**Eric M Roberts, MD PhD**

|  |  |
| --- | --- |
| **Berkeley, California, USA**  **[eroberts330@gmail.com](mailto:eroberts330@gmail.com)** | **[linkedin.com/in/eric--roberts](http://linkedin.com/in/eric--roberts)**  **[vitalita.com/biostatistics](http://vitalita.com/biostatistics)** |

A dedicated scientist committed to applying statistical and machine learning protocols to propel research in genomics and molecular biology. I have an accumulated 21+ years of experience spanning bench scientist support, complex regression modeling, and machine learning research in biotech and public health spaces. I excel in the development of rigorous, original protocols published in peer-reviewed methods journals, and I bring invaluable context with backgrounds in bioinformatics, genomics, clinical medicine, environmental epidemiology, and molecular biology. Working mainly in R and Python, I have flourished both collaborating with and embedded within teams of developers to produce fully functional research pipelines.

***Highlights***

* + *A deep well of experience in statistical methods and applications, including both supervised and unsupervised machine learning in the context of genomics and molecular biology*
  + *Successful track record in development of statistical and data management protocols contributing to company research pipelines*
  + *Customary role in research design and analysis to assure of statistical power and validity*
  + *Fluency in R, Python, and SQL; accustomed to working directly with IT development personnel, including the use of AWS and Docker environments*
  + *Background in clinical medicine with 21+ years of experience in biostatistics*

Qualifications and Skills

|  |  |
| --- | --- |
| **Machine Learning Tools** | Protocols such as generalized linear models with elastic nets, support vector machines, random forests, XGBoost, and asymmetric bagging  Development of original methods in cluster analysis |
| **Genomic Analysis** | Expertise in algorithms, statistics, and logic underlying genomics protocols such as WGCNA, GSA, GSEA, RNASeq2; conceptual familiarity with nucleotide sequence importation, QA/QC, annotation and assembly |
| **Statistical Tools** | Generalized linear modeling, including random and mixed effects, hierarchical, distributed lag non-linear modeling, and survival analysis  Mixture modeling, including independent and hidden Markov approaches  Gibbs and Metropolis-Hastings algorithms for the generation of Markov chains, including probit and auxiliary variable approaches |
| **Data Science** | Principal Component, eigen decomposition, UMAP analysis, and high-level graphics production |
| **Programming and technical skills** | Home languages in R and Python, including a vast array of libraries including scikit-learn and genomics-related packages; oriented to principles enabling contribution to production pipelines, such as modular design, unit testing, and portability within company systems  Comfortable in collaborative environments structured around AWS, Docker, and GitHub  Comfortable with SQL-based database query systems and programming |

Professional Experience

|  |
| --- |
| **Principal Biostatistician, Verge Genomics 2023 – 2024**   * Led research in **supervised machine learning classifiers** for the prediction of gene targets for drug discovery efforts. Authored **Python** code stack managing complex **API queries across company platform** for the assembly of training and annotation data for user-selected disease indications, performance comparisons of a suite of classifiers using nested cross-validation, and downstream analysis of feature weights to inform future data generation efforts * Formulated **novel metrics** for gauging the **replicability** of **gene expression correlations** across cohorts and between human, animal, and in vitro models based on the probability distributions of correlation matrix elements, becoming one of the core protocols for the **company drug discovery platform** used to select disease signatures for neurodegenerative, neuropsychiatric, and autoimmune indications * Formulated a **novel expansion** of Gene Set Analysis (GSA; Efron and Tibshirani (2007), Ann. Appl. Stat. 1(1):107-129) to quantify the **consistency of set-wise gene dysregulation** across experiments and between human, animal, and in vitro models; Python code for this and custom implementation of efficiency-enhanced versions of original GSA calculations was incorporated into the company analytic pipeline, becoming one of the core protocols for **weighing strength of evidence** for the **company drug discovery platform** * Devised **flexible Generalized Linear Mixed Modeling protocol** enabling the efficient analysis and plotting of diverse **microscopy experimental outputs** for bench scientists across the company |
| **Senior Biostatistician, Verge Genomics 2019 - 2022**   * Formulated **novel unsupervised machine learning** approach for the **detection of clusters of co-expressed genes**, similar in spirit to Langfelder Horvath (2008, BMC Bioinformatics 9:559) but based on probability theoretic reasoning for increased validity of output; adopted as one of the **core protocols** for the **company drug discovery platform** used to select disease signatures for neurodegenerative, neuropsychiatric, and autoimmune indications * Conducted **study design**, **power analysis**, and **results reporting** for in vitro and in vivo studies using outcomes including rodent biomarker and behavioral endpoints, histological changes, cell survival, cellular morphological differentiation, gene expression, puncta and stress granule formation, and immunofluorescence in collaboration with bench scientists across the company * Implemented software for the fitting of **random-effects hazard models** for an **automated pipeline** tracking survival and morphology of neurons for in vitro experiments |

