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

PURDUE UNIVERSITY — The Data Mine

Antonio Alejo, Brenden Ziemann, Daniel Tolosa, Filip Krastev, Haarika Raghavapudi, Ju Na Lee, Kendalyn Fruehauf, Naren Ram, Pallavi Joshi, Raj Banda, Ruilin Yu, Sai Mahit Vaddadi

Introduction

About Molecular Stethoscope

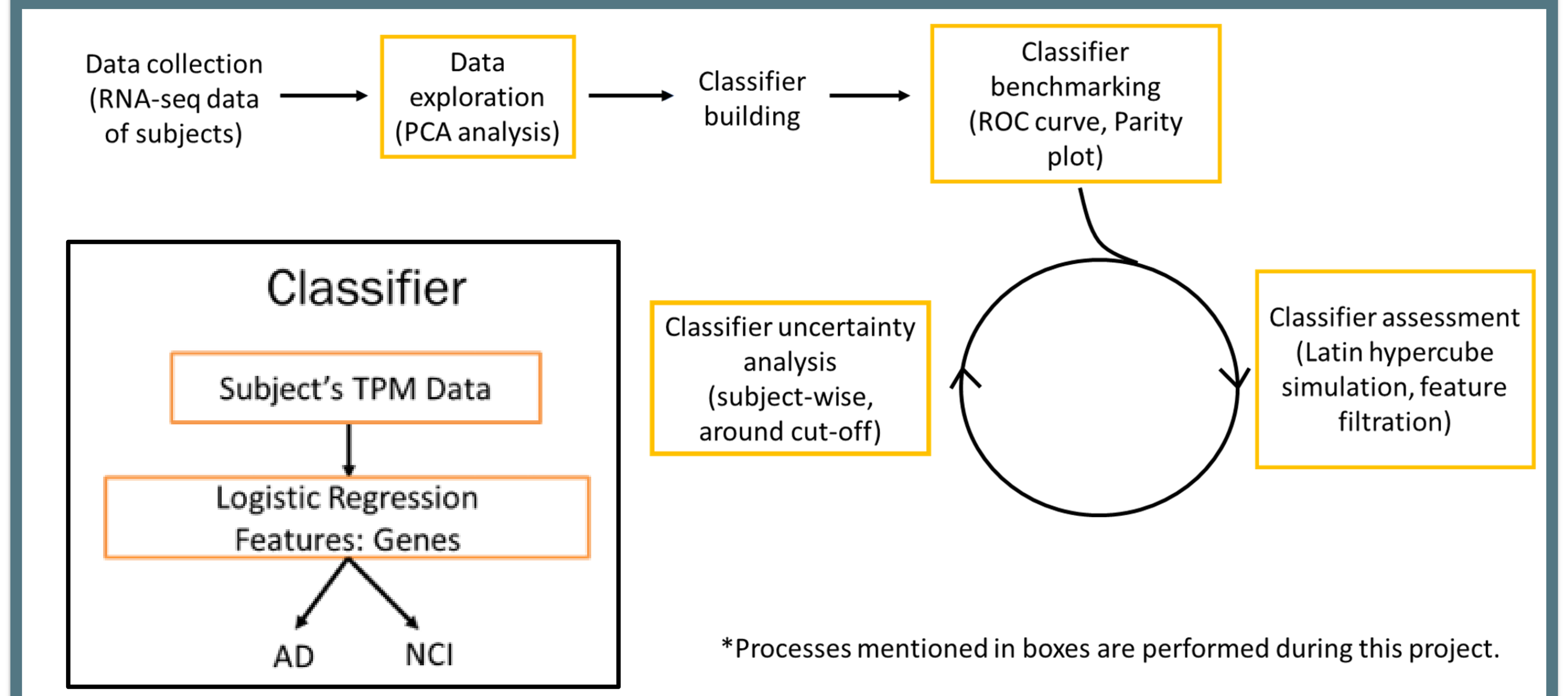
Molecular Stethoscope is a biotechnology company that aims to develop clinical diagnostic assays for Alzheimer's Disease by using machine learning and cell free mRNA seq technology.

