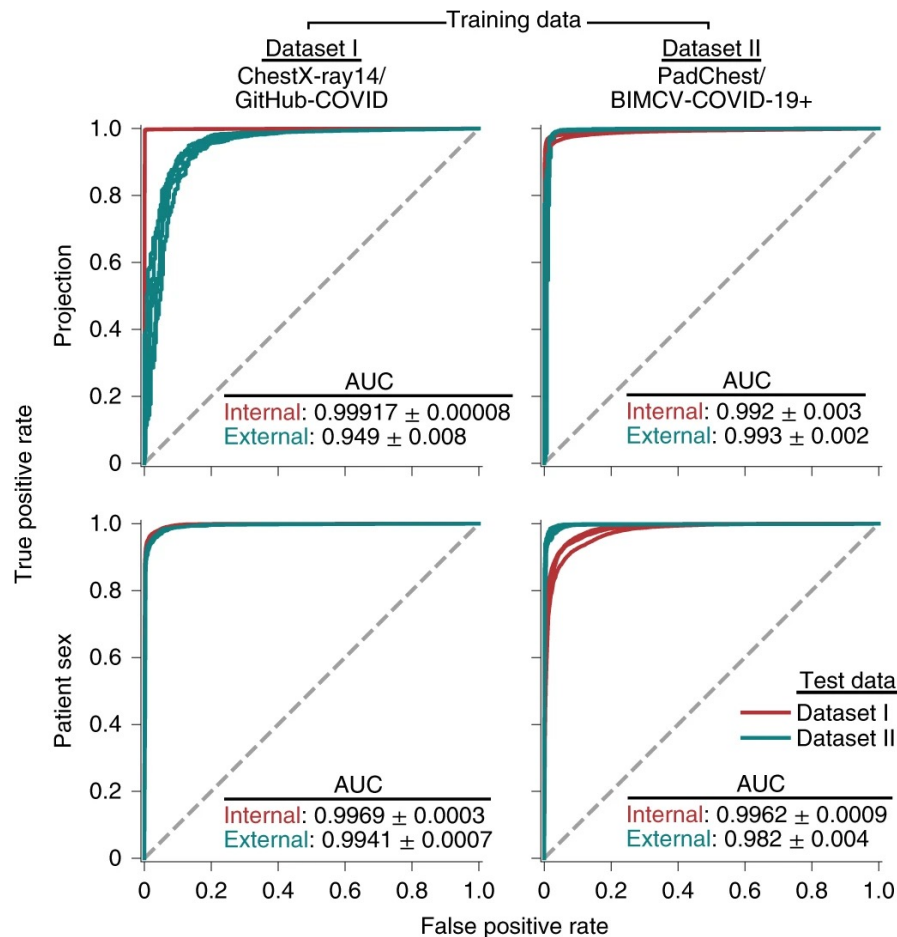
Stage 1: Shortcut learning challenge

- Performance is evaluated on internal test set
 - New, held-out examples from the same data source as the training radiographs
- Performance is evaluated on external test set
 - Radiographs from a new hospital system
- **Generalization gap**
 $|AUC_{\text{internal set}} - AUC_{\text{external set}}|$



Stage 1: Shortcut learning challenge

- Models failed to learn the true underlying disease pathology
- They used **shortcuts**:
 - Spurious associations between the presence or absence of disease
 - Radiographic features that reflect variations in image acquisition

