Haotian(Matthew) Ma

+1-206-696-6842 | matthewma2022@gmail.com | matthew2022.github.io

haotian-matthew-ma | matthewma2022 | googlescholar

Long Island City, New York - 11101, United States

Objective

As a student researcher with a strong foundation in biostatistics, biochemistry, and psychology, I specialize in applying computational methods to complex biomedical challenges. My research is centered on the intersection of natural language processing (NLP) and evidence-based medicine, with a particular focus on bioinformatics and clinical data science.

Professional Experience

- **Weill Cornell Medicine [🤖]** *September 2024 - Present*
  Research Associate New York, NY
  - Employed NLP techniques to analyze variant re-classifications such as breast cancer variants BRCA
  - Implemented TensorFlow, PyTorch, and spaCy for text pre-processing and machine learning model development
  - Processed and analyzed large-scale datasets on sources such as ClinVar and PubMed
  - Built a clinical web interface using HTML, JavaScript, and CSS to provide user-friendly healthcare tools
  - Collaborated with lab members on drafting papers targeting leading conferences such as ACL and MedInfo
- **University of Washington [🧠]** *June 2021 - March 2022*
  Research Assistant Seattle, WA
  - Collected data from weekly monitored python learning advancements of 90+ student subjects
  - Organized and cleaned 3200+ raw data using Python
  - Conducted a series of statistical modelings including residual analysis, A/B Testing, and regression analysis
  - Collaborated with other teams on moderating the experimental design of the study

Education

- **Columbia University** *September 2022 - May 2024*
  Master of Science in Biostatistics New York, NY
  - GPA: 3.72/4.0
- **University of Washington** *September 2018 - June 2022*
  BA in Biochemistry, BS in Psychology Seattle, WA
  - GPA: 3.89/4.0

Projects

- **Dyadic Cluster Analysis For Comorbidity in Psychiatric Disorders in Children (ABCD) Project** *2024*
  Columbia University
  - Performed data-cleaning and exploratory data analysis on ABCD data, encompassing 11,000+ participants and 90 diagnoses in order to investigate comorbidity in psychiatric disorders in children
  - Categorized participants using cluster analysis techniques, including K-means and Latent Class Analysis then investigated cluster differences using One-Way ANOVA and other statistical tests
  - Implemented the Ising model and Lasso-regularized logistic regression to estimate a network of diagnoses, while also performing community detection algorithms and clique percolation methods for graph structure learning
  - Applied Linear Mixed-Effect models, Type III ANOVA, and Estimated Marginal Means to scrutinize T1-weighted and DTI data
- **Breathing Air Project (BAP)** *2023*
  Columbia University [🐙]
  - Executed thorough data management, encompassing raw data collection, cleaning, and preparation for end-to-end processes in R, aiming to investigate the relationship between Air Quality Index and environmentally related diseases
  - Employed Pearson's CHI-Squared Test and Welch Two Sample T -test to analyze the interplay between different environmental factors and their impact on air quality and disease prevalence
  - Leveraged R Shiny to develop an interactive website interface, and R for data visualization to effectively display results on maps

Patents and Publications

C=Conference, J=Journal, P=Patent, S=In Submission, T=Thesis

[C.3] Zihan Xu, **Haotian Ma**, Gongbo Zhang, Yihao Ding, Chunhua Weng, Yifan Peng (2025). **Natural Language Processing in Support of Evidence-based Medicine: A Scoping Review**. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics. Jul 27–Aug 1, 2025, Vienna, Austria. (Findings, accepted/in press). Preprint DOI: 10.48550/arXiv.2505.22280

[C.2] **Haotian Ma**, Zihan Xu, Wendy Chung, Chunhua Weng, Yifan Peng (2025). **A Pilot Meta-Research on Evolving Evidence Behind Genetic Variant (Re)Classification**. In *Proceedings of MEDINFO 2025: 20th World Congress on Medical and Health Informatics* (*Studies in Health Technology and Informatics*, vol. 329), pp. 108–112. IOS Press. Aug 9–13, 2025, Taipei International Convention Center (TICC), Taipei, Taiwan. DOI: 10.3233/SHTI250811

[C.1] Max Lovitt, **Haotian Ma**, Song Wang, Yifan Peng (2024). **Suicide Risk Assessment on Social Media with Semi-Supervised Learning**. In *Proceedings of the 2024 IEEE International Conference on Big Data (BigData 2024)*, pp. 8541–8549. IEEE. Dec 15–18, 2024, Washington, DC, USA. DOI: 10.1109/bigdata62323.2024.10825422

[J.1] Song Wang, Yishu Wei, **Haotian Ma**, Max Lovitt, Kelly Deng, Yuan Meng, Zihan Xu, Jingze Zhang, Yunyu Xiao, Ying Ding, Xuhai Xu, Joydeep Ghosh, Yifan Peng (2025). **A Multi-Stage Large Language Model Framework for Extracting Suicide-Related Social Determinants of Health**. *Communications Medicine* (accepted, in press). Preprint DOI: 10.48550/arXiv.2508.05003.

Skills

- **Programming Languages:** Python, R, Numpy, Pandas, Sklearn, Pytorch, Ten- sorflow, CSS
- **Data Analysis:** RStudio, SAS, MySQL, PostgreSQL, Tableau, Microsoft Office Suite

Professional Memberships

- **American Medical Informatics Association (AMIA)**, Member *July 2025 - Present*

Additional Information

Languages: Chinese(Native), English (Native), Japanese (Fluent)
Interests: Tennis, Badminton, Swimming, Dog-walking, Skiing, Singing