



Palliative & Advanced Illness
Research (PAIR) Center

AI Diagnostic Decision Support for Older Adults in Primary Care

UCSF CODEX Diagnostic Excellence Webinar Series
August 27, 2025

Gary E. Weissman, M.D., M.S.

Assistant Professor of Medicine and Informatics
Core Faculty, Palliative and Advanced Illness Research (PAIR) Center

Disclosures

Support for this project:

- National Academy of Medicine Scholars in Diagnostic Excellence Program
- John A. Hartford Foundation
- Gordon and Betty Moore Foundation
- Council of Medical Specialty Societies (CMSS)

Past project support:

- Penn AlTech Pilot Award (NIH/NIA P30AG073105) for early project development

Other support:

- NIH (NHLBI, NIGMS)
- ARPA-H

Study Overview

Problem

- Outpatient diagnostic error rate ~5% (12 million Americans annually)
- Most diagnostic errors are due to ordering diagnostic tests or breakdowns in history taking and/or physical exam

Objective

Promote diagnostic excellence through AI-facilitated wayfinding in primary care for older adults with a safe, effective, and equitable open-source tool:

INTERLACE (diagNoSTic Excellence foR oLder Adults in primary CarE)

Specific Challenges for Diagnostic Excellence

- Older adults
 - Complexity
 - Frailty
 - Distinct clinical presentations
 - Distinct diagnostic considerations
- Primary care
 - Broad Scope
 - Lack of gold standard labels
 - Team of stakeholders includes patient, caregiver, and clinician

Lessons for Diagnostic Excellence Work

We all are, have been, or could be patients at risk of a diagnostic error

- Diagnostic excellence work is personal

Patient Stories with Dr. Diana Cejas

December 17, 2021



00:00 / 29:05

NAM Scholar in Dx Ex, Dr. Linda Geng, interviews Dr. Diana Cejas about her 5-year cancer misdiagnosis and how the delay led to further complications. Dr. Cejas also shares how her experience of doctors not listening to her concerns has changed the way she interacts with patients and her recommendations to improve the diagnostic process.

Lessons for Diagnostic Excellence Work

Setting and population focus, not disease focus

- How can we build a tool that supports the diagnostic process in primary care (setting) for older adults (population)?

