

# JUNWEI LU

Assistant Professor of Biostatistics

677 Huntington Avenue, 2-409, Boston MA 02115

+1-617-432-4910 | [junweilu@hsph.harvard.edu](mailto:junweilu@hsph.harvard.edu) | [junwei-lu.github.io](https://junwei-lu.github.io)

## EDUCATION

Date	Discipline	Degree	Institution
2018	Operations Research and Financial Engineering	Ph.D.	Princeton University
2012	Mathematics	B.Sc	Fudan University

## ACADEMIC APPOINTMENT

Dates	Title	Department	Institution
2018 - Present	Assistant Professor	Biostatistics	Harvard T.H. Chan School of Public Health

## COMMITTEE SERVICE

### DEPARTMENTAL/SCHOOL AND UNIVERSITY SERVICE

Dates	Name of Committee / Role	Institution
2024 - Present	Undergraduate Summer Research Program / Member	Office of the Vice Provost
2023 - Present	Department computing committee / Chair	Department of Biostatistics
2022 - Present	Department EDIB committee / Faculty representative	Department of Biostatistics
2019 - Present	Data Science in Action committee / Program Director	Harvard Medical School
2019 - Present	CBQG Master program executive committee / Member	Department of Biostatistics
2018 - 2021	Doctoral admissions committee / Member	Department of Biostatistics
2019 - 2020	Department Colloquium committee / Member	Department of Biostatistics

### PROFESSIONAL ACTIVITIES

Dates	Role	Organization/Conference Name
2024/06	Conference Session Chair	ENAR, International Biometric Society
2024/05	Conference Organization Chair	Radcliffe Exploratory Seminar
2024/01	Senior Program Committee	International Conference on Frontiers of Data Science
2023/04	Senior Program Committee	International Forum on Statistics
2022/01	Seminar Organizer	Mathematical Foundation of Data Science
2021/03	Seminar Organizer	National Institute of Statistical Sciences
2020/03	Conference Session Organizer	International Chinese Statistical Association
2019/03	Conference Session Organizer	New England Statistical Society
2018/11	Conference Session Organizer	International Chinese Statistical Association

### EDITORIAL ROLES

Ad hoc reviewer: Science Advances, Annals of Statistics, Journal of the American Statistical Association (JASA), Biometrika, Annals of Applied Statistics, Journal of Machine Learning Research, Journal of Multivariate Analysis, Electronic Journal of Statistics (EJS), Neural Information Processing Systems, International Conference on Machine Learning (ICML).

## FUNDED GRANTS AND PENDING PROJECTS

### ACTIVE GRANTS

2024 – 2027	NSF DMS 2434664 Contact Principal Investigator, \$1,197,523.00 (total award) <i>Collaborative Research: AIMING: AI Tools to Knowledge Discovery and Rigorous Reasoning in Polyhedral Geometry</i> Major Goals: The objective of this proposal is to develop innovative AI tools for generating polyhedral samples, discovering and interpreting conjectures, and rigorous reasoning on polyhedral geometry with applications to open problems like the polynomial Hirsch conjecture.
2024 – 2027	AWS Impact Computing Project at the Harvard Data Science Initiative Co-Principal Investigator, \$379,324.00 (total award)

*Development of a deep learning model to enable serosurveillance of vector-borne diseases and their hosts*

Major Goals: The goal of the proposal is to develop “VectorScan,” a tool that uses a phage display library to detect antibodies against pathogens and vectors like mosquitoes and ticks, in order to study the impact of climate change on vector-borne diseases. By analyzing these antibodies in blood samples, the project aims to create comprehensive data on disease prevalence, which will help in understanding the link between climate change and the spread of diseases such as malaria, Zika, and Lyme.

2024 – 2025

William F. Milton Fund

Principal Investigator, \$98,144.90 (total award)

*Generative Causal AI for Alzheimer’s Disease Therapeutic Management*

Major Goals: The research aims to tackle the complex issues linked to Alzheimer’s Disease (AD) by employing a novel approach that merges generative artificial intelligence (AI) with causal inference techniques. This cutting-edge framework is designed to improve the analysis of diverse clinical data, facilitating more effective causal studies and revealing precise therapeutic avenues for AD.

