Zachary Jacokes

(615) 604-7498 | zjacokes@gmail.com | GitHub | Google Scholar

SUMMARY

Biomedical data scientist with expertise in statistical modeling, machine learning, and reproducible analytical pipelines for high-dimensional, multi-site clinical and neuroimaging datasets. Research focus on patient phenotyping, cohort construction, and identifying stable latent structure across heterogeneous populations. Strong publication record in neurodevelopmental biomarkers and brain-behavior modeling. Committed to reproducible, well-documented, version-controlled science; experienced working collaboratively across clinical, engineering, and academic research teams.

EDUCATION

University of Virginia — Ph.D. in Data Science (Expected May 2026) Fall 2021 – Spring 2026
Member of Inaugural School of Data Science Ph.D. Cohort

Emory University — B.A. in Psychology Fall 2009 – Spring 2013

DOCTORAL RESEARCH

- Developed patient phenotyping and subgroup identification frameworks using unsupervised learning, spectral embedding, and representation learning across multi-site, high-dimensional clinical and neuroimaging datasets (500+ subjects) (publication under review; available on medRxiv).
- Designed and validated statistical models linking patient-level biomarkers to behavioral and clinical outcomes, using nested cross-validation, power analysis, and effect size estimation to ensure findings generalize across populations and data collection environments.
- Applied harmonization methods (ComBat and related approaches) to correct for site and acquisition effects across multi-site datasets, enabling valid cross-cohort comparisons analogous to multi-arm observational study design.
- Built end-to-end reproducible HPC pipelines (Slurm, Docker/Singularity) for large-scale data processing and model experimentation, with version-controlled, documented, modular code designed for reuse and auditability, reducing iteration cycles from days to hours.
- Developed interpretable models integrating multimodal data (neuroimaging, behavioral, clinical) to identify features associated with patient subgroup membership and clinical outcomes, demonstrating experience translating high-dimensional biological signal into actionable findings.

EXPERIENCE

Senior Data Specialist, University of Virginia Fall 2019 – Fall 2021

Clinical Data Infrastructure

- Architected and deployed a multi-site clinical data platform (REDCap) supporting 500+ participants across 5+ research sites and 30+ standardized behavioral instruments, serving as the longitudinal data backbone for a federally funded neurodevelopmental study.
- Designed database schema from scratch; built automated scoring pipelines computing summary metrics, T-scores, and sex-normed clinical scales across instruments, enabling endpoint derivation for downstream statistical and ML analysis.
- Implemented validation and constraint logic (range checks, type enforcement, PHI safeguards) and led audit response efforts to meet HIPAA and NIMH/NDA compliance requirements.
- Authored comprehensive documentation and trained clinical and research staff across sites; managed role-based access controls for clinicians, research assistants, and analysts.

Neuroimaging Pipeline & Reproducible Science

- Designed and implemented an automated end-to-end pipeline for ingestion, de-identification, preprocessing, and BIDS-structured output of multi-modal data (fMRI, DTI, structural MRI, EEG) across heterogeneous acquisition environments.

- Built parallelized HPC workflows enabling petabyte-scale dataset processing with reproducible, version-controlled outputs structured for downstream statistical and ML analysis.
- Contributed to cohort phenotyping by integrating imaging-derived biomarkers with clinical behavioral measures, supporting patient stratification and subgroup analyses.

Mentorship & Collaboration

- Coordinated cross-functional project teams spanning data engineering, quality control, and scientific analysis across multi-site research initiatives.
- Mentored two undergraduate researchers in ML techniques (LDA, logistic regression, tree-based methods, genomic analyses), imaging pipeline development, and research design; both contributed to published and conference work.

Programmer/Analyst, University of Southern California Fall 2015 – Fall 2019

- Coordinated multi-site neuroimaging data collection, harmonization, and distribution for the GENDAAR Research Consortium — a multi-institution longitudinal neurodevelopmental study requiring standardized data governance across sites.
- Developed MRI quality control protocol using factor analysis for cross-site data consistency; presented at OHBM 2017 and 2018 and adopted as lab standard. Designed reproducible statistical analysis workflows adopted as lab standards.
- Published multiple first- and co-authored papers on neuroimaging, multi-site data challenges, and structural brain abnormalities in neurodevelopmental disorders.

Research Assistant, Yerkes National Primate Research Center Summer 2014 – Summer 2015

- Developed Python-based data manipulation and analysis tools; contributed to experimental design and behavioral neuroscience methodology.

TECHNICAL SKILLS

Statistical Modeling & ML: Supervised and unsupervised learning, patient phenotyping, cohort construction, dimensionality reduction, representation learning, causal inference, deep learning (PyTorch/TensorFlow), time-series modeling, experimental design, nested cross-validation, power analysis, harmonization (ComBat)

Programming: Python, R, SQL, Bash; version-controlled collaborative codebases (Git/GitHub); modular, documented, reproducible scientific code

Data Infrastructure: REDCap, relational database design, ETL pipelines, large-scale data validation, HPC (Slurm), containerization (Docker/Singularity), AWS S3, Globus

Neuroimaging & Biomedical: fMRIPrep, AFNI, NiLearn, BIDS, multimodal data integration (fMRI, DTI, EEG, behavioral)

Scientific Communication: 14 peer-reviewed publications, 15+ conference presentations (OHBM 2016–2023), cross-functional stakeholder communication, technical documentation and training

SELECTED PUBLICATIONS

1. Jacokes Z, Beeler-Duden S, Lawson S, et al. Autism Sensory Profiles Predict Stimulus-Evoked Insula Connectivity. MedRxiv (preprint). *Topography-aware brain-behavior modeling; patient subgroup analysis.*
2. Jacokes Z, Adoremos I, Hussain AR, et al. Unsupervised Dimensionality Reduction Techniques for the Assessment of ASD Biomarkers. Biocomputing 2025. World Scientific; 2024:614–630. *Representation learning for phenotyping in high-dimensional clinical data.*
3. Jacokes Z, Jack A, Sullivan CAW, et al. Linear discriminant analysis of phenotypic data for classifying autism spectrum disorder by diagnosis and sex. Front Neurosci. 2022;16:1040085. *ML-based patient classification; generalization across heterogeneous populations.*
4. Ressa HJ, Newman BT, Jacokes Z, et al. Widespread associations between behavioral metrics and brain microstructure in ASD suggest age mediates subtypes. Imaging Neuroscience. 2025;3. *Age-stratified subgroup analysis; phenotype-biomarker association relevant to treatment response modeling.*
5. Newman BT, Jacokes Z, Venkadesh S, et al. Conduction velocity, G-ratio, and extracellular water as microstructural characteristics of ASD. PLoS ONE. 2024;19(4):e0301964. *Multimodal biomarker characterization; interpretable feature-outcome integration.*

6. Gupta R, Audhkhasi K, Jacokes Z, Rozga A, Narayanan S. Modeling Multiple Time Series Annotations as Noisy Distortions of the Ground Truth. IEEE Trans Affective Comput. 2018;9(1):76–89. *EM framework for noisy ground truth inference in longitudinal observational data.*

Full publication list: 14 journal articles, 1 book chapter, 15+ conference abstracts (OHBM 2016–2023)

UNIVERSITY SERVICE & LEADERSHIP

- University of Virginia Raven Society: First inductee from the School of Data Science; member of Selection Committee
- UVA Brain Institute: Consulted on neuroscience funding allocation and contributed to strategic direction discussions
- Neurodata Interest Group: Founding member; led biweekly discussions on seminal publications in neuroscience and data science; practiced and supported scientific presentation skills across the group