DE GRUYTER

Diagnosis 2024; aop



Gary E. Weissman*, Laura Zwaan and Sigall K. Bell

Diagnostic scope: the AI can't see what the mind doesn't know

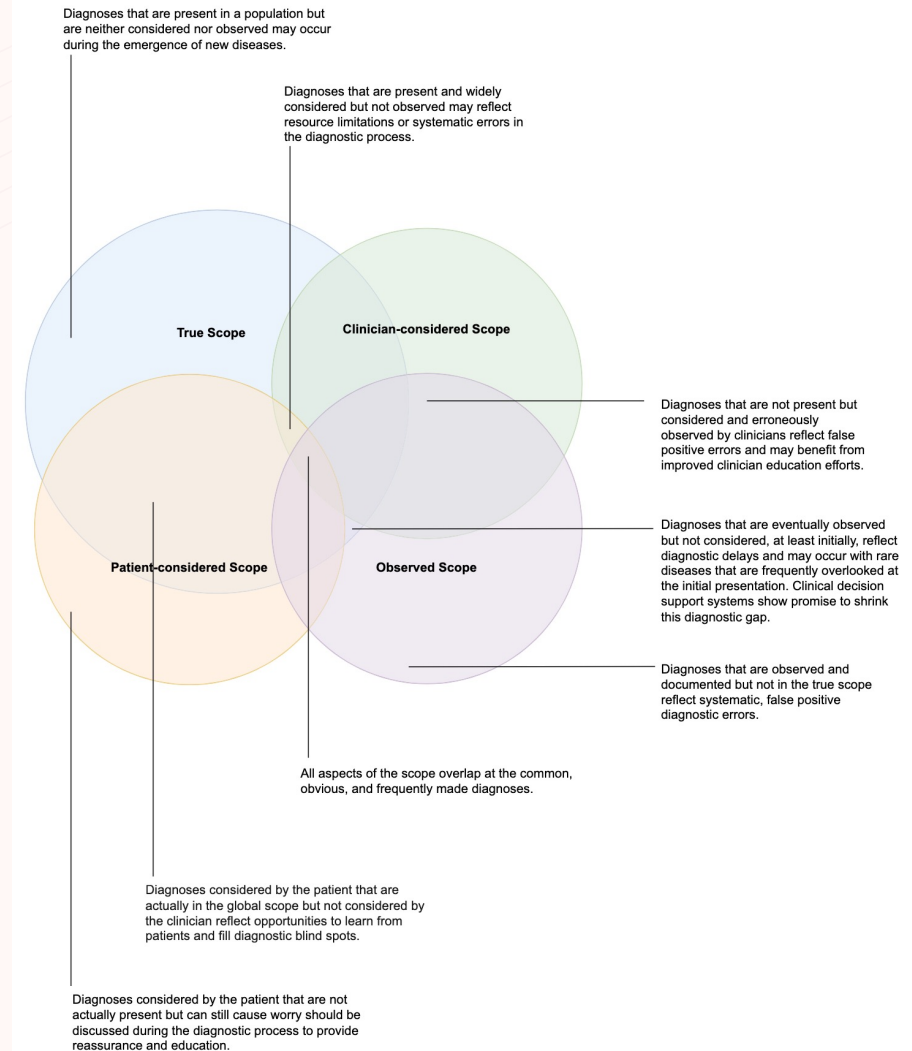


Figure 1: The true diagnostic scope includes the full range of diagnoses present in a particular clinical setting. Patients and clinicians in a clinical setting each may consider a distinct range of diagnoses. The observed diagnostic scope reflects all documented diagnoses, right or wrong, in a specific setting. Each aspect of the diagnostic scope both overlaps and diverges from the others. Importantly, the figure is not drawn to any particular scale.

Lessons for Diagnostic Excellence Work

Lack of gold-standard labels

For a supervised learning process, how do we acquire a “true” label for thousands of diagnoses and tests for hundreds of thousands of primary care encounters to train a large deep neural network model?

Annals of Internal Medicine

IDEAS AND OPINIONS

Chess Lessons: Harnessing Collective Human Intelligence and Imitation Learning to Support Clinical Decisions

Gary E. Weissman, MD, MSHP; Lyle H. Ungar, PhD; and Scott D. Halpern, MD, PhD

Lessons for Diagnostic Excellence Work

Clinically relevant, early phase studies

- Traditional measures of model predictive performance don't equate to even potential clinical effects
- Some signal of safety, acceptability, and/or appropriateness are needed to justify deployment or equipoise for a clinical trial



