There is immense excitement in healthcare about the potential benefits of adopting Artificial Intelligence (AI), which promises billions of dollars in savings through operational efficiencies, improved care, and labor cost reductions [1]. Foundation models, a class of machine learning systems trained on large-scale datasets and adaptable to a wide range of tasks and domains, are revolutionizing the use of AI across multiple industries. In healthcare, foundation models represent a paradigm shift from single-purpose models to generalist AI systems composed of reusable components [2–4]. However, translating foundation models into tangible benefits in healthcare faces challenges [5] and requires developing methods to capture and utilize *tacit knowledge*—expertise that is contextual, embedded in practice, and rarely documented.

Tacit knowledge is a defining feature of multi-stakeholder decision-making processes in medicine. Consider a complex healthcare process such as a tumor board meeting, where the goal is to formulate a personalized treatment plan for a cancer patient by synthesizing the knowledge and preferences of multiple experts. From a modeling standpoint, this requires orchestrating various skills, including data retrieval, summarization, and risk stratification—all of which are deeply informed by tacit knowledge. For instance, determining which experts will attend, identifying which data needs to be reviewed, and knowing where to find specific information are all decisions rooted in the collective experience of the team members. Current development regimes for foundation models and large language models (LLMs) fail to capture these complex healthcare processes [6], providing little guidance on how to choose, adapt, and evaluate models before deployment [7,8]. Critically, developing robust evaluation frameworks and advancing the next generation of healthcare AI requires innovative approaches to studying these tacit knowledge tasks.

My work is uniquely positioned to bridge these gaps with research that draws on expertise in computer science, biomedical informatics, and hospital information systems. In my current role as Senior Research Scientist at Stanford and with the Stanford Health Care data science team, I serve as a vital research bridge—connecting theoretical advancements in AI with the practical knowledge required to successfully train and deploy models in hospitals. Capturing tacit knowledge requires studying feedback loops by deploying models in the settings where they will be used (Figure 1), underlining the need to focus on the delivery science of healthcare models and methods to sustain their routine use [9].

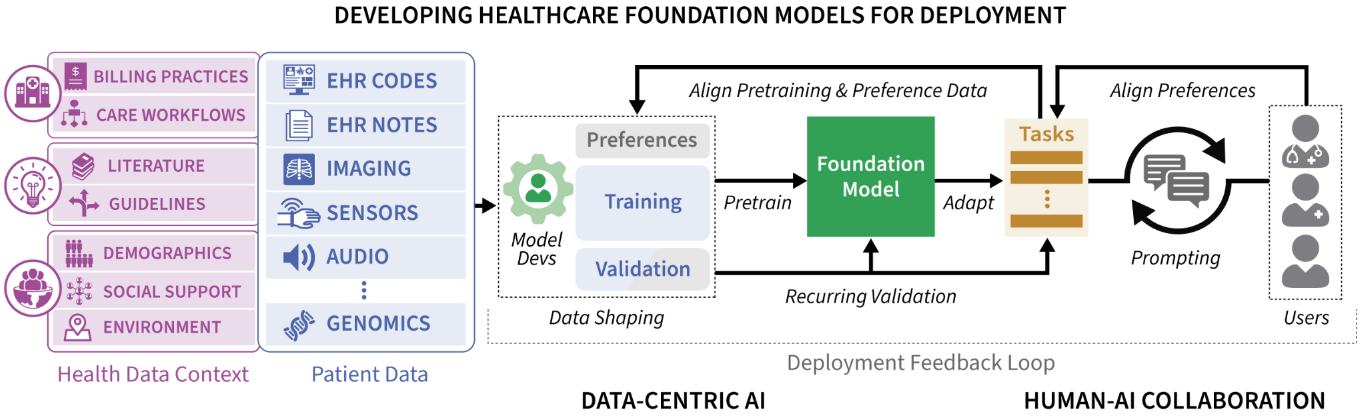

Figure 1. *Deploying healthcare foundation models requires supporting multiple feedback loops that bridge model developers, deployment infrastructure, and end users (e.g., clinicians, patients). Patient data is contextualized for use in model development (**Data-Centric AI**). This is a recurring process that is informed by deployment teams and users to capture "tacit" knowledge (**Human-AI Collaboration**). The core intuition is the need to conduct model development, preference alignment, and validation in a tight deployment feedback loop in the health system.*

SUMMARY OF PRIOR WORK

Healthcare Foundation Models: Electronic health records (EHRs) provide a unique, longitudinal view of patient health and present a vast yet underutilized source of training signals for modeling the progression of diseases. Foundation models that can transform EHR timelines into general-purpose feature embeddings have the potential to significantly accelerate the development of AI-driven clinical applications. Motivated by this, we developed CLMBR, an autoregressive EHR foundation model pretrained on data from 2.8M Stanford Health Care patients [10]. CLMBR has demonstrated several advantages for predictive models, including enhanced robustness to distribution shifts across patient subgroups, time, and geography [11–13], as well as superior few-shot learning performance compared to encoder-based (BERT-style) models [14].

