Jay Jojo Cheng

+1-425-230-2242 | jay.jojo.cheng@gmail.com
linkedin.com/in/jay-jojo-cheng | jay-jojo-cheng.github.io | github.com/jay-jojo-cheng

Education

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

- GPA: 3.92/4.00; Special recognition for Mathematical Statistics Core (achieved highest historical score)

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
Created causal machine learning method balancing flexibility and interpretability. *R*

- Developed first nonlinear causal attribution methodology for high-dimensional, multi-treatment setting by implementing reluctant additive modeling, angle-based learning, and feature pruning.
- Quantified uncertainty of prediction performance by deriving novel interpolator theory results for individualized treatment rule estimation.

Metacoop: distributed learning for optimal treatment allocation Jun 2020 – Jan 2022
Created privacy-preserving causal machine learning method for distributed, heterogenous data. *R, C++, HTCondor*

- Enabled the use of the best theoretical causal-effect attribution methods in real practice settings (e.g. user devices, metaanalysis, federated learning, multi-center trial, etc.) by developing a data-private estimation scheme.
- Created new metric, adaptive information criteria (CIC/VIC), to measure model fit and select hyperparameters.
- Computational features (C++): block coordinate descent; strong rule variable screening; hyperparameter tuning; adversarial robustness; data privacy.
- Statistical features (R): energy weight covariate balancing; crossfitted efficiency augmentation; nonparametric propensity score; double-robustness; graphical 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. (Clinical Trials, 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
Won 3rd place (of 866 participants) prediction of tournament results. [results] *PyStan, scikit-learn*

- Developed Bayesian Bradley-Terry classifier and ensembles of SVM, KNN, and Poisson-Binomial GLM models.
- Created new metrics as inputs to machine learning methods by feature engineering 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*

- Identified 1000+ underdiagnosed patients by creating the first EHR-based classifier for the condition.
- Developed phenotyping algorithm with sensitivities > 96% and specificities > 98% using regular expressions, domain logic, 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*

- Published in top journals in the field by designing data collection and statistical analysis for scientific studies on women's health and air pollution.
- 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*

- Initiated new scientific collaborations with the Division of Intramural Population Health Research at the NIH-NICHD and other departments (Radiology, Systems Engineering, Epidemiology) resulting in two new grants totaling $1.2M by developing, drafting, and revising R-level NIH grants.
- Doubled team size by recruiting, interviewing, hiring, and training 5 new staff.
- Reduced workflow friction by authoring research documentation (IRB applications, patient recruitment, consent forms, study protocols) and onboarding materials from scratch.

Curated Research Datasets

ECHO study dataset replication in eICU Aug 2021

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

- Reproduced the ECHO cohort by writing SQL queries, deriving variables from unstructured data.
- Ideal use cases: metaanalysis, federated learning, censored outcomes, personalized medicine, causal inference.
- Public (credentialed) access via Physionet. [link]

ALINE study dataset replication in eICU Mar 2021

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

- Reproduced the ALINE cohort by writing SQL queries, deriving variables from unstructured data.
- Ideal use cases: metaanalysis, federated learning, censored outcomes, personalized medicine, causal inference.
- Public (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 unstructured data into pivot tables, fixed systematic corruptions in anthropometrics, merged overlapping concepts, expert-labeled 2000 observations.
- 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.
- Authorized 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