Clinician Turing Test: Novel Phase 1b Study Design

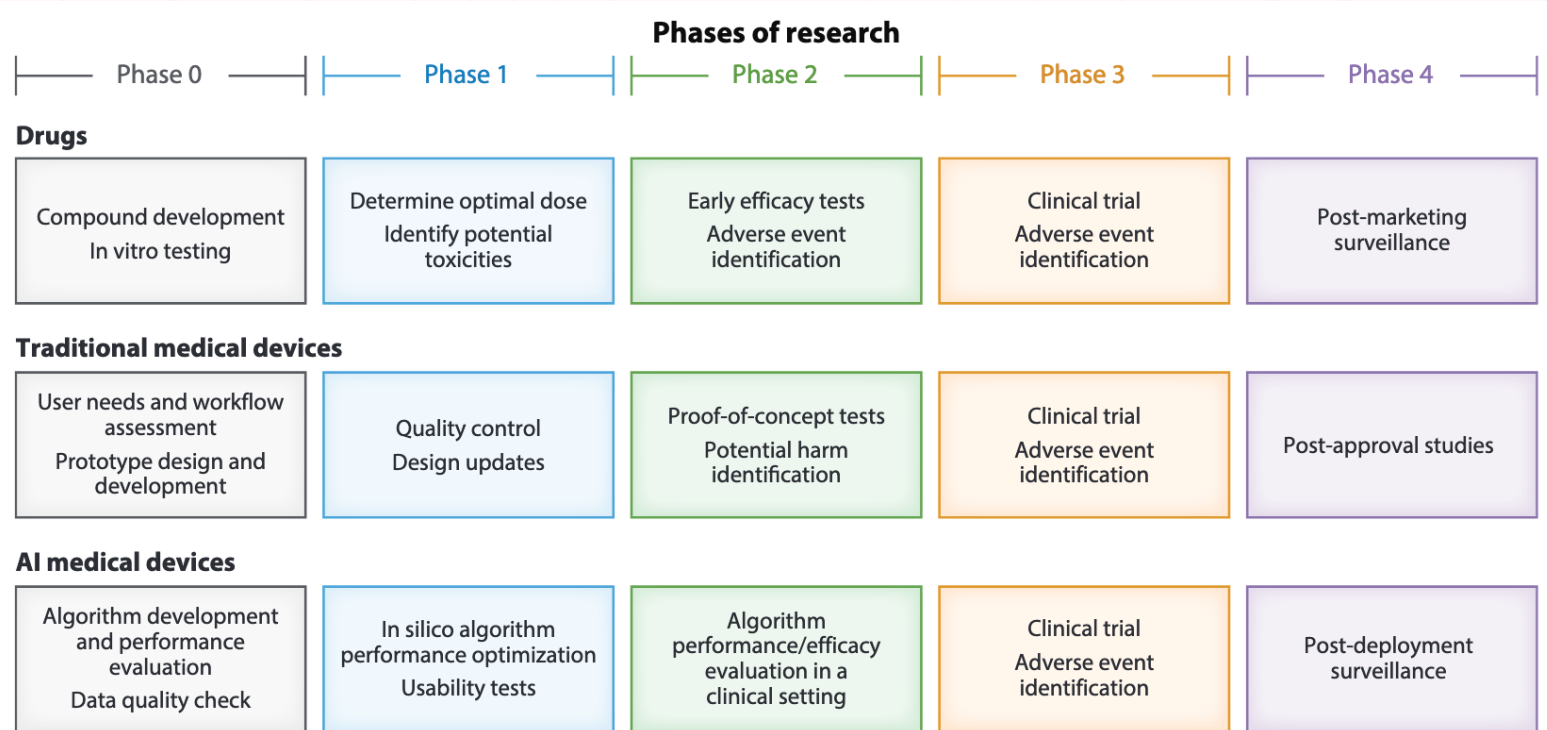


Figure 1

Phases of research and exemplar studies at each phase in the development of drugs, traditional medical devices, and artificial intelligence (AI) medical devices. Because AI medical devices are relatively new, the clinical relevance of the distinction between phases of research may be less familiar. Figure adapted from Reference 75 (CC BY 4.0).

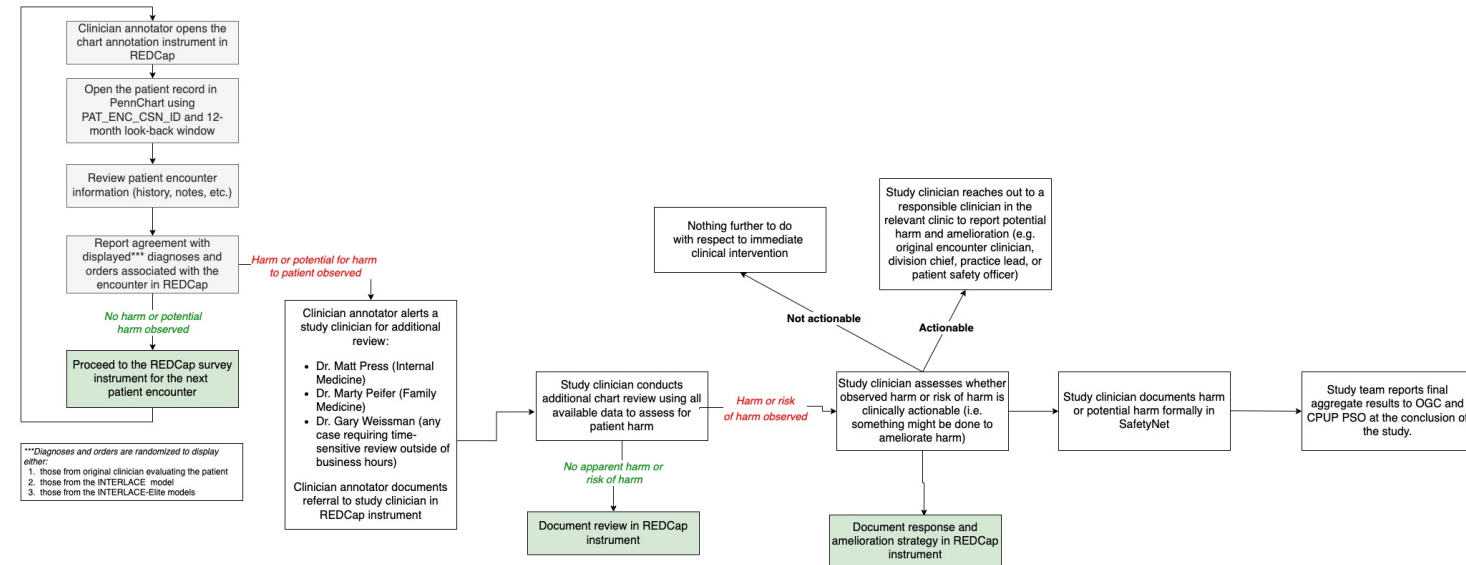
Lessons for Diagnostic Excellence Work

Diagnostic safety as institutional practice

- Office of General Counsel
- Patient Safety Officer
- Cultural shift
- Anticipate real-time discovery of:
 - Potentially actionable patient harms
 - Legal liability

“Wait, you are *trying* to find diagnostic errors?”

