Formulating Impactful ML4H Research Questions

Emily Alsentzer

AHLI Summer Camp – Day 1

Stanford MEDICINE | School of Medicine

Stanford | ENGINEERING
Computer Science

The Taste Gap

Ira Glass,
Creator & Host of
This American Life Podcast

https://www.youtube.com/watch?v=X2wLP0izeJE

The Taste Gap

IRA GLASS

Creator & Host of *This American Life*

All of us who do creative work, we get into it because we have have good taste. But there is this gap. For the first couple years you make stuff, it's just not that good. It's trying to good, it has potential, but it's not. But your taste, the got you into the game, is still killer. And your taste is why work disappoints you. A lot of people never get past this they quit. Most people I know who do interesting, creative went through years of this. We know our work doesn't special thing that we want it to have. Everyone goes

And if you are just starting out or you are still in this phase, gotta know it's normal and the most important thing you is do a lot of work.

Research taste is critical in the era of coding agents

Agent Laboratory: Using LLM Agents as Research Assistants

Samuel Schmidgall[1, 2], Yusheng Su[1], Ze Wang[1], Ximeng Sun[1], Jialian Wu[1], Xiaodong Yu[1], Jiang
Moor[3], Zicheng Liu[1] and Emad Barsoum[1]
[1]AMD, [2]Johns Hopkins University, [3]ETH Zurich

Article | Open access | Published: 25 March 2026

Towards end-to-end automation of AI research

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, Davi
Jeff Clune

Nature **651**, 914–919 (2026) | Cite this article

NEWS | 19 May 2026

Teams of AI agents boost speed of research

Systems can generate hypotheses, interpret data and suggest ways to develop medicines.

Accelerating scientific breakthroughs with an AI co-scientist

February 19, 2025 · Juraj Gottweis, Google Fellow, and Vivek Natarajan, Research Lead

Outline for Today

- Defining research taste and impact
- Common pitfalls in problem formulation
- Questions to ask when beginning a new project
- Walkthrough of example ML4H projects

Research Taste

the tacit knowledge, intuitions, and judgments that guide a researcher's decisions throughout the research process

- Selecting important and tractable problems
- Designing informative experiments
- Recognizing surprising results worth pursuing
- Knowing when to persist, pivot, or quit

"My Research Process: Understanding and Cultivating Research Taste" by Neel Nanda

Research Taste

the tacit knowledge, intuitions, and judgments that guide a researcher's decisions throughout the research process

- Selecting important and tractable problems
- Designing informative experiments
- Recognizing surprising results worth pursuing
- Knowing when to persist, pivot, or quit

What constitutes impactful research?

You must determine what kind of impact you want to have!

Belief

Changes what the field holds to be true

e.g. Shortcut learning research revealed failure mode in medical imaging models

Capability

Creates a method, tool, dataset, or result others can build on

e.g. Shared datasets / benchmarks or an open-source model

Agenda

Reframes what people work on by opening a subfield

e.g. “Datasheets for Datasets” spurred research in dataset documentation

Practice

Changes what someone (e.g., clinician, regulator, etc.) does in practice

e.g. Obermeyer et al research on proxy labels shifted how practitioners considered bias in a deployed system.

Examples of Impactful Research

DOI:10.1145/345872

Documentation to facilitate communication between dataset creators and consumers.

BY TIMNIT GEBRU, JAMIE MORGENSTERN, BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN, HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD

Datasheets for Datasets

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Data Descriptor | Open access | Published: 24 May 2016

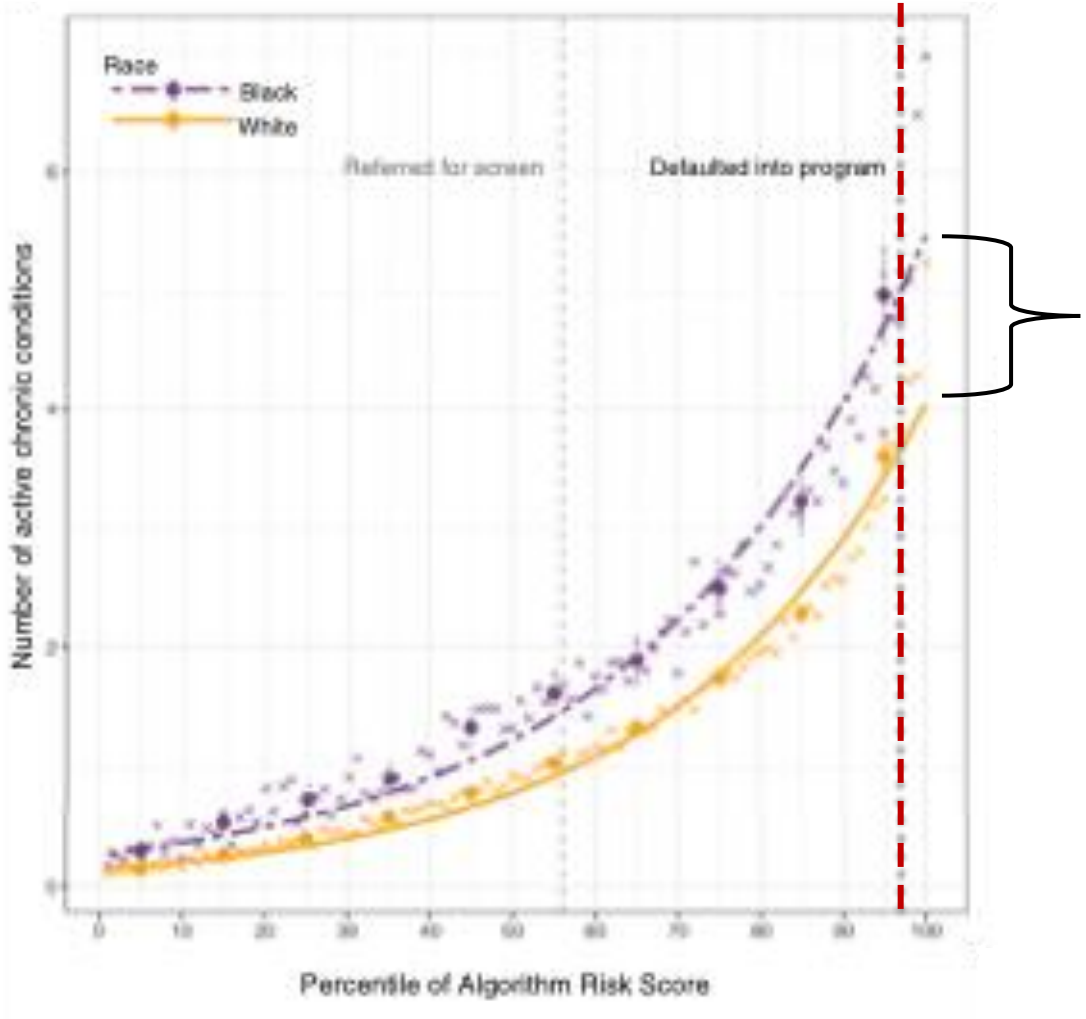
MIMIC-III, a freely accessible critical care database

Alistair E.W. Johnson, Tom J. Pollard ✉, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi & Roger G. Mark

Scientific Data **3**, Article number: 160035 (2016) | Cite this article

377k Accesses | **7026** Citations | **177** Altmetric | Metrics

Healthcare cost as a proxy for health need produces a racially biased algorithm

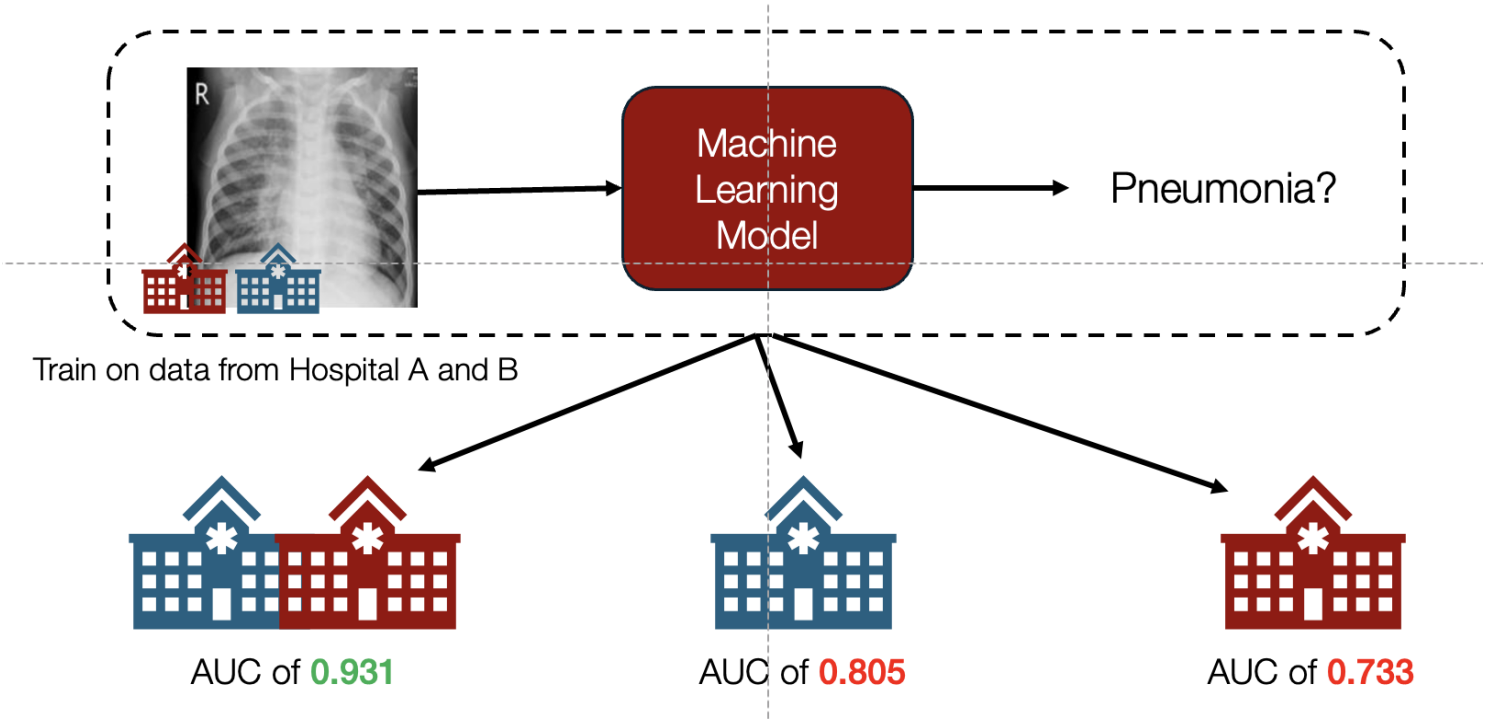


The model learns to predict who can afford to go to get care, not who needs care

At the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites

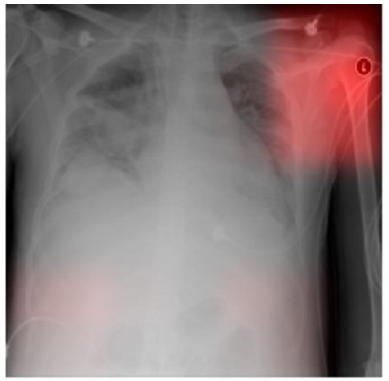
Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019.

