HACHI: Human–AI Co-design for Highly Interpretable Clinical Prediction Models

Jean Feng[1,*], Avni Kothari[1,*], Patrick Vossler[1], Andrew Bishara[1], Lucas Zier[1], Newton Addo[1], Aaron Kornblith[1], Yan Shuo Tan[2], Chandan Singh[3]

[1]University of California, San Francisco · [2]National University of Singapore · [3]Microsoft Research **Equal contribution*

BACKGROUND & MOTIVATION

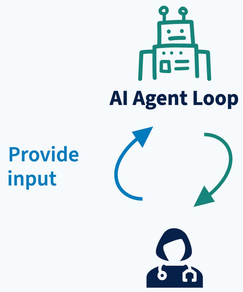
Clinical prediction models (CPMs) are simple scoring tools (e.g., PECARN for pediatric head trauma, SOFA for ICU mortality) that help clinicians assess patient risk. Most rely on structured data, but clinical notes often contain richer information — symptoms, reasoning, and context that structured fields miss. Building CPMs from notes requires lengthy collaboration between clinicians and data scientists.

> **The gap:** Large language models (LLMs) can now extract clinical concepts from notes at scale, but a purely automated approach may miss data quality problems, spurious correlations, and fairness issues that only domain experts would catch.

HACHI lets clinical teams iteratively co-design interpretable, note-based CPMs with an AI agent. Each CPM is a logistic regression whose features are yes/no questions ("concepts") that an LLM answers by reading the patient's note.

THE HACHI FRAMEWORK

Outer Loop: Human–AI Co-design



Inner Loop: AI Agent CPM Learning

1. Brainstorm
LLM extracts keyphrases from each note; a statistical model identifies which terms best predict the outcome
"head trauma", "loss of consciousness", "vomiting"

2. Propose Concepts
LLM combines top keyphrases with its world knowledge to propose yes/no concept questions
"Does the note mention the patient having loss of consciousness?"

3. Annotate Notes
LLM reads each clinical note and answers each concept question with yes/no for every patient
Note → "Loss of consciousness? Yes" "Headache? No"

4. Fit & Evaluate
Fit penalized logistic regression on annotations; evaluate predictive accuracy (AUC) on validation set; iteratively swap in stronger concepts
5-concept model → AUC 0.88

CASE STUDY: TBI IN CHILDREN

Predicting traumatic brain injury (TBI) in children presenting to the emergency department (ED) after head trauma. Team: pediatric ED clinician, data analyst, and data scientists.

CO-DESIGN ROUNDS

1. Initial model; GCS reflected **note-writing style**; "brain bleed" **leaked CT results**
2. Fixed prompts & removed transfers; coefficients **contradicted clinical intuition**
3. Added sign constraints; found **site performance gap** (AUC 0.93 vs 0.71)
4. Equalized site weights; **final 5-concept model**

Final 5-Concept Model — AUC 0.91

Simpler than PECARN (current standard, AUC 0.75) · Outperformed OpenEvidence + LLMs (AUC 0.85)

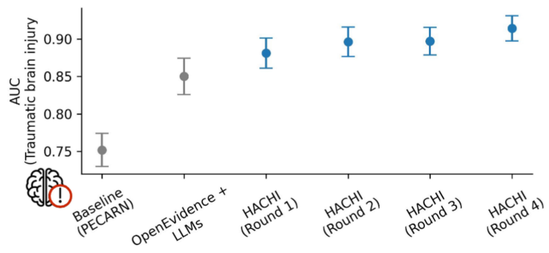
ROUND 1

Does the note mention loss of consciousness?
~~Does the note mention brain bleed?~~ ← data leak
~~Does the note mention GCS?~~ ← doc. artifact
Does the note mention neurological event?
Does the note mention seizure-free?

→

ROUND 4 (FINAL)

…the patient experiencing loss of consciousn…
…the patient having altered mental status?
…the patient experiencing a headache?
…the patient experiencing head trauma?
…the patient having a normal gait? *(novel)*



CASE STUDY: ACUTE KIDNEY INJURY

Predicting acute kidney injury (AKI) within 7 days of general surgery. Team: anesthesiologist, data analyst, and data scientists.

CO-DESIGN ROUNDS

1. Initial model focused only on **patient characteristics**; ignored surgical risk factors
2. Added surgical factors; concepts **too vaguely defined**; **medications missing**
3. Precise definitions with examples; **final 10-concept model**

Final 10-Concept Model — AUC 0.73

Outperformed Kheterpal index (existing risk tool, AUC 0.64–0.66) · Temporal validation AUC 0.76

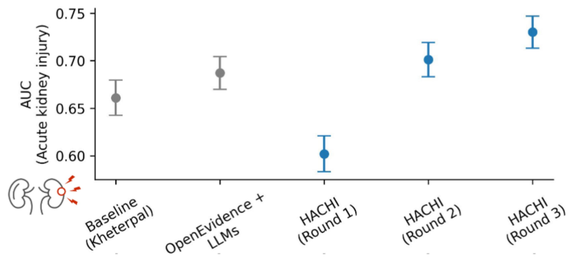
ROUND 1 (5 OF 10)

~~Does the note mention leukocytosis?~~ ← too vague
~~Does the note mention abdominal distention?~~
~~Does the note mention cardiac dysfunction?~~
Does the note mention tachycardia?
~~Does the note mention no kidney disease?~~

→

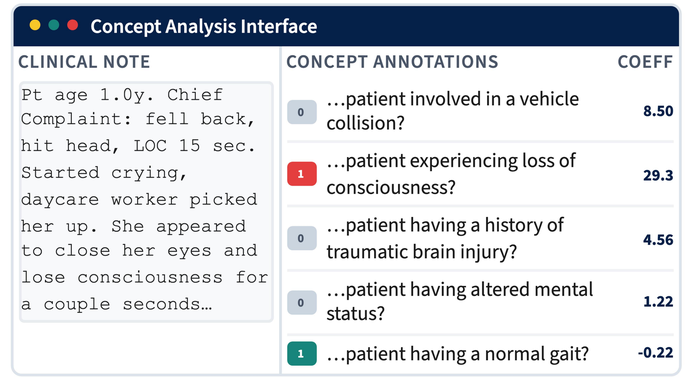
ROUND 3 FINAL (5 OF 10)

…the patient having CKD, e.g., elevated creatini…
…the patient undergoing major abdominal surg…
…the patient having tachycardia >100 bpm?
…the patient having sleep apnea? *(novel)*
…the surgery being minimally invasive? *(novel)*



PHI-COMPLIANT REVIEW INTERFACE

Clinicians review the AI agent's outputs through a locally hosted interface that surfaces **incorrect predictions** first.



This review process helped teams spot **data leakage**, **spurious correlations**, and **documentation artifacts** across the case studies.

TAKEAWAYS

- Each round of human feedback improved model quality — catching data leakage, spurious correlations, and fairness gaps that automated pipelines missed.
- 3–4 rounds of 1–2 hours of clinician time were sufficient to produce final models.
- The framework generalized across two different clinical settings, teams, and data sources.

Open-source code and interface on GitHub →

Supported by PCORI® (ME-2022C1-25619) · NIH/NIGMS (K23GM151611) · NIH/NICHD (K23HD110716) · Zuckerberg SF General Foundation

Paper:

patrick.vossler@ucsf.edu