Zachary Jacokes

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

SUMMARY

Data scientist and clinical research professional with extensive experience designing and operating data systems for longitudinal, multi-site, regulated biomedical studies. Expertise in cohort construction, patient phenotyping, harmonization of heterogeneous clinical data, and building reproducible analytical pipelines that meet regulatory and compliance standards (HIPAA, NIMH/NDA). Deep domain knowledge in neurology and neurodevelopmental disorders; methods experience spanning causal inference, high-dimensional statistical learning, and multimodal data integration. Seeking roles in real-world evidence, clinical data science, or outcomes research where rigorous analysis of complex patient populations drives therapeutic and regulatory decision-making.

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

CORE COMPETENCIES

Clinical & Outcomes Analysis: cohort construction, patient phenotyping, longitudinal data modeling, time-to-event analysis, propensity score methods, comparative effectiveness, observational study design.
Regulatory & Compliance: HIPAA, NIMH/NDA data standards, audit response, PHI handling, multi-site IRB data governance.
Data Systems: REDCap, relational database design, schema design, ETL pipelines, large-scale data validation.
ML & Statistics: supervised and unsupervised learning, dimensionality reduction, causal inference, deep learning.
Infrastructure: Python, R, SQL, Bash, HPC (Slurm), Docker/Singularity, AWS S3, Git and Github.

EXPERIENCE

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

Clinical Data Platform (REDCap)

- Architected and deployed a multi-site clinical data platform across 5+ research sites supporting 500+ participants and 30+ standardized behavioral instruments, functioning as the longitudinal backbone for a federally funded neurodevelopmental study.
- Designed database schema from scratch to align digital capture with validated clinical instruments while meeting NIMH/NDA reporting requirements.
- Built automated scoring pipelines computing summary metrics, T-scores, and sex-normed clinical scales across instruments, enabling endpoint derivation across sites.
- Implemented validation and constraint logic (range checks, type enforcement, PHI safeguards) to ensure data integrity at scale; led audit response efforts and adapted validation practices to meet HIPAA compliance requirements.
- Developed branching logic and conditional workflows to dynamically adapt surveys based on participant responses.
- Managed role-based access controls for clinicians, research assistants, and analysts across sites, supporting secure and compliant data workflows.
- Authored comprehensive documentation and trained clinical and research staff on system use and data standards.

Neuroimaging Data Pipeline (End-to-End Automation)

- Designed and implemented an automated pipeline for ingestion, de-identification, and preprocessing of multi-modal imaging data (fMRI, DTI, structural MRI, EEG), enabling consistent biomarker derivation across heterogeneous acquisition environments.
- Applied ComBat and related harmonization methods to correct for site and scanner effects in multi-site neuroimaging data, ensuring comparability of derived biomarkers across cohort sub-populations.

- Built parallelized HPC workflows (Slurm) enabling petabyte-scale dataset processing with reproducible 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.

Team Leadership & Mentorship

- Coordinated cross-functional project teams spanning data engineering, quality control, and clinical analysis workflows across multi-site research initiatives.
- Mentored two undergraduate researchers in ML techniques, imaging pipeline development, and research design.

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

- Coordinated multi-site neuroimaging data collection and distribution pipelines for the GENDAAR Research Consortium, a multi-institution longitudinal neurodevelopmental study requiring standardized data governance across sites.
- Designed reproducible statistical analysis workflows adopted as lab standards; developed MRI quality control protocol using factor analysis for cross-site data consistency.
- Published multiple first-author and co-authored papers on neuroimaging, multi-site data challenges, and structural brain abnormalities.

DOCTORAL RESEARCH

- Designed machine learning experimentation pipelines for high-dimensional, multi-site, longitudinal datasets (500+ subjects) combining neuroimaging-derived biomarkers with behavioral phenotype data, applying nested cross-validation and power analysis to ensure robust, generalizable findings.
- Developed patient phenotyping frameworks using unsupervised learning and spectral embedding to identify stable clinical subgroups across heterogeneous datasets (directly applicable to patient stratification in outcomes research; publication under review and available on medRxiv).
- Applied harmonization methods (ComBat, nested cross-validation, effect size estimation) across scanners, sites, and populations to enable valid cross-cohort comparisons (analogous to covariate adjustment and site correction in multi-arm observational studies).
- Built interpretable models linking neural biomarkers to behavioral outcomes through topography-aware brain-behavior integration, demonstrating experience translating biological signal into clinically meaningful endpoints.
- Designed and operationalized scalable HPC pipelines for reproducible large-scale analysis, reducing iteration cycles from days to hours.

SELECTED PUBLICATIONS

1. Jacokes Z, Beeler-Duden S, Lawson S, et al. Autism Sensory Profiles Predict Stimulus-Evoked Insula Connectivity. MedRxiv (preprint). *Patient subgroup analysis; brain-behavior endpoint derivation.*
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. *Phenotyping and biomarker identification 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 and generalization across 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.*
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 and interpretability.*
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 data.*

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