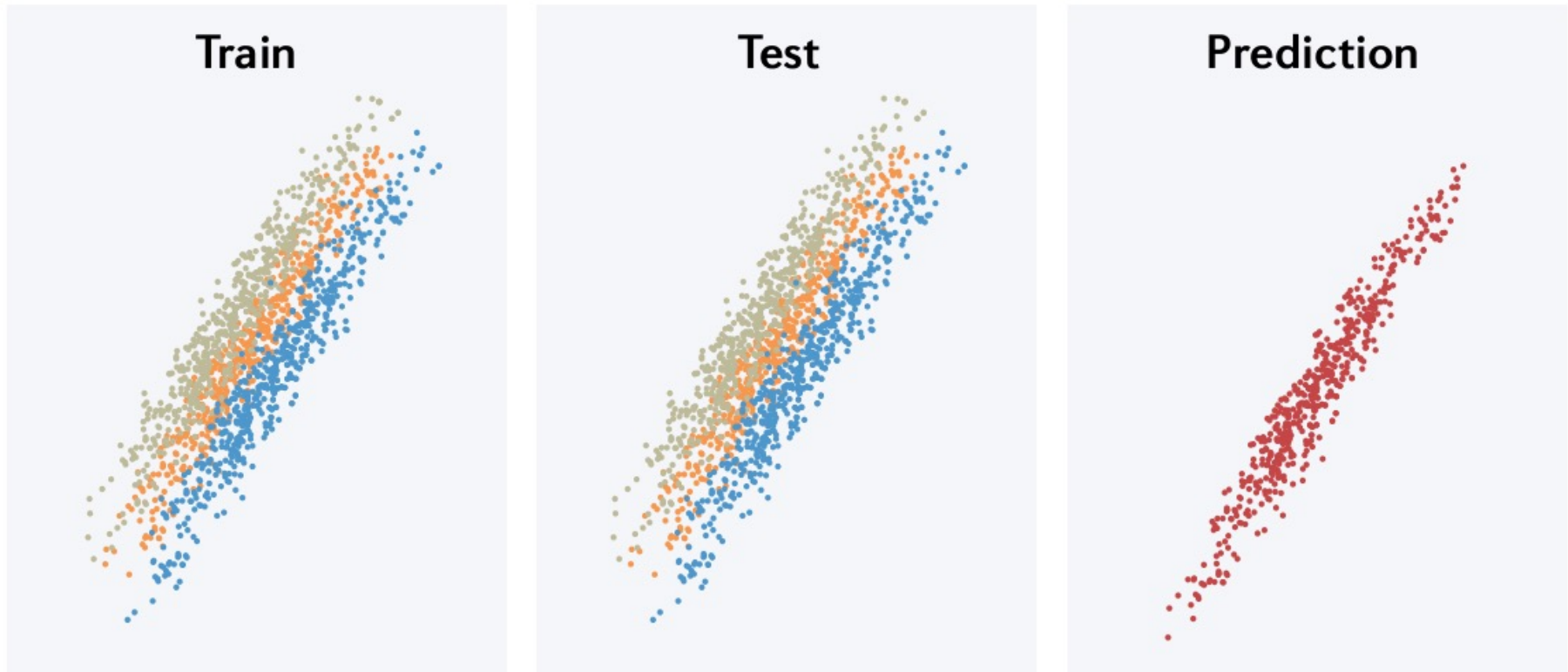


Models accurately predict both the radiographic projection and patient sex for both internal and external test data, which supports that these concepts are easily learned and available to be leveraged as shortcuts

Stage 1: Shortcut learning challenge

- Lessons learned:
 - Seemingly high-performance AI systems may derive the majority of their performance from the exploitation of undesired shortcuts
 - Developers and users of these models need to verify that AI systems rely on the desired signals
 - High data quality is important for robust and useful models

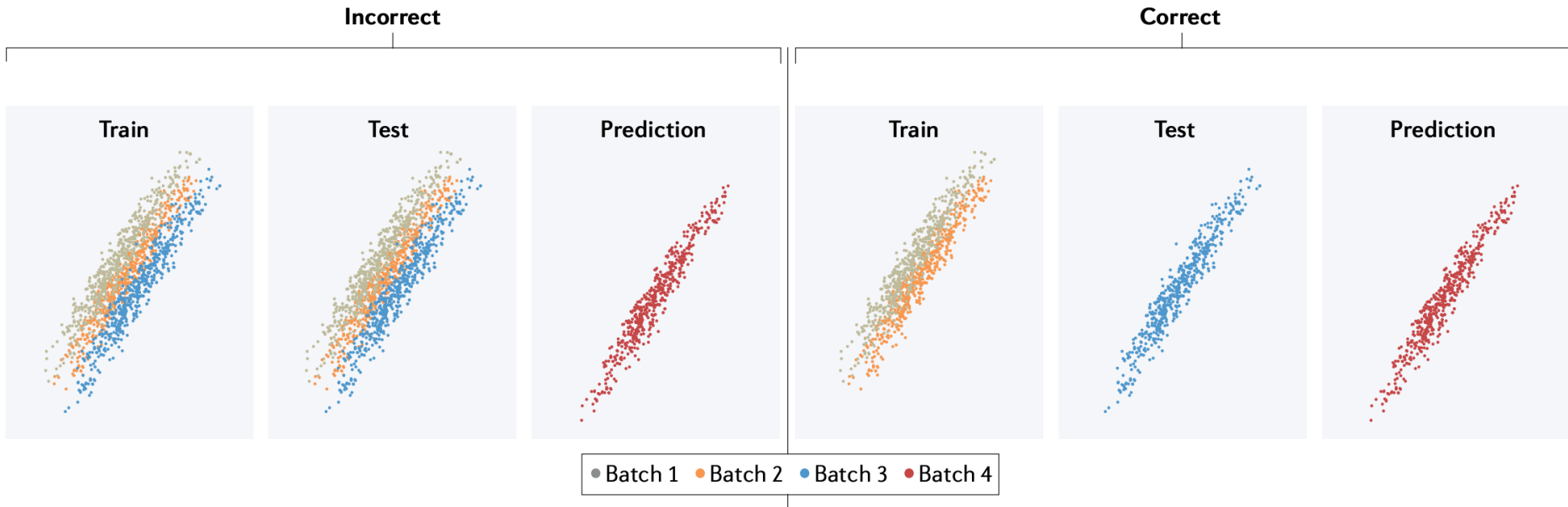
Quick check: What is the problem with this data split?