2024 – 2027

NIH/NINDS R01NS098023

Co-Investigator

*Leveraging electronic health records to optimize treatment selection and response in multiple sclerosis*

Major Goals: The goal of the proposed project is to develop analytical approaches incorporating electronic health records data at the point of care to optimize the selection of disease-modifying treatments and patient outcomes in people with multiple sclerosis.

2022 – 2025

Merck Sharp and Dohme Corp LKR204888

Co-Investigator

*Electronic Health Records (EHR) and Imaging driven Prognostic Modeling of Renal Cell Carcinoma (RCC)*

Major Goals: The overall objective of this project is to understand the disease progression of early-stage (T1 and T2 subgroups) renal cell carcinoma (RCC). HDBMI will construct an electronic health records (EHR) cohort of early-stage RCC patients and develop prognostic models for RCC recurrence or progression using digital imaging linked EHR data.

2022 – 2025

NIH OT2OD032581

Research Associate

*AIM-AHEAD Coordinating Center: Infrastructure 1 and Research Fellowship Administration*

Major Goals: The goal of this project is to develop an adaptable infrastructure that makes AI/ML research accessible to currently underrepresented communities and addresses health inequities, as part of the AIM-AHEAD consortium.

2021 – 2024

NIH/NIEHS R01ES032418

Co-Investigator

*Identifying low dose measurement error corrected effects of multiple pollutants using causal modeling*

Major Goals: The proposed study will determine whether there are causal effects of exposure to air pollutants and temperature on mortality, heart attacks, and strokes and whether those effects persist below current ambient standards. The study will also identify whether there are interactions between the different exposures and whether some particle components are more toxic than others.

2018 – 2024

Department of Veteran Affairs 36C24E18D0048/36C24E20N0129

Co-Investigator

*MAVERIC Project*

Major Goals: Aims to develop and apply bioinformatics approaches for phenotyping and analysis of high dimensional datasets directly translatable to clinical research studies.

2018 – 2025

NIH/NCI R35CA220523

Co-Investigator

*Unraveling the Complexities of Risk and Mechanism in Cancer*

Major Goals: This project will develop advanced computational methods to explore the regulatory networks that govern cancer development and response to therapies, including differences in men and women.

## COMPLETED GRANTS

- 2019 – 2022 NSF DMS 1916211  
Principal Investigator, \$147,258.00 (total award)  
*Combinatorial Inference: Statistical Uncertainty Assessment for Discrete Structures*  
Major Goals: The objective of this proposal aims to develop novel inferential methods for unknown discrete structures arising from statistical models. To assess uncertainty when inferring these discrete structures, we propose to develop a new research field named combinatorial inference. This proposal also aims to train the next generation of data scientists to acquire competitive inferential skills for complicated structures arising from large-scale scientific problems.
- 2017 – 2023 NIH/NCI U01CA209414  
Co-Investigator  
*The Boston Lung Cancer Survival Cohort*  
Major Goals: This study uses the combination of biomarker data, tumor characterization, traditional epidemiologic risk factor data, and electronic medical records will facilitate powerful translational research and provide unique opportunities to further explore predictors of overall survival and treatment outcome.