Models can learn to rely on hospital specific features



Models can learn to detect a metal token that radiology technicians place on the patient in the corner of the image field of view at the time they capture the image

Why? Substantial difference in pneumonia prevalence between hospitals 34.2% vs 1.2%



Zech et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study . BMJ. 2018

What constitutes impactful research?

- **Reach**: how many people or decisions does it affect?
- **Durability**: does it contribute something that lasts the test of time?
- **Counterfactuality**: would this have appeared anyway within a year?
- **Actionability**: does a real decision change downstream?
- **Surprise**: does it challenge priors or confirm what everyone assumed?
- **Timing**: is it completed when it became possible or necessary?

What time horizon are you working towards?

Every project implicitly targets a future state of the world

Projects can optimize for

- Today's healthcare system vs. reimagined future healthcare system
- Available data vs. new data collection paradigms
- Near-term deployability vs. long-term transformation
- Publishable results vs. sustained impact

TODAY
Ambient documentation

FUTURE
Improving patient-clinician shared decision making

Let's Discuss!

- What kind of impact do you want your research to have?
- What time horizon does your work operate under?
- Which papers in the ML4H field do you think have had the most impact? Why?

Common Pitfalls in Problem Formulation

Failures in Problem Formulation & Grounding

KEY QUESTION

Is this a well-formed question grounded in the actual problem, setting, and people it's for?
for?

Projects fail when they:

Ask the wrong question

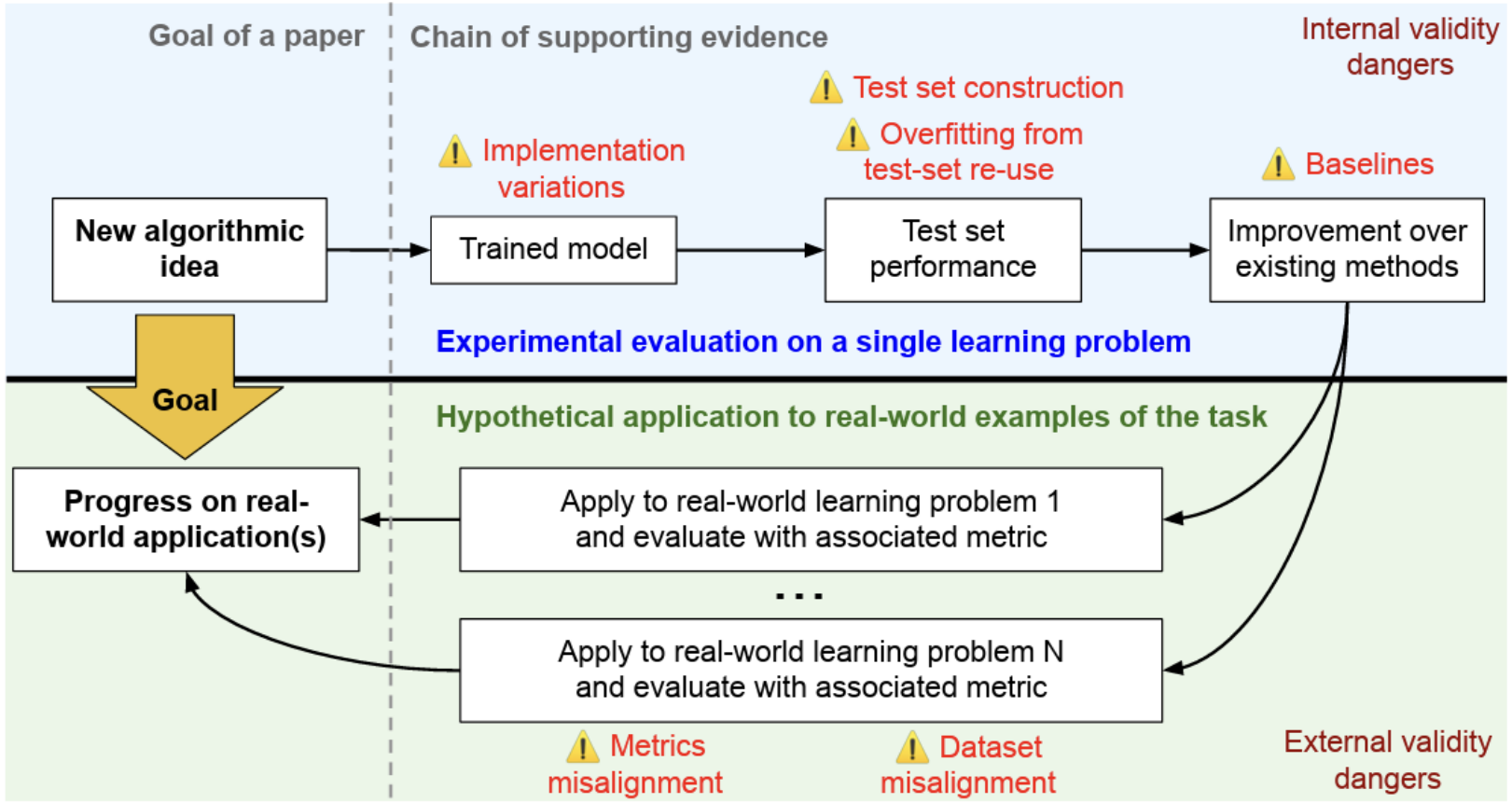
- Tackle a problem that isn't the true bottleneck
- Predict something clinicians already know
- Predict outcomes that aren't linked to concrete actions
- Optimize proxy targets rather than clinically meaningful goals

Don’t consider deployment

- Produce tools that do not fit stakeholder stakeholder workflows
- Ignore stakeholder incentives and implementation constraints
- Leverage data unavailable at deployment or unrepresentative of real-world settings

Failures in Problem Formulation & Grounding

Misalignment between benchmarks & real-world applications can lead to working on ill-posed problems

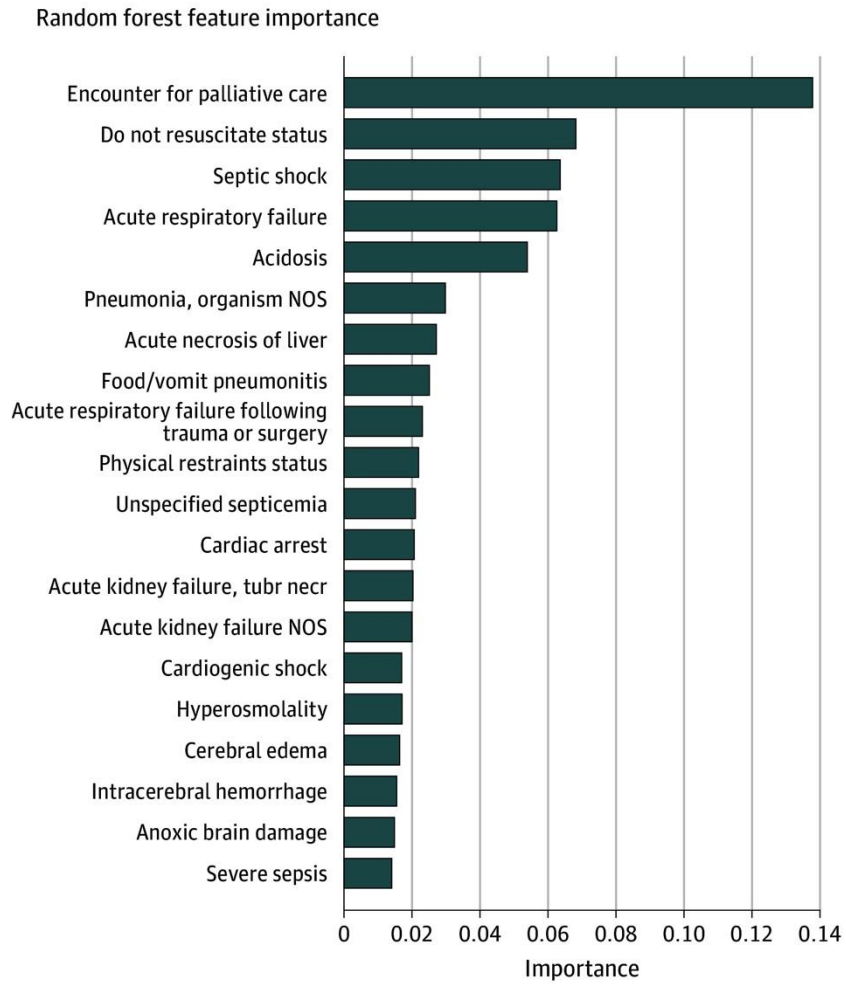


Liao et al. Are We Learning Yet? A Meta-Review of Evaluation Failures Across Machine Learning. NeurIPS 2021.

Failures in Problem Formulation & Grounding

Including variables that are not available at prediction time during deployment can overestimate model performance

One study found that 40.2% of published AI models trained to predict same-admission outcomes used ICD codes as features