● Batch 1 ● Batch 2 ● Batch 3 ● Batch 4

Join at [slido.com](https://www.slido.com) with #033364

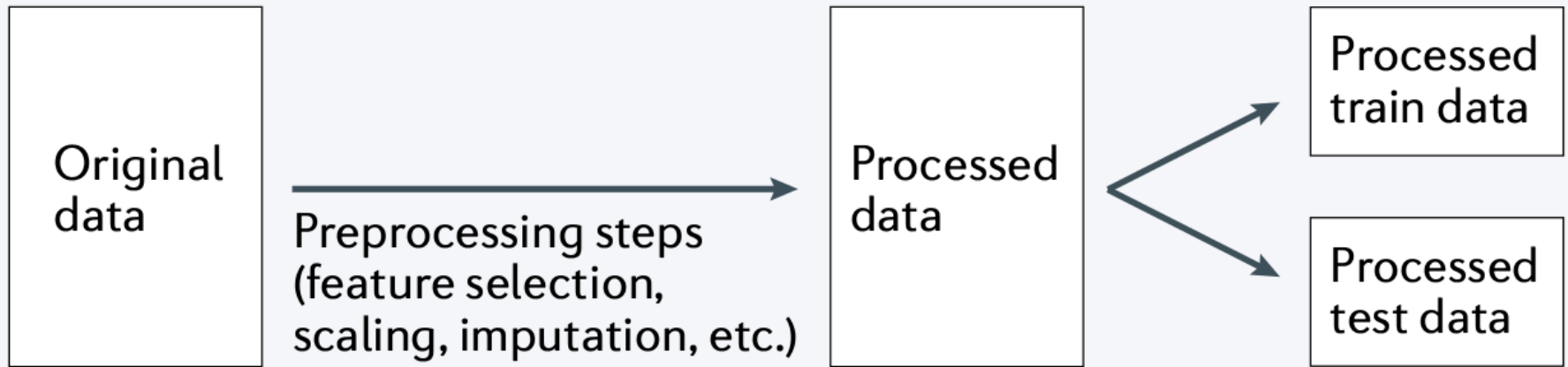
Quick check: What is the problem with this data split?



Distributional differences can arise from various sources, such as batch effects. If training and test sets are a mixture of examples from every batch (left), performance on the test set will be much higher than on a new batch

To fit a model that will generalize to new batches, training and test sets should be composed of different batches (right)

Quick check: What is the problem with this ML workflow?

