Yimeng Shang

(+1) 646-704-5390 | yqs5519@psu.edu | ys3298.github.io

Innovative and analytical thinker dedicated to public health, with extensive research skills and problem-solving experience.

Education

Pennsylvania State University *Hershey, PA*
Ph.D. in Biostatistics (GPA: 4.0/4.0) *2021.08 - 2025.04 (Expected)*
Advisor: Dr. Lan Kong

Columbia University *New York, NY*
M.S. in Biostatistics (GPA:4.0/4.0) *2019.08 - 2021.06*

East China Normal University *Shanghai, China*
B.S. in Mathematics (GPA:3.5/4.0) *2015.09 - 2019.06*

Work Experience

Pennsylvania State University, College of Medicine *Hershey, PA*
Statistical Consultant and Data Analyst *2021.09 - Present*
- Collaborated with physicians and biomedical researchers to develop statistical plans, conduct statistical analysis, and write research output.
- Utilized increasing mixed chimerism data to aid in understanding and predicting leukemia relapse survival outcome in oncology patients.
- Analyzed longitudinal data from randomized controlled trials using linear mixed-effects models to compare a physical activity treatment intervention with a delayed treatment control condition.
- Investigated the relationship between short sleep combined with sleep apnea and cardiovascular diseases using UK Biobank accelerometer data.

Merck & Co., Inc. *Upper Gwynedd, PA*
Biostatistics Intern, BARDS *2024.06 - 2024.08*
- Examined the impact of varying covariate overlap across diverse trial populations on indirect treatment comparison (ITC) methods, including the Bucher method, Simulated Treatment Comparison (STC), and Matching-Adjusted Indirect Comparison (MAIC), to analyze longitudinal outcomes using comprehensive simulation studies.
- Proposed Arm-based MAIC to preserve the balance between arms in the reweighted population, which showed more accurate and precise estimation, better-controlled Type I error, and greater statistical power, compared to conventional study-based MAIC.
- Implemented and evaluated various ITC methods to compare the effect of pneumococcal vaccines V114 and PCV20.

Cytel, Inc. *Boston, MA*
Strategic Consulting / Biostatistics Intern *2022.06 - 2022.08*
- Proposed a predictive variable/biomarker selection algorithm for subgroup identification using knockoff filters to control for multiple comparisons.
- Built an interactive Shiny app to facilitate the use of the proposed algorithm.
- Supported early-phase dose escalation and cohort expansion simulations and prepared the statistical analysis plan for FDA submission.

Eli Lilly & Co., Inc. *Shanghai, China*
Data Science & Solution Intern *2018.09 - 2019.06*
- Assisted with data management in clinical trials, including data cleaning and addressing missing data queries, under the supervision of the China DSS team.
- Conducted quantitative analysis and developed an automated Shiny app for reproducible monthly analysis to enhance efficiency.

Research Experience

Estimating Per-Protocol Effects in Randomized Controlled Trials with Survival Outcomes and Competing Events: Addressing Non-Adherence

Penn State University

Supervised by Dr. Yu-Han Chiu *2024.08 - Present*

- Conducted statistical analysis for the COSMOS trials to estimate the per-protocol and intent-to-treat effects of cocoa flavanol supplementation on preventing cardiovascular disease (CVD) events, accounting for non-CVD deaths as competing events.
- Developed and applied an inverse probability weighting (IPW) estimator to address censoring, non-adherence, and competing events when estimating per-protocol effects.
- Utilized the parametric g-formula with time-varying covariates for robust estimation in the presence of censoring, non-adherence, and competing risks.
- Evaluated different methods for handling competing events, including total effect and direct effect approaches, to ensure accurate estimation of causal effects.

A Latent Variable Approach for Causal Effect Estimation under Misclassified Treatment Assignment

Penn State University

Supervised by Dr. Lan Kong *2024.03 - 2024.12*

- Proposed a latent variable approach that treats true treatment assignment as a latent variable for causal effect estimation, accounting for potential misclassification of treatment assignments.
- Decomposed the complete likelihood function into three components: the propensity score model, measurement error model, and outcome model, with parameters estimated using the expectation–maximization (EM) algorithm.
- Incorporated validation data and machine learning approach to enhance the measurement error modeling and doubly-robust estimation for propensity score model and outcome model.
- Demonstrated the superiority of the proposed framework in reducing the bias caused by misclassification, especially when utilizing a machine learning algorithm for the measurement error model, through simulation studies.

Robust Propensity Score Estimation via Loss Function Calibration

Penn State University

Supervised by Dr. Lan Kong *2023.03 - 2024.03*

- Proposed robust propensity score (PS) estimation method under model misspecification by incorporating covariate imbalance into loss function of multiple machine learning methods
- Conducted simulation studies with various model specifications to compare causal effect estimation using different propensity score methods and causal estimators (e.g., Horvitz-Thompson(HT), Hájek, doubly robust).
- Validated the robustness of the proposed method against both correctly specified and misspecified propensity score models, demonstrating a significant reduction in bias and RMSE.