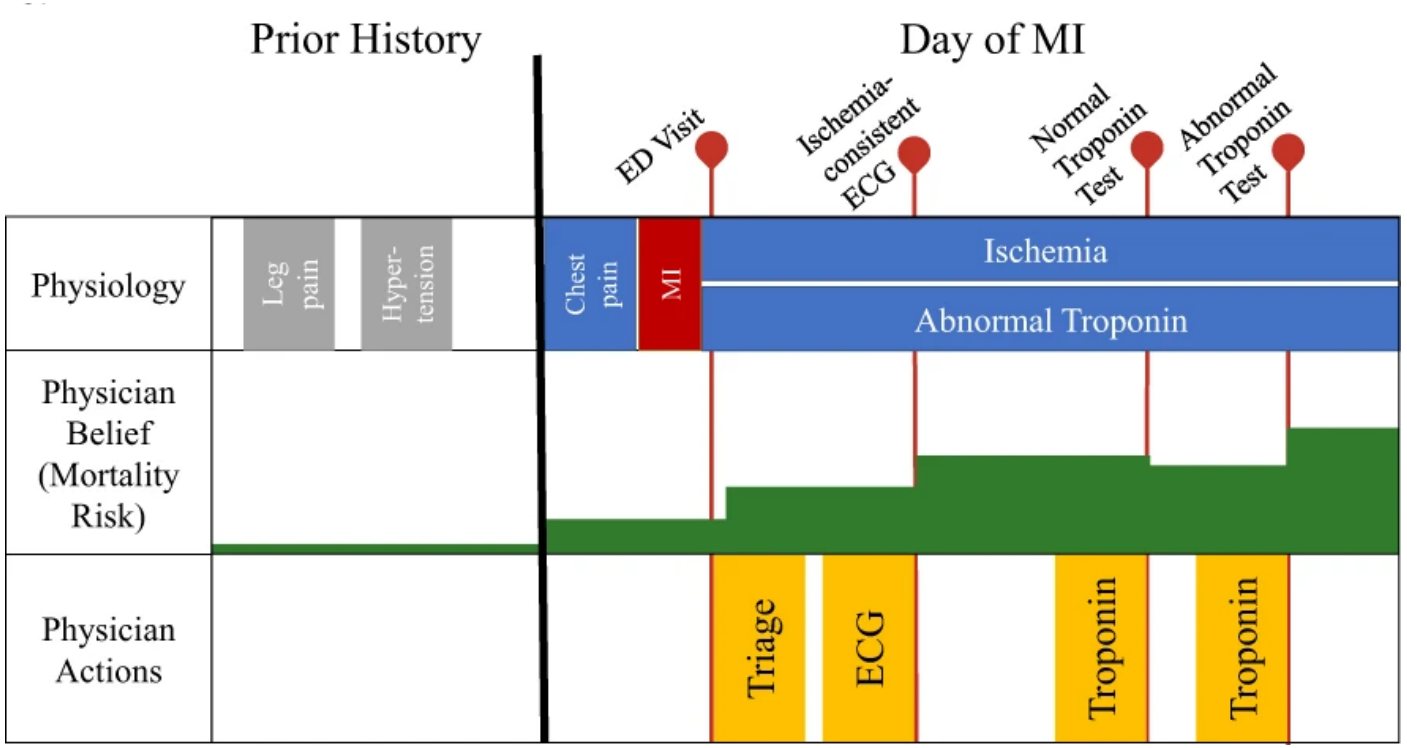


Ramadan et al. Diagnostic Codes in AI Prediction Models and Label Leakage of Same-Admission Clinical Outcomes. JAMA Network Open. 2025.

Failures in Problem Formulation & Grounding

Predict whether a patient will have a myocardial infarction based on data from the ED

Models trained on EHR data can learn to recognize clinical suspicion rather than contributing novel insights



Beaulieu-Jones et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? Npj Digital Medicine. 2021.

ARTIFICIAL INTELLIGENCE

Hundreds of AI tools have been built to catch covid. None of them helped.

Some have been used in hospitals, despite not being properly tested. But the pandemic could help make medical AI better.

By Will Douglas Heaven July 30, 2021

technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/
Wynants et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020.

Why?

Model developers made incorrect assumptions about the data, which meant that the trained models did not work as claimed.

Driggs highlights the problem of what he calls Frankenstein data sets, which are spliced together from multiple sources and can contain duplicates. This means that some tools end up being tested on the same data they were trained on, making them appear more accurate than they are.

It also muddies the origin of certain data sets. This can mean that researchers miss important features that skew the training of their models. Many unwittingly used a data set that contained chest scans of children who did not have covid as their examples of what non-covid cases looked like. But as a result, the AIs learned to identify kids, not covid.

Driggs's group trained its own model using a data set that contained a mix of scans taken when patients were lying down and standing up. Because patients scanned while lying down were more likely to be seriously ill, the AI learned wrongly to predict serious covid risk from a person's position.

In yet other cases, some AIs were found to be picking up on the text font that certain hospitals used to label the scans. As a result, fonts from hospitals with more serious caseloads became predictors of covid risk.

Failures in Significance

KEY QUESTION

If the research succeeds, what do we know that we didn't before and how much does that knowledge move the field?

Projects can have limited significance when they:

- Focus solely on building a model that beats other models on a benchmark benchmark
- Show an improvement that is too small to change anyone's conclusions
- Improve on a baseline no one uses
- Demonstrate something others can't reproduce or extend
- Over-optimize to a narrowly defined setting and fail to generalize

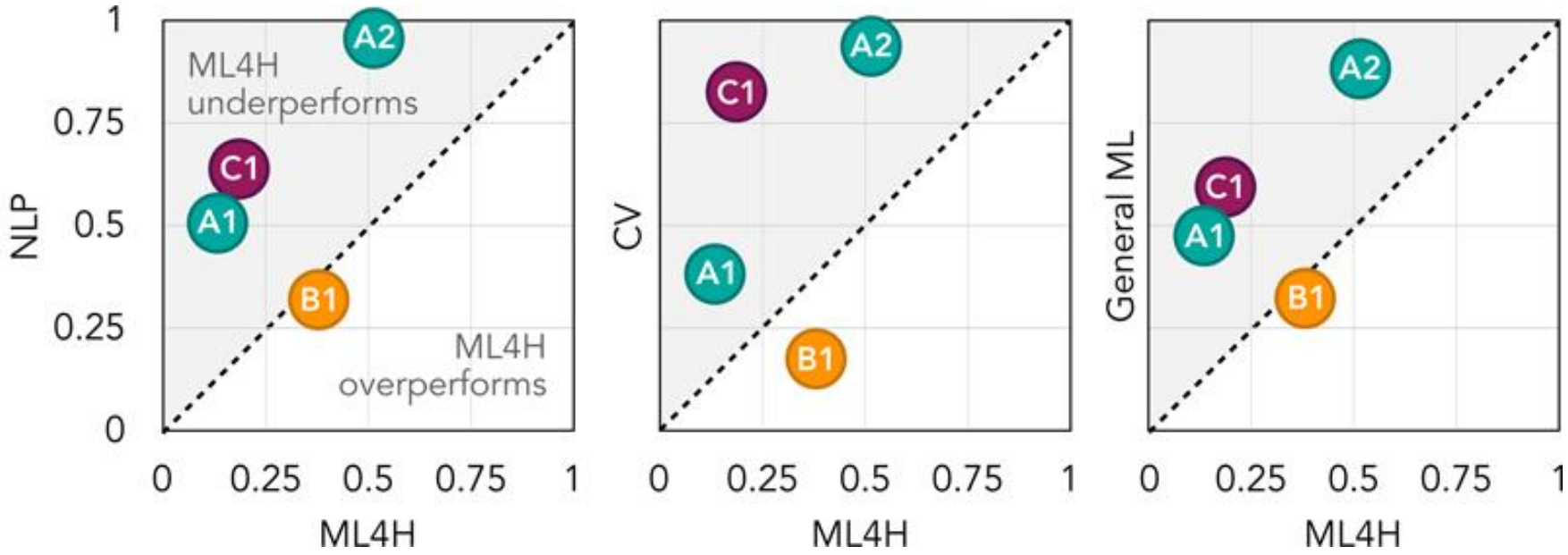
Failures in Significance

Evaluation Metrics:

A. Technical replicability
   1. Code available
   2. Public dataset

B. Statistical replicability
   1. Variance reported

C. Conceptual replicability
   1. Multiple datasets



Fraction of papers satisfying certain conditions by ML field

McDermott et al. Reproducibility in Machine Learning and Healthcare: How far do we have to go? Sci. Transl Med. 2021.

Failures in Timing

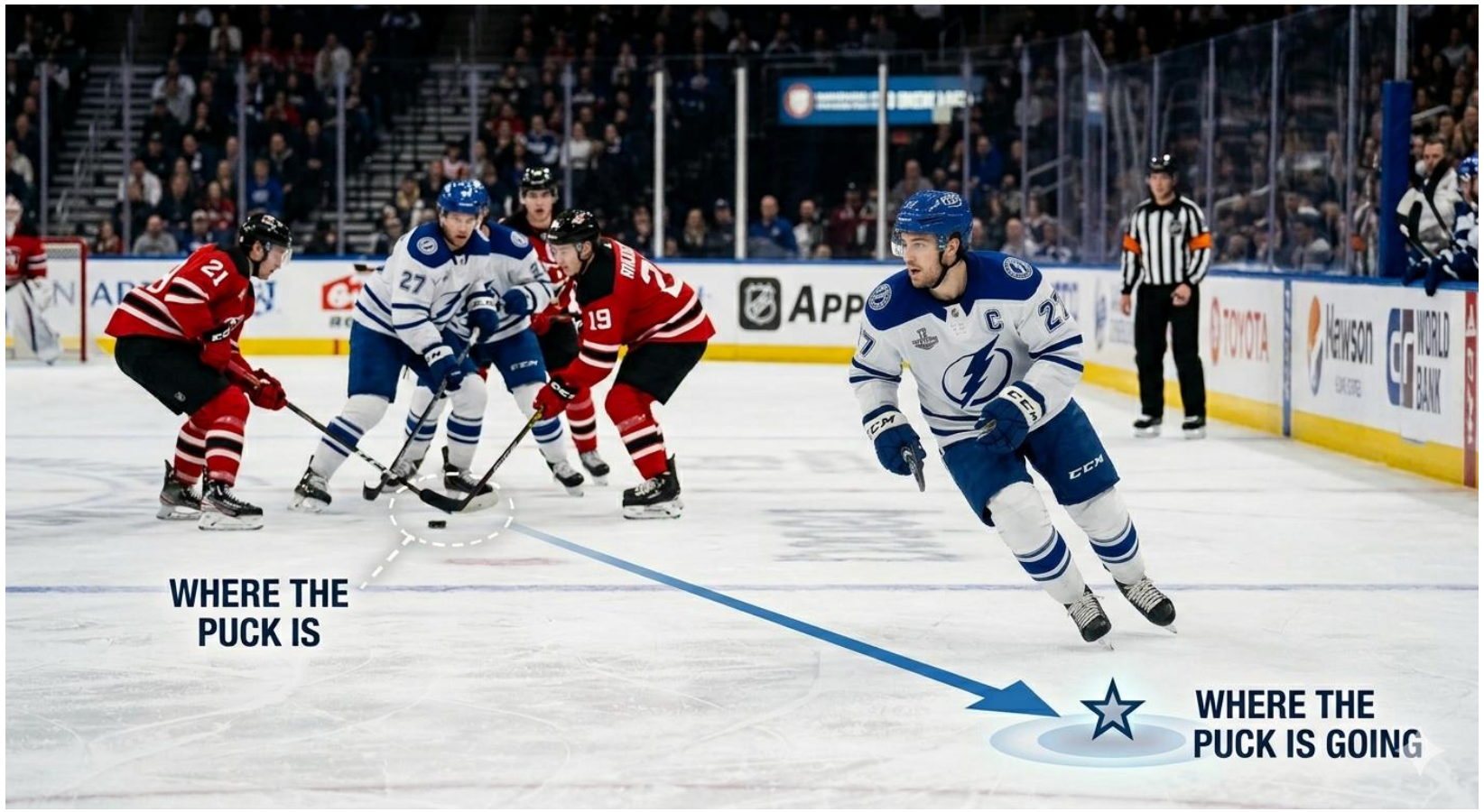
KEY QUESTION

Why now? Will the outcome still matter when you're done?

Projects fail when:

- A general advance will solve the problem as a byproduct before you finish
- An external change (e.g., new data source, guideline, regulation) obviates the question while it's in progress
- The contribution has a short shelf life. it won't hold or matter beyond a brief window
- Someone else completes your project before you do

Avoiding Failures in Timing



"Skate to where the puck is going, not where it has been." - Wayne Gretzky

Failures in Feasibility

KEY QUESTION

Can the project be completed now and can you do it?

