INTERLACE: Advancing diagnostic excellence for older adults through collective intelligence and imitation learning

Christopher D. Streiffer, MD, MS[1, 3], Matthew J. Press, MD, MSc[2, 3], Nicholas S. Bishop[1], Benjamin Schmid, MS[1], Lyle H. Ungar, PhD[4], MaryAnne Peifer, MD, MSIS[2, 5], Gary E. Weissman, MD, MSHP[1, 3]

[1]Palliative and Advanced Illness Research Center, [2]Leonard Davis Institute of Health Economics, [3]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, [4]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, [5]Department of Family Medicine and Community Health, Perelman School of Medicine, University of Pennsylvania

Background

- Diagnostic errors affect approximately **5% of U.S. outpatients annually**; older adults are disproportionately impacted likely from complexity and cognitive limitations
- Existing clinical decision support systems (CDSS) often fall short due to:
  - Narrow diagnostic focus
  - Lack of generalizability
  - Dependence on imperfect or unavailable gold-standard labels
- **Objective:** To develop and evaluate a neural-network-based CDSS that uses imitation learning to replicate collective physician decision-making and improve diagnostic accuracy in older adults

Methods

Data Source

- Over 920,000 primary care visits for adults aged 65 and older
- Integration of structured (labs, vitals, diagnoses, demographics) and unstructured (clinical notes) EHR data

Model Development

- **General model:** trained on all eligible encounters
- **Elite model:** trained on encounters from peer-nominated expert clinicians
- Social network analysis to identify elite diagnosticians using PageRank and HITS algorithms

Model Inputs

- **Text Data:** Vectorized using BioClinicalBERT and MetaMap to extract features
- **Historical Structured Data:** Labs, medications, previous visit diagnoses, previous symptoms
- **Current Visit Data:** Vitals, symptoms

Model Outputs

- Output labels were mapped from >1,600 ICD-10 codes to 669 diagnoses and 1,000 orders

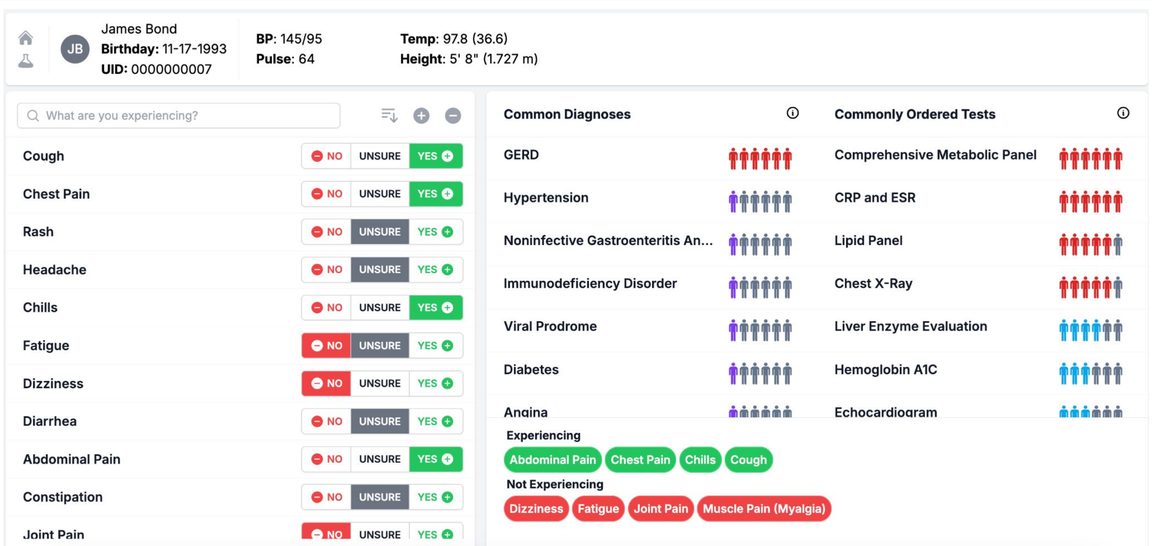

Figure 1. INTERLACE user interface. The CDSS integrates input from clinicians, computer scientists, and HCI experts to support collaborative diagnostic decision-making. Patients enter symptoms on the left, which are combined with historical data to generate real-time diagnostic and order recommendations on the right. Visual "person charts" display the collective intelligence-derived likelihood, promoting shared exploration and patient engagement.

