BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Kim, Yejin

eRA COMMONS USER NAME (credential, e.g., agency login): YEJINKIM

POSITION TITLE: Assistant Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Pohang University of Science and Technology, South Korea	B.S.	02/2012	Industrial Engineering
University of California San Diego, CA	Visiting scholar	08/2016	Biomedical informatics
Pohang University of Science and Technology	Ph.D.	08/2017	Computer Science (machine learning)
Pohang University of Science and Technology	Postdoctoral	12/2017	Computer Science (machine learning)

A. Personal Statement

Innovating biomedical research via machine learning requires blending biological/clinical insight into algorithmic thinking. My background in machine learning and biomedical informatics gives me a unique perspective for discovering, defining, and solving biomedical challenges. Currently, I am an Assistant Professor in health informatics since 2018 and am a founding member of the Center for Secure Artificial Intelligence for Healthcare in the School of Biomedical Informatics (SBMI) at the University of Texas Health Science Center at Houston (UTHealth). My research areas are disease progression modeling for neurodegenerative diseases and counterfactual outcome modeling for treatment development. My innovative counterfactual federated model to harmonize fragmented clinical trial data is funded by NIA R01 as PI (R01AG082721). My integrative models of causal inference in real-world data and biological knowledge-base was under consideration for the R56 award as PI. My pioneering computational phenotyping model was funded by the Robert Woods Johnson foundation as PI. My tensor factorization models for treatment effect significantly contributed to my team obtaining NIA R01 (Co-I) and CPRIT Rising Star awards (Co-I). I led my international team to rank top 2 (out of 54) in the DREAM Challenge competition for drug repurposing. I authored 50+ articles with 600+ citations as a primary author. I have strong publication records from prestigious data science conferences (such as KDD, CIKM, and IJCAI for novel methodology development) and high-impact translational journals (such as Bioinformatics, JBI, JAMIA, PLoS Digital Health); my PLoS one publication was selected as the top 10% most cited article in 2016. My publication covers diverse scientific inquiries for neurodegenerative disease therapy (e.g., drug candidate ranking, emulating clinical trial), data modality (e.g., clinical trial data, claim data), and methodology (e.g., treatment effect estimation, federated learning) to address critical biomedical challenges. I am mentoring 8 pre- and postdoctoral trainees from various backgrounds such as mathematics, computer science, biostatistics, pharmaceutics, and pharmacology. For professional service, I am the organizer of the SBMI Machine Learning Datathon series at Texas Medical Center since 2019. I have served on a program committee of prestigious data science and/or medical informatics conferences such as AAAI, IJCAI, KDD Data Science workshop, WWW, AMIA, IEEE-BCB, and IEEE-BIBM.

Ongoing Research Support

R01AG082721 (MPIs **Kim**, Jiang) Role: contact PI 07/01/23 – 06/30/28
Harmonizing multiple clinical trials for Alzheimer's disease to investigate differential responses to treatment via federated counterfactual learning
The project aims at developing novel machine learning solutions to harmonize multiple clinical trial data for aggregated analysis on individual treatment effect

R01AG066749-03S1 (MPIs Jiang, **Kim**) Role: MPI 09/01/22 – 08/31/23
Ethically optimize machine learning models with real-world data to improve algorithmic fairness
The project aims at developing novel machine learning solutions to mitigate algorithmic unfairness by addressing data biases

U24NS107322-S (PI Savitz) Role: Co-Investigator 08/01/21-07/31/23
Stroke impact on progression of Alzheimer's (SIPA)
The goal of this study is to determine the impact of acute ischemic stroke on progression of Alzheimer's Disease. These studies will use hyperacute MRI to bring new insights that will broaden our understanding of how cerebrovascular disease contributes to neurodegeneration in patients with Alzheimer's Disease.

Bosarge Family Foundation (PI Savitz) Role: Co-Investigator 12/01/21-05/31/23
Stem Cell Post-COVID19 Trial
The goal is to determine treatment effect of repurposable drugs on post COVID19 sequelae. Computational drug repurposing model will guide the prioritization of candidates.

RR180012 (Jiang) Role: Co-Investigator 05/01/18-04/30/23
CPRIT
CPRIT Rising Stars Award
The goal is to develop novel computational phenotyping methods for cancer patients to study how they respond to different drug combination therapy.