Projects fail when they:

- Depend on data that doesn't exist, isn't accessible, or can't be obtained at the scale and quality needed
- Require methods that aren't mature or available yet
- Ignore why prior work failed
- Don’t align with your key advantages (data, knowledge, partnerships)
- Have no viable minimum version

Let’s Discuss!

- What kind of failure modes have your research projects been susceptible to?
- What's your unfair advantage (e.g., data, expertise, collaborators) and what questions does it uniquely position you to ask?

Questions to ask when beginning a new project

Project Planning Questions

Problem: What's the problem? Why is it important? Why is it challenging?

Status Quo: What has been done previously to address the problem?

Gap: What is the gap in existing solutions / approaches? Why hasn't this gap been addressed yet?

Innovation: What novel idea, approach, or insight are we proposing? Why is this innovation expected to overcome existing limitations or challenges?

Impact: What broader impacts will our solution have if successful? Who will benefit from this work, and in what ways? How will this solution influence future research or practice?

Project Planning Questions

Methods: How specifically will we implement or test this innovation? What methods or techniques will be used, and why are they appropriate?

Data: What dataset / cohort will we use to implement or test the innovation?

Expected Contributions: What are the anticipated outcomes or deliverables of this work? How will these contributions advance the field?

Evidence/Experiments Needed: What experiments or evidence are required to validate our approach? What criteria will we use to determine if we have succeeded?

Risk & Mitigation: What potential challenges or limitations can we anticipate? What contingency plans or alternative strategies are available? How will we assess whether these mitigation strategies are effective?

Example ML4H Projects

Readmission prediction

Failure to consider downstream action

The Problem

Hospital A's readmission rates are rising. Readmissions are disruptive to patients and caregivers, increase costs, and raise the risk of hospital-acquired complications. They are also financially penalized. CMS reduces Medicare reimbursement for hospitals with higher-than-expected 30-day readmission rates.

Initial Research Question

Can we predict 30-day readmission for patients discharged from hospital A?

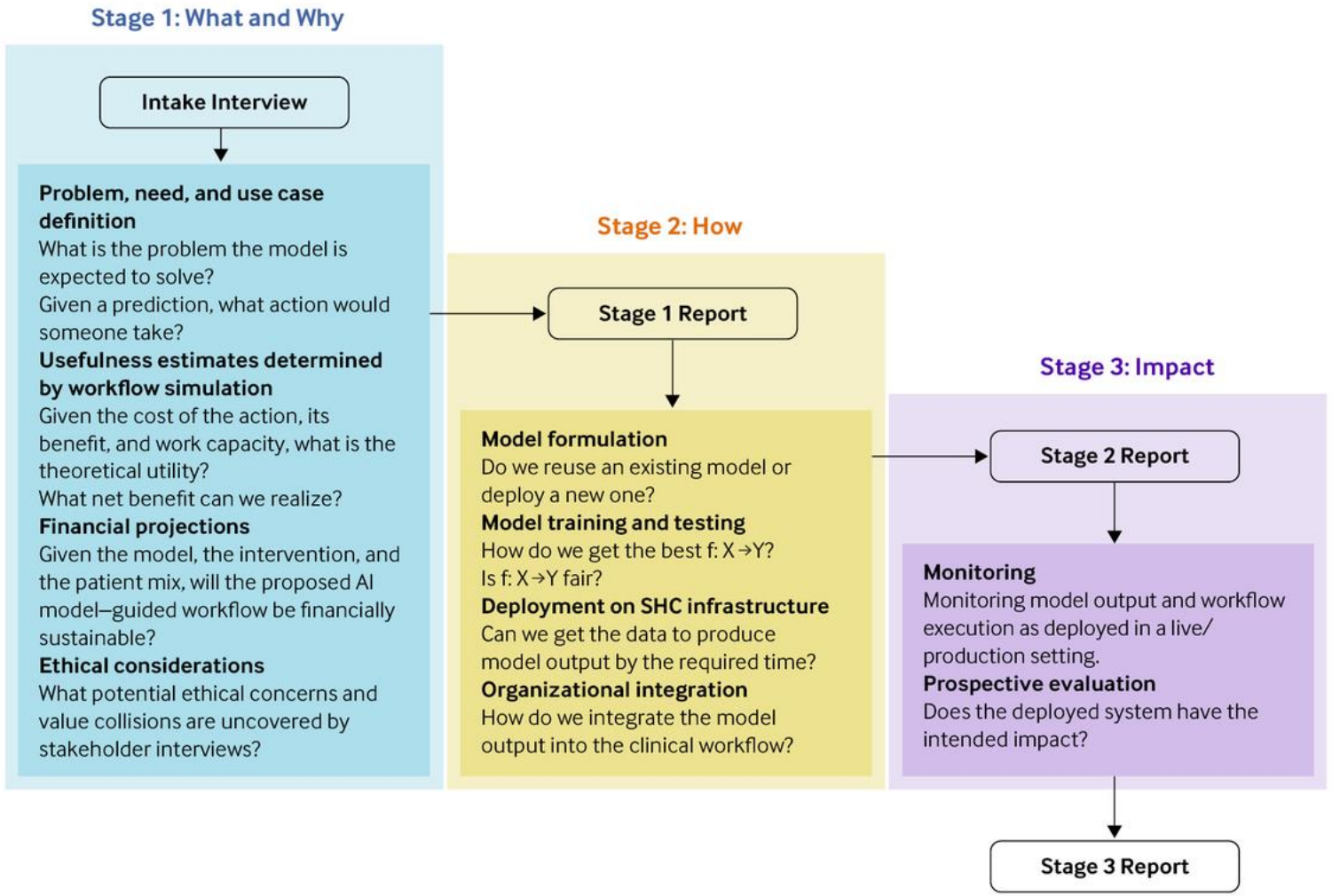
The Challenge

Preventing readmissions requires identifying which patients benefit from specific interventions

Revised Research Question

Among patients discharged from hospital A with markers of care fragmentation, can we predict 30-day readmission?

At Stanford, AI models undergo Fair, Useful, and Reliable Model (FURM) Assessments



- 21 FURM assessments conducted since 2022
- 16/21 models were deployed
- Example: Epic's V2 readmission model

Callahan et al. Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems. NEJM Catalyst. 2024.

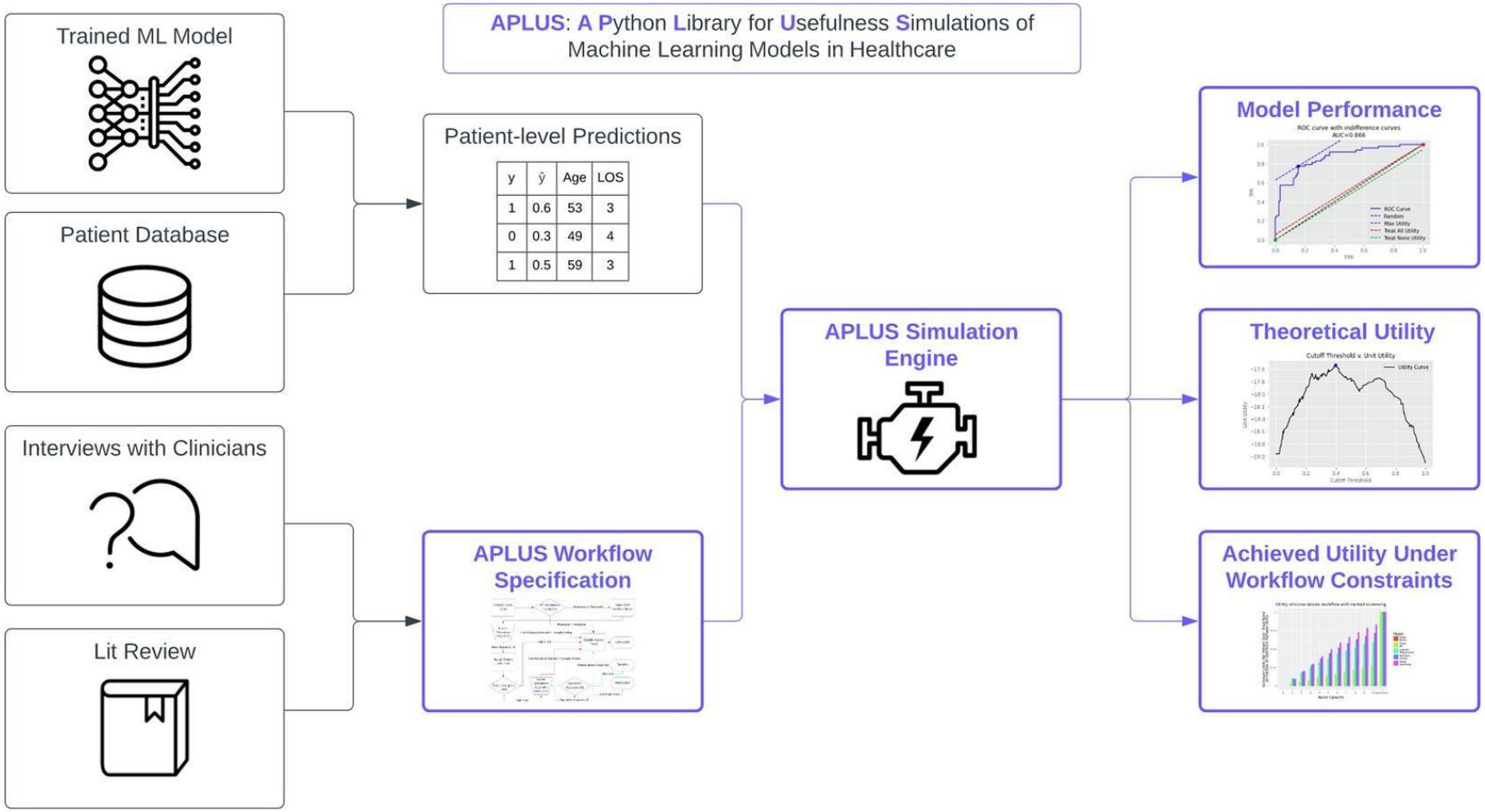
Key Stakeholders

Role	Name
Sponsors	Readmission Sponsors
Operational Stakeholders	• ACQO • CNIO • …
Clinical Stakeholders	• Director of Transitional Care • Assistant Manager of Employer Based Clinics • …
Technical Stakeholders	• Director, Inpatient Portfolio • Team Lead, IT Clinical Applications - Inpatient
Medical Informatics	• Inpatient & Ambulatory • …
Nursing Informatics	Sr. Program Manager