However, current autoregressive models for disease prognosis are limited in their ability to effectively learn long-term temporal patterns. To address this limitation, we further developed MOTOR, a novel time-to-event (TTE) foundation model for outcome prediction [15]. The core idea behind MOTOR is to capture long-term temporal dynamics by using clinical event time distributions from EHRs for pretraining, enabling it to predict both the probability and timing of clinical events. TTE pretraining increases prognostic accuracy for diseases like pancreatic cancer compared to CLMBR and achieves a 4.6% performance gain in time-dependent C-statistics over state-of-the-art TTE models. Furthermore, we have extended this approach to cross-modal supervision by using our TTE objective to improve prognosis prediction with 3D imaging data [16], showcasing the potential of longitudinal EHR data to enhance multimodal predictive models.

Benchmarking Medical AI: Evaluating foundation models and LLMs in healthcare is complicated by the limited availability of diverse EHR benchmarks and a lack of realistic tasks [17], with only 5% of recent research studies evaluated using EHR data [6]. EHR foundation models with private weights create additional barriers to reproducible science. To address these challenges, we have focused on three core contributions. First, all of our EHR foundation model weights are made available via Hugging Face, the first lab to release large-scale EHR models for use by the research community. Second, we have released three new longitudinal EHR benchmarks built on Stanford Health Care data: EHRSHOT [14], a few-shot benchmark of 6,749 patients; INSPECT [18], a multimodal dataset of 19,402 patient EHRs, CT scans, and paired radiology reports; and MedAlign [19], a clinician-generated instruction-following benchmark. MedAlign underscores the need for realistic tasks to assess technical progress as all medical-pretrained LLMs we evaluated performed worse on MedAlign compared to their base versions, despite reporting improvements on multiple-choice benchmarks. Finally, as part of an international, multi-institutional collaboration, we co-developed the Medical Event Data Standard (MEDS), a machine learning standard to facilitate an interoperable ecosystem of tools for EHR foundation model training and evaluation [20,21].

Data-Centric AI & Human-AI Collaboration: High-quality training data is a key driver of recent advances in AI. However, manually collecting such data in medicine is expensive and time-consuming. Mitigating these costs motivated our development of Snorkel, a framework for principled, programmatic generation of training data. Here experts iteratively encode domain insights and other forms of tacit knowledge as programmatic labeling functions, which are probabilistically modeled to correct for errors and generate cleaned training data [22]. We have worked closely with clinicians to apply Snorkel across many modalities in medicine, including clinical notes [23–25], cardiac MRIs [26], and wearable sensors [27]. In the LLM era, these ideas have evolved into natural language prompting. We have studied programmatic supervision as a prompt-based ("no-code") method, enabling data scientists to train high-performance, low-cost models with little to no manual curation [19]. Our work anticipated programmatic labeling as central to human-AI collaboration and large-scale data curation as essential for modern LLM development in industry [28].

FUTURE WORK

Resilient Systems for Capturing Tacit Knowledge: The next generation of healthcare foundation models need to optimize for effective human-AI teaming. Central to this goal is the development of methods and infrastructure to study AI feedback loops between users and models, enabling the capture of tacit knowledge embedded in healthcare systems. This requires new preference alignment frameworks that can reconcile conflicting priorities among diverse stakeholder groups, enabling medical experts to better guide, interrogate, and correct model behaviors. Achieving this requires resilient AI infrastructure and innovative data integration methods to capture the wealth of data available to practicing clinicians within hospitals. The need to improve interoperability and support no-code and low-code technologies for health data ecosystems is reflected in recent funding calls (ARPA-H-SOL-24-103). Harnessing tacit knowledge through feedback mechanisms also aligns with PRECISE-AI (ARPA-H-SOL-25-113), which emphasizes self-correcting techniques to continuously monitor and maintain the performance of AI decision-support tools. This requires building close partnerships with hospital data teams. My research will focus on building these bridges and developing methods for capturing feedback loops and accelerating human-AI teaming.

