Assessment of AUC and fold change for precision medicine

Kenneth Smith[1] and Sharlee Climer[1*]
[1]University of Missouri – St. Louis, Missouri, USA
*email: climer@umsl.edu

Abstract

Precision medicine is advancing patient care for complex human diseases. Discovery of biomarkers to diagnose specific subtypes within a heterogeneous diseased population is a key step towards realizing the benefits of precision medicine. However, popular statistical methods for evaluating candidate biomarkers, fold change and AUC, were designed for homogeneous data and we evaluate their performance here. In general, these metrics overlook nearly 'ideal' biomarkers when they represent less than half of the diseased population. We introduce a new metric to address this shortfall and run a series of trials comprised of simulated and biological data.

Introduction

Advances in precision medicine (PM) for cancer patients is extending the healthspan for countless lives by tailoring treatments to heterogeneous cancer subtypes. PM utilizes specific biomarker information, such as genetics and levels of proteins being produced by the patient, to diagnose their specific subtype of the disease and enable tailored treatments, prognoses, and monitoring. An additional benefit of PM is that it facilitates understanding of underlying biological mechanisms by teasing apart biomarkers into subtype groups. Knowledge of distinct biomarkers associated with each subtype empowers drug discovery as well as selections of individuals for drug trials. Many complex diseases are heterogeneous and hold potential to benefit from PM. For example, heterogeneous subtypes of late-onset Alzheimer disease (AD) are exhibited by the spectrum of genetic risk factors and clinical outcomes observed for this enigmatic disease. Efforts are underway to enable PM for AD, including the Accelerating Medicines Partnership® for AD 2.0[1], which began in 2021, and Alzheimer Precision Medicine Initiative[2], which began in 2016.

The realization of successful PM can only be attained by identifying disease subtypes and developing practical methods to diagnose and treat each subtype. A common approach is to use statistical methods to test associations of candidate biomarkers with the disease. Candidate biomarkers include genetics, demographics, and a host of observations such as imaging data, body mass index, and omics data, which includes levels of gene expression, proteins, lipids, and metabolites. Different statistics are used for categorical, ordinal, and numerical data types. Herein we focus on numerical data types, which includes omics data, imaging data (such as PET amyloid load), and other observations. Common statistics for this domain include fold change (FC) of levels between diseased cases and normal controls and area under the receiver operating characteristic curve (AUC)[3]. In addition to association tests on each biomarker, patterns comprised of multiple biomarkers are frequently identified by modeling the data as a network in which each biomarker is represented by a node and correlations between pairs of biomarkers are represented by edges between the corresponding node pairs[4,6,8,10].

This nascent research field faces challenges due to multiple issues, such as the need for large sample sizes for research studies to elicit power to sift out a subtype that may only represent a small fraction of the diseased cases. Another major challenge is that traditional statistical methods can be inappropriate when heterogeneity exists, and analytical approaches need to be reevaluated for use in this challenging domain. We previously reported that the use of standard correlation metrics in network modeling leads to increased type II errors in the presence of subtype groups[5,7,9,11]. All the correlation measures that we have examined, including Pearson's correlation coefficient[12], *r*-squared[13], dot product[14], and mutual information[15], return single scalar values that are crippled by heterogeneity[5]. They are universal measures, in that individuals in the entire group are viewed as a whole and thus subtle but crucial subgroup structures, which manifest heterogeneity of the individuals in the group, are ignored. If two analytes are highly correlated for a subset of individuals but not at all correlated for the others, the correlation value is reduced due to the latter individuals, thereby contributing to false negative signals[5,11].

In this manuscript, we examine the use of FC and AUC when subtype groups exist. A traditional approach for identifying differentially expressed (DE) analytes involves calculating the FC of the analyte expression levels between two groups, where FC = (level in diseased cases) / (level in normal controls). If the FC is above or below a certain threshold, the analyte is considered differentially expressed. A single value representing the expression level of the analyte is required for each group: usually the median or mean. Often a threshold of >2 is used to indicate significant up-expression in the diseased cases group and a threshold of <0.5 for down-expression. In order to better visualize up- and down-regulated analytes, the log2FC is often employed, where log2FC = $\log_2$((level in diseased cases) / (level in normal controls)), providing a threshold of abs(log2FC) > 1 tests for both up- and down-regulation[16]. FC calculations are unstable when the expression levels are near the noise level of the measurement system. This can lead to false positives at low intensity levels. On the other hand, FC is also biased against samples that have high expression levels, but small differences between two groups[17]. Mariani *et. al.* report that high FC thresholds are needed for low intensity genes and lower thresholds are needed for high intensity genes. They introduced a variable FC threshold-based approach that uses LOESS to estimate a variance based on expression intensity, thereby alleviating the bias at both high and low intensity levels[17]. Despite these improvements to the FC calculation, there is a fundamental problem with this metric: Use of the mean or median in the presence of heterogeneity tends to miss subgroup signals.