Intended use of a model in a workflow shapes evaluation

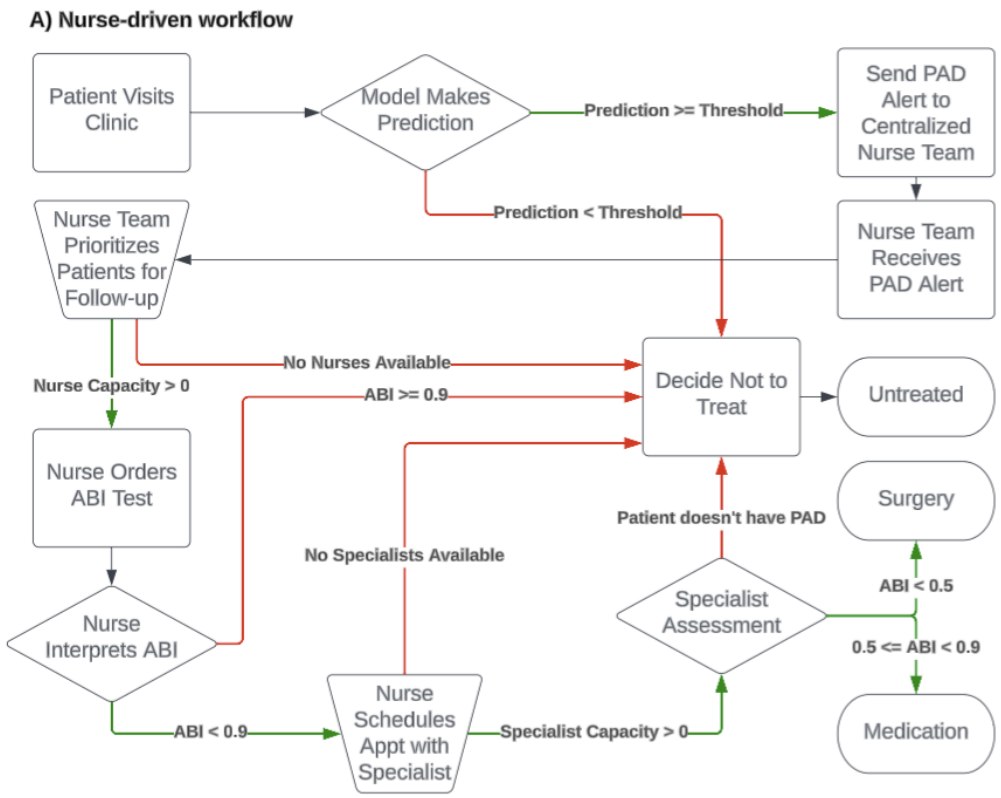
	Option 1	Option 2
Workflow	High risk patients receive 1:1 support from case manager	High risk patients are included in priority queue for follow up appointments
Metric	Precision at *k* Where *k* is determined by capacity of case managers	Recall at *k* Where *k* is determined by follow up throughput

Assessing usefulness via workflow simulations

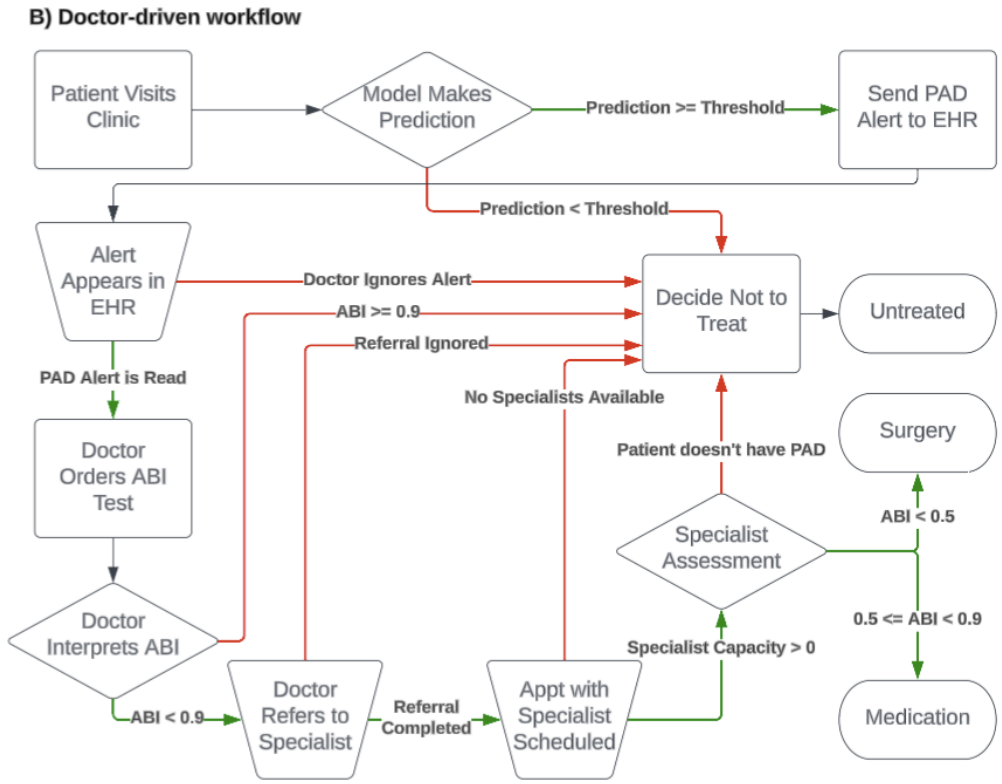


Wornow et al. APLUS: A Python library for usefulness simulations of machine learning models in healthcare. J Biomed Inform. 2023.

Assessing Nurse vs Doctor-driven Workflows for Peripheral Artery Disease (PAD) Screening



Centralized team of nurses reviewing the PAD model's predictions for all patients



PAD model's predictions appear as a real-time alert in a patient's EHR during their visit to the clinic

https://aplusml.readthedocs.io/en/latest/usage/tutorial_pad.html

Generating discharge summaries

Failure to understand clinical workflows

The Problem

Discharge summaries summarize the events that happen to a patient during an inpatient stay. Generating discharge summaries is time consuming for clinicians and can delay discharge.

Initial Research Question

Can we use LLMs to generate the full discharge summary right before the patient is discharged?

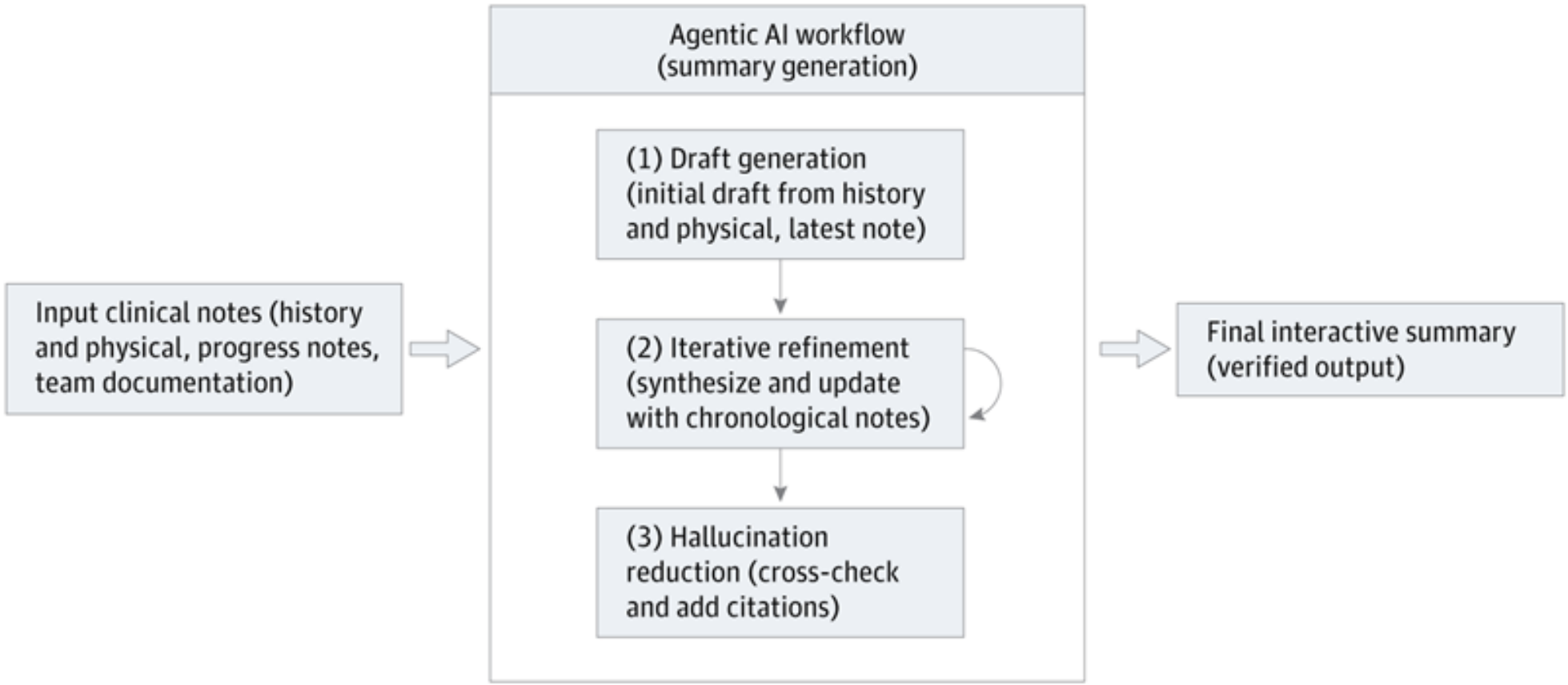
The Challenge

Most of the discharge summary is auto-generated from structured data! Clinicians draft the discharge summary throughout the patient's hospital stay

Revised Research Question

Can we use LLMs to generate the brief hospital course section of the discharge summary? Can we generate interim discharge summaries every day?

Generating Discharge Summaries



Grolleau et al. Physician-Reported Safety Outcomes of AI-Generated Hospital Course Summaries. JAMA Network Open. 2026.

Be wary of incumbents if you are aiming for deployment

Artificial Intelligence (AI)

Roadmaps give you the big picture about key features that are already available and our broad vision for where we're headed. Many of our unreleased features are still being designed and do not yet have a planned release date. It's possible features might take a different approach than originally described here, might not be released for a while, or might be deprioritized.