INTERLACE CHART ANNOTATION PROCEDURES



INTERLACE Model Development

Input data: EHR data including labs, demographics, vitals, diagnoses, medications, utilization, clinical text

Targets: 669 common and do-not-miss diagnoses collapsed from ~1,700 ICD codes and 1,000 most commonly ordered tests (labs, imaging, referrals)

Population: 707,598 primary care encounters at Penn Medicine for patients >= 65 from 2015-2023

Model Training and Selection: 156 deep neural network architectures evaluated and best model chosen based on validation performance

Model Evaluation: per-diagnosis, macro-averaged, micro-averaged measures of calibration and discrimination

Model Performance in the Held-out Test Set

Category	Measure	Median	Q1	Q2
Diagnoses	C-statistic	0.97	0.85	0.99
	PPV_25	0.85	0.70	0.94
	PPV_50	0.96	0.88	0.98
	Scaled Brier	0.22	0.0004	0.67
Orders	R ²	0.25	0.001	0.67
	C-statistic	0.85	0.78	0.91
	PPV_25	0.16	0.000	0.34
	PPV_50	0.29	0.000	0.51
	Scaled Brier	0.001	-0.0004	0.08
	R ²	0.0019	0.00025	0.017

INTERLACE Model Fine-Tuning on Elite Diagnosticians

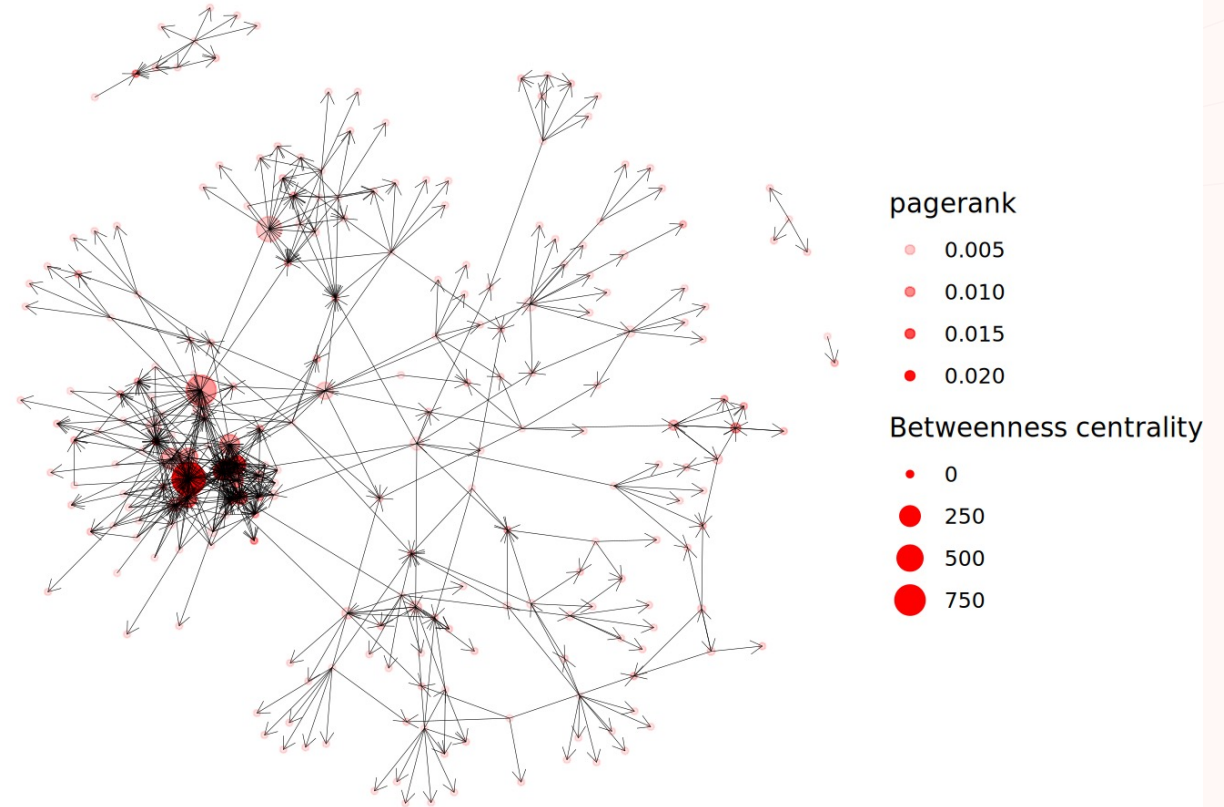
Peer Nomination Survey: Which of your peers would most reliably make the right diagnosis in a patient presenting with an uncertain constellation of symptoms?

