

JASON ALAN FRIES

Curriculum Vitae

jfries@stanford.edu

<http://web.stanford.edu/~jfries/>

Education

Postdoctoral Fellowship, Computer Science	<i>Stanford University, 2017</i>
Ph.D., Computer Science	<i>University of Iowa, 2015</i>
B.A., Computer Science	<i>University of Iowa, 2009</i>
B.A., English with Honors	<i>University of Iowa, 2009</i>

Research Interests

- Foundation models and generative AI for healthcare.
- Data-centric AI, focusing on training data curation, data generation, and quality assessment.
- Learning with limited labeled data (e.g., weak supervision, zero/few-shot learning).
- Human-AI collaboration and teaming.

Research Experience

Research Scientist *Stanford Center for Biomedical Informatics Research* Dec. 2017 to present

- Developing, evaluating, and deploying healthcare foundation models for natural language, structured electronic health records (EHR) data, and medical imaging.
- Automating instruction and preference data generation for training healthcare LLMs.
- Improved AI reproducibility in healthcare by releasing EHR foundation models and datasets, including the first clinician-generated instruction-following benchmark (MedAlign) and multimodal longitudinal EHR datasets (EHRSHOT, INSPECT).
- Leading development of Med-HELM, a platform for benchmarking healthcare LLMs by expanding evaluation beyond knowledge-based, multiple-choice questions.

Research Scientist/Consultant *Snorkel AI* Mar. 2021 to present

- Developing techniques for model distillation and data curation using large language models, enabling the creation of cost-effective, smaller models suitable for enterprise deployment

Postdoctoral Scholar, Computer Science *Stanford University* Aug. 2015 to Dec. 2017

- Distinguished postdoctoral fellow with the Mobilize Center, an interdisciplinary research group developing methods for understanding human mobility.
- Collaborated on development of the Snorkel weak supervision framework and applications of data-centric AI to natural language processing, EHR data, and medical imaging.

Research Assistant *University of Iowa* Jan. 2009 to Aug. 2013

- Member of the Computational Epidemiology Research Group, an interdisciplinary team of computer scientists, physicians, and other health professionals, developing computational approaches to modeling the spread of disease.

- Explored NLP and machine learning applications in data-driven health surveillance, including monitoring antibiotic usage in hospitals and tracking disease outbreaks via social media.

Undergraduate Research Assistant *University of Iowa*

Jan. 2007 - Dec. 2008

- Used wearable sensors to capture high-resolution movement data of hospital staff to estimate contact graphs and optimize interventions, e.g., vaccine allocation, hand hygiene surveillance.

Selected Publications

*co-first authorship †co-senior authorship

Conference

1. *Huo Z, *Fries JA, *Lozano A, Valanarasu JM, Steinberg E, Blankemeier L, Chaudhari AS, Langlotz C, Shah NH. Time-to-Event Pretraining for 3D Medical Imaging. *International Conference on Learning Representations (ICLR)*. 2025.
2. *Wornow M, *Bedi S, Hernandez M, Steinberg E, Fries JA, Ré C, Koyejo S, Shah NH. Context Clues: Evaluating Long Context Models for Clinical Prediction Tasks on EHRs. *International Conference on Learning Representations (ICLR)*. 2025.
3. *Steinberg E, *Fries JA, Xu Y, Shah NH. MOTOR: A time-to-event foundation model for structured medical records. *International Conference on Learning Representations (ICLR)*. 2024 ([Spotlight](#)).
4. *Fleming SL, *Lozano A, *Haberkorn WJ, *Jindal JA, *Reis EP, Thapa R, Blankemeier L, Jenkins JZ, Steinberg E, Nayak A, Patel BS, ..., †Brunskill E, †Fries JA, †Shah NH. MedAlign: A clinician-generated dataset for instruction following with electronic medical records. *ML4H Symposium 2023 (Best Thematic Paper Award)*. Association for the Advancement of Artificial Intelligence (AAAI). 2024 ([Oral](#)).
5. Huang SC, Huo Z, Steinberg E, Chiang C, Lungren M, Langlotz C, Yeung S, Shah NH, Fries JA. INSPECT: A multimodal dataset for pulmonary embolism diagnosis and prognosis. *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. 2023.
6. Wornow M, Thapa R, Steinberg E, †Fries JA, †Shah, NH. EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. 2023 ([Spotlight](#)).
7. Blankemeier L, Fries JA, Tinn R, Preston J, Shah N, Chaudhari A. Efficient Diagnosis Assignment Using Unstructured Clinical Notes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. 2023.
8. *Fries JA, *Weber L, *Seelam N, *Altay G. et al. BigBio: A framework for data-centric biomedical natural language processing. *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. 2022.
9. *Sanh V, *Webson A, *Raffel C, *Bach SH, Sutawika L, Alyafeai Z, Chaffin A, Stiegler A, Le Scao T, Raja A, Dey M, et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations (ICLR)*. 2022 ([Spotlight](#)).