Problem - Alzheimer's Disease (AD) affects more than 40 million people worldwide. Current diagnostic tools are inaccurate & invasive.

Motivation – We hope to develop a non-invasive, clinical-grade diagnostic tool for AD, using a multi-analyte classifier on cell-free mRNA data.

Goal - Using this classifier, **can we estimate inherent measurement uncertainties in the classifier**?

Process Flowchart



*Processes mentioned in boxes are performed during this project.

Acknowledgments

We would like to thank Dr. John Sninsky, Dr. David Ross, Dr. Sarah Wang, Dr. Jerome Braun and Dr. Samantha Khoury from Molecular Stethoscope for making this project possible and their guidance. Thanks to Dr. Ward, Maggie Betz, Deb, and the rest of the Data Mine team for their constant support. Special thanks to Dr. Marko Samara for leading the ASU team.

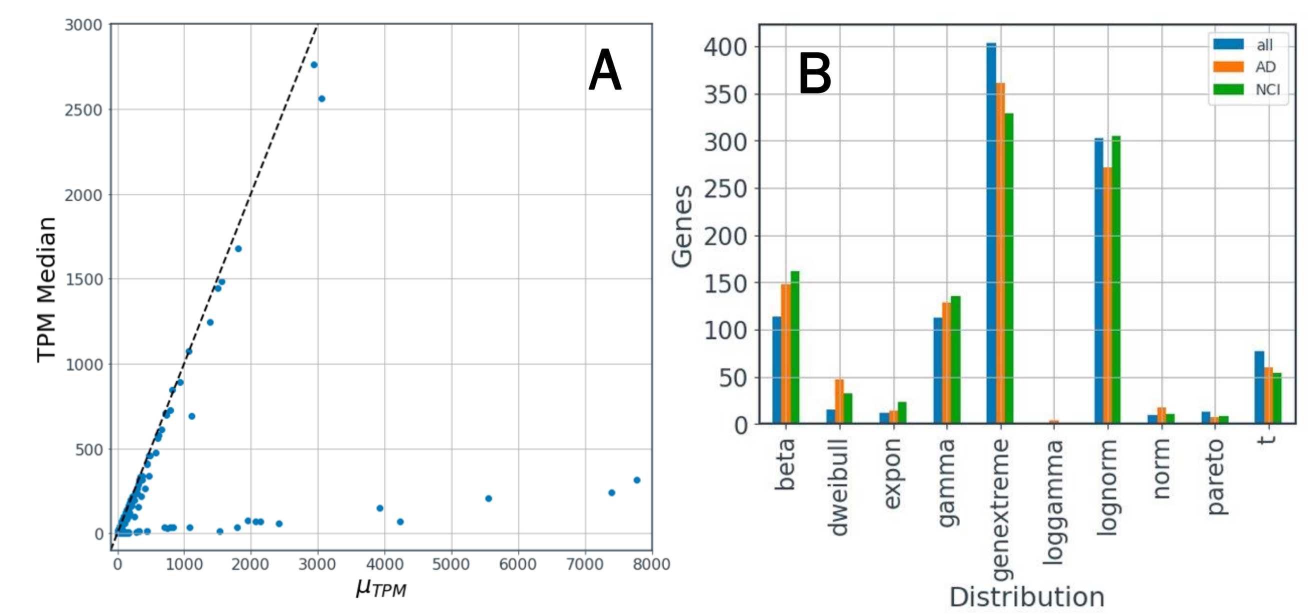
References

- Beaver et al. An FDA Perspective on the Regulatory Implications of Complex Signatures to Predict Response to Targeted Therapies. ClinCancer Res. 2017 Mar 15;23(6):1368-1372
- 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

Data Exploration

We explored the original Toden et. al. dataset by investigating the distribution of inherent data along with using Principal Component (PCA) analysis to investigate genes that may affect Alzheimer's diagnosis.

