Uncertainty Analysis of Alzheimer's Disease Cell-Free mRNA Assay

PURDUE UNIVERSITY | The Data Mine

Adam Barmash Rubinchik , Alex McQuade, Ali Eman, Frederick Mildenhall, Jerlson Sik, Joylyn Pan, Katelynn Brandwick, Kunj Patel, Pranav Jadhav, Shrinand Perumal, Siddharth Dsilva, Taylor Stafford, Teresa Nguyen, Yin Tingyu, Yi Lin Yang

SUPERFLUID

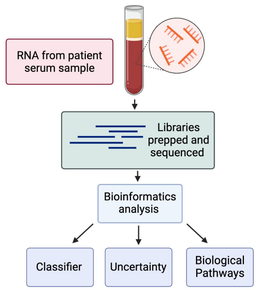
Introduction

- **Alzheimer's Disease** is the most common cause of dementia, affecting more than 40 million people around the world.
- **Noninvasive assays** can accelerate the development of therapeutic strategies and clinical diagnosis and prognosis for AD.
- **Superfluid DX** is a biotechnology company that aims to develop clinical diagnostic assays for Alzheimer's Disease by using machine learning and mRNA seq technology.

Objectives

- Generate a classifier with a high Youden's index.
- Identify differences in biological pathway enrichment between Alzheimer's and healthy patients
- Investigating Alzheimer's Gene Expression using different distributions, simulating and exploring correlation-based aggregation for uncertainty

Background



- Cell free RNA (cf-RNA) is found in the bloodstream
- cf-RNA can be extracted without invasive assays
- cf-RNA contains biomarkers for detecting key diseases through changes in gene abundance

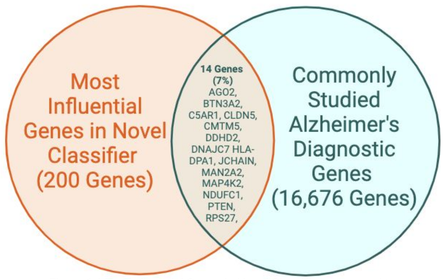


Figure 1. Most Influential Genes in Novel Classifier compared to Commonly Studied Alzheimer's Diagnostic Genes. List of commonly studied genes was developed using NeuroPro and Agora gene databases. Graphic developed using Bio Render. Genes filtered to the top 100 most and bottom 100 least influential genes in the classifier for Alzheimer's diagnosis.

Genes associated with Alzheimer's in the cf-RNA have limited overlap with established diagnostic genes for Alzheimer's

Classifier

Classifier Dev. Process

Preprocessing: Robust Scaler → Logistic regression with L2 regularization → Hyperparameter tuning: random search → Evaluation by performance on rule in vs rule out test

Classifier trained solely on cf-mRNA data (262 participants)

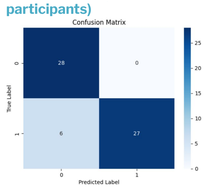


Fig. 2 Youden's index (optimal threshold):
Sensitivity + Specificity - 1
*0.81 + 1.00 - 1 = **0.81***
PPV = 1.00, NPV = 0.82

Classifier trained on cf-mRNA and APOE data (87 participants)

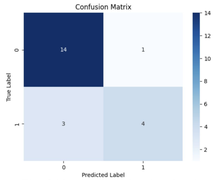


Fig. 3 Youden's index (optimal threshold):
Sensitivity + Specificity - 1
*0.57 + 0.93 - 1 = **0.5***
PPV = 0.80, NPV = 0.82

Dataset

We developed multiple classifiers using three datasets: APOE odds ratios, cf-mRNA data, combined dataset of cf-mRNA data and APOE.

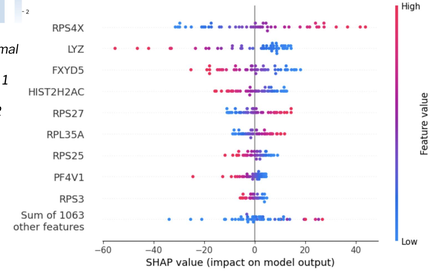


Fig. 4 SHAP values of the top 10 genes (features with the highest weight).

Performance on Rule in vs Rule out Test

The first classifier has a high PPV value which indicates that it can confidently *rule in* AD.

Limitations

The size of the dataset results in a small set of test data. Uncertain how results would scale.

Uncertainty

Analyzing Gene Expression

The ridge plot visualizes individual gene expression distributions, facilitating comparison across genes. The scatter plot contrasts log-transformed expression means with Kolmogorov-Smirnov test results.