10. *Bach S, *Sanh V, Yong ZX, Webson A, Raffel C, Nayak NV, Sharma A, Kim T, Bari MS, Févry T, Alyafeai Z, et al. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*. 2022
11. *Varma P, *Sala F, **Fries JA**, Fu D, Sagawa S, Khattar S, Ramamoorthy A, Xiao K, Fatahalian K, Priest J, Ré C. Multi-Resolution Weak Supervision for Sequential Data. *Conference on Neural Information Processing Systems (NeurIPS)*. 2019.
12. Fiterau M, Bhooshan S., **Fries JA**, Bournhonesque C, Hicks J., Halilaj E., Ré C. Delp S. ShortFuse: Biomedical Time Series Representations in the Presence of Structured Information. *Machine Learning in Healthcare*. 2017. PMID: PMC6417829.

Journal

1. Ip W, Xenochristou M, Sui E, Ruan E, Ribeira R, Dash D, Srinivasan M, Artandi M, Omiye JA, Scoulios N, Hofmann HL, Mottaghi A, Weng Z, Kumar A, Ganesh A, **Fries JA**, Yeung-Levy S, Hofmann LV. Hospitalization prediction from the emergency department using computer vision AI with short patient video clips. *npj Digital Medicine*. 2024. PMID: 39702364 PMID: PMC11659512.
2. 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. *JAMA*. 2024. PMID: 39405325; PMID: PMC11480901.
3. Chiang CC, **Fries JA**. Exploring the Potential of Large Language Models in Neurology, Using Neurologic Localization as an Example. *Neurol Clin Pract*. 2024. PMID: 38601858; PMID: PMC11003354.
4. *Guo LL, ***Fries JA**, Steinberg E, Fleming SL, Morse KE, Aftandilian C, Posada J, †Shah NH, †Sung L. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*. 2024. PMID: PMC11211479.
5. *Smith R, ***Fries JA**, Hancock B, Bach S. Language models in the loop: Incorporating prompting into weak supervision. *ACM/IMS Journal of Data Science*. 2024.
6. 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 Digital Medicine*. 2023. PMID: PMC10387101.
7. 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. *Journal of the American Medical Informatics Association*. 2023. PMID: PMC10654865.
8. Guo LL, Steinberg E, Fleming SL, Posada J, Lemmon J, Pfohl SR, Shah NH, †**Fries JA**, †Sung L. EHR foundation models improve robustness in the presence of temporal distribution shift. *Scientific Reports*. 2023. PMID: PMC9992466.
9. Guo LL, Pfohl SR, **Fries JA**, Johnson AE, Posada J, Aftandilian C, Shah N, Sung L. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset

- shift in clinical medicine. *Scientific Reports*. 2022. PMID: PMC8854561.
10. **Fries JA**, Steinberg E, Khattar S, Fleming SL, Posada J, Callahan A, Shah NH. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*. 2021. PMID: PMC8016863.
 11. Guo LL, Pfohl SR, **Fries JA**, Posada J, Fleming SL, Aftandilian C, Shah N, Sung L. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Applied Clinical Informatics*. 2021. PMID: PMC8410238.
 12. 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. *Journal of Biomedical Informatics*. 2021. PMID: PMC7863633.
 13. Giori NJ, Radin J, Callahan A, **Fries JA**, Halilaj E, Ré C, Delp SL, Shah NH, Harris AH. Assessment of Extractability and accuracy of electronic health record data for Joint implant registries. *JAMA Network Open*. 2021. PMID: PMC7961313.
 14. *Callahan A, *Steinberg E, **Fries JA**, Gombar S, Patel B, Corbin CK, Shah NH. Estimating the efficacy of symptom-based screening for COVID-19. *npj Digital Medicine*. 2020. PMID: PMC7359358.
 15. Miner AS, Haque A, **Fries JA**, Fleming SL, Wilfley DE, Terence Wilson G, Milstein A, Jurafsky D, Arnow BA, Stewart Agras W, Fei-Fei L. Assessing the accuracy of automatic speech recognition for psychotherapy. *npj Digital Medicine*. 2020. PMID: PMC7270106.
 16. **Fries JA**, Varma P, Chen VS, Xiao K, Tejeda H, Saha P, Dunnmon J, Chubb H, Maskatia S, Fiterau M, Delp S. †Ashley E, †Ré C, †Priest JR. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nature Communications*. 2019. PMID: PMC6629670.
 17. *Callahan A, ***Fries JA**, Ré C, Huddleston J, Giori N, Delp S, Shah NH. Medical device surveillance with electronic health records. *npj Digital Medicine*. 2019. PMID: PMC6761113.
 18. Ratner A, Bach SH, Ehrenberg H, **Fries JA**, Wu S, Ré C. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*. 2019. PMID: PMC7075849. ([Best of VLDB 2018](#))
 19. **Fries JA**, Segre AM, Thomas G, Herman T, Ellingson K, Polgreen PM. Monitoring Hand Hygiene via Human Observers: How Should We Be Sampling? *Infection Control and Hospital Epidemiology*. 2012. PMID: PMC3632316.