Standard 2 x 2 contingency tables are commonly used to assess predictive accuracy using various statistics, such as Fisher's Exact Test and Youden's J index[19]. Note that Youden's J definition can be rearranged to produce a simpler interpretation:

J = (% of disease cases correctly predicted) – (% of normal controls incorrectly predicted)

A key benefit of utilizing Youden's J is that subgroups are captured, rather than being lost in a summary statistic, as is done with FC. However, without other information, small subgroups may be lost when compared to a moderate case/control model with the same J value, just lower TPR and FPR, i.e. J = 0.70 - 0.51, vs. J = 0.20 - 0.01.

When testing numerical values, FC and 2 x 2 contingency tables require the selection of a threshold to separate diagnostic classifications. AUC originated as a tool for radar receivers and became a prevalent tool for the diagnostic ability of a biomarker without reliance upon a specified threshold to predict

diseased cases vs. normal controls. For AUC, a plot of the true positive rate (TPR) vs. false positive rate (FPR) is constructed by sweeping through all possible threshold values and the area under the curve is returned as the AUC value[18]. If no discrimination is provided by the biomarker, the plot would be a diagonal line from (0,0) to (1,1) with an AUC value of 0.5. AUC is a widely accepted statistic for evaluating biomarkers[3,20,21]. This metric is equally weighted on the TPR and FPR. When a given biomarker represents a single subtype, we expect the TPR to be relatively low as only a subset of AD cases is associated. Consequently, we hypothesized that screening based on AUC may discard valuable subtype biomarkers. Using simulated tests mimicking nearly 'ideal' biomarkers for subsets of disease cases, we demonstrate the failure of AUC to capture their significance.

We introduce a new metric for evaluating biomarker diagnostic accuracy which sweeps across threshold values while evaluating the Youden J index. Plots are drawn for each of TPR and FPR and the area between the plots is multiplied by 100 to produce the index value. We compute the values for this new statistic, J Sweep (JS), for the synthetic datasets previously noted as well as gene expression data generated from autopsied brain samples for AD cases and normal controls.

Methods

Biological data trials. We utilized gene expression data derived from human cortex tissue for 8,560 genes for 176 AD cases and 188 controls using Sentrix HumanRef-8 Expression BeadChip[22]. Expression data are available on NCBI's Gene Expression Omnibus (GEO), Accession GSE15222. We used these data to estimate expected parameters.

Simulated data trials. Samples were drawn from one of two normal distributions in this study: N_1~N(0.03, 0.0016) and N_2~(0.40, 0.0256). The mean and standard deviations were derived from analysis of highly DE proteins from a COVID-19 study. The size of the subtype, as a percentage of the cases, was varied over six scenarios from 5% to 50%. In each scenario, the cases in the subtype group were sampled from N_2 and the remaining cases, along with all controls, were sampled from N_1. A total of 1000 cases and 1000 controls were simulated in each scenario. Each scenario was simulated 1000 times.

New statistic. Assuming data that has been scaled and shifted as needed to provide a range from zero to one, we propose a new statistic as follows:

$$JS = 100 \times \int_{thresh=0}^{thresh=1} TPR - FPR$$

The threshold for JS is swept across all possible values and at each point it is equal to the Youden J index.

Results

Results for the simulated and biological data trials are shown in **Table 1**. The median values over 8,560 genes are shown in the biological data row. We expect only a small fraction of these genes to be differentially expressed between AD cases and normal controls, so these median values provide a baseline. All four methods performed poorly for the simulated trials with a subgroup comprised of only 5% of the diseased cases. The AUC values were higher than the baseline for subsets of size 30% and

higher, yet the values are low considering the simulated data represented nearly 'ideal' scenarios. JS exhibited the best performance, albeit the 10% subset score would not be considered to be significant. However, subsets of size 20% to 50% are all significant. AUC and JS plots for several of the trials are shown in **Figure 1**.