High-dimensional Propensity Score Estimation via Outcome-Assisted Variable Selection for Real World Data (RWD)

Penn State University

Supervised by Dr. Lan Kong *2023.06 - 2024.09*

- Extended the *loss function calibration* method to a high-dimensional setting by incorporating outcome-assisted variable selection for propensity score model.
- Extracted cohorts with high-dimensional baseline covariates to emulate clinical trial data using real-world data from the MarketScan Claims Database.
- Conducted plasmode simulations with the extracted real-world data to evaluate the proposed high-dimensional method.
- Demonstrated that the proposed method outperforms others (outcome adaptive LASSO, hdCBPS) in providing unbiased causal effect estimation.

Non-Parametric Analysis of Transient Data: a Pseudo-Competing Event Approach

Penn State University

Supervised by Dr. Shouhao Zhou *2022.10 - 2023.12*

- Proposed a novel non-parametric approach to enhance estimation and hypothesis testing for transient survival data by conceptualizing state transitions as pseudo-competing events and reframing the analysis as a competing events problem.
- Calibrated the cumulative incidence function by inverse probability weighting to eliminate systematic bias from the pseudo-competing transition risks.
- Demonstrated unbiased estimation with accurate type I error control and robust statistical power by simulation studies.
- Developed a Shiny app and associated software paper for its application.

Statistical Analysis of High Dimensional Metabolomics Data in Autism Spectrum Disorder (ASD) *Columbia University*

Supervised by Dr. Xiaoyu Che *2020.06-2020.10*

- Constructed logistic regression and Cox hazard model to estimate the effect size for each biomarker; adjusted for multiple comparisons using Hochberg step-up method; conducted power analysis to compare the models and did sensitivity analysis by adjusting for potential confounding variables and testing interaction terms.
- Applied Bayesian generalized linear models to calculate credible intervals and select analytes with large Bayesian factors.
- Implemented Adaptive LASSO, Random Forest, and XGBoosting algorithms as feature selection methods with Bootstrap for a robust predictive model.

Publications

Shang Y, Chiu Y, Kong L. "Robust Propensity Score Estimation via Loss Function Calibration". *Statistical Methods in Medical Research*, in press. 2024

Shang Y, Ning J, Minagawa K, Zhou S. "Non-Parametric Analysis of Transient Data: a Pseudo-Competing Event Approach". *Statistics in Medicine* (Under Review). 2024

Shang Y, Chiu Y, Kong L. 2024. "A Latent Variable Approach for Causal Effect Estimation under Misclassified Treatment Assignment". (Ready to submit).

Shang Y, Kim Y, Mt-Isa S, Li J. 2024. "Assessing the performance of indirect treatment comparison methods for longitudinal outcomes". (In preparation).

Zhang R, **Shang Y**, Cioccio J... "Sensitivity and specificity of chimerism tests in predicting leukemia relapse using increasing mixed chimerism". *The Journal of Molecular Diagnostics*, Volume 26, Issue 12, 1159 - 1170

Slobodanka P, **Shang Y**, Alexandors V ... "C-reactive protein improves the ability to detect hypertension and insulin resistance in mild-to-moderate obstructive sleep apnea: age effect". *Journal of Sleep Research*, 2024; e14386

Vgontzas A, **Shang Y**, He F... 0392 "Insomnia with Short Sleep Duration Is Associated with Heart Disease and Stroke: Evidence from the UK Biobank Cohort". *Sleep*. 2024 May 1;47(Supplement_1):A168-9.

Che X,..., **Shang Y**, Zhang K, Susser E, Fiehn O, & Lipkin W I. "Metabolomic analysis of maternal mid-gestation plasma and cord blood in autism spectrum disorders". *Molecular psychiatry*, 2023; 28(6):2355–2369

Endres KM, Kierys K, **Shang Y**... A Multicenter Retrospective Evaluation of Specialized Laboratory Investigations in the Workup of Pediatric Patients With New-Onset Supraventricular Tachycardia. *J Emerg Nurs*. 2022;48(6):678-687

Abdalla M, Chiuzan C, **Shang Y**... Factors Associated with Insomnia Symptoms in a Longitudinal Study among New York City Healthcare Workers during the COVID-19 Pandemic. *Int J Environ Res Public Health*. 2021;18(17):8970

Shechter A, Chiuzan C, **Shang Y**, et al. Prevalence, Incidence, and Factors Associated with Posttraumatic Stress at Three-Month Follow-Up among New York City Healthcare Workers after the First Wave of the COVID-19 Pandemic. *Int J Environ Res Public Health*. 2021;19(1):262

Skills

Statistics	Causal inference, Clinical trials, Survival analysis, Variable selection, Measurement error, Machine learning, Medical collaborative data analysis, Real-world data/evidence (Claims/EHR data).
Programming	R (base R, Tidyverse, ggplot, RShiny, ggsurvfit, parallel computing), Python (Pytorch), SAS, STATA, Bash, Linux.

Awards

Spring 2024 Travel Award	Penn State College of Medicine
Scored the highest in the Biostatistics Ph.D. qualifying exam	Penn State College of Medicine