## PENDING GRANTS

- 2025 – 2030 NIH/NIA Pending R01  
Principal Investigator  
*Scalable statistical relational learning with Multi-institutional EHR on Alzheimer's Disease*  
Major Goals: As the leading cause of dementia and chronic neurological disability, AD is the ideal test case since people with AD experience highly variable rates of disease progression and trajectories of decline. The methodological aims include leveraging longitudinal EHR data and directional KG with embedding of continuous variables to improve the prediction of clinically relevant outcomes in AD.
- 2025 – 2030 NSF DMS Pending  
Principal Investigator  
*CAREER: Structural Inference and Statistical Decision-Making for Generative AI Methods*  
Major Goals: This proposal is constructed in the context of a new generation of generative AI methods, developed to handle the increasing complexities of modern data, focusing on enhancing inferential capabilities in scientific applications.
- 2025 – 2030 NIH/NIA Pending R01  
Multiple Principal Investigator  
*Probabilistic Multi-Scale Alignment Networks using Temporal High-Dimensional Proteomics Data in Early Prediction of Alzheimer's Disease in Women and Minority Populations*  
Major Goals: The proposed research addresses the urgent need for advanced computational methods to utilize multi-scale biological information to predict Alzheimer's disease (AD) in women and minority populations, enhancing accuracy and causal inference of AD predictions.
- 2025 – 2030 NIH/NCI Pending U01  
Multiple Principal Investigator  
*Large Scale Distributed Integrative Knowledge Extraction from Multiple EHR Systems*  
Major Goals: The proposal advocates for utilizing large-scale EHR to address the limitations of traditional cancer research methods, focusing on enhancing generalizability and overcoming small sample sizes through detailed, longitudinal clinical data from diverse populations.
- 2025 – 2030 NIH/NCI Pending R01  
Multiple Principal Investigator  
*Multi-Site Collaborative Learning Methods for Cancer Research using Longitudinal Observational Data*  
Major Goals: We aim to develop collaborative learning across multiple healthcare institutions for more reliable and dynamic clinical decision-making in the realm of lung cancer care, ensuring both data security and efficiency. The proposed research addresses the pressing need for the generalizable and transferable statistical methods analyzing the multi-institutional and longitudinal NSCLC treatment datasets and proposes to solve the three key questions in precision oncology: the transferable prediction of recurrence, the characterization patients' clinical variability, and the treatment decision making for immunotherapy.

2025 – 2028 NSF DMS Pending  
Principal Investigator  
*FDT-BioTech: Spatially and Temporally Multi-Scale Aging Clock in silico with Uncertainty Quantification and Automated Hypotheses Discovery*  
Major Goals: This project aims to develop the next generation of aging clock by leveraging the digital twin integrating the multi-omic data, providing a more holistic view of the aging process.

## TEACHING AND TRAINING

---

### TEACHING IN HARVARD CHAN SCHOOL COURSES

Dates	Course number / Title	Responsibility
Fall 2023	BST235 / Advanced Regression & Statistical Learning	Instructor
Fall 2022	BST235 / Advanced Regression & Statistical Learning	Instructor
Fall 2021	BST235 / Advanced Regression & Statistical Learning	Instructor
Fall 2020	BST235 / Advanced Regression & Statistical Learning	Instructor
Spring 2020	BST263 / Statistical Learning	Instructor

### TEACHING IN OTHER COURSES

Dates	Course Name	Responsibility
2019-2024	Harvard Data Science in Action Summer School	Instructor/Mentor
2024/07	Harvard StatStart Summer School	Instructor
2024/11	Harvard Data Adventure Day	Instructor

