Portfolio of Medical Foundation Models to Accelerate Clinical and Translational Research

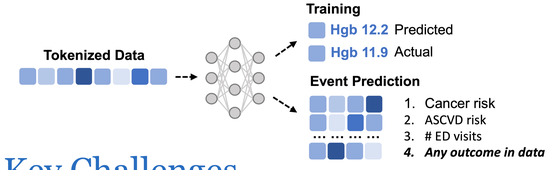
Andrew J. Loza[1,2], Weipeng Zhou[1], Yountian Liu[1], Erica Stutz[1], Ikgyu Shin[1], Qingyu Chen[1], Hua Xu[1], Daniella Meeker[1], Lucila Ohno-Machado[1]

1. Department of Biomedical Informatics and Data Science (BIDS), Yale University School of Medicine
2. Department of Pediatrics, Yale University School of Medicine

Yale SCHOOL OF MEDICINE

Background

- Real-World Data holds significant promise to advance personalized medicine through improved prediction of clinical events, diagnoses, and treatment responses.[1]
- Existing analytical frameworks fall short: limited use of data modalities, single-purpose models, susceptible to overfitting.
- Medical foundation models, such as our recent collaborative work on **Comet** (1 billion parameter model trained in Epic Cosmos) address these limitations.[2]
- Inspired by large language models: instead of predicting text, predict sequence of clinical observations, actions, and passage of time.[2]
- Once trained, can predict any outcome represented in the training data.



Key Challenges

Adapting methods to clinical data has nuance and many open questions:

- Healthcare data is multimodal and differs from internet data.[3,4] What is the best tokenization technique? Do we adapt or rebuild existing models?
- There are different medical data "dialects".[5] How can we plan for interoperability to support federated learning and collaboration?
- Data is high-risk and protected.[6] How can we ensure patient protection? How can we still benchmark new methods in a reproducible manner?
- Researchers will have a variety of use cases: event prediction, target trial emulation, patient matching.[7] How can we support all of these?
- Data collected from one health-system often is incomplete.[8] How can we best handle and understand missing data?

Aim: Accelerate Research

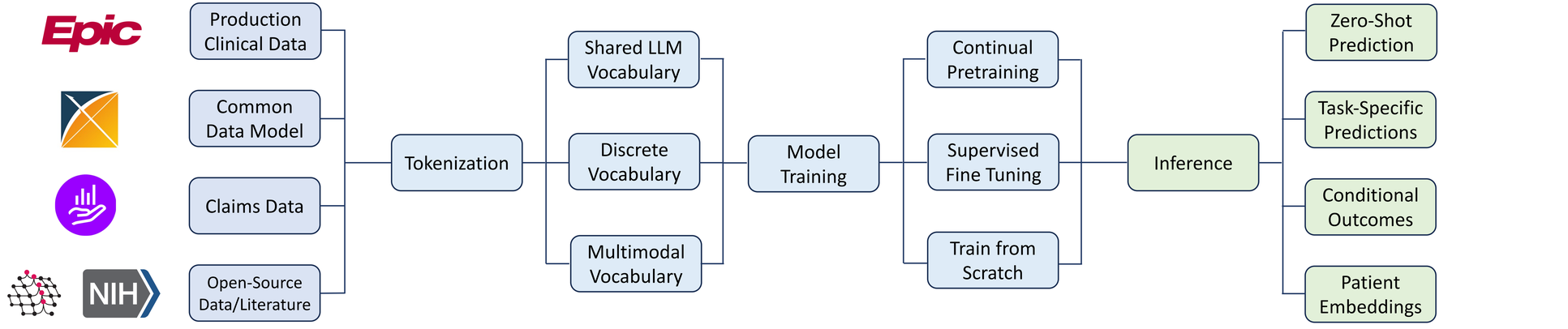
Hub initiative: create a portfolio of medical foundation models embedded in a secure computational pipeline to address key knowledge gaps and meet a variety of use cases.

- Unified framework for leveraging the full spectrum of health record data with state-of-the art performance.
- Task agnostic: predictions for any task represented in the data.
- "Round trip" of predictions back to source data format to use existing analysis pipelines.
- Common data model and open-source formats to facilitate interoperability and model design benchmarking.

Key References:

1. Concato J, Jacqueline C. "Real World Evidence – Where are we now?" *NEJM*. 2022
2. Waxler S, et al. "Generative Medical Event Models improve with Scale *arXiv preprint*. 2025
3. Loza AJ, et al. "multivariateGPT: a decoder-only transformer for multivariate categorical and numeric data" *arXiv preprint*. 2025
4. Cao, Z, et al. "The development landscape of large language models for biomedical applications." *Ann. Rev. Biomed Data Science*. 2025.
5. – 9. Please see QR code for full reference list

Acknowledgements: We are grateful for support from the CTSA Grant UL1TR001863 including KL2 support for AJL from KL2TR001862; AJL has received additional funding from The Hartwell Foundation.