- **Genes accounting for the highest variation of subjects are related to brain function, inflammation, cell signaling.**
- **The approximations of normal distributions are probably adequate for many of the genes, though less so for genes with lower average counts.**
- **While the genes with the most explained variance have biological significance in Alzheimer's disease, the reverse is not true.**



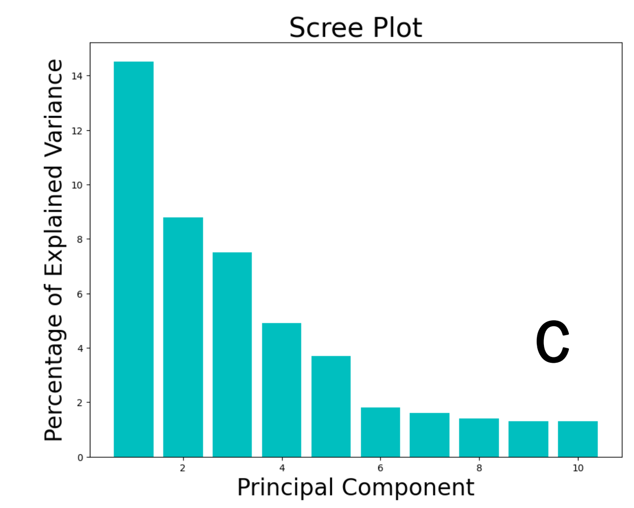
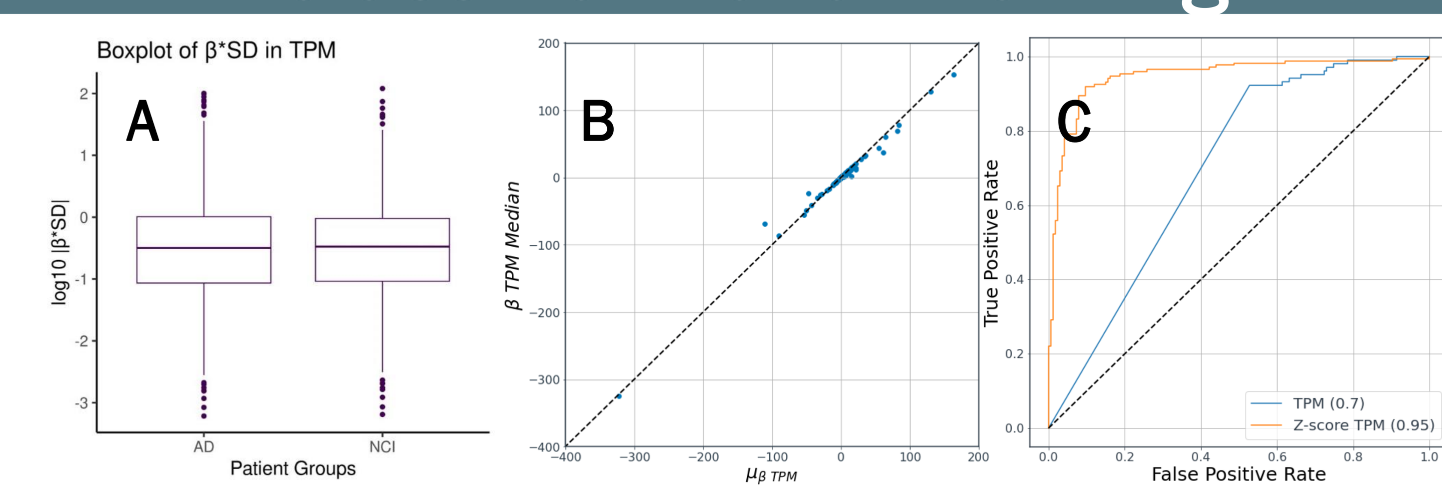


Fig 1: Exploration of Toden et al. data: (A) Parity Plot of Mean and Median TPM counts and (B) count of fitted distributions for each gene in all patients in the Toden et. al. dataset. Distributions were also fitted on subsets of the data in (B). (C) PCA scree plot of top 10 principal components containing genes which have highest variations across samples.

Gene (PCs 1 & 2)	Biological relevance in Alzheimer's Disease
BIN2, LRBA, RGS2	AD marker
CMTM5, PRUNE1	Neural abnormalities
RGS6, ADGRE5	Abnormal cellular signal transduction
TNFSF4, VEPH1, MNDA	Inflammation
NAMPT, SELL, UBC	AD progression

Table 1: Genes in PC1 and PC2 with highest variations across samples: The genes are tabulated with their biological relevance in Alzheimer's Disease.