Want to learn more? Related Documents

Ready for Broad Adoption

Staying Current (with Upgrade)

- In Basket Art: Draft Responses to Patient Messages
- Outpatient Insights
- AI Text Assistant
- Draft Hospital Course in Discharge Summary
- Actionable Follow-Ups for Radiology
- Ambient Voice Documentation for Outpatient Notes
- Patient-Facing Inpatient Instructions
- Draft Answers for Prescription Prior Authorizations
- Automatic Adenoma Detection Rate
- Automatic Sessile Serrated Lesion Detection Rate
- Extracted Sig Details
- Discrete Data Extraction for Synoptic Forms
- Transplant Evaluation Insights
- Anesthesia Pre-Procedure Insights
- Preprocedural Screening Insights
- Dialysis Rounding Insights
- Draft Invasive Procedure Narrative

Staying Current (Generally Available)

- Ambient Voice Recognized Ordering
- Ambient Voice Documentation for Inpatient Notes
- Ambient Voice Documentation for Problem-Oriented Charting
- Inpatient Insights
- In Basket Art: Draft Patient-Facing Result Comments

Pioneering

Pioneering (Includes Fit Considerations)

- Ask Art
- Ambient Voice Documentation for Nursing
- Synopsis Insights
- Draft Patient Instructions from Notes
- Extract Social Drivers of Health from Notes
- ED Facility Charge Edge Case Detection
- Post-Procedure Diagnosis Suggestions
- Cancer Staging Extraction from Notes
- Draft Therapy Progress Note
- Initial Patient Presentation Insights
- Actionable Follow-Ups for Structural Heart Patients
- Prior Study Comparisons
- Nursing Assignment Agent

Available for Early Feedback to Limited Sites

- In Basket Art: Translations
- Context-Driven In Basket Insights
- Ambient Voice Documentation for OASIS Forms

In Development and Future Plans

In Development

- Visit Agenda Agent
- Ambient Voice Documentation for Tooth Charting
- Ambient Voice Documentation for Orthopaedic Clinic Procedures
- Ambient Voice Documentation for Endoscopic Findings
- Ambient Voice Documentation for Skin Exams
- Ambient Voice Documentation for Home Health and Hospice
- Suggested Encounter Actions for Patient Messages
- Colorectal Cancer Screening Follow-up Agent
- Diagnosis Checker
- Detect Sensitive Content in Notes
- Extract Orders and Tasks from Notes
- Draft Translations for AVS Instructions
- Medication Usage Insights
- Medication Verification Insights
- Oncology Insights
- Pregnancy Insights
- Surgical Consult Insights
- Surgical Case Length Estimation
- Patient-Reported Outcome Insights
- Group Session Insights
- Extract Treatment History for Behavioral Health Notes
- "How do I..." Search in Epic Assistant
- Chronic Care Management Insights

https://roadmap.epic.com/roadmaps/839_CognitiveComputingGenAI

Predicting Postpartum Hemorrhage

Failure to align prediction task with meaningful clinical actions

The Problem

PPH is the leading cause of maternal death worldwide, affecting over million women annually and accounting for ~27% of maternal deaths PPH rates are rising at Stanford.

Initial Research Question

Can we predict which patients will experience PPH ***one hour prior to delivery?***

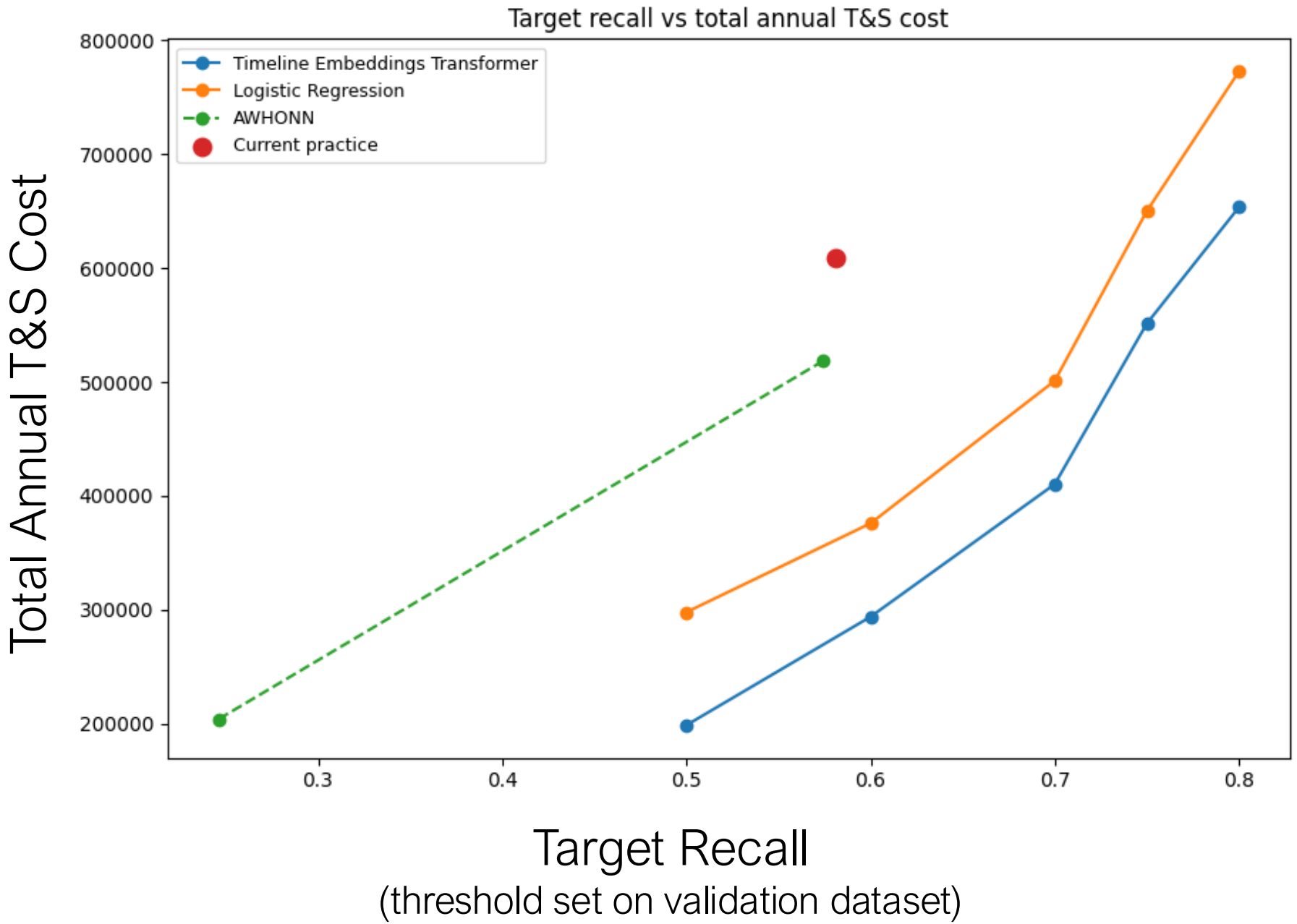
The Challenge

Most interventions only help *after* PPH onset. There's limited evidence for for prophylactic drug administration, though more targeted type & screen screen offers some cost savings.

Revised Research Question

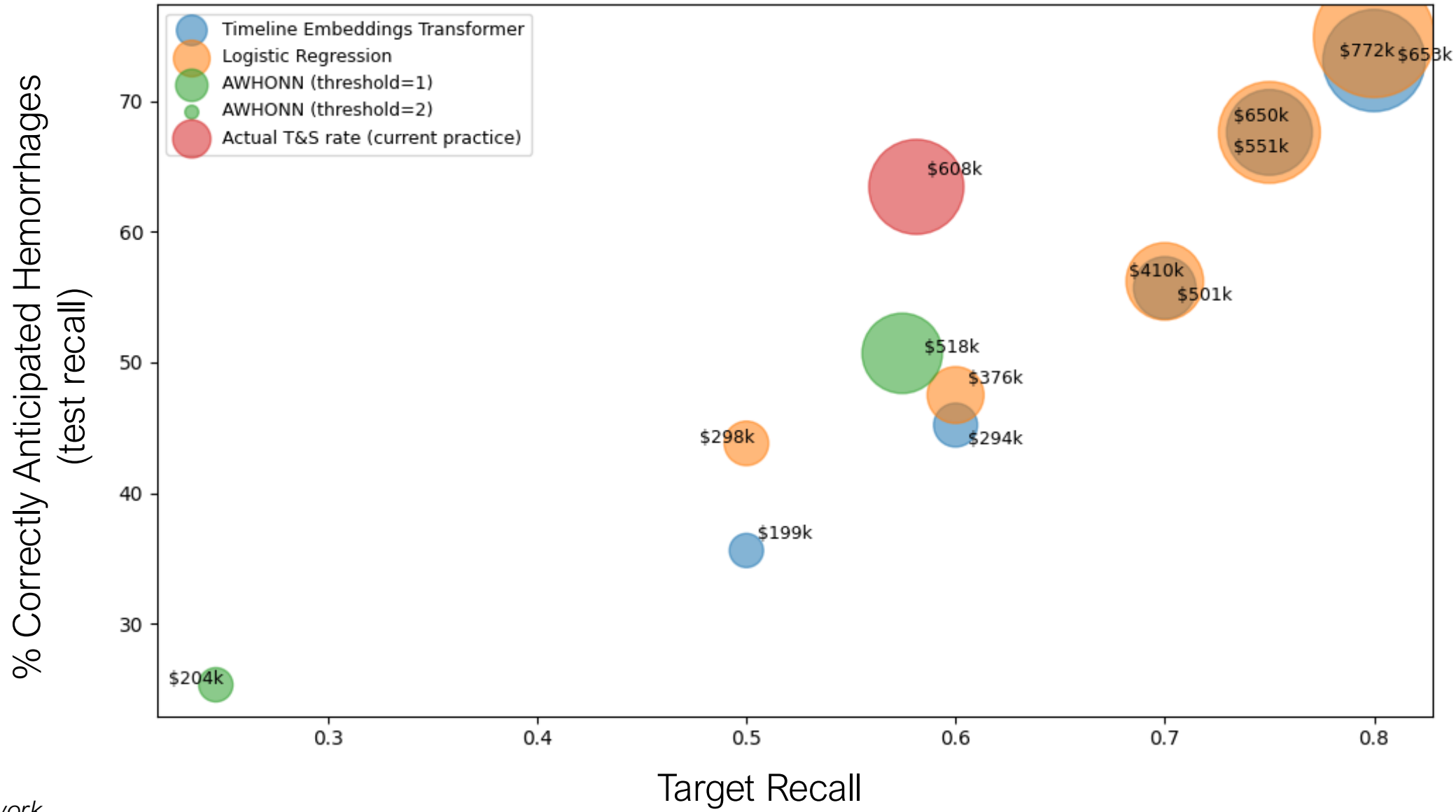
Can we predict PPH **throughout the pregnancy course?** Can we identify the factors driving rising rates?

AI models seem to be more cost effective?!



unpublished preliminary work

Problem: Target recall doesn't hold on the test dataset



unpublished preliminary work

Information extraction from clinical notes

Failure to imagine future state

The Problem

Substantial information is only found in clinical notes. Extracting this information in a standardized way can enable downstream real world evidence research and clinical decision support.

Initial Research Question

Can we develop tools to extract information at scale from longitudinal clinical notes?