Identification of Elite Diagnosticians: Top 25 based on 3 graph measures (In-degree, PageRank, and Betweenness Centrality)

Population: Penn Medicine clinicians in internal medicine, geriatrics, and family medicine

Model Fine-tuning: Froze all but last layer of model and re-trained with small learning rate on visits from “elite” diagnosticians

Model Evaluation: Same as general model



Randomized, Phase 1b Study: A Clinician Turing Test

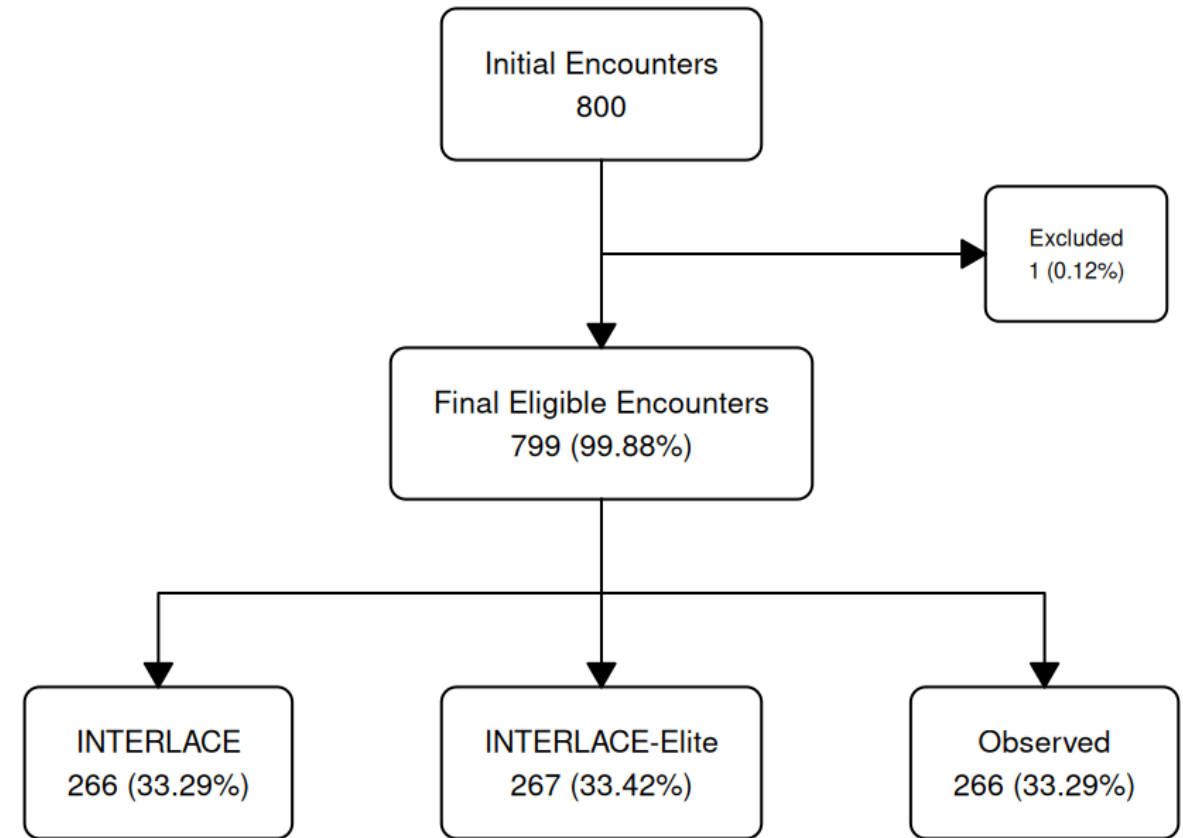
Clinician Annotators: Physicians and advanced practice providers recruited from primary care practices at Penn Medicine

Task: Each annotator reviews 80 encounters in the EHR then sees a list of suggested diagnoses and another list of suggests tests

Annotation: For each encounter, i) agree or disagree with each suggestion, ii) add additional important diagnoses or tests not present in list of suggestions

Randomization: Each encounter was randomly assigned to present i) what was actually recorded in the EHR (control), ii) suggestions from INTERLACE, or iii) suggestions from INTERLACE-elite

