

Joshua Lee

Atlanta, GA 30030 | 706-325-8278 | jlee05132000@gmail.com | linkedin.com/in/joshua-j-lee

EDUCATION

Georgia Institute of Technology - M.S. Computer Science (AI concentration), May 2026

GPA: 3.79 / 4.0

Relevant coursework: Natural Language Processing; Operations Research and Optimization (LP/MIP); Simulation Modeling

Georgia Institute of Technology - B.S. Computer Science, 2018 - 2022

PROFESSIONAL SUMMARY

Data scientist with 3+ years applying statistical modeling, machine learning, and NLP to large-scale clinical, claims, and EHR data using Python, R, SQL, and AWS-hosted infrastructure. Three peer-reviewed publications. Experienced owning projects end-to-end from ambiguous problem definition through validation, building production data pipelines with quality monitoring, and communicating findings to technical and non-technical stakeholders. M.S. in Computer Science (Georgia Tech, AI concentration).

EXPERIENCE

Research Data Analyst II / Statistician, UCLA Health

January 2023 - Present

End-to-end data science workflows supporting clinical research, operational analytics, and peer-reviewed publications across oncology, transplant medicine, pediatric mental health, and homeless health care (HIPAA-regulated environment).

- Designed and independently implemented statistical and machine learning models (multivariable regression, generalized linear models across distributions, random forest classification) using scikit-learn, selecting covariates in partnership with clinicians and producing publication-grade effect estimates (odds ratios, confidence intervals, adjusted means, AUC) across 18 studies and surveys over 3 years.
- Contributed to a peer-reviewed AI and NLP study applying an open-source large language model (Llama 3.3-70B) with systematic prompt engineering and Pydantic schema validation to classify clinical text, achieving AUROC 0.935 to 0.980 against human annotation.
- Analyzed CCAE and MarketScan commercial claims data (2009 to 2024) to construct multi-year longitudinal chronic pain cohorts of approximately 5.2 million patients, reconciling ICD-9 and ICD-10 coding schemas and working with CPT and claims code sets.
- Built interactive dashboards (ADRC, STI programs) using Git-based version control and Claude Code for clinical and research leadership including the ADRC Director, tracking 20+ years of data and 10,000+ patients; optimized load time approximately 21% via parallel processing.
- Designed matched-cohort comparison studies (nearest-neighbor and exact matching, age and sex caliper controls) with Kruskal-Wallis and pairwise Wilcoxon hypothesis testing as a quasi-experimental approach to quantify group differences.
- Developed SQL views, stored procedures, and analytic tables against an enterprise data warehouse (OMOP CDM) to support cohort construction.
- Implemented automated data quality monitoring across production pipelines, preventing 50+ downstream errors through early detection of upstream integrity issues.
- Developed distributed analytics pipelines on AWS-hosted Databricks and Spark and PySpark, reducing runtime for large-scale analyses by approximately 15%.
- Led REDCap training for research staff and mentored a student worker on dashboard development.

Research Computer Scientist, U.S. Department of Veterans Affairs

October 2025 - Present

CDA-funded research program developing a clinically deployable absolute-risk prediction model for second primary lung cancer in VA patients with a prior lung cancer diagnosis (HIPAA-compliant secure environment).

- Building and evaluating a CatBoost gradient-boosted competing-risks survival model against a penalized regression approach on a 75,000+ patient cohort, using calibration and decision-curve net benefit as formal acceptance criteria.
- Designed and maintain the SQL data pipeline, views, stored procedures, and analytic tables on VA VINCI/CDW (on-premise SQL Server), covering demographics, comorbidities, staging, area deprivation index, and mortality.
- Drafted the statistical analysis plan under iterative PI review, partnering with an external statistical consultant to surface and resolve open methodological decisions.

PEER-REVIEWED PUBLICATIONS

Detection of Youth Suicide Interventions in Clinical Record Text using an Open-Source Language Model (middle author). AMIA Annual Symposium Proceedings, 2026.

Applied Llama 3.2-3B and 3.3-70B with prompt engineering and Pydantic schema validation; AUROC 0.935 to 0.980 across four intervention types on 6,687 clinical notes.

Alterations in Circulating Measures of Th2 Immune Responses Pre-Lung Transplant Associates with Reduced Primary Graft Dysfunction (middle author). Journal of Heart and Lung Transplantation, 2024.

Independently implemented multivariable logistic regression on 211 multicenter lung transplant patients (odds ratios, 95% confidence intervals).

Electronic Health Record-Integrated Legal Documentation to Measure Involuntary Mental Health Detention of Children (middle author). JAACAP Open, 2024.

Random forest classifier (AUC = 0.83) on 3,440 pediatric ED visits, integrating EHR data with neighborhood-level vulnerability indices.

SKILLS

Languages and Tools: Python, R, SQL, Spark, PySpark, Databricks, scikit-learn, CatBoost, PyTorch (exploratory), Pandas, RShiny, Streamlit, Docker, Git, Claude Code

Methods: predictive modeling, generalized linear models, survival and competing-risks analysis, matched-cohort and quasi-experimental design, hypothesis testing, NLP and LLM prompt engineering, longitudinal data analysis

Certifications: AWS Certified Solutions Architect - Associate; CITI Data or Specimens Only Research