Model Hubs for Medical AI: How Far is Our 😆 Moment?

Jason Fries, PhD  Research Scientist, Center for Biomedical Informatics Research
The State of AI in Healthcare

- Cost to prototype a single model: $200,000

- Models are rarely deployed: 593 COVID-19 models; virtually none were deployed.

- Medical data are noisy, replete with errors, biases, missingness.

- Most AI is trained and tested on cleaned data.

AI Chasm

Healthcare AI Research

Improve Clinical Outcomes
Reduce Costs & Burnout
Improve Patient Lives
Healthcare AI Research

DEPLOYMENT

AI Chasm

Improve Clinical Outcomes
Reduce Costs & Burnout
Improve Patient Lives
Generative AI
Breaks into the Mainstream

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 to

…and their many issues
Foundation Models and AI’s “Industrial Age”

Bommasani et al. “On the Opportunities and Risks of Foundation Models”
Foundation Models and AI’s “Industrial Age”

Healthcare Data → Medical Foundation Model → Task Adaptation

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

Natural Language Interaction → Human-AI Collaboration
Exploring Novel Architectures
Pretraining Objectives
Capturing Multimodality

Systematic methods for data curation, evaluation, and generation
• Choosing Training Examples
• Automating Evaluation
• Weakly Supervised Model Training

Aligning Models with Real Needs
Optimizing Feedback Loops
Improving Model Trust
Model Hubs for Medicine

Transforming AI

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

When is medicine’s “Hugging Face” moment?
Building Foundation Models for Healthcare
Medical Foundation Models

- Natural Language
- 2D Imaging
Foundation Models and AI’s “Industrial Age”

- **Electronic Health Records**

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

**HEALTHCARE DATA**

**REUSABLE COMPONENTS**

**TASK ADAPTATION**

**HUMAN-AI COLLABORATION**
Electronic Health Record (EHR) Data is Multimodal

Contains multiple types of data ordered by time
AI to Enhance Medical Decision Making

What Occurred in the Past?
- Chart Summarization
- Cohort Construction

What is Occurring Now?
- Identify blood clots in lung CT scans
- Identify cancerous cells in pathology slides

Predict Future Risks & Intervention Benefits
- Will patient develop nephritis?
- Will patient develop chronic pulmonary hypertension?

Patient EHR Timeline

Example ML Applications
- Whether to Treat
- How to Treat
- subject to
  - Policy
  - Capacity to Act
  - Intervention Properties
Language Modeling 101

\[ S = \text{Where are we going} \]

\[
P(S) = P(\text{Where}) \times P(\text{are} \mid \text{Where}) \times P(\text{we} \mid \text{Where are}) \times P(\text{going} \mid \text{Where are we})
\]

\[
P(w_1, w_2, \ldots, w_n) = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2) \ldots p(w_n \mid w_1, w_2, \ldots, w_{n-1})
\]

\[
= \prod_{i=1}^{n} p(w_i \mid w_1, \ldots, w_{i-1})
\]
Structured EHRs form a “Language”

\[ S = \text{ICD10/R634}, \text{ICD10/E78.2}, \text{ICD10/Z95.2}, \text{CPT4/99214} \ldots \text{ICD10/I63} \]

Map all medical codes to a finite symbol vocabulary (i.e., analog of words)

\[
P(S) = \prod_{i=1}^{N} P(w_i | w_1 \ldots w_{i-1})
\]
Self-Supervised Learning with EHRs

Entire Patient Population

Pretrain

FOUNDATION MODEL

Adapt

Small Labeled Set

Transfer Learning: Assumes Shared Structure
Autoregressive Language Modeling with Codes

**Key Intuition:** In medicine, accurately generating future health states captures many use cases of AI.
Our EHR Foundation Model Work

**Methods Development**

- **Structured Data**
- **Knowledge Graphs**

**CLMBR:** Clinical language modeling-based representations

**MOTOR:** Many Outcome Time Oriented Representations

**Autoregressive**

**Time-to-Event**

**2021**

**2023**
Key Intuition: We often don’t just want to know if something will happen, but also when it will happen.

Time-to-Event (or Survival) Modeling
Use a Different Pretraining Objective

- Massive (8,192 tasks) time-to-event pretraining
- Larger scales of pretraining data (2M to 55M patients) for EHR and insurance claims data
- 19 evaluation tasks

Celiac, Lupus, Pancreatic Cancer, NAFLD, Stroke, Lab Value Prediction, Radiological Findings
Outperforms autoregressive pretraining across all time horizons

Reduces requirements for labeled examples for adaptation by up to 95%
Evaluating Foundation Models for Healthcare
Goals for our Medical Foundation Models

- Better Performance with Less Data
- Robustness to Distribution Shifts
- Cross-Site Adaptability
Benefits of EHR Foundation Models

Better Performance with Less Data

Robustness to Distribution Shifts

Cross-Site Adaptability

- +3.5 to 19% increase in AUROC [1]
- Match SOTA w/ 95%+ less training data [4,5]
- Improved temporal robustness (+43%) [2]
- Improved subgroup performance [3]

Cross-Site Adaptation of EHR Foundation Models
Cross-Site Adaptation of EHR Foundation Models

- Out-of-Box Performance
- Continued Pretraining
- Label Efficiency
## Overall Performance

Out-of-the-box CLMBR **outperformed GBM** in 5/8 tasks (SK), 6/8 tasks (MIMIC). **DAPT** improved CLMBR performance in all tasks (SK), and in 7/8 tasks (MIMIC).

