Why significant variables aren’t automatically good predictors

Adeline Lo, Herman Chernoff , Tian Zheng and Shaw-Hwa Lo

Chiraag Limaye

Sriranga Ramakrishna

Problem Statement

- Inability to use the results of the identified statistically significant variables
- “Why Significant Variables not leading to good predictions of the outcome?”

Road Map

- Introduction
- Highly Significant v/s Highly Predictive
- Three Examples
- Analyzing the Real Breast Cancer Data
- Conclusion

Introduction

- Prediction was important
- Newly Identified variables
- GWAS Study
- Size of the Data
  - Variable Selection
  - Variable Prediction

Highly Significant vs Highly Predictive Variables

- Two popular Concepts
  1. Significance : statistical Inference
  2. Prediction : Identifying future behavior
- Key Difference : Underlying Distribution
- $P(T_n >= t_n)$
- $\sum_{x:f_D(x)<f_H(x)} f_D(x)$ and $\sum_{x:f_D(x)\geq f_H(x)} f_H(x)$.

  $$\text{prediction rate} = 0.5 \sum_x \max(f_D(x),\ f_H(x)).$$

- Highly significant : uses assumption, but no knowledge of exact distributions
- Highly Predictive uses knowledge of both f_h and f_d

Example 1

Hypothesis H

- For Variable X:
  - Mean = 0 , SD = 1
  - $0<a(x)<1$
  - Error rate = e(c,H)
- For Variable Y:
  - Mean = 0 , SD = 1

Hypothesis K

- For Variable X:
  - Mean = 3 , SD = 1
  - a(x) = random values
  - Error rate = e(c,K)
- For Variable Y:
  - Mean = 0 , SD = 0.05