Discussion

Precision medicine is based upon the assumption that different subtypes exist for the given disease. We show here that popular statistics used for assessing biomarkers, fold change and AUC, generally perform suboptimally when heterogeneity exists. We also provide a new metric, JS, that appears to hold some promise in this domain. More generally, it is important that we scrutinize statistical methods that we apply to heterogeneous data to ensure progress in biomarker discovery.

Acknowledgements

We thank Jamie Lea for insights regarding the Youden J statistic. This research was funded by National Institute on Aging (NIA) grants 1RF1AG053303-01 and 3RF1AG053303-01S2.

Table 1 Median values for the biological and simulated datasets.

Trial	Fold Change	FC p-value	Max J index	AUC	JS
Biological	***1.03***	***0.00034***	***0.204***	***0.610***	***3.41***
Sim_5%	1.02	0.05757	0.056	0.525	1.51
Sim_10%	**1.04**	**0.00022**	0.101	0.548	**3.61**
Sim_20%	**1.09**	**3.11E-14**	0.194	0.598	**8.61**
Sim_30%	**1.15**	**1.29E-29**	**0.288**	**0.646**	**14.31**
Sim_40%	**1.26**	**1.04E-51**	**0.383**	**0.695**	**20.90**
Sim_50%	**1.52**	**2.82E-79**	**0.477**	**0.743**	**28.25**