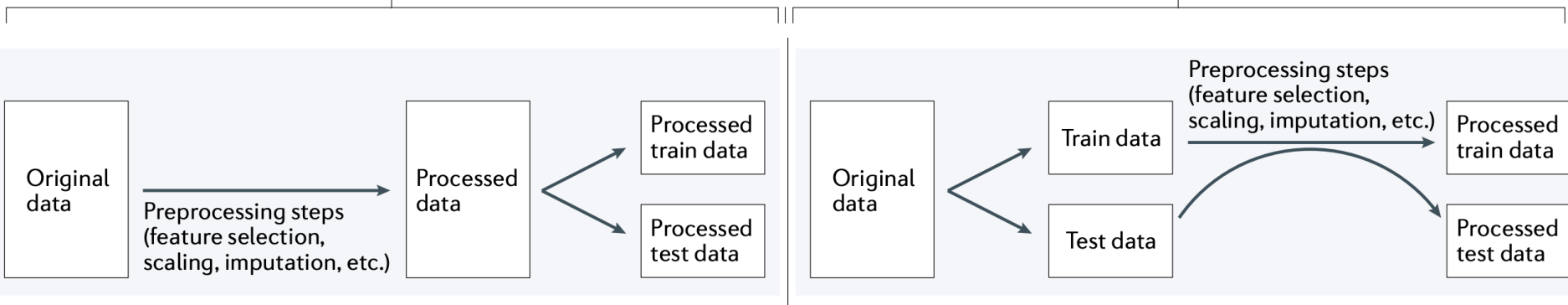


Join at [slido.com](https://www.slido.com) with #033364

Quick check: What is the problem with this ML workflow?

Incorrect

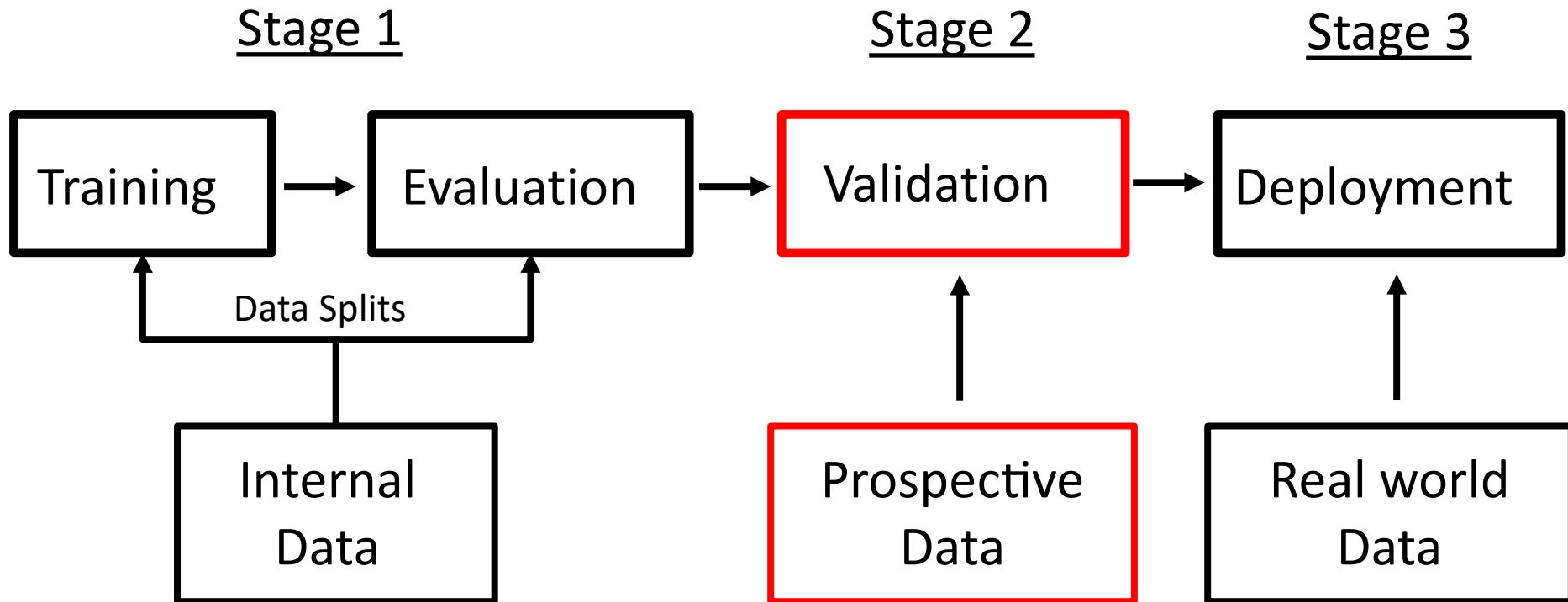
Correct



Information leakage can happen when information is leaked from the test set into the training as a result of the training and test sets being preprocessed together (left)

