Improving AI/ML-Readiness of Data Generated from HABLE or Other NIH-Funded Research

PI: Sid O'Bryant **Speaker: Fan Zhang**

Institute for Translational Research

Department of Family Medicine

UNT Health Science Center

Summary of the project

Currently, data collected and shared in the biorepository of our parent project HABLE continues to expand rapidly. Artificial Intelligence and Machine Learning (AI/ML) cannot make these data valuable to biomedical research until these data are AI/ML-ready. Therefore, there is an urgent need to develop effective AI/ML readiness for HABLE and other NIH-funded data-sharing projects.

The proposal will focus on three critical and common areas to improve the AI/ML-readiness of data generated from our parent HABLE project: missing data imputation, feature selection and outlier removal, and data readiness report.

Project goals

- Aim 1) Develop a Machine Learning Based Multiple Imputation Method for Handling Missing Data;
- Aim 2) Develop a Recursive Feature Elimination and Cross-Validation (RFE-CV) Algorithm for Feature Selection and Outlier removal ;
- Aim 3) Develop an Integrated Tool to Report Data Readiness.

The administrative supplement project will benefit not only the parent HABLE project but also all other NIH-funded data-sharing projects. We expect that with the development of the algorithms and tools we will complete high data readiness in the HABLE project, which will eventually make HABLE more innovative in developing state of the art methods for AD clinical trials, leading in the development of effective personalized treatments which slow the progression of, and prevent, AD.

Highlights of Outputs

- Zhang, F., Petersen, M., Johnson, L., Hall, J. & O'Bryant, S.E. A Machine Learning-Based Multiple Imputation Method for the Health and Aging Brain among Latino Elders Study. AAIC 2022 poster https://alz.confex.com/alz/2022/meetingapp.cgi/Paper/63234
- Zhang, F. Handling Missing Data. Lecture Series in TCU on Oct 19, 2022
- Zhang, F., Petersen, M., Johnson, L., Hall, J. & O'Bryant, S.E. Hyperparameter Tuning with High Performance Computing Machine Learning for Imbalanced Alzheimer's Disease Data. in Applied Sciences Vol. 12 (2022). https://doi.org/10.3390/app12136670
- Zhang, F., Petersen, M., Johnson, L., Hall, J. & O'Bryant, S.E. Combination of Serum and Plasma Biomarkers Could Improve Prediction Performance for Alzheimer's Disease. in Genes Vol. 13 (2022). https://doi.org/10.3390/genes13101738
- Zhang, F., Petersen, M., Johnson, L., Hall, J. & O'Bryant, S.E. Data Readiness Reporting for the Health and Aging Brain among Latino Elders Study. AAIC 2023 poster (in preparation)

Challenges

- Automatic and dynamic
  missing data imputation
  feature selection and outlier removal

- Real time dynamic data readiness reporting system

Future Work

- semi-automated user interface data readiness reporting system
  model-based approach
  human-machine interface
  real-time dynamic data readiness reporting

Thank you!