Clinical Data Sources

Clinical and Translational research has varied needs. By building a portfolio of foundation models, we can meet these needs and learn how data sources complement each other.

Production Clinical Data:
- Most granular clinical data for research questions from physiology to operations
- Epic Clarity/Caboodle data from YNHHS
- 92M encounters across 4.8M patients

Common Data Model:
- Facilitates federated learning and health-system collaboration due to shared data definitions
- Full version of YNHH data in OMOP format

Claims Data:
- Less granular but more complete longitudinal record
- Truven Health MarketScan® with 124M patient records

Open-Source Data and Literature:
- Reproducible benchmarking for novel method development
- MIMIC-IV: 364k patients for benchmarking novel medical foundation models.
- LMOD[9] (Large Multimodal Ophthalmology Dataset)

Learning Patient Trajectories

Secure, High-Performance Computing is Essential for Safe Cutting-Edge Model Development

Hopper: NIST 800-171/HIPAA compliant secure computing environment
- 32 NVIDIA H200 GPUs, 60 NVIDIA H100 GPUs, 112 other GPUs

Bouchet: higher performance cluster for non-sensitive data
- 80 NVIDIA H200 GPUs, 48 NVIDIA RTX 5000 ADA GPUs

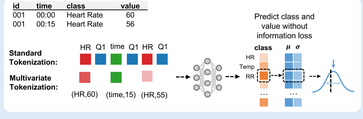
Tokenization: Process of converting raw data into discrete elements for deep learning models.
- How to handle numbers, passage of time, images?
- Can we leverage trained LLMs?

Shared Vocabulary: tokenize data by converting tables into flat text representations for frontier LLMs (Llama, gpt-oss, etc.).

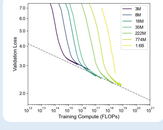
```
Patient timeline: Age 56; order placed: CBC, priority
stat; lab result: hemoglobin 12.3, units mg/dL
```

Joint Image-Text Models: adapt existing models that combine image and text to perform well in the medical domain.

Discrete vs Multimodal Vocabulary for Data: can also learn a new vocabulary custom to the data where whole data elements or parts of data elements become the "words".
- **Discrete:** uses quantile bins for numeric values, lossy.
- **Multimodal:** novel method developed by our team to handle numeric data without losing information.[3]



Training: medical foundation models are trained by predicting the next data element in a sequence.
- Self-supervised: no manual annotation, can scale to billions of data points.
- Multi-task: next token prediction is inherently multi-task and leads to models robust to overfitting.
- Performance scaling: larger models trained on more data continue to improve.



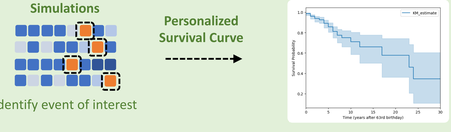
Training from scratch: for models with a custom vocabulary, the Hopper cluster allows for efficient training of 1.6 billion parameter models from scratch.

Continual pretraining: for models using a shared vocabulary with a frontier LLM like Llama-3 120b, can use the fully trained model and resume training on clinical data in a secure environment.
- Leverages the "knowledge" in existing language models.

Supervised fine tuning: after pretraining, models can be specialized on a single task using labeled data. This requires fewer samples and achieves better performance than other task-specific approaches.

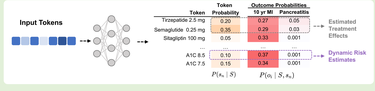
Downstream Applications

Zero-Shot Prediction: models can generate predictions on any task represented in the training data through Monte-Carlo sampling (from scratch models) or prompting (LLM based models).



Task Specific Predictions: supervised fine tuning produces models that can generate predictions without simulation trading off flexibility for speed while still deriving robustness from pretraining.

Conditional Outcomes: we have developed an extension to the next token prediction task which allows the model to also predict the probability of outcomes, conditional on each *potential* next token.
- Creates a unified propensity and outcome model.
- Supports Target Trial Emulation with Conditional and Average Treatment Effect Estimation.



Embeddings: in addition to predictions, models can be used to generate embeddings of longitudinal patient data. These can support a range of applications:
- Patient matching: can be used to find which patients in a health system are most closely matched to a candidate patient for trial recruitment and design.
- Phenotyping: can be used to identify subpopulations.