Better Benchmarking. My research aims to establish more rigorous evaluation regimes for foundation models in healthcare. Recent failures of continued pretraining to improve medical-specialized LLMs over base models [29] underscore the need for robust evaluation frameworks to guide model development. Current healthcare benchmarks disproportionately focus on simplified, knowledge-based tasks such as multiple-choice exams. These tasks offer limited practical guidance for deploying LLMs in healthcare systems [7,8] and contrast with realistic applications, such as tumor board meetings, which involve coordinating multiple complex subtasks (e.g., retrieving data, summarizing patient histories, assessing risks). Deployed LLMs are also increasingly recognized as compound systems composed of multiple interconnected components, rather than single monolithic models [30]. Benchmarking efforts must therefore evaluate these systems holistically, incorporating components beyond the model itself. Thus fostering partnerships with hospital data science teams and other healthcare entities is essential to establishing recurring validation processes that capture the dynamic nature of deployed models [31]. Finally, benchmarks must follow evaluation best practices, reporting performance across protected subgroups, temporal distribution shifts, and geographic contexts, utilizing statistical rigor that is standard in medical research but not yet common in LLM benchmarking [32].

Multimodal Foundation Models in Healthcare. The next generation of healthcare foundation models must integrate and contextualize medical data from a diverse array of modalities—such as imaging, omics, and wearables—alongside external knowledge sources like medical literature. Current self-supervised learning methods focus on paired data (e.g., image/text) with short temporal contexts, reducing the ability of models to learn features that correlate with long-term disease progression. My research is predicated on the idea that, just as natural language semantically grounds vision-language models, longitudinal EHRs ground multimodal data in the temporal progression of health. Building on the success of CLMBR and MOTOR, I aim to continue developing methods for transforming EHR timelines into novel sources of training supervision for multimodal foundation models. These efforts will build on my expertise in data-centric AI and methods for modeling imperfect information to generate high-quality training data at scale.

In conclusion, data is essential for advancing healthcare foundation models, requiring new methods to maintain feedback loops and capture tacit medical knowledge for training and aligning models. As private and open-source models grow more capable, the need for reliable data integration and evaluation strategies in healthcare systems will only increase.

References

*co-first † co-senior

1. Sahni NR, Carrus B. Artificial intelligence in U.s. health care delivery. N Engl J Med. 2023;389: 348–358.

2. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21: 1470–1480.

3. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision-language foundation model for diverse biomedical tasks. Nat Med. 2024; 1–13.

4. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616: 259–265.

5. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: The potentials and pitfalls : A narrative review: A narrative review. Ann Intern Med. 2024;177: 210–220.

6. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, **Fries JA**, Wornow M, Swaminathan A, Lehmann LS, Hong HJ, Kashyap M, Chaurasia AR, Shah NR, Singh K, Tazbaz T, Milstein A, Pfeffer MA, Shah NH. Testing and evaluation of health care applications of large language models: A systematic review: A systematic review. JAMA. 2024 [cited 24 Nov 2024]. doi:10.1001/jama.2024.21700

7. Bedi S, Jain SS, Shah NH. Evaluating the clinical benefits of LLMs. Nat Med. 2024;30: 2409–2410.

8. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. 2023;330: 866–869.

9. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med. 2020;3: 107.

10. Steinberg E, Jung K, **Fries JA**, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. J Biomed Inform. 2021;113: 103637.

11. Lemmon J, Guo LL, Steinberg E, Morse KE, Fleming SL, Aftandilian C, Pfohl SR, Posada JD, Shah N, **Fries JA**, Sung L. Self-supervised machine learning using adult inpatient data produces effective models for pediatric clinical prediction tasks. J Am Med Inform Assoc. 2023. doi:10.1093/jamia/ocad175

12. Guo LL, Steinberg E, Fleming SL, Posada J, Lemmon J, Pfohl SR, Shah N, †**Fries JA**, † Sung L. EHR foundation models improve robustness in the presence of temporal distribution shift. Sci Rep. 2023;13: 3767.

13. *Guo LL, ***Fries JA**, Steinberg E, Fleming SL, Morse K, Aftandilian C, et al. A multi-center study on the adaptability of a shared foundation model for electronic health records. NPJ Digit Med. 2024;7: 171.