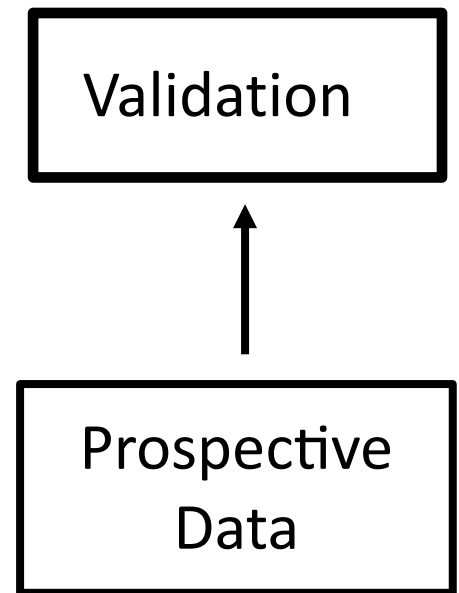
Instead, the raw data should be split into training and test sets with preprocessing performed separately (right)

Roadmap for ML development



Stage 2: Prospective validation

- Stage 2 is focused on prospectively validating the model from Stage 1 on “live” data coming in under a controlled setting
- Often **follows a trial format** with pre-registered endpoints and set for a fixed amount of time
- Goal is to show the model performs well and generalizes to real-world data



Stage 2: Cardiovascular AI trial

- Accurate quantification of cardiac function is necessary for disease diagnosis, risk stratification and assessment of treatment response
 - **Left ventricular ejection fraction (LVEF)** is routinely used to guide clinical decisions regarding patient appropriateness for a wide range of medical and device therapies as well as interventions, including surgeries
- **Clinical guidelines:**
 - When assessing LVEF based on echocardiography, the measurements are performed repeatedly over multiple cardiac cycles to improve precision and account for arrhythmic or hemodynamic sources of variation
- **Practice:**
 - Repeated human measurements are rarely done in practice given logistical constraints present in most clinical imaging laboratories and single tracings or a visual estimation of LVEF is often used as a pragmatic alternative
 - Such an approach is suboptimal for detection of subtle changes in LVEF, which is needed for making important therapeutic decisions (for example, eligibility for continued chemotherapy or defibrillator implantation)

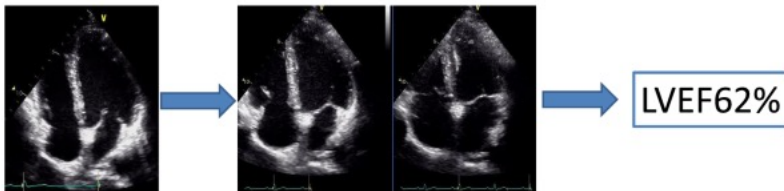
Blinded, randomized trial of sonographer versus AI cardiac function assessment, Nature, 2023

Stage 2: Cardiovascular AI trial

- Many AI models are developed with the goal of automating assessment of LVEF in real-world patient care settings
- These AI models demonstrated improved precision on retrospective datasets
- No current cardiovascular AI technologies were validated in blinded, randomized clinical trials
- In addition, human-computer interaction and the effect of AI prompting on clinical interpretations is underexplored in clinical studies

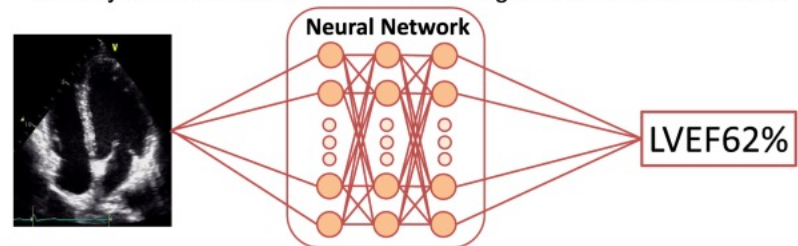
Conventional AI

- Learn the location of the heart and endocardial borders.
- Measure the LVEF with tracking the endocardial borders.



New AI: Deep Learning

- Directly estimate the LVEF without tracking the endocardial borders.



Blinded, randomized trial of sonographer versus AI cardiac function assessment, Nature, 2023

Stage 2: Cardiovascular AI trial

Blinded, randomized non-inferiority clinical trial to prospectively assess the effect of initial assessment by AI versus conventional initial assessment by a sonographer on final cardiologist interpretation of LVEF

- **Primary endpoint:** Change in the LVEF between initial AI or sonographer assessment and final cardiologist assessment, evaluated by the proportion of studies with substantial change (more than 5% change)
- **Results:** Proportion of studies substantially changed was 16.8% in the AI group and 27.2% in the sonographer group
- **Results:** Mean absolute difference between final cardiologist assessment and independent previous cardiologist assessment was 6.29% in the AI group and 7.23% in the sonographer group