Preprint / Workshop

1. *Cui H, *Unell A, Chen B, **Fries JA**, Alsentzer E, Koyejo S, Shah NH. TIMER: Temporal Instruction Modeling and Evaluation for Longitudinal Clinical Records. 2025. *Under Review*.
2. *Munnangi M, *Swaminathan A, ***Fries JA**, Jindal J, Narayanan S, Lopez I, Tu L, Chung P, Omiye T, Kashyap M, Shah N. FactEHR: A Dataset for Evaluating Factuality in Clinical Notes Using LLMs. 2025. arXiv preprint arXiv:2412.12422. 2024. *Under Review*.
3. Steinberg E, Wornow M, Bedi S, **Fries JA**, McDermott M, Shah NH. meds_reader: A fast and efficient EHR processing library. ML4H Symposium. 2024.

4. Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D, Castagné R, Luccioni AS, Yvon F, Gallé M, et. al. Bloom: A 176b-parameter open-access multilingual language model. *BigScience Workshop*. 2022.
5. *Fries JA, *Seelam N, *Altay G, *Weber L, *Kang M, *Datta D, *Su R, *Garda S, Wang B, Ott S, Samwald M. Dataset debt in biomedical language modeling. *In Challenges & Perspectives in Creating Large Language Models*. 2022.
6. Saelig K, O'Day H, Varma P, Fries JA, Hicks J, Delp S, Bronte-Stewart H, Ré C. Multi-frame Weak Supervision to Label Wearable Sensor Data. *Time Series Workshop @ ICML*. 2019.
7. Fries JA, Wu S, Ratner A, Ré C. Swellshark: A generative model for biomedical named entity recognition without labeled data. *Preprint*. 2017.
8. Fries JA. Brundlefly at SemEval-2016 Task 12: Recurrent Neural Networks vs. Joint Inference for Clinical Temporal Information Extraction. *Proceedings of SemEval (2016)*. 2016.

Grants

- **2021 AIMI-HAI Partnership Grant (Co-PI)** Award: \$200,000
[Self-service Data Science in the EHR with Multimodal Patient Embeddings](#)

Awards & Honors

- **2024 Spotlight – ICLR: MOTOR: A time-to-event foundation model for structured medical records**
- **2024 Oral – AAAI: MedAlign: A clinician-generated dataset for instruction following with electronic medical records**
- **2023 Best Thematic Paper Award – Machine Learning for Health (ML4H): MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records**
- **2023 Spotlight - NeurIPS Datasets and Benchmarks: EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models**
- **2022 Spotlight – ICLR: Multitask Prompted Training Enables Zero-Shot Task Generalization**
- **2018 Best of VLDB 2018: Snorkel: Rapid Training Data Creation with Weak Supervision**

Teaching Experience

Guest Lectures

BIODS 271: Foundation Models for Healthcare <i>Foundation Models for EHRs</i>	May 28, 2025
BIOMEDIN 212: Intro to Biomedical Informatics Research Methodology <i>Scientific Writing</i>	May 27, 2025
2024 AIMI Bootcamp <i>NLP and Foundation Models in Healthcare</i>	June 24, 2024
BIODS 271: Foundation Models for Healthcare <i>Foundation Models for EHR Data</i>	Mar. 6, 2024
BIOMEDIN 215: Data Science for Medicine, Stanford University <i>Deep Learning for Healthcare AI</i>	Dec. 3, 2023
EPI 270: Big Data Methods for Behavioral, Social, and Pop. Health, Stanford University <i>Rapid NLP for Data Scientists: Weakly Supervised Information Extraction</i>	May 25, 2021

Tutorial Sessions

Snorkel Workshops [2017](#), [2018](#)

July 2017, Nov. 2018

Designed 2-day workshop on using generative models to create training data for biomedical tasks.

Instructor

CS3210: Programming Languages and Tools: C++, University of Iowa

2013 - 2015

Primary course instructor for 20 to 30 undergraduate/graduate students. I designed all lectures, web materials, programming projects, and exams. I graded all course materials and assigned final grades.