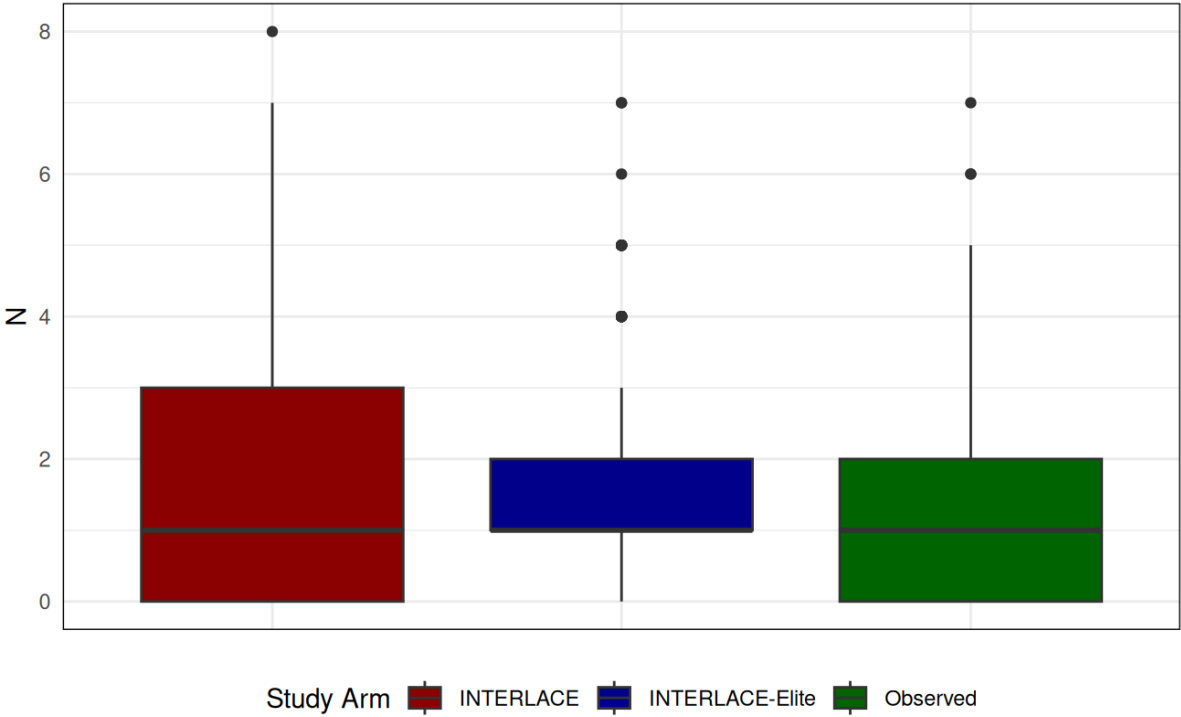
Blinding: Annotators do not know the source of the suggestions to which they are assigned