Example 1 continued

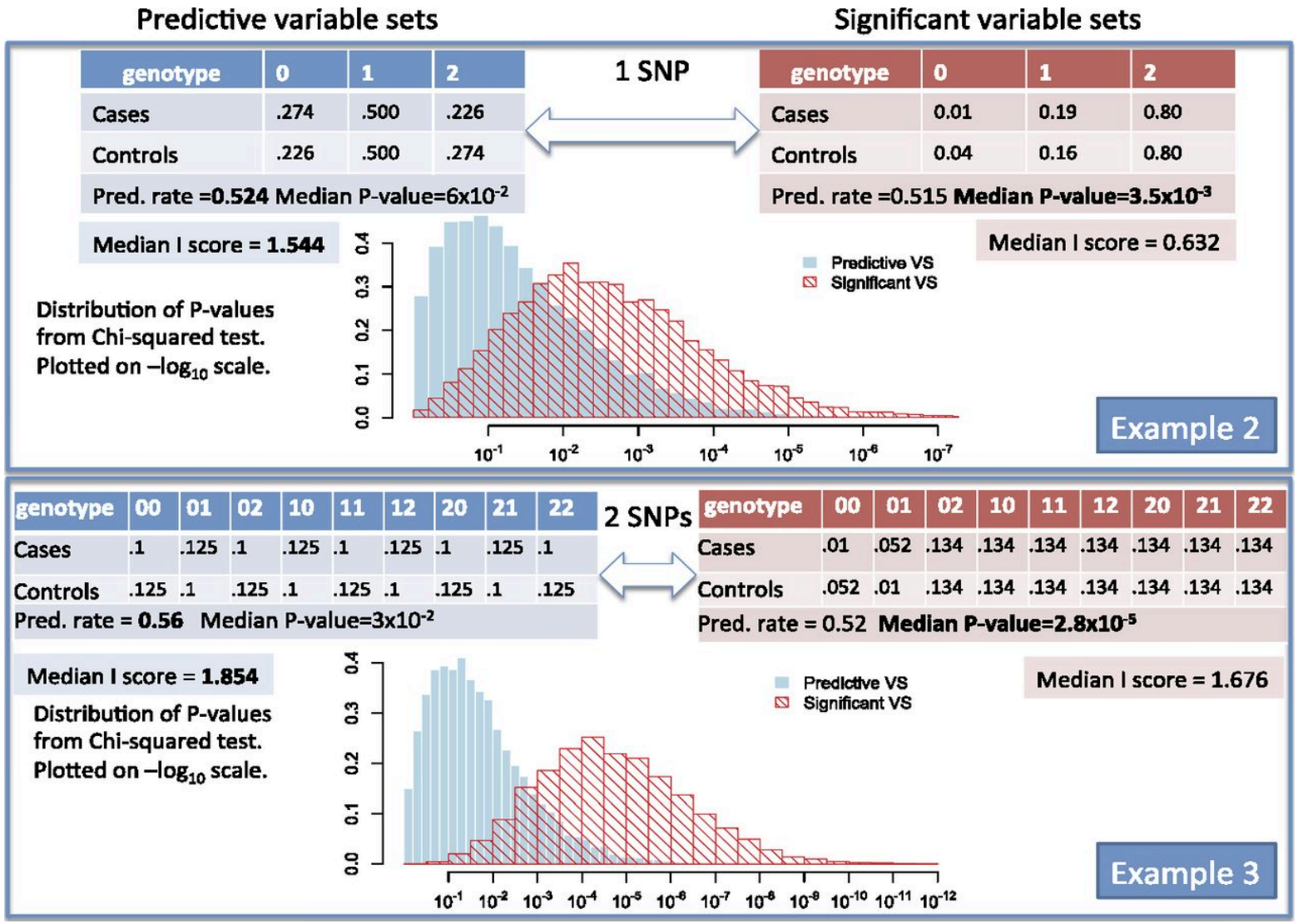
Variable X

- e_x= 0.174
- s_x= 0.0014
- Predictivity = 1-e_x = 0.826

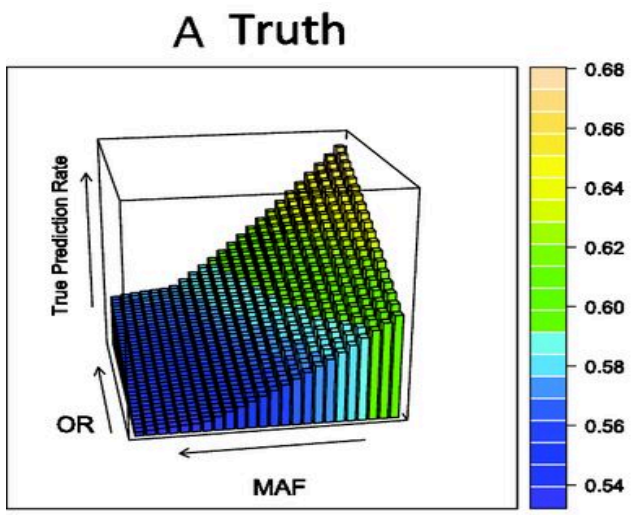
Variable Y