Classifier Benchmarking



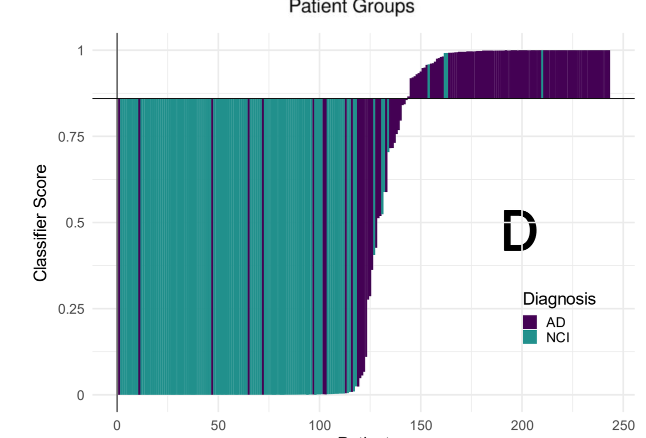


Fig 2: Benchmarking of Toden et. al. Classifier: (A) Box plot of the classifier coefficient (β) * TPM Standard Deviation of genes in all patient subgroups. (B) Parity Plot of Mean and Median β*TPM counts in all genes. (C) ROC Curve of Toden et. al. classifier on raw TPM data and TPM data normalized to a Z-score based on each gene. (D) Waterfall Plot of patient classifier score data.

We benchmarked Toden et. al.'s data and classifier by investigating the effect of the trained coefficients, normalization, and consistency of data.

- **Z-score normalization of data was used to develop classifier and performance decreases if TPM are used.**
- **Variation in data along each gene are consistent regardless of subset.**
- **Classifier is accurate in its predictions aside from a few patients near cut-off score.**

Classifier Assessment

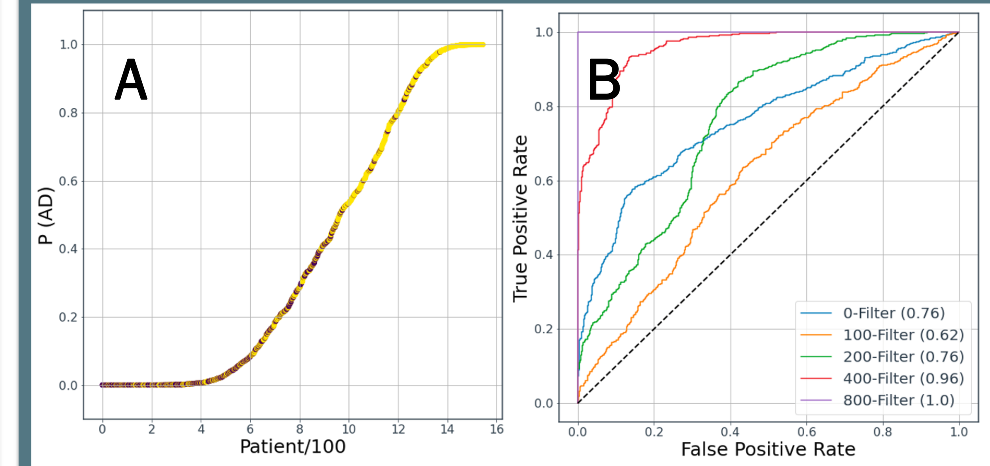


Fig 3: Assessment of Toden et al. classifier on simulated data: Plots on 1,566 simulated patient data classifier scores using MC (A) (B) ROC curves of Toden et. al. classifier simulated patient data via MC that were both unfiltered and filtered for lowest β* Mean TPM count genes.

We assessed the original Toden et. al. classifier by determining its performance on simulated patient data using Monte Carlo (MC) and Latin hypercube sampling (LHS).