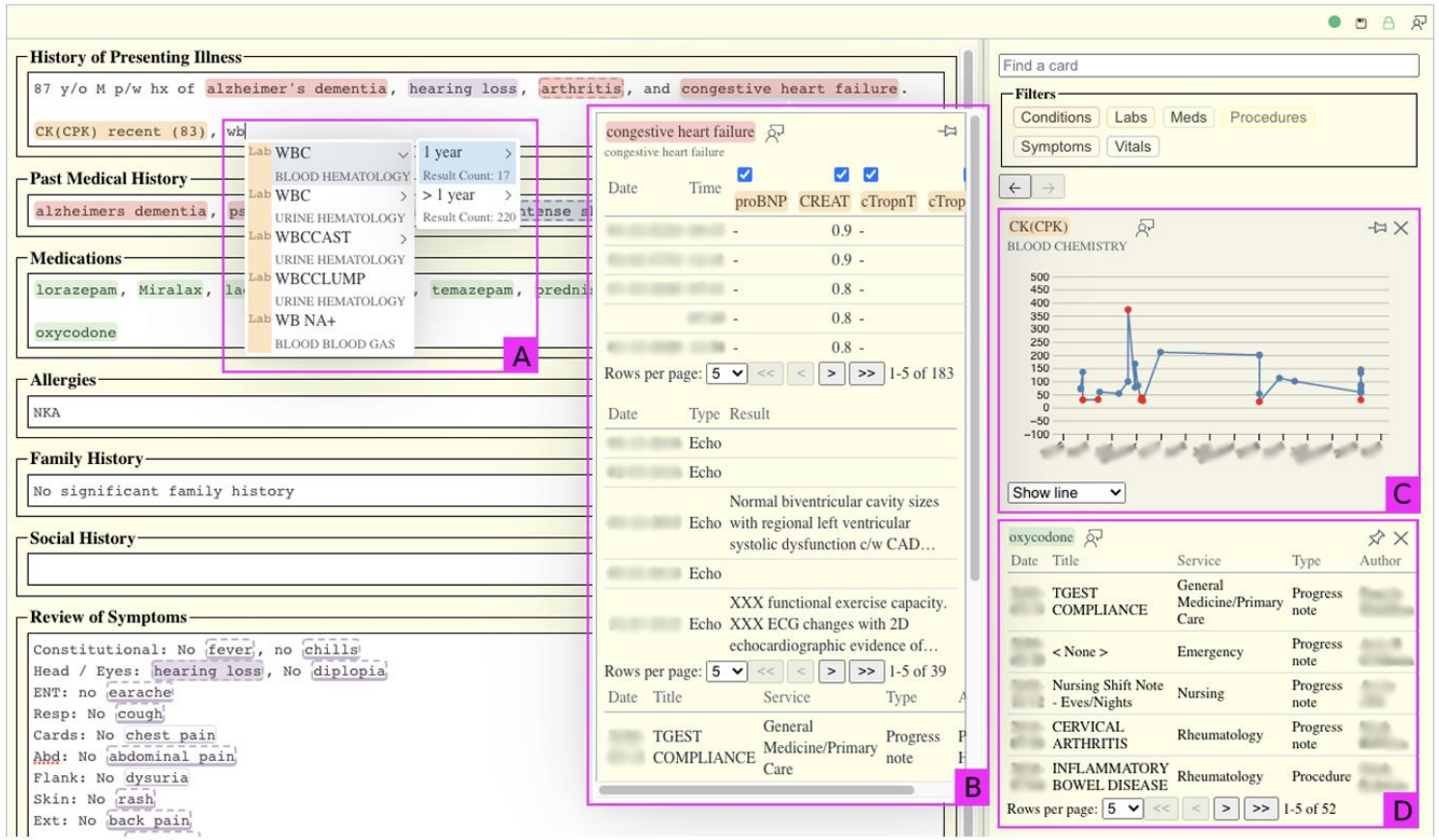
The Challenge

Research question is designed as a band aid to current EHR workflows, rather than reimaging better workflows

Revised Research Question

Can we redesign EHR systems to automatically collect this structured data at data entry?

MedKnowts: Unified Documentation and Information Retrieval for Electronic Health Records



Murray et al. MedKnowts: Unified Documentation and Information Retrieval for Electronic Health Records. UIST 2021.

Rare Disease Diagnosis

Failure to identify true bottleneck

The Problem

There are ~300 million individuals with a rare disease. Diagnosis is extremely challenging. There is substantial heterogeneity in clinical presentations, and clinicians have limited experience due to the low prevalence of each disease

Identifying the molecular diagnosis requires integrating multiple sources of evidence to sort through thousands of candidate variants

Initial Research Question

Can we leverage AI to shorten the diagnostic odyssey for rare disease patients?

The Challenge

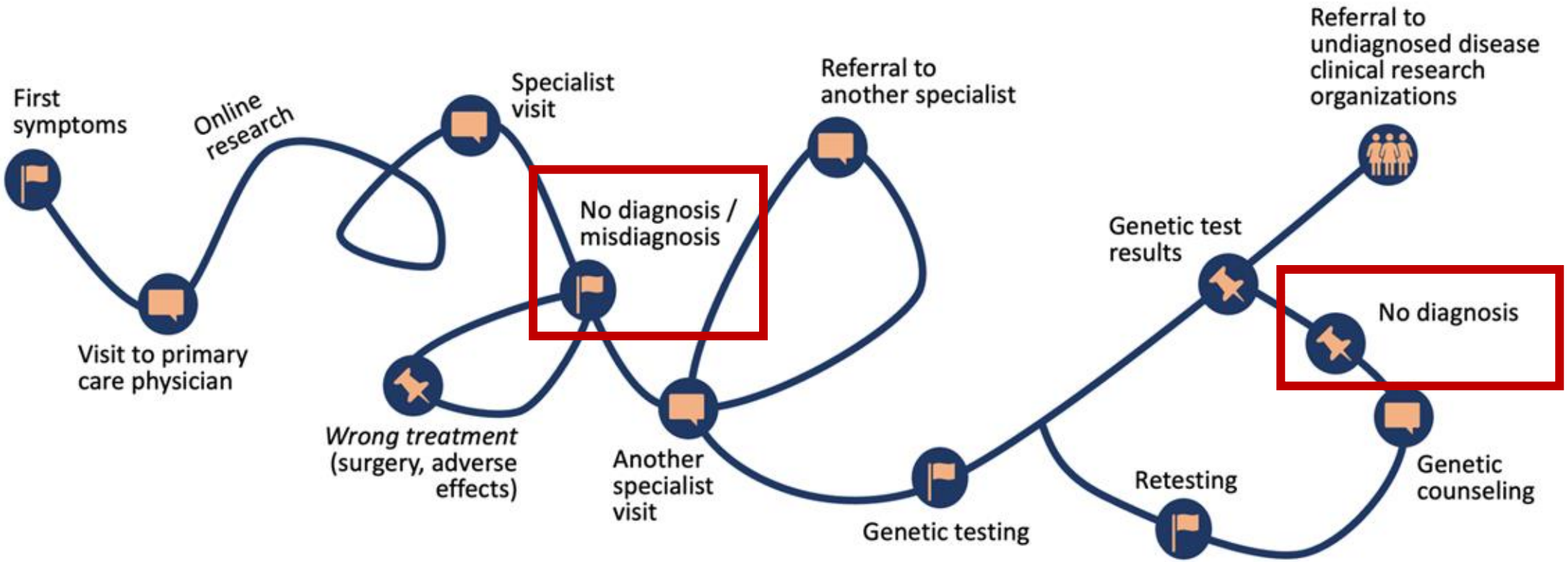
Where in the diagnostic odyssey to intervene? What data is available at that time?

Revised Research Question

Can we identify patients with undiagnosed rare diseases who need genetic workup? Can we identify “causal” variants for patients with genetic testing but no confirmed diagnosis?

Rare Disease Diagnostic Odyssey

Where is the true bottleneck to diagnosis?

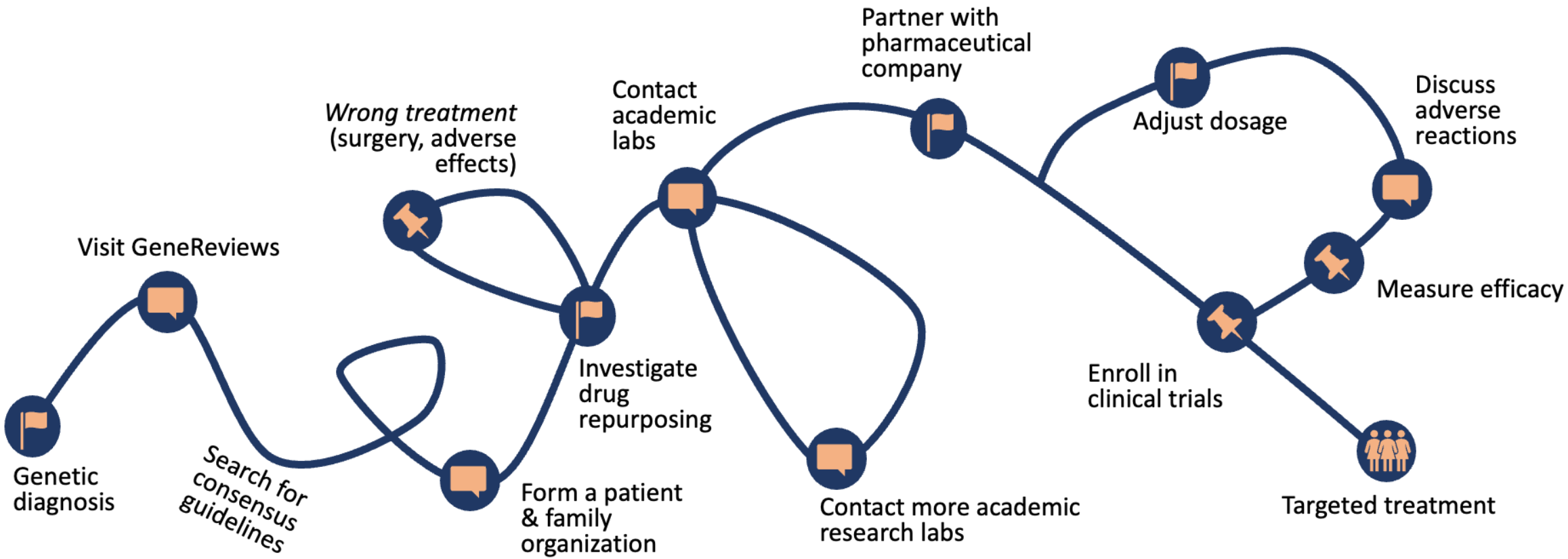


Adapted from The Marfan Foundation

Average time to diagnosis: ~5-7 years (if ever)
Number of specialists seen: >7
Diagnostic costs per patient: $100,000+ (including misdiagnoses and unnecessary testing)
Emotional and psychological toll on patients and families

Slide from Sam Finlayson

Rare Disease Therapeutic Odyssey



Therapeutic options for rare diseases are limited.
Targeted treatments exist for ~5% of rare genetic diseases.*

*LAM Braga, CGC Filho, FB Mota (2022) Ther Adv Rare Dis.
Slide from Sam Finlayson