Randomized, Phase 1b Study: A Clinician Turing Test

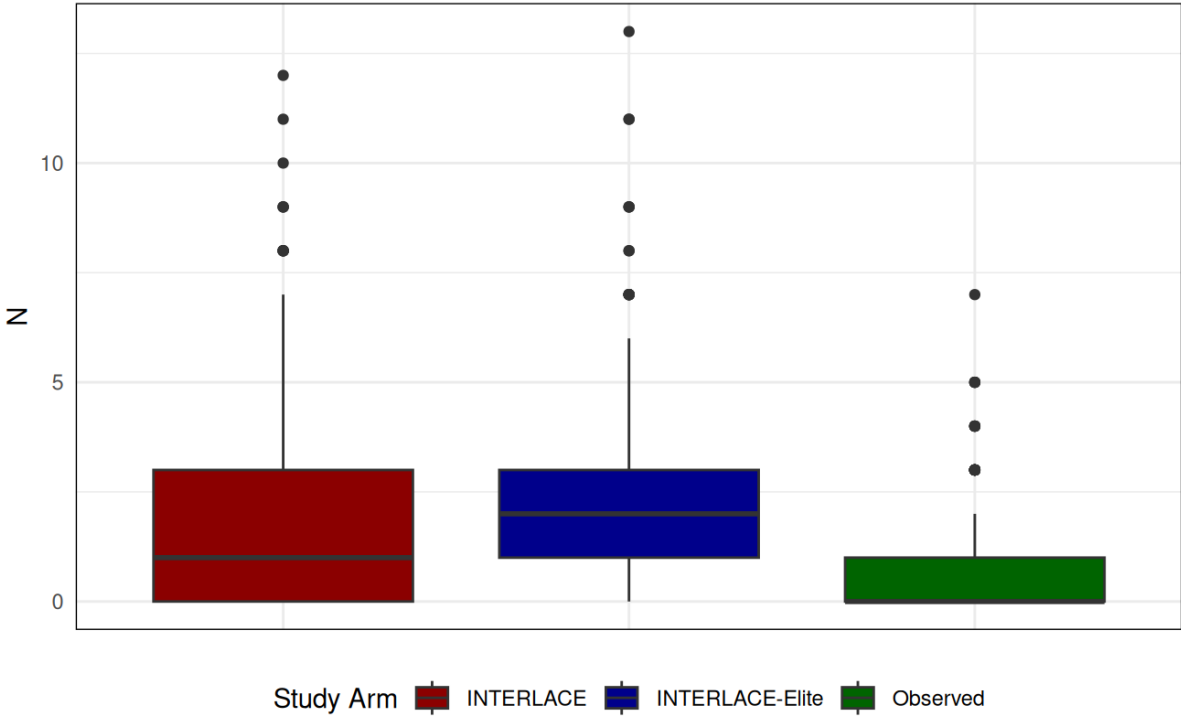
Composite Diagnosis Disagreement

Count



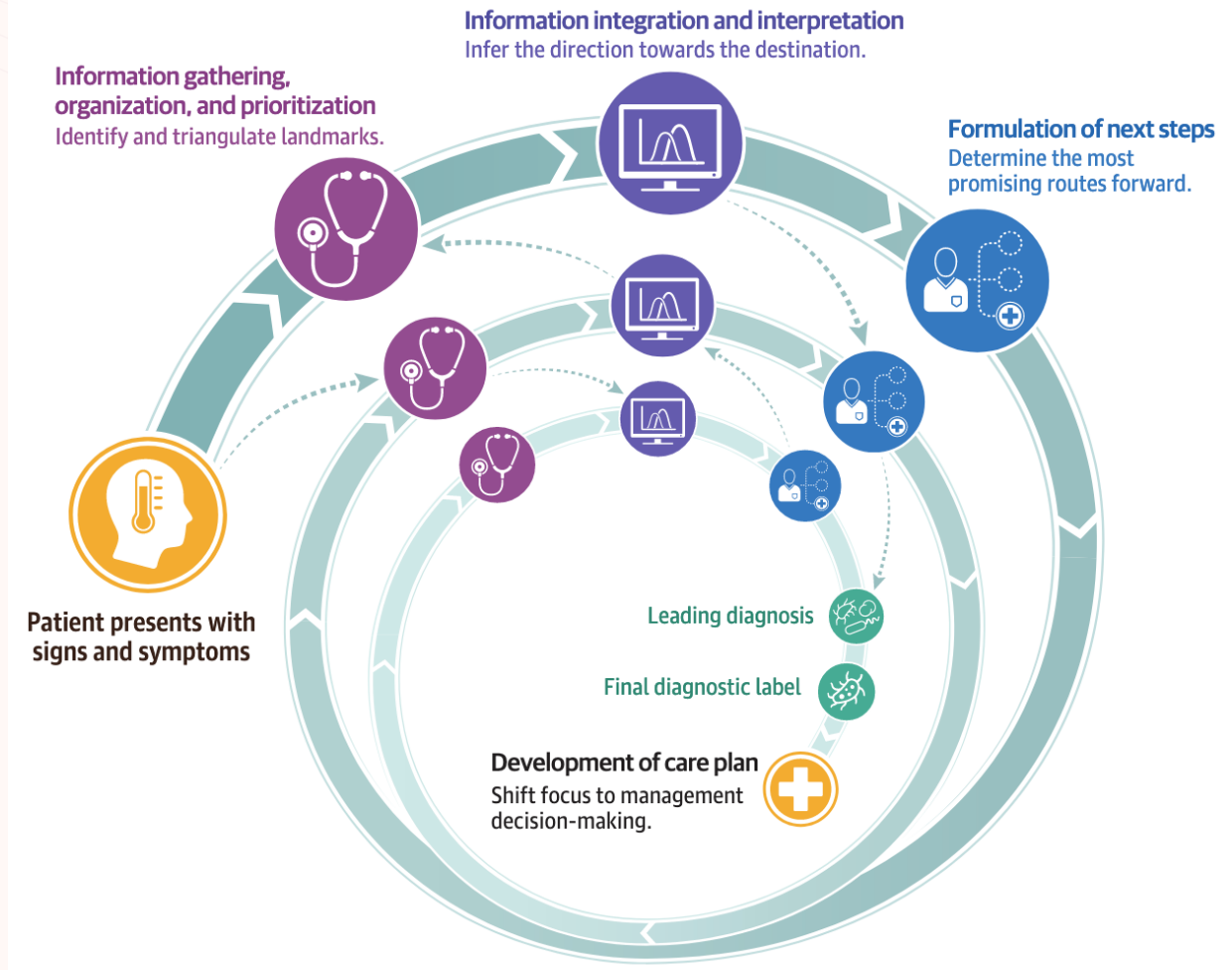
Composite order Disagreement

Count





Facilitating Wayfinding

Figure. The Dynamic Diagnostic Refinement Process



Real-time, Cooperative Decision Support



James Bond

Birthday: 11-17-1993




UID: 0000000007























BP: 145/95


Pulse: 64

Temp: 97.8 (36.6)


Height: 5' 8" (1.727 m)




Cough	<div> NO</div>	UNSURE	<div>YES </div>
Chest Pain	<div> NO</div>	UNSURE	<div>YES </div>
Rash	<div> NO</div>	UNSURE	<div>YES </div>
Headache	<div> NO</div>	UNSURE	<div>YES </div>
Chills	<div> NO</div>	UNSURE	<div>YES </div>
Fatigue	<div> NO</div>	UNSURE	<div>YES </div>
Dizziness	<div> NO</div>	UNSURE	<div>YES </div>
Diarrhea	<div> NO</div>	UNSURE	<div>YES </div>
Abdominal Pain	<div> NO</div>	UNSURE	<div>YES </div>
Constipation	<div> NO</div>	UNSURE	<div>YES </div>
Joint Pain	<div> NO</div>	UNSURE	<div>YES </div>

Common Diagnoses 


GERD




Hypertension




Noninfective Gastroenteritis An...




Immunodeficiency Disorder




Viral Prodrome



Diabetes



Angina



Experiencing

Abdominal Pain

Chest Pain

Chills

Cough


Not Experiencing

Dizziness


Fatigue

Joint Pain


Muscle Pain (Myalgia)

Commonly Ordered Tests 


Comprehensive Metabolic Panel




CRP and ESR




Lipid Panel




Chest X-Ray




Liver Enzyme Evaluation



Hemoglobin A1C



Echocardiogram



Limitations

1. If it wasn't documented it didn't happen
2. Many of the same patients in train/val/test: tradeoff between information leakage vs real-world usage
3. Binary classification constrained to same threshold (10%) for all outcomes and categories which may not be clinically optimal