14. Wornow M, Thapa R, Steinberg E, †**Fries JA**, †Shah N. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. Adv Neural Inf Process Syst. 2023 [cited 9 Nov 2023]. Available: https://openreview.net/pdf?id=CsXC6IcdwI

15. *Steinberg E, ***Fries JA**, Xu Y, Shah N. MOTOR: A Time-to-Event Foundation Model For Structured Medical Records. The Twelfth International Conference on Learning Representations. 2023. Available: https://openreview.net/pdf?id=NialiwI2V6

16. *Huo Z, ***Fries JA**, *Lozano A, Valanarasu JMJ, Steinberg E, Blankemeier L, et al. Time-to-event pretraining for 3D medical imaging. arXiv [cs.CV]. 2024. Available: https://arxiv.org/abs/2411.09361.

17. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, **Fries JA**, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. NPJ Digit Med. 2023;6: 135.

18. Huang SC, Huo Z, Steinberg E, Chiang CC, Lungren MP, Langlotz C, Yeung S, Shah N, **Fries JA**. INSPECT: A

Multimodal Dataset for Pulmonary Embolism Diagnosis and Prognosis. Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track. 2023. Available: https://openreview.net/pdf?id=3sRR2u72oQ

19. Fleming SL, Lozano A, Haberkorn WJ, Jindal JA, Reis EP, Thapa R, Blankemeier L, Genkins JZ, Steinberg E, Nayak A, Patel BS, Chiang CC, Callahan A, Huo Z, Gatidis S, Adams SJ, Fayanju O, Shah SJ, Savage T, Goh E, Chaudhari AS, Aghaeepour N, Sharp C, Pfeffer MA, Liang P, Chen JH, Morse KE, †Brunskill EP, †**Fries JA**, †Shah NH. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. AAAI Press; 2023. Available: http://arxiv.org/abs/2308.14089

20. Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health. ICLR 2024 Workshop on Learning from Time Series For Health. 2024. Available: https://openreview.net/pdf?id=IsHy2ebjIG

21. Steinberg E, Wornow M, Bedi S, **Fries JA**, McDermott MBA, Shah NH. meds_reader: A fast and efficient EHR processing library. arXiv [cs.LG]. 2024. Available: http://arxiv.org/abs/2409.09095

22. Ratner A, Bach SH, Ehrenberg H, **Fries JA**, Wu S, Ré C. Snorkel: Rapid Training Data Creation with Weak Supervision. Proceedings VLDB Endowment. 2017;11: 269–282.

23. **Fries JA**, Wu S, Ratner A, Ré C. SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data. arXiv [cs.CL]. 2017. Available: http://arxiv.org/abs/1704.06360

24. *Callahan A, ***Fries JA**, Ré C, Huddleston JI 3rd, Giori NJ, Delp S, et al. Medical device surveillance with electronic health records. NPJ Digit Med. 2019;2: 94.

25. **Fries JA**, Steinberg E, Khattar S, Fleming SL, Posada J, Callahan A, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. Nat Commun. 2021;12: 2017.

26. **Fries JA**, Varma P, Chen VS, Xiao K, Tejeda H, Saha P, et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. Nat Commun. 2019;10: 3111.

27. Varma P, Sala F, Sagawa S, **Fries JA**, Fu D, Khattar S, et al. Multi-Resolution Weak Supervision for Sequential Data. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. pp. 192–203.

28. Nvidia, Adler B, Agarwal N, Aithal A, Anh DH, Bhattacharya P, et al. Nemotron-4 340B Technical Report. arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2406.11704

29. Jeong DP, Garg S, Lipton ZC, Oberst M. Medical adaptation of large language and vision-language models: Are we making progress? arXiv [cs.CL]. 2024. Available: http://arxiv.org/abs/2411.04118

30. Gupta R, Matei Zaharia, Omar Khattab, Lingjiao Chen, Jared Quincy Davis, Heather Miller, Chris Potts, James Zou, Michael Carbin, Jonathan Frankle, Naveen Rao, Ali Ghodsi. The Shift from Models to Compound AI Systems. In: The Berkeley Artificial Intelligence Research Blog [Internet]. [cited 25 Nov 2024]. Available: http://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/

31. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. Nat Med. 2023. doi:10.1038/s41591-023-02540-z

32. Miller E. Adding error bars to evals: A statistical approach to language model evaluations. arXiv [stat.AP]. 2024. Available: http://arxiv.org/abs/2411.00640