Pending

(**Kim**) Role: PI 09/01/23-08/31/28
NIH/NIA R01
Network and Population Based Combinatorial Drug Repurposing for Alzheimer's Disease and Related Dementia
Project Goal: To identify repurposable drug combinations by developing novel informatics models to harmonize heterogeneous and complementary data/knowledge, including human interactome networks, transcriptomes, clinical observation, and trial data into a coherent machine learning framework.

(Kim, Jiang) Role: PI 07/01/23-06/30/28
NIH/NIA R01
Subpopulations with different Alzheimer's disease risk and progression for targeted treatment
Project Goal: To identify subpopulations/subtypes of AD/ADRD in real world and clinical trials retrospectively.

Selected Completed Research Support

Robert Woods Johnson Foundation (MPIs, **Kim**, Jiang) Role: PI 09/15/19 - 09/14/21
Computational Phenotyping to Better Understand Obesity
This project will develop a dynamic pattern mining model to identify trends/patterns of weight gain (temporal phenotyping) that is most likely to be diabetic.

NIGMS R01GM124111 (PI Qi) Role: Co-Investigator 07/01/17 - 06/30/21
Privacy-preserving methods and tools for handling missing data in distributed health data networks
The goal of this study is to develop privacy-preserving distributed strategies and tools for handling missing data in distributed health data networks.

B. Positions and Honors

Positions and Employment

2020- Member, Institute for Stroke and Cerebrovascular Diseases, UTHealth, Houston, TX
2018- Assistant Professor, School of Biomedical Informatics, UTHealth, Houston, TX
2018 Machine Learning Scientist, Advanced Recommendation Team, Kakao Corp, Seoul, Korea
2017 Postdoc Researcher, Pohang University of Science and Technology, Pohang, Korea
2015-16 Visiting Research Scholar, Division of Biomedical Informatics, San Diego, CA
2013-16 Research Assistant, Department of Medical Informatics, Catholic University, Seoul, Korea
2012-17 Graduate Research Assistant, Pohang University of Science and Technology, Pohang, Korea

Other Experience and Professional Memberships

2020 Grant Review Panel, Luxembourg National Research Fund Industrial Fellowships Program
2020- Program Committee, International Joint Conference on Artificial Intelligence (IJCAI)
2020- Program Committee, AMIA Informatics Summit
2020- Program Committee, IEEE International Conference on Healthcare Informatics (ICHI)
2019- Program Committee, Association for the Advancement of Artificial Intelligence Conference (AAAI)
2019- Program Committee, IEEE International Conference on Bioinformatics and Biomedicine (BIBM)
2019- Program Committee, ACM SIGKDD Workshop on Applied Data Science for Healthcare (KDD)
2019- Member, American Medical Informatics Association (AMIA)
2019- Co-organizer, UTHealth SBMI Machine Learning Datathon Series
2019 Program Committee, Cerner/SBMI National Data Science Challenge
2018 Program Committee, World Wide Web Conference (WWW)
2017- Program Committee Member, ACM-BCB 2017 Conference on Bioinformatics, Computational Biology, and Health Informatics
2017- Program Committee, Annual Symposium of American Medical Informatics Association (AMIA)

2017- Reviewer, Cell Reports, Briefing in Bioinformatics, JAMA Network Open, J. of Biomedical Informatics, J. Am. Medical Informatics, npj Scientific Reports, Cerebral Cortex

Honors

2020 Ranked 2nd place out of 54 international teams in DREAM challenge for Drug Repurposing
2020 Selected as top 10% most cited PLoS One article in 2016
2020 Selected as Special Lightning Round Session Poster at Alzheimer's Association International Conference 2020 Technology and Dementia
2008-12 University full scholarship, Pohang University of Science and Technology
2010-17 Member, Young Engineer Honor Society of Korea (National Academy of Engineering of Korea)