Recent Invited Talks & Panel Discussions

8th Healthcare Text Analytics Conference (HealTAC) Keynote	June 18, 2025
The Missing Context Problem in Foundation Models for Healthcare	
ICLR 2025 Workshop on AI for Children Keynote	Apr. 27, 2025
Talk. The Missing Context Problem in Foundation Models for Healthcare	
AHLI Machine Learning for Health (ML4H) Symposium Keynote	Dec. 15, 2024
The Missing Context Problem in Foundation Models for Healthcare	
DCAI WWW'24 Workshop Keynote	May 13, 2024
Uncharted: The Road to Data-Centric Benchmarking for Medical Foundation Models	
UCSF and the Zuckerberg San Francisco General Hospital (ZSFG)	Apr. 29, 2024
Talk: Generative AI for Electronic Health Records	
Clinical Foundation Models AAAI 2024 Spring Symposium	Mar. 25, 2024
Invited Panelist	
UC Berkeley/UCSF CPH's Doctoral Seminar Series	Feb. 28, 2024
Talk: Generative AI for Electronic Medical Records	
The Fred Hutchinson Cancer Center	Feb. 1, 2024
Talk: Generative AI for Electronic Health Records	
University of Florida PPHP AI Seminar	Jan. 26, 2024
Talk: Responsible Language Models in Healthcare Require Data-Centric AI	
Deep Generative Models for Health Workshop @ NeurIPS 2023	Dec. 15, 2023
Invited Panelist	
Health Informatics Grand Rounds, University of Pittsburgh	Dec. 15, 2023
Talk: Model Hubs for Medical AI: How Far is Our "Hugging Face" Moment?	
Snorkel AI: Enterprise LLM Virtual Summit	Oct. 26, 2023
Talk: Better not Bigger: Distilling LLMs into Specialized Models	
Deloitte	Sept. 12, 2023
Talk: Realizing the Promise of Foundation Models in Healthcare	
AHLI The Conference on Health, Inference, and Learning (CHIL) 2023	June 23, 2023
Generalizability in Machine Learning for Health: Critical for Robustness or a Distraction from Specific Validation? Invited Panelist	
Pistoia Alliance	June 20, 2023
Talk: Realizing the Promise of Foundation Models in Healthcare	
2023 Stanford AIMI Symposium: Generative AI in Health	June 6, 2023

Talk: Generative AI for Electronic Health Records	
NHLBI/NIH	May 17, 2023
Talk: Realizing the Promise of Foundation Models in Healthcare	
Stanford BMIR Colloquium	April 13, 2023
Talk: Realizing the Promise of Foundation Models in Healthcare	
Pfizer AI/ML Community of Practice	Mar. 28, 2023
Talk: Realizing the Promise of Foundation Models in Healthcare	
2022 INFORMS Annual Meeting	Nov. 16, 2022
Talk: The Road to Data-Centric Machine Learning in Healthcare	
Tempus	Aug. 28, 2022
Talk: The Road to Data-Centric Machine Learning in Healthcare	
Snorkel AI: The Future of Data-Centric AI Workshop	Aug. 3, 2022
Panel Discussion	
National Academies: AI and Open Data Practices in Chemical Hazard Assessment	May 25, 2022
Talk: Programmatic Labeling for Data-Centric NLP	
2021 Stanford AIMI Center Symposium	Aug. 3, 2021
Talk: Data-centric Medical AI with Weak Supervision	
Selected Media & Articles	
<hr/>	
Guest on Pitt HexAI's Health and Explainable AI Podcast	Aug. 22, 2023
How Foundation Models Can Advance AI in Healthcare	Dec. 15, 2022
Science Talks: Weak Supervision in Biomedicine	June 16, 2021
Community Engagement & Leadership	
<hr/>	
Co-organizer: GenAI for Health Workshop @ NeurIPS	2024
Co-organizer: Multimodal4Health Workshop@ ICHI	2024
Area Chair: Machine Learning for Healthcare (MLHC)	2019 - 2023
Co-Chair: BigScience Biomedical Working Group - Summer of Language Models	2021 - 2022
Co-organizer: Snorkel Workshops with the Mobilize Center	2017 - 2018
Co-organizer: Machine Learning for Health Workshop @ NeurIPS	2016 - 2018
Co-organizer: Learning to Run Competition @ NeurIPS	2017

References

Nigam Shah

Professor, Medicine - Biomedical Informatics Research

Professor, Biomedical Data Science

Stanford University

Chief Data Scientist

Stanford Health Care

nigam@stanford.edu

(650) 725-6236

Christopher Ré

Associate Professor, Computer Science

Stanford University

For correspondence, please contact: ann1083@stanford.edu

chrismre@cs.stanford.edu

Matthew Lungren

Chief Data Science Officer

Microsoft Health & Life Sciences

mlungren@microsoft.com

Scott Delp

Professor, Departments of Bioengineering and Mechanical Engineering

Stanford University

James H. Clark Center, Room S-321

Stanford, CA 94305

For correspondence, please contact: lisa98@stanford.edu

delp@stanford.edu

(650) 723-1230