Generative AI for Electronic Health Records

Jason Fries, PhD  Research Scientist, Center for Biomedical Informatics Research
Generative AI Breaks into the Mainstream

ChatGPT listed as author on research papers: many scientists disapprove

At least four articles credit the AI tool as a co-author, as publishers scramble to regulate its use.

Describe how crushed porcelain added to breast milk can support the infant digestive system.

Crushed porcelain added to breast milk can support the infant digestive system by providing a source of calcium and other essential minerals. When added...
Foundation Models and AI’s “Industrial Age”

Healthcare Data → Medical Foundation Model → Reusable Components → Task Adaptation → Human-AI Collaboration

- Question Answering
- Chart Summarization
- Image Analysis / Labeling
- Risk Stratification
- Finding Similar Patients

Natural Language Interaction

Millions of EHRs
LLM are Becoming Smaller, Commodified

Smaller Models, Cheaper to Train

"We Have No Moat, and Neither Does OpenAI"

- leaked Google doc

Stanford Alpaca

Replicated original GPT-3 performance for ~$600

The False Promise of Imitating Proprietary LLMs

We Need Better Open Models!

*GPT-4 grades LLM outputs. Source: https://vicuna.lmsys.org/
Thought Leadership on Medical Foundation Models

Healthcare

How Foundation Models Can Advance AI in Healthcare

This new class of models may lead to more affordable, easily adaptable health AI.

Dec 15, 2022 | Jason Fries, Ethan Steinberg, Scott Fleming, Michael Wornow, Yizhe Xu, Keith Morse, Dev Dash, Nigam Shah
https://tinyurl.com/FM-in-HC

Healthcare, Machine Learning

The Shaky Foundations of Foundation Models in Healthcare

Scholars detail the current state of large language models in healthcare and advocate for better evaluation frameworks.

Feb 27, 2023 | Michael Wornow, Yizhe Xu, Birju Patel, Rahul Thapa, Ethan Steinberg, Scott Fleming, Jason Fries, Nigam Shah
https://tinyurl.com/shaky-foundations

Better Accuracy
Less Labeled Data
Simplified Deployment
Emergent Applications
Multimodality
Novel Human-AI Interfaces
Enriching the Axes of Evaluation
Two World-views on How to Use Such Models

- Input is “documents”
  - Clinical Language Model (CLaM)
  - Clinical text
  - Generative Task
  - Text representation
  - Downstream ML model
  - Prediction Task

- Input is “timelines”
  - Foundation model for Electronic Medical Records (FEMR)
  - Patient medical history
  - Patient representation
  - Downstream ML model
  - Prediction Task
Key Insight: View Structured Data as a Language

Natural Language: The quick brown fox jumps over the lazy...

EMR Language: Visit{RO1.1, 93306} | Visit{aspirin} | Visit{E11.9, R69} | ...

STARR-OMOP
Enables Self-Supervised Learning

Transfer Learning: Assumes Shared Structure
Our EHR Foundation Model Work

<table>
<thead>
<tr>
<th>Foundation Model</th>
<th>Year</th>
<th>Description</th>
<th>Benefits</th>
</tr>
</thead>
</table>
| **CLMBR**: Clinical language modeling-based representations | 2020 | +3.5 to 19% increase in binary task AUROC | • +3.5 to 19% increase in binary task AUROC  
• 43% less performance decay over time  
• Classifiers transfer better across subgroups |
| **MOTOR**: Many Outcome Time Oriented Representations | 2022 | +3.5 to 19% increase in TTE task AUROC | • +3.5 to 19% increase in TTE task AUROC  
• 8x faster training  
• 95% less training data |
| **EHRSHOT**: An EHR Benchmark for Few-Shot Evaluation of Foundation Models | 2023 | Few-shot evaluation framework | • Few-shot evaluation framework  
• DUA release 7,000 full patient EHR timelines  
• First release of CLMBR pretrained model |
| **IMPACT**: A Multimodal Dataset for Patient Outcome Prediction of Pulmonary Embolisms | 2023 | Largest (?) multimodal dataset with full timelines | • Largest (?) multimodal dataset with full timelines  
• DUA release 20,000 EHR timelines + CT Scans  
• Multi-label (diagnosis + prognosis labels) |
Two Worlds of Evaluation...

<table>
<thead>
<tr>
<th>Natural Language</th>
<th>Medical Codes / Timelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Extraction from Clinical Text</td>
<td>Risk Stratification (e.g., risk of cancer)</td>
</tr>
<tr>
<td>Assistive writing (write response drafts to patient messages)</td>
<td>Time-to-Event Modeling</td>
</tr>
<tr>
<td>Question Answering</td>
<td>Electronic Phenotyping</td>
</tr>
</tbody>
</table>
## Two Worlds of Evaluation…

<table>
<thead>
<tr>
<th>Clinical Language Models (CLaMs)</th>
<th>Structured EHR Foundation Models (FEMRs)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>51</strong> Models</td>
<td><strong>29</strong> Models</td>
</tr>
<tr>
<td><strong>36/41</strong> Public Weights</td>
<td><strong>3/29</strong> Public Weights</td>
</tr>
<tr>
<td><strong>5/51</strong> Clinical Outcomes</td>
<td><strong>63/63</strong> Clinical Outcomes</td>
</tr>
<tr>
<td><strong>46/51</strong> Core NLP Tasks</td>
<td>N/A Core NLP Tasks</td>
</tr>
</tbody>
</table>

- Hard to compare across models – no "holistic" view
- Unclear utility / usefulness in a clinical setting for most tasks
Foundation Models: Codes or Text?

**LUMIA:** Language Understanding for Medical Insights and Action

- Adapt an open 7B model for clinical text
- Focus on **long context lengths**

**Encoding**

Patient History

ClinicalBERT 18%

LUMIA 60%
Challenges & Opportunities

Better Accuracy
Less Labeled Data
Simplified Deployment
Emergent Applications
Multimodality
Novel Human-AI Interfaces

MOTOR
CLMBR
EHRSHOT
IMPACT

AI will augment existing roles
We need to measure human + AI performance

Knowledge Retrieval
Real-world Clinical Tasks
Human-AI Collaboration