<table>
<thead>
<tr>
<th></th>
<th>SickKids</th>
<th>MIMIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GBM</td>
<td>CLMBR</td>
</tr>
<tr>
<td><strong>In-hospital</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>0.893</td>
<td>0.941</td>
</tr>
<tr>
<td></td>
<td>[0.815, 0.953]</td>
<td>[0.902, 0.971]</td>
</tr>
<tr>
<td><strong>Long LOS</strong></td>
<td>0.866</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>[0.853, 0.879]</td>
<td>[0.8, 0.83]</td>
</tr>
<tr>
<td><strong>30-day Readmission</strong></td>
<td>0.783</td>
<td>0.774</td>
</tr>
<tr>
<td></td>
<td>[0.755, 0.809]</td>
<td>[0.747, 0.799]</td>
</tr>
<tr>
<td><strong>Hypoglycemia</strong></td>
<td>0.88</td>
<td>0.915</td>
</tr>
<tr>
<td></td>
<td>[0.83, 0.924]</td>
<td>[0.879, 0.945]</td>
</tr>
<tr>
<td><strong>Hyponatremia</strong></td>
<td>0.783</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>[0.579, 0.957]</td>
<td>[0.88, 0.963]</td>
</tr>
<tr>
<td><strong>Hyperkalemia</strong></td>
<td>0.749</td>
<td>0.793</td>
</tr>
<tr>
<td></td>
<td>[0.687, 0.808]</td>
<td>[0.743, 0.838]</td>
</tr>
<tr>
<td><strong>Thrombocytopenia</strong></td>
<td>0.953</td>
<td>0.962</td>
</tr>
<tr>
<td></td>
<td>[0.928, 0.975]</td>
<td>[0.946, 0.975]</td>
</tr>
<tr>
<td><strong>Anemia</strong></td>
<td>0.919</td>
<td>0.918</td>
</tr>
<tr>
<td></td>
<td>[0.898, 0.938]</td>
<td>[0.897, 0.937]</td>
</tr>
</tbody>
</table>
Continued Pretraining

Incorporate additional hospital-specific pretraining data

Require **60-90% less** pretraining data
Improved Few-Shot Performance

Improved Label Efficiency

Require less than 1% of training examples to match performance of XGBoost
Benefits of EHR Foundation Models

Better Performance with Less Data

- +3.5 to 19% increase in AUROC [1]
- Match SOTA w/ 95%+ less training data [4,5]

Robustness to Distribution Shifts

- Improved temporal robustness (+43%) [2]
- Improved subgroup performance [3]

Cross-Site Adaptability

- Transfer pretrained models across hospitals
- Require 60-90% less pretraining data [4]

Reproducibility of Foundation Model Research
Releasing New Medical Datasets

**EHRSHOT**: An EHR Benchmark for Few-Shot Evaluation of Foundation Models

- **2023**
- **6,739** Patients
- **Tabular**

**INSPECT**: A Multimodal Dataset for Patient Outcome Prediction of Pulmonary Embolisms

- **2023**
- **19,402** Patients
- **EHR**
- **CT Scans**
- **Tabular**
- **Radiology Notes**

NeurIPS 2023

**MedAlign**: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records

- **2023**
- **267** Patients
- **EHR**
- **Tabular**
- **All Clinical Notes**

ML4H Symposium 2024

- **BEST THEMATIC PAPER**

To Appear **AAAI 2024**
Open & Accessible Model Weights

Sharing pre-trained model

Initially we really hoped to share our models but unfortunately, the pre-trained models are no longer sharable. According to SBMI Data Service Office: "Under the terms of our contracts with data vendors, we are not permitted to share any of the data utilized in our publications, as well as large models derived from those data."

https://github.com/ZhiGroup/Med-BERT

Transfer learning is the primary value prop of foundation models!

Foundation Models Risk Increasing our Reproducibility Crisis
Enabling Open Science

Our first model hub release!

- Gated model on Hugging Face
- Requires CITI ethics training
- Non-commercial use only

Margaret Mitchell
Chief AI Ethics Scientist, Hugging Face
DRAFT PROPOSAL - Data Schema for ML Developers

Bert Arnrich, Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, Robin van de Water

https://github.com/Medical-Event-Data-Standard/meds
Open Weights are Critical to Fair & Secure Models

Why Anthropic and OpenAI are obsessed with securing LLM model weights

**Transparency** (training data, model weights) is critical for fair and secure models!

“Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, **dataset construction**, training method, or similar.”
Closing Thoughts
Opportunities Moving Forward

- Model Development
- Data-Centric AI
- Human-AI Collaboration
Opportunities Moving Forward

Data-Centric AI
- Data Cleaning (medical data is messy!)
- Model-guided Training (Weak Supervision? “Superalignment”)
- Informed Data Selection & Curation

Human-AI Collaboration
- Feedback loops and decision making
- Optimizing evidence synthesis
- Capturing preferences to guide model improvement
- Building trust in models
Opportunities Moving Forward

Model Development

Data-Centric AI

Human-AI Collaboration
Calls for the Academic Community

Smaller Models, Cheaper to Train

- LLaMA-13B
- Alpaca-13B
- Vicuna-13B
- Bard
- ChatGPT

Reimagine Model Evaluation

- Knowledge Retrieval
- Real Clinical Workflows
- Human-AI Collaboration

Lead Building Open, Reproducible Medical Base Models

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

jason-fries@stanford.edu

I’m on the academic job market this season – don’t hesitate to reach out!

Emperor Kuzco!
NeurIPS 2023 “Keep AI Open”