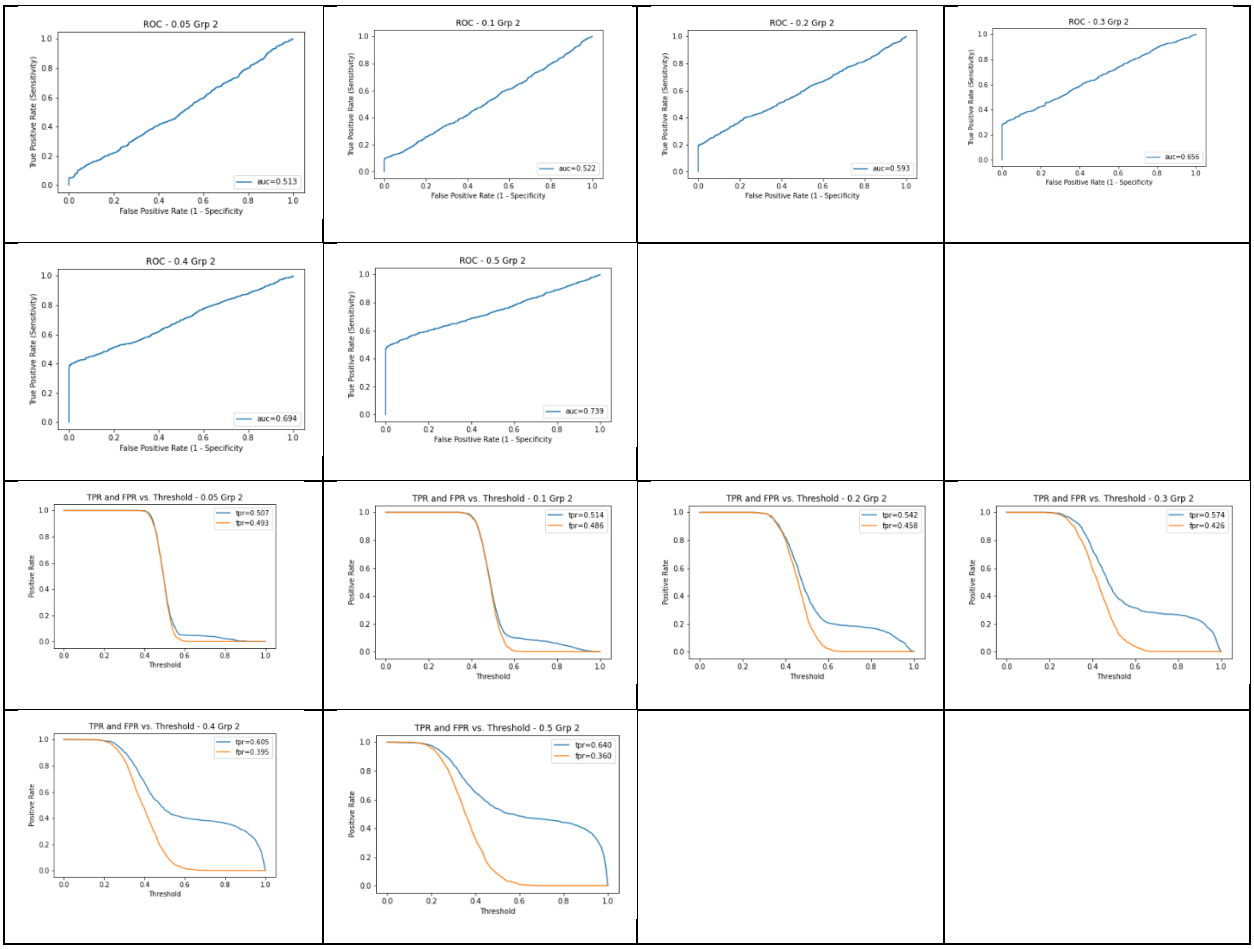

Figure 1 AUC plots (top) and JS plots (bottom) for subtype groups of 5%, 10%, 20%, 30%, 40%, and 50% of the diseased cases.

References

1. Accelerating Medicines Partnership® Program for Alzheimer's Disease (AMP® AD 2.0) | National Institute on Aging https://www.nia.nih.gov/research/amp-ad-second-iteration.

2. Hampel, H., Vergallo, A., Perry, G., and Lista, S. (2019). The Alzheimer Precision Medicine Initiative. J. Alzheimer's Dis. *68*, 1–24.

3. Xia, J., Broadhurst, D.I., Wilson, M., and Wishart, D.S. (2013). Translational biomarker discovery in clinical metabolomics: An introductory tutorial. Metabolomics *9*, 280–299.

4. Fuller, T., Langfelder, P., Presson, A., and Horvath, S. (2011). Review of Weighted Gene Coexpression Network Analysis. Handb. Stat. Bioinforma., 369–388.

5. Climer, S. (2021). Connecting the dots: The boons and banes of network modeling. Patterns *2*, 100374.

6. Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. PLoS Comput. Biol. *4*, e1000117.

7. Climer, S., Yang, W., de Las Fuentes, L., Dávila-Román, V.G., and Gu, C.C. (2014). A custom correlation coefficient (CCC) approach for fast identification of multi-SNP association patterns in genome-wide SNPs data. Genet. Epidemiol. *38*, 610–621.

8. Choi, J.K., Yu, U., Yoo, O.J., and Kim, S. (2005). Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics *21*, 4348–4355.

9. Climer, S., Templeton, A.R., and Zhang, W. (2014). Allele-Specific Network Reveals Combinatorial Interaction That Transcends Small Effects in Psoriasis GWAS. PLoS Comput. Biol. *10*.

10. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

11. Climer, S., Templeton, A.R., Garvin, M., Jacobson, D., Lane, M., Hulver, S., Scheid, B., Chen, Z., Cruchaga, C., and Zhang, W. (2020). Synchronized genetic activities in Alzheimer's brains revealed by heterogeneity-capturing network analysis. bioRxiv, 2020.01.28.923730.

12. Rodgers, J.L., and Nicewater, W.A. (1988). Thirteen Ways to Look at the Correlation Coefficient. Am. Stat. *42*, 59–66.

13. Hill, W.G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. *38*, 226–231.

14. Ogita, T., Rump, S.M., and Oishi, S. (2005). Accurate sum and dot product. SIAM J. Sci. Comput. *26*, 1955–1988.

15. Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top. *69*, 16.

16. Pacholewska, A. (2017). "Loget" - a Uniform Differential Expression Unit to Replace "logFC" and "log2FC." Matters.

17. Mariani, T.J., Budhraja, V., Mecham, B.H., Gu, C.C., Watson, M.A., and Sadovsky, Y. (2003). A variable fold change threshold determines significance for expression microarrays. FASEB J. *17*, 321–323.

18. Pepe, M.S. (2000). Receiver Operating Characteristic Methodology. J. Am. Stat. Assoc. *95*, 308–311.

19. Youden, W.J. (1950). Index for rating diagnostic tests. Cancer *3*, 32–35.

20. Zweig, M.H., and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. *39*, 561–577.

21. Sø, K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. J. Clin. Pathol. *62*, 1–5.

22. Webster, J.A., Gibbs, J.R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., et al. (2009). Genetic control of human brain transcript expression in Alzheimer disease. Am. J. Hum. Genet. *84*, 445–58.