C. Contributions to Science

1. **Multimodal progression phenotyping for neurological disorders** is an active area of my research. Many neurologic disorders, such as Alzheimer's, Parkinson's disease, epilepsy, stroke, are clinically very heterogeneous, varying between patients in terms of cognitive symptoms, test findings, and rates of progression. Several recent treatment trials have shown efficacy in a subset of patients, but not all patients, which implies that there are subsets of patients who respond differently to treatments. This motivated us to develop data-driven phenotyping models using multimodal source of data (e.g., clinical presentation, neuroimaging). I have derived computational phenotypes based on multimodal matrix factorization for longitudinal progression of cognitive impairments in Alzheimer's disease (Kim et al Sci. Rep. 2020), causal inference for clinical pathway from transient ischemia attack to dementia (Kim et al., PLoS Digital Health 2022), deep learning model for EEG signals after epileptic seizure (organizer of the Texas Medical Center 2019 Datathon), Markov decision process modeling for disease state changes in Parkinson's disease (Kim et al., Sci. Rep. 2021), and customized tensor factorization for transition patterns from epilepsy to Alzheimer's (Kim et al. JBI 2020) and discriminative sleep patterns of Alzheimer's patients (Kim et al. AMIA 2020). My prior studies and track records on disease phenotyping positioned me well for proposed project to quantify heterogeneous treatment effect in this neurological disorder.

a. **Kim, Y**., Suescun, J., Schiess, M. C., & Jiang, X. (2021). Computational medication regimen for Parkinson's disease using reinforcement learning. *Scientific reports*, *11*(1), 1-9.
b. **Kim, Y.**, Jiang, X., Giancardo, L., Pena, D., Bukhbinder, A., Amran, A., & Schulz, P.E. (2020). Multimodal Phenotyping of Alzheimer's Disease with Longitudinal Magnetic Resonance Imaging and Cognitive Function Data, Scientific Reports, doi.org/10.1038/s41598-020-62263-w
c. **Kim Y**, Lhatoo S, Zhang GQ, Chen L, Jiang X. Temporal phenotyping for transitional disease progress: An application to epilepsy and Alzheimer's disease. J Biomed Inform. 2020 Jul;107:103462. doi: 10.1016/j.jbi.2020.103462. Epub 2020 Jun 18. PMID: 32562896; PMCID: PMC7374015.

2. **Causal inference** to understand disease progression is another research area. Many critical diseases are multifactorial diseases that combinations of factors contribute to confounding effects. Causal inference is a principled approach to taking the multifactorial effect into consideration. It takes other confounders into account to determine causal effects solely from one risk factor of interest. From an algorithm development perspective, my team is actively working on developing a scalable constraint-based causal inference model based on the PC model. We flipped the conditional inference test's order in reverse to increase the efficiency in parallel computing. From an application perspective, I worked with vascular disease experts to identify race-specific transition patterns from vascular risk factors to AD and related dementia; derive a causal relationship between influenza vaccination and AD risk using nationwide healthcare claim data. I collaborated with psychiatrists to derive a causal structure to investigate the multifactorial effects of personality traits on behavioral addiction; this study was selected as the top 10% most cited paper in PLoS One.
    a. **Kim, Y**., Zhang, K., Savitz, S., Chen, L., Schulz, P., Jiang, X. (2022). Counterfactual Analysis of Differential Comorbidity Risk Factors in Alzheimer's Disease and Related Dementias. PLoS Digital Health 1(3), e0000018.
    b. Ling, Y., Upadhyaya, P., Chen, L., Jiang, X., & **Kim, Y**. (2022). Emulate randomized clinical trials using heterogeneous treatment effect estimation for personalized treatments: Methodology review and benchmark. Journal of biomedical informatics, 137, 104256. PMID: 36455806.
    c. Upadhyaya, P., Zhang, K., Li, C., Jiang, X., & **Kim, Y**. (2023) Scalable Causal Structure Learning: Traditional and Deep Learning Algorithms and New Opportunities in Biomedicine. Journal of Medical Internet Research Medical Informatics.