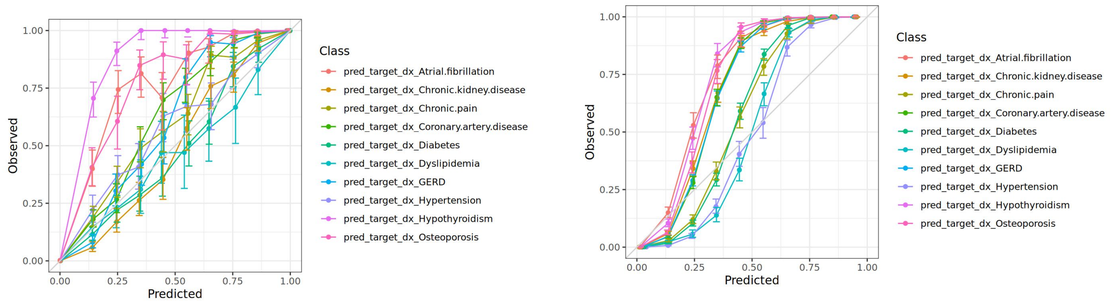

Figure 2. Diagnosis performance. Calibration curves for the general model (left) shows a near-linear relationship between predicted and observed probabilities, with under-prediction for Hypothyroidism, Osteoporosis, and Atrial Fibrillation. Elite model (right) performance follows a more sigmoid pattern compared to the general model, consistent with possible overfitting of the classifier. However, the curves are notably improved for the diagnosis of Hypothyroidism, Osteoporosis, and Atrial Fibrillation.

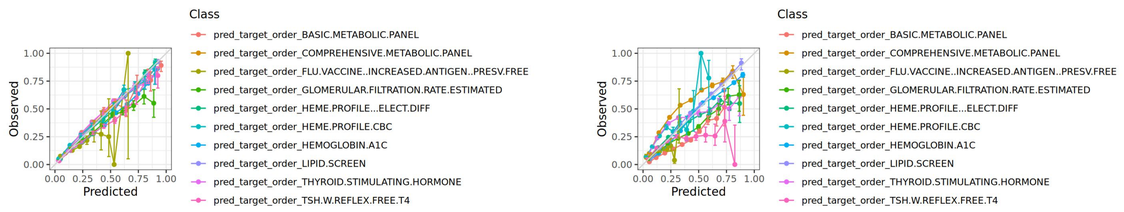

Figure 3. Order performance. Calibration curves for general model (left) display a near-linear relationship for all orders except the Flu Vaccine. Differences in calibration curves likely reflect the wide variability in prevalence for each class. Elite model (right) performance follows a similar pattern with increased variance compared to the general model.

Diagnosis Performance	General Model	Elite Model
C-statistic	0.92	0.8
PPV (25% Sensitivity)	0.76	0.82
Calibration Slope	1.02	1.45
Order Performance	**General Model**	**Elite Model**
C-statistic	0.84	0.61
PPV (25% Sensitivity)	0.23	0.25
Calibration Slope	0.99	0.45

Table 1. Performance metrics. Macro-averaged results for the general and elite models looking at C-statistic, positive predictive value (PPV), and calibration slope.

Results and Discussion

- **General model** achieved strong diagnostic performance (c-statistic 0.92, PPV 0.76) with moderate order prediction performance (c-statistic 0.84, PPV 0.23)
- **Elite model** improved diagnostic precision (PPV 0.82) but showed signs of overfitting and lower calibration
- Best performance seen in **common chronic conditions** (e.g., hypertension, diabetes); struggled with rare diagnoses and low-frequency orders
- **Symptom-driven interface** enables real-time, collaborative exploration of diagnostic and test options
- Demonstrates potential for **scalable, trustworthy CDSS** tailored to older adults, balancing generalizability and clinician precision

Conclusion

- Imitation learning and collective intelligence can enable flexible, accurate CDSS tailored to the complexities of primary care
- The general model achieved superior overall performance, while the elite model offered improved precision at the cost of generalizability
- This approach aligns with real-world diagnostic reasoning by offering symptom-based, probabilistic guidance rather than fixed predictions

Acknowledgements: This work is supported by PennAITech Pilot Grant (NIH/NIA P30AG073105) and the National Academy of Medicine (NAM) of the National Academy of Sciences (SCON-10001137). Dr. Weissman is a current NAM Scholar in Diagnostic Excellence, a program administered by the NAM in partnership with the Council of Medical Specialty Societies (CMSS) and funded by the Gordon and Betty Moore Foundation, with additional support from the John A. Hartford Foundation.