Yu Deng

CONTACT INFORMATION — Feinberg School of Medicine, 303 E Superior street, Chicago, IL — Email:yudeng2015@u.northwestern.edu

RESEARCH INTEREST

I am a PhD student majoring in biomedical informatics. My focus is on applying supervised and unsupervised learning methods for clinical research. I have applied various unsupervised learning techniques to find disease subtypes and novel risk factors. I have used joint modeling techniques to improve prediction performance using longitudinal data as well as using Natural Language Processing (NLP) for computational phenotyping.

EDUCATION

Northwestern University, Chicago, IL, USA
- PhD Candidate, biomedical informatics, Feinberg School of Medicine, expected graduation June 2020
- Major GPA: 3.79/4.00
- Coursework: Programming for Big Data, Advanced Biostatistics, Deep Learning From Scratch

SKILLS

Programming Skills
- Python (sklearn, numpy, pandas), R, SQL

Machine Learning Algorithms
- Classical & Penalized Regression Methods (LASSO, Ridge), SVM, Random Forest, K-nearest Neighbors, Adaboosting, Deep Learning (MLP, RNN), Cox regression
- K-means, hierarchical clustering, Partition Around Medoid (PAM), Latent Class Analysis (LCA), non-Negative Tensor Factorization (NTF), non-negative Matrix Factorization (NMF)
- Feature Engineering (e.g. PCA, tensor factorization, forward/backward selection), Regularization
- Experimental Design, Hypothesis Testing, A/B Testing

Tools
- SQLServer, Spark, Tensorflow

WORK EXPERIENCE

Center for Health Information Partnerships, Northwestern University, Chicago
- Graduate Research Assistant, July 2016-Present
- Used SQL to extract clinical data. Developed various algorithms to model disease progression and subtypes

Bioinformatics Laboratory, Tsinghua University, Beijing, China
- Research Assistant, Dec 2014-August 2015
- Developed pipeline for genetic data analysis (i.e.RNA-seq, DNA-seq) on computing cluster

HONOURS AND DISTINCTIONS
- **Honorable Mention in Student Poster Competition,** International Chinese Statistical Association (ICSA), 2018
- **First Prize in Student Poster Competition,** Northwestern Biomedical Informatics Day, 2017
- Driskill Graduate Program Full Scholarship, Northwestern University 2015 – 2017
- Outstanding Study Abroad Undergraduate Scholarship, Chinese Scholarship Council, 2013

SELECTED PUBLICATIONS/CONFERENCES

Use of Clinical Phenotypes and Non-negative Tensor Factorization for Heart Failure Prediction
Deng Y., Furmanchuk A., Chen R., Ahmad F., Sun J., Kho A.(2017). *AMIA Annual Symposium. Podium Presentation*

Use of Repeated Measurements for Cardiovascular Disease Prediction: the Application of Joint Model
Deng Y., Zhong Y., Kho A., Zhao L.(2019). ENAR 2019. Poster presentation

Identification of Systemic Lupus Erythematosus Subtypes using Latent Class Analysis
Deng Y., Ghosh A., Luo Y., Kho A., Goldman R., Walunas T.(2019). Medinfo. Podium Presentation (submitted)

Natural Language Processing for EHR-Based Computational Phenotyping
Zeng, Z., **Deng, Y**., Li, X., Naumann, T., & Luo, Y. (2018). Natural Language Processing for EHR-Based Computational Phenotyping. arXiv preprint arXiv:1806.04820.

Characterizing Design Patterns of EHR-Driven Phenotype Extraction Algorithms
Zhong, Y., Rasmussen L., **Deng Y**., Pacheco J., Smith M., Starren J., Wei W., Walton N., Hripcsak G., Chute C., Luo Y. Characterizing Design Patterns of EHR-Driven Phenotype Extraction Algorithms. IEEE on Bioinformatics and Biomedicine (paper accepted)

Measuring Nascent Transcripts by Nascent-seq. Methods in Molecular Biology
Chen, F. X., Marshall, S. A., **Deng, Y**., & Tianjiao, S. (2018). Measuring Nascent Transcripts by Nascent-seq. In Next Generation Sequencing (pp. 19-26). Humana Press, New York, NY.

SELECTED RESEARCH PROJECTS

Use of Clinical Phenotypes and Non-negative Tensor Factorization for Heart Failure (HF) Prediction, 2017
- Performed NTF on large scale, sparse medical record data; Generated latent clusters
- Performed dimension reduction including NTF, MTF, PCA on medical record data to get important features
- Compared model performance between dimension reduction techniques and random forest using the top features

Developed/customized joint multivariate Bayesian Model for Cardiovascular Disease (CVD) Prediction, 2017
- Built joint multivariate Bayesian model using longitudinal data to predict CVD time-to-event
- Imputed missing data using multivariate imputation by chained equation
- Evaluated model performance using AUC and NRI. Improved AUC from 0.85 to 0.87 with statistical significance.

Using Deep Learning for Post-Operative Atrial Fibrillation (AF) prediction, 2018 -
- Extracted data from Northwestern Memorial Hospital data warehouse using SQL
- Trained supervised learning models including LASSO, Random Forest, deep learning (on tensorflow)
- Trained and fine-tuned deep learning model by changing number of epochs, number of hidden layers, learning rate