COMPLETED

Safety and Efficacy Study of AI LVEF (EchoNet-RCT)

ClinicalTrials.gov ID NCT05140642

Sponsor Cedars-Sinai Medical Center

Information provided by David Ouyang, Cedars-Sinai Medical Center (Responsible Party)

Last Update Posted 2022-07-05

+ Expand all content - Collapse all content

Study Details Researcher View No Results Posted Record History

On this page

- Study Overview
- Contacts and Locations
- Participation Criteria
- Study Plan
- Collaborators and Investigators
- Publications
- Study Record Dates
- More information

Study Overview

Brief Summary

To determine whether an integrated AI decision support can save time and improve accuracy of assessment of echocardiograms, the investigators are conducting a blinded, randomized controlled study of AI guided measurements of left ventricular ejection fraction compared to sonographer measurements in preliminary readings of echocardiograms.

Official Title

Blinded Randomized Controlled Trial of Artificial Intelligence Guided Assessment of Cardiac Function

Conditions

Heart Failure, Systolic Heart Failure, Diastolic

Intervention / Treatment

- Other: Automated annotation of the left ventricle through deep learning
- Other: Sonographer Measurement of LVEF

Other Study ID Numbers

- STUDY00001707

Study Start (Actual) 2022-04-01

Primary Completion (Actual) 2022-06-29

Study Completion (Actual) 2022-06-29

Enrollment (Actual) 3495

Study Type Interventional

Phase Not Applicable

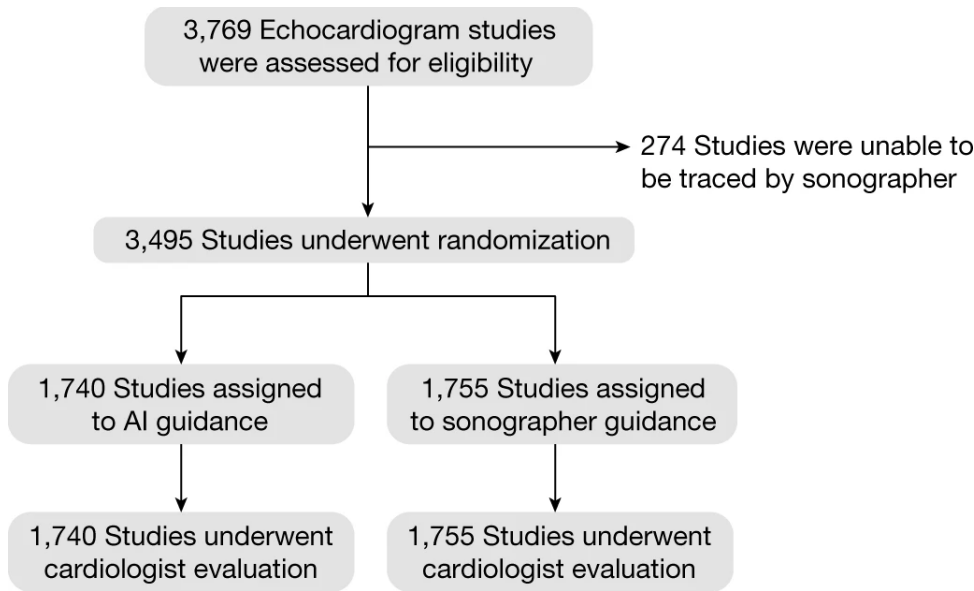
Resource links provided by the National Library of Medicine NIH NLM

ClinicalTrials.gov ID: NCT05140642

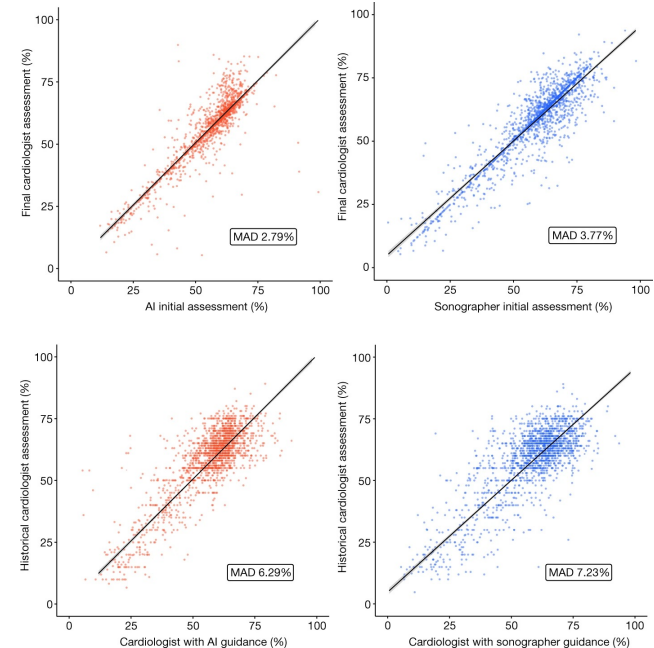
Blinded, randomized trial of sonographer versus AI cardiac function assessment, Nature, 2023