### ADVISORY AND SUPERVISORY RESPONSIBILITIES

Dates	Name of Trainee	Type of Supervision	Current Position
2023-Present	Dominic DiSanto	Ph.D. Thesis co-advisor	Ph.D. candidate
2019-Present	Zebin Wang	Ph.D. Thesis advisor	Ph.D. candidate
2023-Present	Parker Knight	Ph.D. Dissertation committee	Ph.D. candidate
2023-Present	Carmen Rodriguez	Ph.D. Dissertation committee	Ph.D. candidate
2023-Present	Lola Wang	MS Thesis advisor	Master student
2023-Present	Yibin Xiong	MS Thesis advisor	Master student
2024-Present	Yinuo Cheng	Academic mentor	Master student
2024-Present	Junxi Feng	Academic mentor	Master student
2024-Present	Yaoxuan Ju	Academic mentor	Master student
2024-Present	Xi Li	Academic mentor	Master student
2024-Present	Bowen Ma	Academic mentor	Master student
2024-Present	Nick Rezaee	Academic mentor	Master student
2024-Present	Raunak Vijay	Academic mentor	Master student
2024-Present	Yibin Xiong	Academic mentor	Master student
2024-Present	Xinran Yu	Academic mentor	Master student
2024-Present	Nadia Zhao	Academic mentor	Master student
2023-Present	Chris Wang	Academic mentor	Master student
2023-Present	Yuelai Wang	Academic mentor	Master student
2021-2024	Jerry Chang	Dissertation committee	Industry
2019-2024	Yue Liu	Ph.D. Thesis co-advisor	Industry
2019-2023	Shuting Shen	Ph.D. Thesis co-advisor	Postdoc, Duke U.
2019-2023	Patrick Nnamdi	Ph.D. Thesis co-advisor	Postdoc, HSPH
2019-2022	Lu Zhang	Ph.D. Dissertation committee	Industry
2019-2022	Matthew Quinn	Ph.D. Dissertation committee	Industry
2019-2022	Harrison Reeder	Ph.D. Dissertation committee	Instructor, HMS
2019-2021	Molei Liu	Ph.D. Dissertation committee	Faculty, Columbia U.
2020-2021	Shaoling Han	MS Thesis advisor	Industry
2020-2021	Wenjie Gu	MS Thesis advisor	Industry
2021-2022	Raphael Kim	Academic mentor	Industry
2021-2022	Junlun Li	Academic mentor	Industry
2021-2022	Yujie Zhang	Academic mentor	Industry
2021-2022	Guangya Wan	Academic mentor	Industry
2021-2022	Xin Xiong	Academic mentor	Industry

**INVITED PRESENTATIONS**

<b>Dates</b>	<b>Venue</b>
2024/09	Foundations of Data Science and Machine Learning, Davis, CA
2024/09	Penn Conference on Big Data in Biomedical and Population Health Sciences, Philadelphia, PA
2024/07	Joint Conference on Statistics and Data Science, Kunming, CN
2024/07	International Conference on Frontiers of Data Science, Hangzhou, CN
2024/04	McGill University, Department of Epidemiology, Biostatistics and Occupational Health
2024/04	Annual Conference on Information Sciences and Systems, Princeton, NJ
2024/03	ENAR Spring Meeting, Baltimore, MA
2023/11	Duke University, Department of Biostatistics and Bioinformatics
2023/11	Princeton Day of Statistics, Princeton, NJ
2023/05	The Eleventh International Conference on Learning Representations, Kigali, RW
2023/05	ICSA Applied Statistics Symposium, Ann Arbor, MI
2022/11	Million Veteran Program Annual Conference, St. Petersburg, FL
2019/11	University of Chicago, Booth School of Business
2019/11	ICSA China Statistics Symposium, Hangzhou, CN
2019/10	INFORMS Annual Meeting, Seattle, WA
2019/07	New England Statistics Symposium, Hartford, CT
2019/04	University of Washington, Department of Biostatistics
2018/10	Harvard University, Department of Statistics
2018/08	Fudan University, School of Data Science
2018/03	University of Auckland, Department of Statistics
2018/02	University of California, Irvine, Department of Statistics
2018/02	University of Michigan, Department of Biostatistics
2018/02	Florida State University, Department of Statistics
2018/01	Duke University, Fuqua School of Business
2018/01	University of Florida, Department of Statistics
2018/01	University of British Columbia, Department of Statistics
2018/01	University of Minnesota, Department of Industrial and Systems Engineering
2017/11	Duke University, Department of Biostatistics and Bioinformatics
2017/11	Georgia Institute of Technology, School of Industrial and Systems Engineering
2017/10	INFORMS Annual Meeting, Houston, TX
2017/08	Joint Statistical Meetings, Baltimore, MD
2016/08	Joint Statistical Meetings, Chicago, IL
2015/08	Joint Statistical Meetings, Seattle, WA
2014/05	New England Statistical Symposium, Boston, MA
2013/08	Joint Statistical Meetings, Montreal, QC.