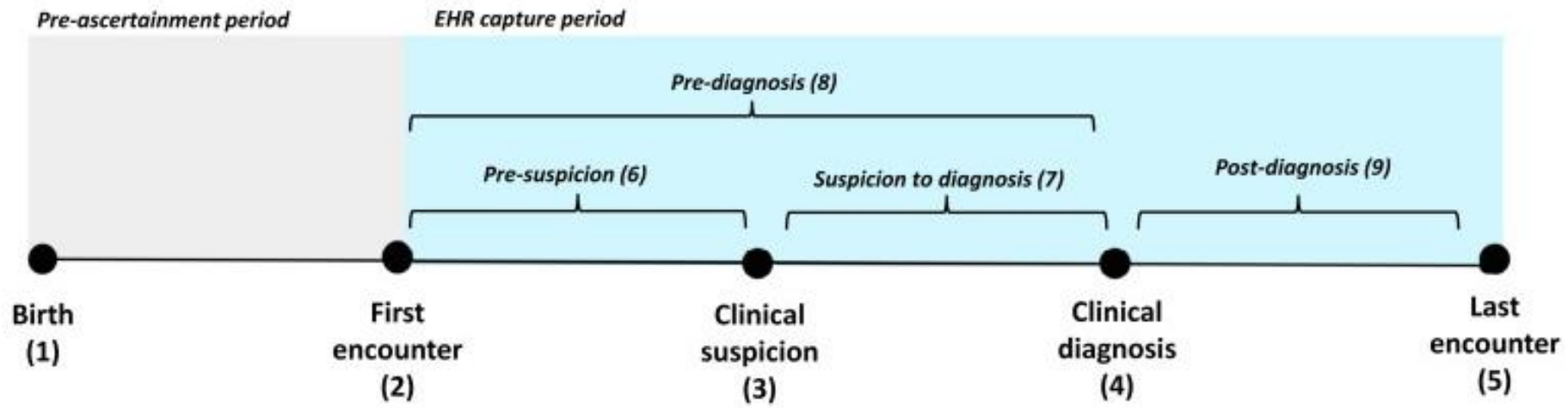
Make sure your model setup aligns with the problem formulation

Goal: Identify patients with undiagnosed rare diseases before they are referred to a specialist

Initial Setup: Train a model that takes longitudinal EHR data as input and outputs whether a patient has a particular rare disease *d*

Challenge: Model learns features from clinical suspicion period. Many rare diseases have the same downstream action.

Revised Setup: Train a model that inputs data prior to diagnosis and outputs whether a patient needs a genetic testing workup



Tinker et al. Phenotypic presentation of Mendelian disease across the diagnostic trajectory in electronic health records. Genet Med. 2023

Leveraging synthetic data for model training training

Failure to leverage methods with principled motivation

The Problem

Many clinical tasks have limited data available for model training. Synthetic data can improve the performance and robustness of models

Initial Research Question

Can generative models (e.g., GANs) be trained on our dataset to generate synthetic samples that augment the training data and improve model performance?

The Challenge

A GAN does not introduce new relationships beyond those present in the training data. Training on its outputs may act as a form of regularization, but it does not add new information.

Revised Research Question

Can external knowledge be incorporated into a generative model to produce synthetic data beyond what is contained in the original dataset?

How do your beliefs about the future state of the world shape what you focus on?

Hypothesis: Clinical data will remain siloed

Research: Focus on federated learning or model merging approaches

Hypothesis: General domain models will quickly be able to perform your clinical tasks without much domain specialization

Research: Focus on evaluation, deployment, human-AI interaction

JUNE2026

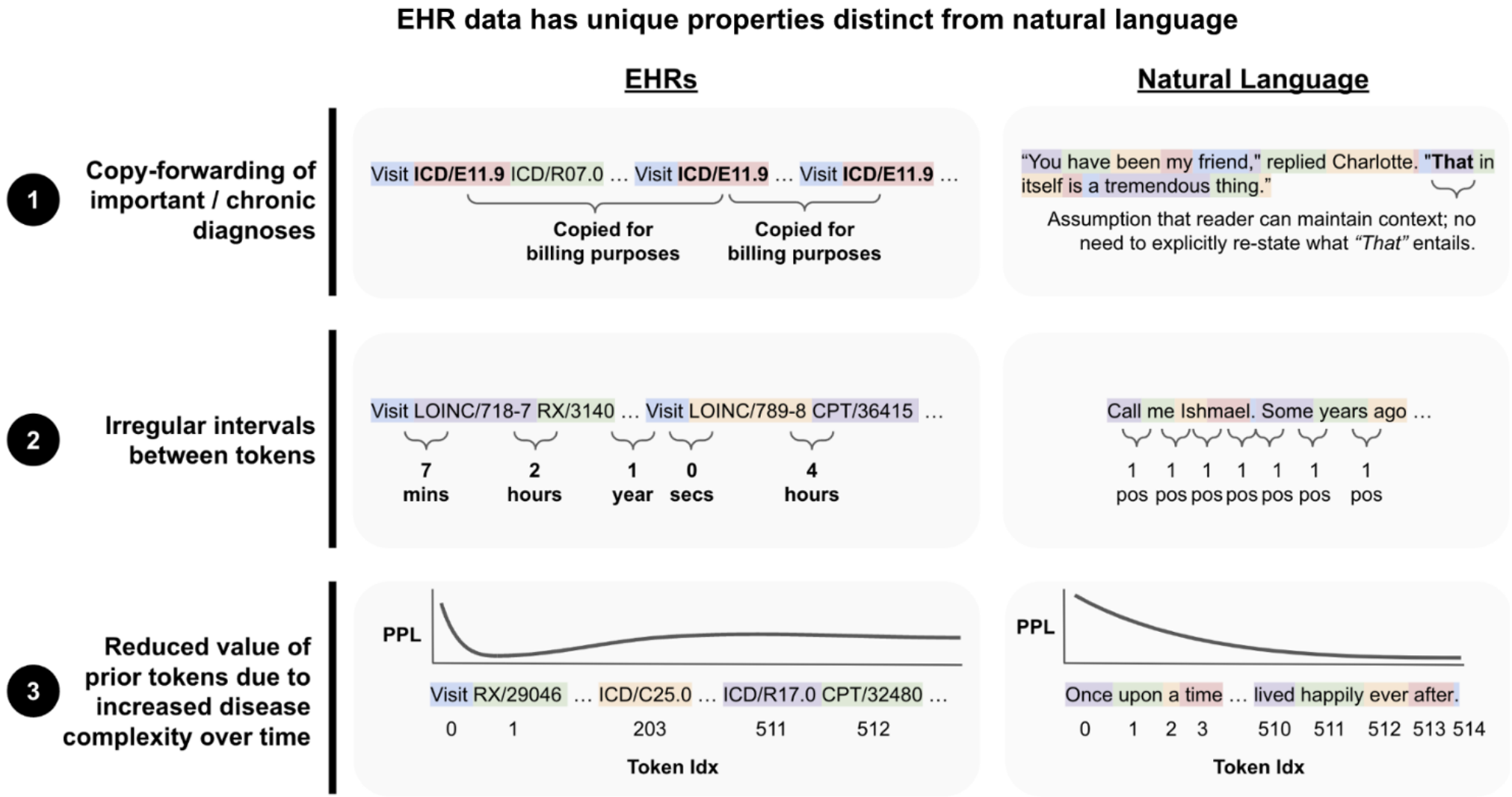
ARTICLE

Domain-Specific AI Should Focus on Workflows Rather Than Modeling

GENERATIVE AI HEALTH

I spent the past few years working on AI for health. Starting with custom multimodal encoders, post-training, and sophisticated multi-agent architecturs, I now see modeling work becoming less and less important for domain-specific applications. I believe that researchers should embrace this change and focus on owning the application, workflows, and evals instead.

Having worked on AI for health in the past few years, contributing to Med-Gemini and AMIE, I have witnessed a quite significant shift in the type of research that is required. Typically, domain-specific AI research such as AI for health needed a few years to adopt state-of-the-art models. Of course there are some notable exceptions where adoption happened in the reverse direction, such as U-Net. But usually, modeling research resulting in new architectures or methods was primarily done on general computer vision or natural language processing tasks before slowly getting adopted in health or other domains.



Wornow et al. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHRs. ICLR, 2025.

Let’s Discuss!

- What failure modes do you commonly see in the ML4H literature?
- What training do ML4H researchers need to avoid these common pitfalls?
- What bets are you making about the future in the work you are pursuing?

The Elephant in the Room: Incentives

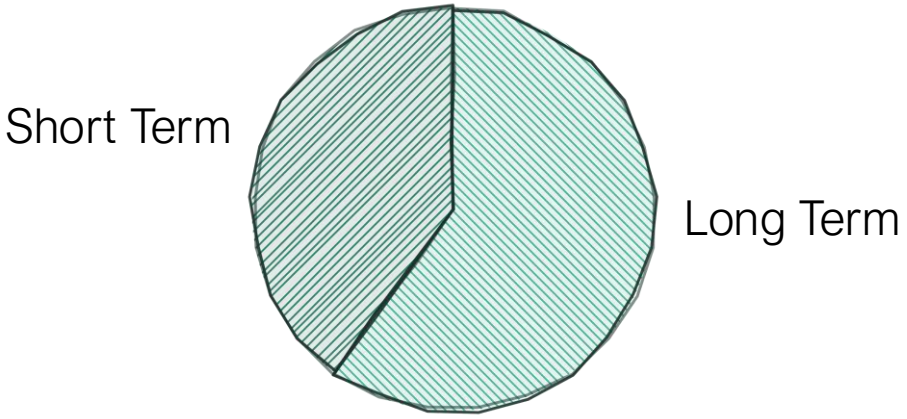
The reward structure of ML/CS publishing optimizes for:

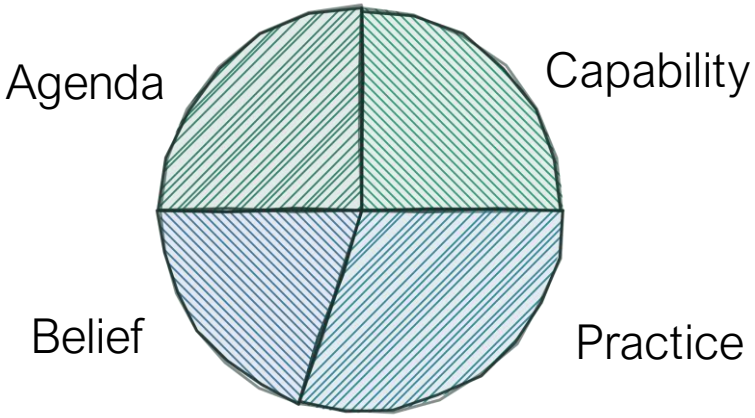
- Novelty over correctness
- Speed over depth
- Benchmarks over downstream impact
- Complexity over simplicity

Be cognizant of whether you are producing work that aligns with the impact you want to have

Formulating a Research Agenda

Cultivating a diverse research agenda with a mix of short-term and longer-term projects can balance immediate relevance with the possibility of larger future impact





Other Resources

- "You and Your Research" by Richard Hamming
- "Research Taste Exercises" by Chris Olah
- "My Research Process: Understanding and Cultivating Research Taste" by Neel Nanda

Questions?

ealsentzer@stanford.edu

I’m hiring postdocs!