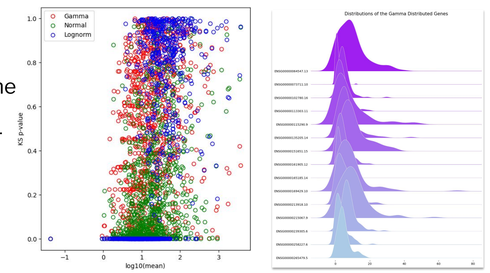


Figure 5 & 6: Kolmogorov-Smirnov Statistic Values & Gamma Distribution Ridge Plot

Simulating Different Distributions on Alzheimer's Genes Data

We primarily utilize statistical tests to identify the most fitting distribution pattern. We simulated gene samples with Gamma distributions to better understand the genetic variability associated with Alzheimer's Disease.

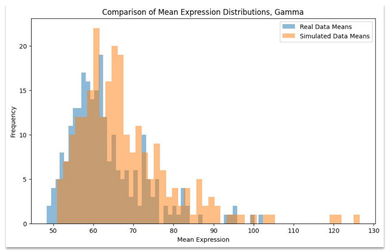


Figure 7: Comparison of Real Data Mean and Simulated Data Mean

Biological Insights

Gene Ontology Analysis for Genes Associated with Alzheimer's

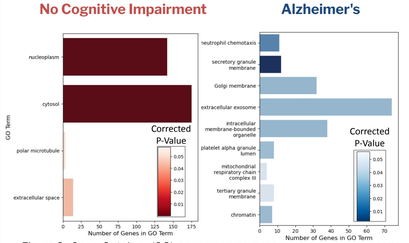


Figure 8 : Gene-Ontology (GO) analysis using classifier gene lists. GO terms identified from genes classifier determined to be predictive of Alzheimer's or non cognitively impaired patients. GO terms represent biological processes.

Pathway enrichment diverges between genes associated with Alzheimer's or control patients.

Disease State-informed PCA

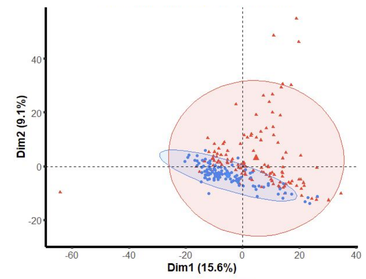


Figure 9: A Principal Component Analysis was conducted over 242 individuals and 967 genes. Both logistic regression coefficients and TPM were leveraged within the analysis. The findings demonstrate that multiple genes of interest account for a hefty portion of the overall variability across both genetic profiles.

Conclusion

- Classifier accuracy relied more on dataset size than a minimal increase in dimensionality (with the addition of APOE odds ratios).
- Minimal agreement between established AD genes and influential genes in determination of AD through classifier
- Still under work: Simulated Data Sets based on Gamma Distribution are not very accurate.

Future Work

- Explore feature interaction and the impact of multicollinear genes on the classifier.
- Further exploration of alzheimer's patient cell free RNA profile compared to established transcriptomic profile
- Expanding on the simulated data sets and exploring the Causal Relations in the Data.

Acknowledgments

We would like to thank Dr. John Sninsky, Dr. David Ross, Dr. Sarah Wang, and Dr. Jerome Braun from Superfluid DX for making this project possible and their guidance. Thanks to Dr. Ward, Maggie Betz, Kendalyn, Arjav, and the rest of the Data Mine team for their constant support. Special thanks to Dr. Marko Samara and Dr. Steffen Eikenberry for leading the ASU team.

References

Health, Center for Devices and Radiological,FDA. Ovarian Adnexal Mass Assessment Score Test System - Class II Special Controls Guidance for Industry and FDA Staff. February 27, 2020.Kallner, et al. Expression of Measurement Uncertainty in Laboratory Medicine; Approved Guideline. CLSI. 2012;32(4):EP29-A. Theodorsson E., Uncertainty in Measurement and Total Error: Tools for Coping with Diagnostic Uncertainty. Clinics in Laboratory Medicine. 2017;37(1):15-34. DOI:https://doi.org/10.1016/j.cll.2016.09.002. Toden, et al. Noninvasive characterization of Alzheimer's Disease by circulating, cell-free messenger RNA next–generation sequencing. Sci. Adv. 2020;6: eabb 1654. DOI:https://doi.org/10.1126/sciadv.abb1654