|  |  |
| --- | --- |
| **Research Scientist, Public Health Institute 2005 – 2019** | |
| **Novel Methods Development**  **(see publications)** | Application of t-distributed Heckman selection models for multiply missing-not-at-random (MNAR) data with complex sampling design  Quantified time-dependent vulnerability to exposures with multi-hierarchy Bayesian modeling using a Metropolis-within-Gibbs algorithm |
| **Technical Domains** | Generalized linear mixed models, Poisson hidden Markov models, dependent mixture models, multivariate t-distributed models, Heckman selection models, spatial conditional autoregressive modeling, hierarchical Bayesian modeling, Integrated Nested Laplacian Approximation (INLA), geodatabases, spatial Scan statistics, meta-analysis |
| **Project Topics** | Emergency department visits by patients with sickle cell disease, Parkinson’s disease incidence and prevalence, amyotrophic lateral sclerosis incidence, pediatric lead exposure, environmental health surveillance database design, breast cancer geography, exposure risk analysis, drinking water safety |
| **Special Roles** | Principal investigator, leader of statewide stakeholder-driven science education program, shepherded of applications through Committee for the Protection of Human Subjects, navigated bureaucracies across multiple state agencies for project adherence to institutional standards for confidentiality protection |
| **Research Manager, Impact Assessment, Inc. 2002 – 2005** | |
| **Technical Domains** | Real-world data, non-parametric locally-weighted spatial estimation |
| **Project Topics** | Environmental health surveillance, pediatric asthma, maternal and infant health, pesticide exposure, autism spectrum disorders |
| **Special Roles** | Principal Investigator, program evaluation, convened multi-region task force for stakeholder-led evaluation of public health surveillance activities |

Education and Training

|  |
| --- |
| **Johns Hopkins University program in Sequence Analysis and Genomics 2019-2021**  **coursework:**  Introduction to Programming using Java  Introduction to Bioinformatics  Practical Computer Concepts in Bioinformatics  Next Generation DNA Sequencing and Analysis  **Postdoctoral fellowships in Health Services Research and Pediatrics 2003**  University of California at San Francisco  **Residency in Pediatrics 2000**  Tufts University/New England Medical Center  **Doctor of Medicine 1997**  University of Illinois at Urbana-Champaign  **PhD in Social Work 1995**  University of Illinois at Urbana-Champaign    **BA in Chemistry 1989**  University of Chicago |

Clinical experience

|  |
| --- |
| * Five years of ambulatory and inpatient pediatrics, including teaching and mentoring of medical students and residents |

Awards and Honors

|  |
| --- |
| **Resident Research Award**, American Academy of Pediatrics, for “Racial and Ethnic Disparities in Childhood Asthma: The Role of Clinical Findings” **2000**  **Daniel K. Bloomfield Fellowship Award for outstanding contributions to the Medical Scholars Program**, University of Illinois at Urbana-Champaign **1996** |

Selected Publications

|  |
| --- |
| *Selected from > 40; h-index 23*  Willis AW, **Roberts EM**, Beck JC et al. Incidence of Parkinson disease in North America. NPJ Parkinson’s Disease, 2022;8:170.  Marras C, Beck JC, Bower JH, **Roberts EM**, Ritz B, Ross GW, Abbott RD, Savica R, Van Den Eeden SK, Willis AW, Tanner CM; Parkinson’s Foundation P4 Group. Prevalence of Parkinson’s disease across North America. NPJ Parkinson’s Disease, 2018;4:21.  **Roberts EM**, Madrigal D, Valle J, King G, Kite L. Assessing Child Lead Poisoning Case Ascertainment in the US, 1999-2000. Pediatrics, 2017;139(5):e20164266.  **Roberts EM**, English PB. Analysis of multiple-variable missing-not-at-random survey data for child lead surveillance using NHANES. Statistics in Medicine, 2016;35(29):5417-5429.  Valle J, **Roberts EM**, Paulukonis S, Collins N, English P, Kaye W. Epidemiology and surveillance of amyotrophic lateral sclerosis in two large metropolitan areas in California. Amyotrophic lateral sclerosis & frontotemporal degeneration. 2015;16(3-4):209-15.  **Roberts EM**, English PB. Bayesian modeling of time-dependent vulnerability to environmental hazards: an example using autism and pesticide data. Statistics in Medicine. 2013;32(13):2308-19.  **Roberts EM**, English PB, Grether JK, Windham GC, Somberg L, Wolff C. Maternal Residence Near Agricultural Pesticide Applications and Autism Spectrum Disorders Among Children in the California Central Valley. Environmental Health Perspectives, 2007;115(10):1482-9.  Full publication list: <https://scholar.google.com/citations?user=gGnNs-sAAAAJ&hl=en> |