**COMMUNITY SERVICE ACTIVITIES**

<b>Date</b>	<b>Type of activity</b>	<b>Organization</b>
2024/06	IMS Session Chair	Institute of Mathematical Statistics
2023/05	ASA SLDS Paper Award Committee	American Statistical Association
2020/03	ICSA Session Chair	International Chinese Statistical Association
2021/03	Seminar Organizer	National Institute of Statistical Sciences
2019/03	NESS Session Chair	New England Statistical Society
2018/11	ICSA Session Chair	International Chinese Statistical Association

**BIBLIOGRAPHY**

(Note: \* indicates as co-senior author, † indicates lead author was student or advisee at time of writing)

**Statistical Theory & Methods**

- Shen, S.†, Chen, X., Fang, E. X., & Lu, J. (2024). Combinatorial inference on the optimal assortment in multinomial logit models. *Operations Research*, Under revision.
- Xu, Z.†, Gan, Z., Zhou, D., Shen, S., Lu, J.\*, & Cai, T.\* (2024). Inference of dependency knowledge graph for electronic health records. *Journal of the Royal Statistical Society Series B: Statistical Methodology*,

*Under revision.*

- Shen, S.†, Lu, J., & Lin, X. (2024). Fast distributed principal component analysis of large-scale federated data. *Journal of the American Statistical Association*, *Under revision*.  
- won JSM Student Paper Award in ASA SLDS Session in 2023
- Cai, T., & Lu, J. (2024). Knowledge graph embedding with electronic health records data via latent graphical block model. *Journal of Machine Learning Research*, *Under revision*.
- Sonabend A.†, Zhang J.†, Schwartz J., Coull B., & Lu, J. (2024). Scalable Gaussian Process Regression via Median Posterior Inference for Estimating the Health Effects of an Environmental Mixture. *Manuscript*.
- Emedom-Nnamdi, P.†, Smith, T. R., Onnela, J.-P., & Lu, J. (2024). Nonparametric additive value functions: Interpretable reinforcement learning with an application to surgical recovery. *Annals of Applied Statistics*, 0(0), 1-25.
- Zhang, L.†, & Lu, J. (2024). StarTrek: Combinatorial variable selection with false discovery rate control. *The Annals of Statistics*, 52(1), 78-102. doi:10.1214/23-AOS2296.
- Zhou, D., Zhang, Y., Sonabend-W, A., Wang, Z., Lu, J.\*, & Cai, T.\* (2024). Federated offline reinforcement learning. *Journal of the American Statistical Association*, 1–12. doi:10.1080/01621459.2024.2310287.
- Zhou, D., Cai, T., & Lu, J. (2023). Multi-source learning via completion of block-wise overlapping noisy matrices. *Journal of Machine Learning Research*, 24(221), 1–43.
- Zhou, X.†, Tan, K. M., & Lu, J. (2023). Inferring differential hub nodes on differential Gaussian graphical models. *Statistica Sinica*, 35(4).
- Shen, S.†, & Lu, J. (2023). Combinatorial-probabilistic trade-off: P-values of community property test in the stochastic block models. *IEEE Transactions on Information Theory*, 69(10), 6605–6618. doi:10.1109/TIT.2023.3283172.  
- won spotlight paper in International Conference on Learning Representations in 2023  
- won WNAR Best Student Paper Award in 2022
- Lin, Y., Li, S.†, Xu, J., Xu, J., Huang, D., Zheng, W., Cao, Y., & Lu, J. (2023). Graph over-parameterization: Why the graph helps the training of deep graph convolutional network. *Neurocomputing*, 534, 77–85. doi:10.1016/j.neucom.2023.02.054.
- Liu, Y.†, Fang, E. X., & Lu, J. (2023). Lagrangian inference for ranking problems. *Operations Research*, 71(1), 202–223. doi:10.1287/opre.2022.2313.
- Reeder, H. T.†, Lu, J., & Haneuse, S. (2022). Penalized estimation of frailty-based illness–death models for semi-competing risks. *Biometrics*, 79(3), 1657–1669. doi:10.1111/biom.13761.
- Cao, K.†, Chen, Y., Lu, J., Arechiga, N., Gaidon, A., & Ma, T. (2021). Heteroskedastic and imbalanced deep learning with adaptive regularization. *International Conference on Learning Representations*.
- Lu, J., Han, F., & Liu, H. (2021). Robust scatter matrix estimation for high dimensional distributions with heavy tail. *IEEE Transactions on Information Theory*, 67(8), 5283–5304. doi:10.1109/TIT.2021.3088381.
- Ma, C., Lu, J., & Liu, H. (2021). Inter-subject analysis: A partial Gaussian graphical model approach. *Journal of the American Statistical Association*, 116(534), 746–755. doi:10.1080/01621459.2020.1841645.
- Tan, K. M., Lu, J., Zhang, T., & Liu, H. (2020). Estimating and inferring the maximum degree of stimulus-locked time-varying brain connectivity networks. *Biometrics*, 77(2), 379–390. doi:10.1111/biom.13297.
- Jin, Y.†, Wang, Z., & Lu, J. (2020). Computational and statistical tradeoffs in inferring combinatorial structures of Ising model. *Proceedings of the 37th International Conference on Machine Learning*, 119, 4901–4910.