3. **Novel computational phenotyping** is the main area of my research. A phenotype allows data to speak for themselves instead of a historically and clinically driven description of each disease. Electronic health records (EHRs) are increasingly used as a source of the data-driven phenotype. Current approaches for translating EHRs into useful phenotypes are typically slow, manually intensive, limited in scope, and require domain expert knowledge. Computational phenotyping based on machine learning has been proposed to facilitate the extraction of meaningful phenotypes automatically from EHRs without human supervision. My research is based on tensor factorization, which can generate interpretable latent medical concepts using the interaction between components from multiple information sources. Based on the characteristics of biomedical data, I developed novel methods to derive phenotypes that stratify patients' health risks and reflect ICD9 ontology structures or patient demographic differences. I also developed federated tensor factorization to derive phenotypes without sharing patients' sensitive data. A scientific impact that this phenotyping model has is that it translates millions of health records into a handful of phenotypes so that clinicians can understand the massive data at a glance.
    a. **Kim, Y**., Sun, J., Yu, H., & Jiang, X. (2017, August). Federated tensor factorization for computational phenotyping. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 887-895). ACM.
    b. **Kim, Y**., El-Kareh, R., Sun, J., Yu, H., & Jiang, X. (2017). Discriminative and distinct phenotyping by constrained tensor factorization. *Scientific reports*, *7*(1), 1114.
    c. Choi, J., **Kim, Y**., Kim, H. S., Choi, I. Y., & Yu, H. (2018). Phenotyping of Korean patients with better-than-expected efficacy of moderate-intensity statins using tensor factorization. *Plos one*, *13*(6), e0197518.

4. **Sequential decision making for biomedical research** to derive optimal strategy is an active and ongoing area of my research. The sequential decision making in biomedical research includes finding best dynamic strategy in immunohistochemistry profiles test (e.g., which tests should be done in which order for differential diagnosis), batch-wise adaptive cell line experiments (e.g., update drugs to screen based on previous batch of drug efficacy), and medication regimen (e.g., personalized timing and dosage). Reinforcement learning (RL) is an area of machine learning for sequential decision-making problems, but the pure RL methods are sometimes not directly applicable to biomedical problems because it does not consider contextual challenges such as multi-tasking of tests or insurance coverage policy. These real-world challenges motivated the development of customized sequential decision-making framework to facilitate the biomedical research. For example, I have developed a computational medication regimen for Parkinson's disease. I also developed computationally optimal differential diagnosis policy to detect lymphoid neoplasm using immunohistochemistry staining and published an online calculator.
    a. **Kim, Y**., Suescun, J., Schiess, M. C., & Jiang, X. (2021). Computational medication regimen for Parkinson's disease using reinforcement learning. *Scientific reports*, *11*(1), 1-9.
    b. **Kim, Y**., Choi, J., Chong, Y., Jiang, X., & Yu, H. (2017, November). DiagTree: Diagnostic Tree for Differential Diagnosis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)* (pp. 1179-1188). ACM.
    c. Chong, Y., Thakur, N., Lee, J. Y., Hwang, G., Choi, M., **Kim, Y**., ... & Cho, M. Y. (2021). Diagnosis prediction of tumours of unknown origin using ImmunoGenius, a machine learning-based expert system for immunohistochemistry profile interpretation. Diagnostic pathology, 16(1), 1-9.

5. **Computational drug development** is my latest research area. Drug development is an expensive and time-consuming process. It takes many years and costs billions of dollars for a drug to be approved. Systematic integration of previous results and knowledge might change the game by identifying highly promising drugs and their combinations to save cost and speedup discovery. I have developed drug repurposing pipeline using heterogeneous network and counterfactual inference in real-world patient data, drug repurposing based on large-scale electronic health records of AD patients with causal analysis and survival analysis, anti-cancer drug's drug sensitivity data with deep collaborative filtering. I have co-organized a machine learning Datathon for drug repurposing in Texas Medical Center in 2019. My team and I are ranked top 2 (out of 54) in CTD2 BeatAML DREAM Challenge.
    a. Amran, A., Lin, Y., **Kim, Y**., Bernstam, E., Jiang, X., & Schulz, P. E. (2020). Influenza vaccination is associated with a reduced incidence of Alzheimer's disease: Epidemiology/Risk and protective factors in MCI and dementia. *Alzheimer's & Dementia*, *16*, e041693.
    b. **Kim, Y**., Zheng, S., Tang, J., Jim Zheng, W., Li, Z., & Jiang, X. (2021). Anticancer drug synergy prediction in understudied tissues using transfer learning. *Journal of the American Medical Informatics Association*, *28*(1), 42-51.
    c. Hsieh, K., Wang, Y., Chen, L., Zhao, Z., Savitz, S., Jiang, X., Tang, J., & **Kim, Y.** (2021). Drug repurposing for COVID-19 using Graph Neural Network with Multiple Evidence. Scientific Reports 11:23179