Zachary Jacokes, Ph.D.

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

SUMMARY

Data scientist and systems engineer who designs and operates data systems in messy, multi-site, regulated biomedical environments. Ten years of experience turning heterogeneous clinical, neuroimaging, and behavioral data into reproducible analytical infrastructure and decision-relevant evidence. Expertise spans the full stack: cohort construction, patient phenotyping, harmonization across sites and instruments, statistical modeling and machine learning, and containerized HPC pipelines that meet HIPAA and NIMH/NDA standards. Strong publication record in neurodevelopmental biomarkers and brain-behavior modeling. Practiced collaborator across clinical, engineering, and research teams; experienced translating complex analyses to non-technical audiences.

EDUCATION

University of Virginia — Ph.D. in Data Science *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, endpoint derivation.
- **Machine Learning & Modeling:** supervised and unsupervised learning, classification, regression, clustering, dimensionality reduction, representation learning, spectral embedding, time-series modeling, deep learning (PyTorch/TensorFlow), neural networks from first principles (NumPy).
- **Statistics & Causal Inference:** hypothesis testing, mixed-effects models, nested cross-validation, power analysis, effect size estimation, harmonization (ComBat), causal inference for observational and experimental data, A/B testing methodology, multiple-comparisons correction.
- **Regulatory & Compliance:** HIPAA, NIMH/NDA data standards, audit response, PHI handling, multi-site IRB data governance, role-based access control.
- **Data Systems & Engineering:** REDCap, relational database design, schema design from scratch, ETL/ELT pipelines, large-scale data validation, branching logic, derived-metric and scoring pipelines.
- **Infrastructure & Programming:** Python (advanced), R (advanced), SQL, Bash; HPC (Slurm); containerization (Docker, Singularity); AWS S3; Globus; Git/GitHub; reproducible, modular, documented scientific code.
- **Neuroscience & Signal Analysis:** fMRI/DTI preprocessing, EEG analysis, connectivity modeling, BIDS, fMRIPrep, AFNI, NiLearn, multimodal data integration, scanner/site harmonization.
- **Scientific Communication:** 14 peer-reviewed publications, 15+ conference presentations (OHBM 2016–2023), cross-functional stakeholder communication, statistical analysis plans, technical documentation and training.

EXPERIENCE

Doctoral Researcher | University of Virginia School of Data Science *2021 – 2026*

- 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, power analysis, and effect size estimation to ensure robust, generalizable findings.
- Developed patient phenotyping and stratification frameworks using unsupervised learning, spectral embedding, and representation learning to identify stable clinical subgroups across heterogeneous datasets (publication under review; available on medRxiv).

- Applied harmonization methods (ComBat, nested cross-validation) across scanners, sites, and populations to enable valid cross-cohort comparisons — directly analogous to covariate adjustment and site correction in multi-arm observational studies.
- Built interpretable multimodal models linking neural biomarkers to behavioral outcomes through topography-aware brain-behavior integration, translating high-dimensional biological signal into clinically meaningful endpoints.
- Designed and operationalized scalable, containerized HPC pipelines (Slurm, Docker/Singularity) for reproducible large-scale analysis and model experimentation, with version-controlled, modular, documented code; reduced iteration cycles from days to hours.
- Translated statistical methods and results to non-technical audiences through conference presentations, peer-reviewed publications, and cross-functional stakeholder communication.

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 relational database schema from scratch to align digital capture with validated clinical instruments while meeting NIMH/NDA reporting requirements.
- Built automated scoring and ETL pipelines computing summary metrics, T-scores, sex-normed clinical scales, and endpoint aggregations across instruments, enabling downstream statistical and ML analysis.
- Implemented validation and constraint logic (range checks, type enforcement, referential integrity, 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 distributed sites, supporting secure and compliant data workflows.
- Authored comprehensive documentation and trained clinical and research staff on system use, data standards, and quality requirements.

Neuroimaging & Analytical Pipeline Engineering

- Designed and implemented an end-to-end automated pipeline for ingestion, de-identification, preprocessing, and BIDS-structured output of multi-modal imaging data (fMRI, DTI, structural MRI, EEG) across heterogeneous acquisition environments.
- Applied ComBat and related harmonization methods to correct for site and scanner effects, ensuring comparability of derived biomarkers across cohort sub-populations.
- Built parallelized HPC workflows (Slurm) 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.

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 (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 — Lab of Neuroimaging *Fall 2015 – Fall 2019*

- Coordinated multi-site neuroimaging data collection, harmonization, 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 in Python, R, and SPSS, adopted as lab standards; developed MRI quality control protocol using factor analysis for cross-site data consistency (presented at OHBM 2017 and 2018; adopted as lab standard).
- Published multiple first- and co-authored papers on neuroimaging, multi-site data challenges, and structural brain abnormalities in neurodevelopmental disorders.
- Presented analytical results to interdisciplinary scientific audiences including clinicians, engineers, and academic collaborators.

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.

Research Assistant | Georgia Institute of Technology *Summer 2013 – Spring 2014*

- Designed and implemented experimental research protocols; built foundational programming and analysis skills in scientific computing environments.

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; multimodal 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. *— Representation learning and 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 and 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.*
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: An Expectation-Maximization Approach. IEEE Trans Affective Comput. 2018;9(1):76–89. *— EM framework for noisy ground truth inference in longitudinal time-series data.*

Full publication list: 14 peer-reviewed 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.