- Sonabend, A.<sup>†</sup>, Lu, J., Celi, L. A., Cai, T., & Szolovits, P. (2020). Expert-supervised reinforcement learning for offline policy learning and evaluation. *Advances in Neural Information Processing Systems*, 33, 18967-18977.
- Ji, D.<sup>†</sup>, Lu, J., Zhang, Y., Gao, S., & Zhao, H. (2020). Inference of dynamic graph changes for functional connectome. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 3230–3240.
- Lu, J., Kolar, M., & Liu, H. (2020). Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association*, 115(532), 2084–2099. [doi:10.1080/01621459.2019.1689984](https://doi.org/10.1080/01621459.2019.1689984).
- Tan, K. M., Lu, J., Zhang, T., & Liu, H. (2019). Layer-wise learning strategy for nonparametric tensor product smoothing spline regression and graphical models. *Journal of Machine Learning Research*, 20(119), 1–38.
- Li, X., Lu, J., Arora, R., Haupt, J., Liu, H., Wang, Z., & Zhao, T. (2019). Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6), 3489–3514. [doi:10.1109/TIT.2019.2898663](https://doi.org/10.1109/TIT.2019.2898663).
- Neykov, M., Lu, J., & Liu, H. (2019). Combinatorial inference for graphical models. *The Annals of Statistics*, 47(2), 795–827. [doi:10.1214/17-AOS1650](https://doi.org/10.1214/17-AOS1650).
- Battey, H., Fan, J., Liu, H., Lu, J., & Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics*, 46(3), 1352–1382. [doi:10.1214/17-AOS1587](https://doi.org/10.1214/17-AOS1587).
- Lu, H., Cao, Y., Yang, Z., Lu, J., Liu, H., & Wang, Z. (2018). The edge density barrier: Computational-statistical tradeoffs in combinatorial inference. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (pp. 3247–3256). Proceedings of Machine Learning Research.
- Sun, W., Lu, J., & Liu, H. (2018). Sketching method for large scale combinatorial inference. *Advances in Neural Information Processing Systems* (Vol. 31).
- Sun, W., Lu, J., Liu, H., & Cheng, G. (2016). Provable sparse tensor decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3), 899–916. [doi:10.1111/rssb.12190](https://doi.org/10.1111/rssb.12190).