Stage 2: Cardiovascular AI trial

- 1 AI-guided workflow saved time for sonographers and cardiologists
- 2 Cardiologists were not able to distinguish between the initial assessments by AI versus sonographers
- 3 For patients undergoing echocardiographic quantification of cardiac function, initial assessment of LVEF by AI was non-inferior to assessment by sonographers



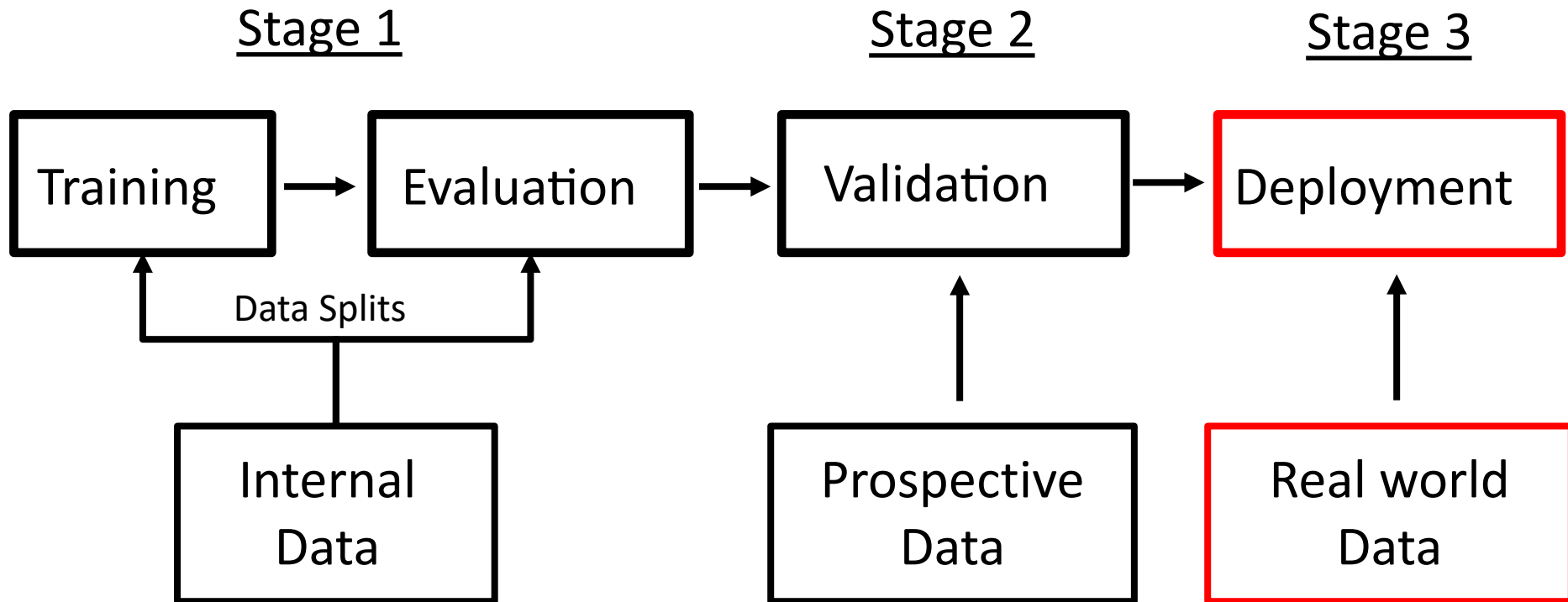
Screening, randomization and follow-up



Comparison of AI versus sonographer guidance on cardiologist assessment and difference between final versus previous cardiologist assessments. Dots represent individual studies and lines represent the lines of best fit. MAD, mean absolute difference

