J. Jojo Cheng

jay.jojo.cheng@gmail.com | jay-jojo-cheng.github.io | github.com/jay-jojo-cheng

Education

University of Wisconsin - Madison Sep 2018 - Jun 2023 (expected)
PhD candidate, Biostatistics and Medical Informatics — Advisor: Guanhua Chen *Madison, WI*

Princeton University Sep 2011 - Jun 2015
AB, Mathematics — Senior Thesis Advisor: Charles Fefferman *Princeton, NJ*

Projects

Multiarm: optimal individualized multi-treatment assignment (ongoing) Dec 2021 – Present
Learning nonlinear multi-category treatment rules that balance flexibility and interpretability. *R*

- Developed statistical methodology and implementation.
  - Statistical features: reluctant additive modeling, angle-based learning, interpolator theory for individualized treatment rule estimation.

Metacoop: distributed learning for individualized treatment rules Jun 2020 – Jan 2022
Causal inference and optimal treatment allocation on distributed, heterogenous data. *R, C++, HTCondor*

- Developed statistical methodology, optimization, and computational implementation for distributed learning when underlying populations are different (e.g. metaanalysis, federated learning, multi-center trial, etc.).
  - Computational features: block coordinate descent; strong rule variable screening; hyperparameter tuning with adaptive information criteria (CIC/VIC); data-private optimization.
  - Statistical features: energy weight covariate balancing; crossfitted efficiency augmentation; nonparametric propensity score; double-robustness; adversarial robustness; graphical utilities for model checking.
- Publications: [Pr1], Software: `metacoop` R library

Research Breadth Rotations (Clinical Trials, Deep Learning) Jan 2019 – Dec 2019
Semester-long applied research rotations advised by Rick Chappell and Vikas Singh. *quantreg, TensorFlow 1*

- Modeled Alzheimers disease progression with quantile regression. (Quantile Regression, Time Series)
- Implemented and trained a modified BERT model for American Family Insurance data. (Transformers, NLP)

Random Matrix Theory Image Denoiser Jul 2019
Image denoising via Marchenko-Pastur Law spectral truncation. [repo] *MATLAB*

- Implemented algorithm for estimating parameters of the Marchenko-Pastur Law and denoising images.
- Developed tutorial and exercises for teaching the theory and methodology of the approach.

Google Cloud & NCAA Machine Learning Kaggle Competition Mar 2019
3rd place (of 866 participants) prediction of tournament results. [results] *PyStan, scikit-learn*

- Develop Bayesian Bradley-Terry model and ensembles of SVM, KNN, and Poisson-Binomial GLM models.
- Feature engineering of advanced box score statistics, differential box scores, and 'crunch time' box score statistics.

Polycystic Ovary Morphology Classifier Jan 2017 - Dec 2017
Tool for multiclass disease classification on 39,000 radiology reports. *NLP, Feature Engineering*

- Developed classifier achieving sensitivities > 96% and specificities > 98% using regular expressions and XGBoost.
- Publications: [4, 7]

Work Experience

Statistical Consultant Jan 2017 – Present
Long-term collaboration with Harvard and Boston University epidemiologists. *Survival Analysis, Survey Design*

- Prepared statistical design and analysis for scientific studies on women's health and air pollution data.
- Publications: [1, 2, 3, 5, 6]

Clinical Research Grants Manager, Boston University Medical Center May 2016 – May 2018
Department of Obstetrics & Gynecology *Project Management, Technical Writing*

- Drafted and revised R level NIH grants with principal investigators.

- Interviewed, hired, and trained 5 new staff; developed onboarding materials from scratch.
- Accomplished planning, compliance, and reconciliation for research and grant funding of over $200,000.
- Authored research documentation (IRB applications, patient recruitment, consent forms, study protocols).
- Initiated new scientific collaborations with the Division of Intramural Population Health Research at the NIH-NICHD and other departments (Radiology, Systems Engineering, Epidemiology).

Curated Research Datasets

ECHO study replication dataset Aug 2021

Dataset for studying the effect of transthoracic echocardiography on ICU patient recovery. [repo] *PostgreSQL*

- Reproduced the ECHO cohort dataset/analysis in the eICU database by writing new SQL queries from scratch and deriving new variables from unstructured data.
- Ideal use cases: metaanalysis, federated learning, censored outcomes, personalized medicine, causal inference.
- Credentialed access via Physionet. [link]

ALINE study replication dataset Mar 2021

Dataset for studying the use of indwelling arterial catheters in hemodynamically stable patients. [repo] *PostgreSQL*

- Reproduced the ALINE cohort dataset/analysis in the eICU database by writing new SQL queries from scratch and deriving new variables from unstructured data.
- Ideal use cases: metaanalysis, federated learning, censored outcomes, personalized medicine, causal inference.
- Credentialed access via Physionet. [link]

PCOM ultrasound dataset Apr 2017

Dataset for detecting polycystic ovary morphology within the Boston University Medical Center EHR. *R*

- Wrangled raw unstructured data into pivot tables, fixed systematic corruptions in anthropometrics, merged overlapping concepts.
- Used externally by studies at Harvard University and the University of Chicago.
- Ideal use cases: NLP, longitudinal analysis/time series, air pollution studies, ethnic disparities in diagnosis.
- Access via request to Boston University IRB.

Preprints/Publications

Pr1. **Cheng JJ**, Huling JD, Chen G. Distributed Learning of Individualized Treatment Rules via Sign-Coherency. 2021+. Submitted.

7. Fruh V, Mahalingaiah S, **Cheng JJ**, Aschengrau A, Lane KJ. Fine Particulate Matter and Polycystic Ovarian Morphology. *Environmental Health*. 2022 January 21. Accepted.
6. Mahalingaiah S, **Cheng JJ**, Winter M, Rodriguez E, Fruh V, et al. Multimodal Recruitment for an Internet-Based Pilot Study of Ovulation and Menstruation (OM) Health. *Journal of Medical Internet Research*. 2021 April 16. doi: 10.2196/24716
5. Mahalingaiah, S, Cosenza, C, **Cheng, JJ**, Rodriguez, E, and Aschengrau, A (2020). Cognitive testing of a survey instrument for self-assessed menstrual cycle characteristics and androgen excess. *Fertility Research and Practice*. 2020 December; 6(1). 10.1186/s40738-020-00088-x
4. **Cheng JJ**, Mahalingaiah S. Data mining polycystic ovary morphology in electronic medical record ultrasound reports. *Fertility Research and Practice*. 2019 December 1; 5(13). doi: 10.1186/s40738-019-0067-7
3. Mahalingaiah S, Lane KJ, Kim C, **Cheng JJ**, Hart JE. Impacts of Air Pollution on Gynecologic Disease: Infertility, Menstrual Irregularity, Uterine Fibroids, and Endometriosis: a Systematic Review and Commentary. *Current Epidemiology Reports*. 2018 September; 5. doi: 10.1007/s40471-018-0157-9
2. Mahalingaiah S, Missmer SE, **Cheng JJ**, Chavarro J, Laden F, et al. Perimenarchal Air Pollution Exposure and Menstrual Disorder. *Human Reproduction*. 2018 March 1; 33(3). doi: 10.1093/humrep/dey005
1. Mahalingaiah S, Sun F, **Cheng JJ**, Chow ET, Lunetta KL, et al. Cardiovascular risk factors among women with self-reported infertility. *Fertility Research and Practice*. 2017 April 11; 3(7). doi: 10.1186/s40738-017-0034-0