## Health & Medical Sciences

- Gan, Z., Zhou, D., Rush, E., Panickan, V. A., Ho, Y.-L., Ostrouchov, G., Xu, Z., Shen, S., Xiong, X., Greco, K. F., Hong, C., Bonzel, C.-L., Wen, J., Costa, L., Cai, T., Begoli, E., Xia, Z., Gaziano, J. M., Liao, K. P., Cho, K., Cai, T., & Lu, J. (2025). ARCH: Large-scale knowledge graph via aggregated narrative codified health records analysis. *Journal of Biomedical Informatics*, To appear.
- Wen, J., Xue, H., Rush, E., Panickan, V. A., Cai, T., Zhou, D., Ho, Y.-L., Costa, L., Begoli, E., Hong, C., Gaziano, J. M., Cho, K., Liao, K. P., Lu, J.\*, & Cai, T.\* (2025). DOME: Directional Medical Embedding Vectors from Electronic Health Records. *Journal of Biomedical Informatics*, To appear.
- Mathis, C.<sup>†</sup>, Wang, L.<sup>†</sup>, Hou, J., Wen, J., Bhattacharya, R., Wang, Z., Sweet, S. M., Elfiky, A., Wang, L. S., Srivastava, R., Turzhitsky, V., McKay, R. R., Jayram, G., Sundaram, M., Cai, T., Cai, T., & Lu, J. (2024). Automating radiomic feature interpretation through residual regression for language models. *Manuscript*.
- Wang, L.<sup>†</sup>, Mathis, C.<sup>†</sup>, Hou, J., Wen, J., Bhattacharya, R., Wang, Z., Sweet, S. M., Elfiky, A., Wang, L. S., Srivastava, R., Turzhitsky, V., McKay, R. R., Jayram, G., Sundaram, M., Cai, T., Cai, T., & Lu, J. (2024). Comprehensive CT imaging pipeline for automated tumor segmentation. *Manuscript*.
- Wen, J., Hou, J., Bonzel, C.-L., Zhao, Y., Castro, V. M., Gainer, V. S., Weisenfeld, D., Cai, T., Ho, Y.-L., Panickan, V. A., Costa, L., Hong, C., Gaziano, J. M., Liao, K. P., Lu, J., & Cho, K., Cai, T. (2024). LATTE: Label-efficient incident phenotyping from longitudinal electronic health records. *Patterns*, 5(1), 100906. [doi:10.1016/j.patter.2023.100906](https://doi.org/10.1016/j.patter.2023.100906).