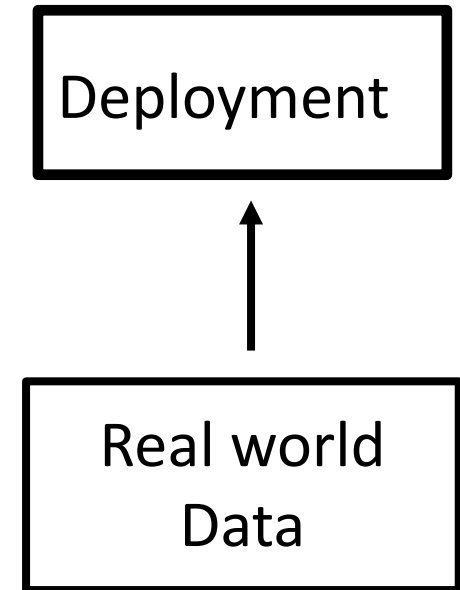
Blinded, randomized trial of sonographer versus AI cardiac function assessment, Nature, 2023

Roadmap for ML development



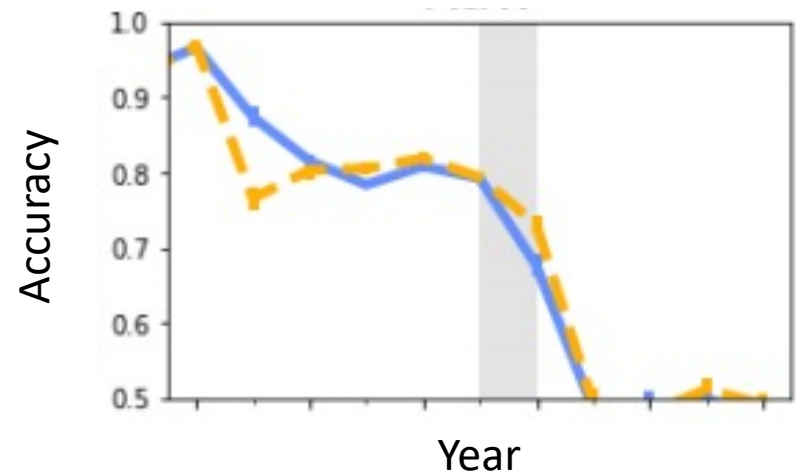
Stage 3: Deployment

- Stage 3 has likely been the ultimate goal
- ML model is implemented in a biomedical or clinical settings and used to guide experiments in the laboratory or provide decision support



Stage 3: Example challenge

- ML model has been extensively validated and shown to be very accurate
- We implement the model and suddenly notice a huge drop in performance during an audit
- What could be going on?

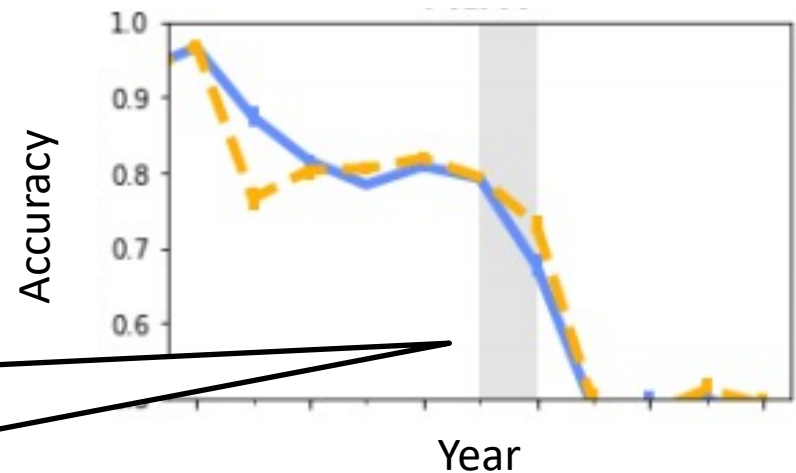


Nestor et al, 2018

Stage 3: Example challenge

- ML model has been extensively validated and shown to be very accurate
- We implement the model and suddenly notice a huge drop in performance during an audit
- What could be going on?

Answer: Your hospital updated its EHR to a new version. Your AI system was completely tied to the old way the data were recorded and now no longer works.



Nestor et al, 2018

Outline for today's class

- ✓ 1. Overview of syllabus
- ✓ 2. What makes biomedical data unique
- ✓ 3. Motivation for machine learning
- ✓ 4. Roadmap for responsible AI