- **If there is too much noise in the data, then it becomes very difficult to develop a classifier.**
- **More assumed noise decreases accuracy in the classifier's predictions.**
- **Important to use normalization in model development.**

Sampling Technique	Description	Pros	Cons
Monte Carlo	Repeatedly sampling a random set of results based on a pre-defined distribution.	• Simple implementation. • Implementable on any statistical distribution.	• Inefficient at high-dimensions. • Error increases in high dimensional sampling.
Latin Hypercube Sampling	Randomly sampling from a set equally spaced grids along a domain.	• Samples broader sample space • Distribution of data not assumed.	• Samples can cluster. • No guarantee samples are independent of one another.

Uncertainty Analysis

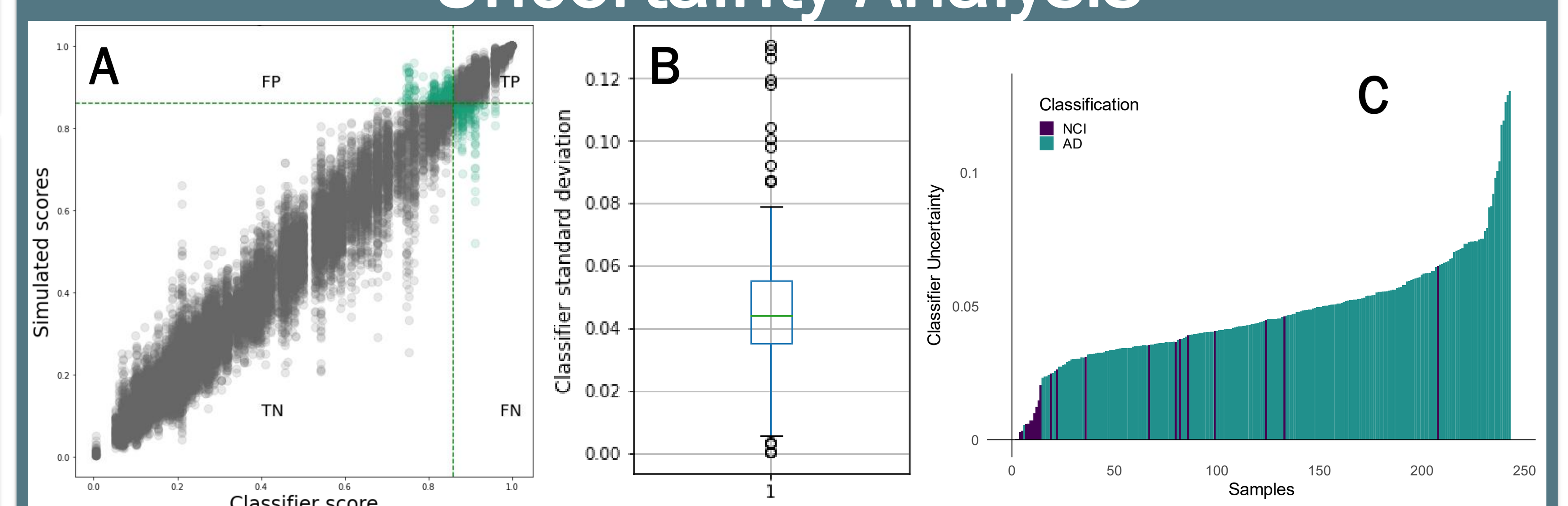


Fig 4: Analysis of Uncertainty of Toden et. al. classifier: (A) Scatter plot of classifier scores on simulated versus original data (n = 100)(B). Box plot and (C) waterfall plot of uncertainty in classifier score on original data. All plots are based on data simulated with 50% feature variation.

- **A high level of uncertainty in TPMs results in greater false positives and negatives in the classifier's predictions, especially around the threshold.**
- **Uncertainty in the classifier score is sensitive to the selection of the threshold.**
- **Highest uncertainty is observed for patients with scores closer to the threshold.**

Conclusions

- **This is the first in-depth study of uncertainty in a high-dimensional RNA-Seq classifier for clinical diagnosis as per FDA-recommended guidelines.**
- **Genes with maximum variation across samples are biologically relevant.**
- **Uncertainty impacts misclassification predominantly at threshold.**
- **Future studies should explore more nuanced individual gene-based variation to model uncertainty.**