- Hou, J., Wen, J., Bhattacharya, R., Wang, Z., Sweet, S. M., Elfiky, A., Wang, L. S., Srivastava, R., Lu, J., Turzhitsky, V., McKay, R. R., Jayram, G., Sundaram, M., Cai, T., & Cai, T. (2024). Identify cancer recurrence in renal cell carcinoma (RCC) in electronic health records – a deep learning approach. *European Society of Medical Oncology (ESMO) Congress*.
- Lu, K.<sup>†</sup>, Potash, P., Lin, X., Sun, Y., Qian, Z., Yuan, Z., Naumann, T., Cai, T., & Lu, J. (2023). Prompt discriminative language models for domain adaptation. In T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, & A. Rumshisky (Eds.), *Proceedings of the 5th Clinical Natural Language Processing Workshop* (pp. 247-258). Toronto, Canada: Association for Computational Linguistics. doi:10.18653/v1/2023.clinicalnlp-1.30.
- Wen, J., Zhang, X., Rush, E., Panickan, V. A., Li, X., Cai, T., Zhou, D., Ho, Y. L., Costa, L., Begoli, E., Hong, C., Gaziano, J. M., Cho, K., Lu, J., Liao, K. P., & Zitnik, M. (2023). Multimodal representation learning for predicting molecule–disease relations. *Bioinformatics*, 39(2), btad085. doi:10.1093/bioinformatics/btad085.
- Nishino, M., Lu, J., Hino, T., Vokes, N. I., Jänne, P. A., Hatabu, H., & Johnson, B. E. (2023). Prediction model for tumor volume nadir in EGFR-mutant NSCLC patients treated with EGFR tyrosine kinase inhibitors. *Journal of Thoracic Imaging*, 38(2), 82–87. doi:10.1097/RTI.0000000000000615.
- Efsthadiadis, G.<sup>†</sup>, Emedom-Nnamdi, P., Kolbeinsson, A., Onnela, J.-P., & Lu, J. (2023). STASIS: Reinforcement learning simulators for human-centric real-world environments. In H. Chen & L. Luo (Eds.), *Trustworthy Machine Learning for Healthcare, Lecture Notes in Computer Science*, 13932, 85-92. Springer. doi:10.1007/978-3-031-39539-0\_8.
- Zhou, D., Gan, Z., Shi, X., Patwari, A., Rush, E., Bonzel, C.-L., Panickan, V. A., Hong, C., Ho, Y.-L., Cai, T., Costa, L., Li, X., Castro, V. M., Murphy, S. N., Brat, G., Weber, G., Avillach, P., Gaziano, J. M., Cho, K., Liao, K. P., Lu, J.\*, & Cai, T.\* (2022). Multiview incomplete knowledge graph integration with application to cross-institutional EHR data harmonization. *Journal of Biomedical Informatics*, 133, 104147. doi:10.1016/j.jbi.2022.104147.
- Hong, C., Rush, E., Liu, M., Zhou, D., Sun, J., Sonabend, A., Castro, V. M., Schubert, P., Panickan, V. A., Cai, T., Costa, L., He, Z., Link, N., Hauser, R., Gaziano, J. M., Murphy, S. N., Ostrouchov, G., Ho, Y.-L., Begoli, E., Lu, J., Cho, K., Liao, K. P., & Cai, T. (2021). Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *NPJ Digital Medicine*, 4(1), 151.
- Nishino, M., Lu, J., Hino, T., Vokes, N. I., Jänne, P. A., Hatabu, H., & Johnson, B. E. (2021). Tumor growth rate after nadir is associated with survival in patients with EGFR-mutant non–small-cell lung cancer treated with epidermal growth factor receptor tyrosine kinase inhibitor. *JCO Precision Oncology*, 5, 1603-1610. doi:10.1200/PO.21.00172.
- Hino, T., Hida, T., Nishino, M., Lu, J., Putman, R. K., Gudmundsson, E. F., Hata, A., Araki, T., Valtchinov, V. I., Honda, O., Yanagawa, M., Yamada, Y., Kamitani, T., Jinzaki, M., Tomiyama, N., Ishigami, K., Honda, H., Estepar, R. S. J., Washko, G. R., Johkoh, T., Christiani, D. C., Lynch, D. A., Gudnason, V., Gudmundsson, G., Hunninghake, G. M., & Hatabu, H. (2021). Progression of traction bronchiectasis/bronchiolectasis in interstitial lung abnormalities is associated with increased all-cause mortality: Age Gene/Environment Susceptibility-Reykjavik Study. *European Journal of Radiology*, 8, 100334. doi:10.1016/j.ejro.2021.100334.
- Hida, T., Hata, A., Lu, J., Valtchinov, V. I., Hino, T., Nishino, M., Honda, H., Tomiyama, N., Christiani, D. C., & Hatabu, H. (2021). Interstitial lung abnormalities in patients with stage I non-small cell lung cancer are associated with shorter overall survival: The Boston lung cancer study. *Cancer Imaging*, 21, 1–7.
- Hida, T., Nishino, M., Hino, T., Lu, J., Putman, R. K., Gudmundsson, E. F., Araki, T., Valtchinov, V. I., Honda, O., Yanagawa, M., Yamada, Y., Hata, A., Jinzaki, M., Tomiyama, N., Honda, H., Estepar, R. S. J., Washko, G. R., Johkoh, T., Christiani, D. C., Lynch, D. A., Gudnason, V., Gudmundsson, G., Hunninghake, G. M., & Hatabu, H. (2020). Traction bronchiectasis/bronchiolectasis is associated with interstitial lung abnormality mortality. *European Journal of Radiology*, 129, 109073. doi:10.1016/j.ejrad.2020.109073.