4. No assessment of degree or potential impact of errors (although preliminary review suggests these are minor, e.g. recommending *both* basic and comprehensive metabolic panels)

Planned next steps

1. Algorithmic equity audit and recalibration for group-wise optimality
2. Comprehensive clinician annotator disagreement analysis
3. In-person pilot feasibility study in 40 primary care encounters among older adults
4. Iterative improvements to model, interface, and optimal thresholds based on pilot study findings
5. Extramural grant application to support a large-scale, pragmatic trial of the INTERLACE tool (multi-site collaborators welcome!)
6. Adaptation to other care environments (e.g. home care, tele-health)

Summary

1. Older adults are at especially high risk of diagnostic errors in the outpatient setting and lack tailored tools
2. The breadth and content of the diagnostic scope should be accounted for in the development of meaningful diagnostic support systems
3. Imitation learning and collective intelligence can provide meaningful suggestions in the absence of gold-standard diagnostic labels
4. Real time, cooperative, diagnostic decision support may facilitate wayfinding through the diagnostic process and include patients, caregivers, and clinicians
5. Open-source AI tools promote transparency, reproducibility, and access
6. Clinical trials are needed to evaluate the effectiveness, safety, and equity of clinical AI tools prior to widespread adoption

Acknowledgements



NAM Scholars in Dx Ex
IN PARTNERSHIP WITH
THE COUNCIL OF MEDICAL SPECIALTY SOCIETIES



The
John A. Hartford
Foundation



The INTERLACE Team:



Matt Press



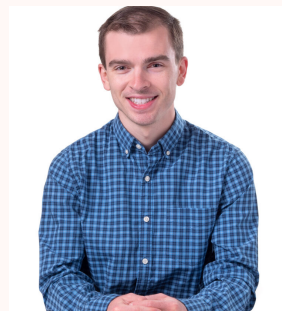
Lyle Ungar



Marty Peifer



Nick Bishop



Benjamin Schmid



Alyssa Sliwa



Anu Vyas



Chris Streiffer