- e_y= 0.06
- s_y= 0.5
- Predictivity = 1-e_y = 0.94

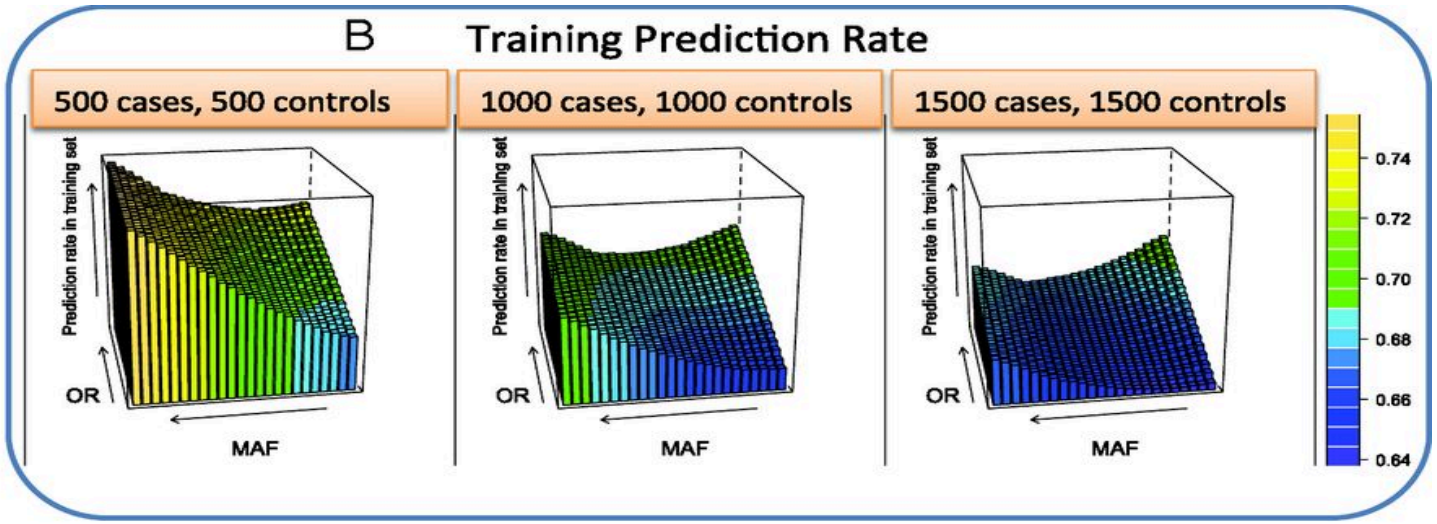
Examples 2 and 3

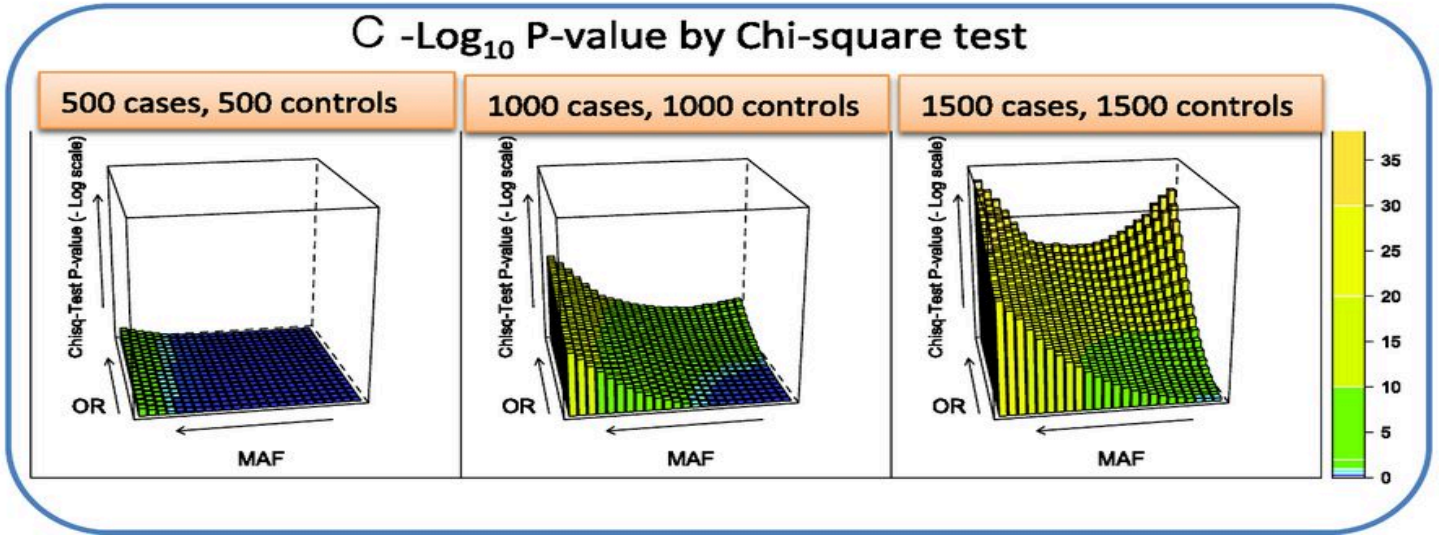


Comparing Significance test with I score

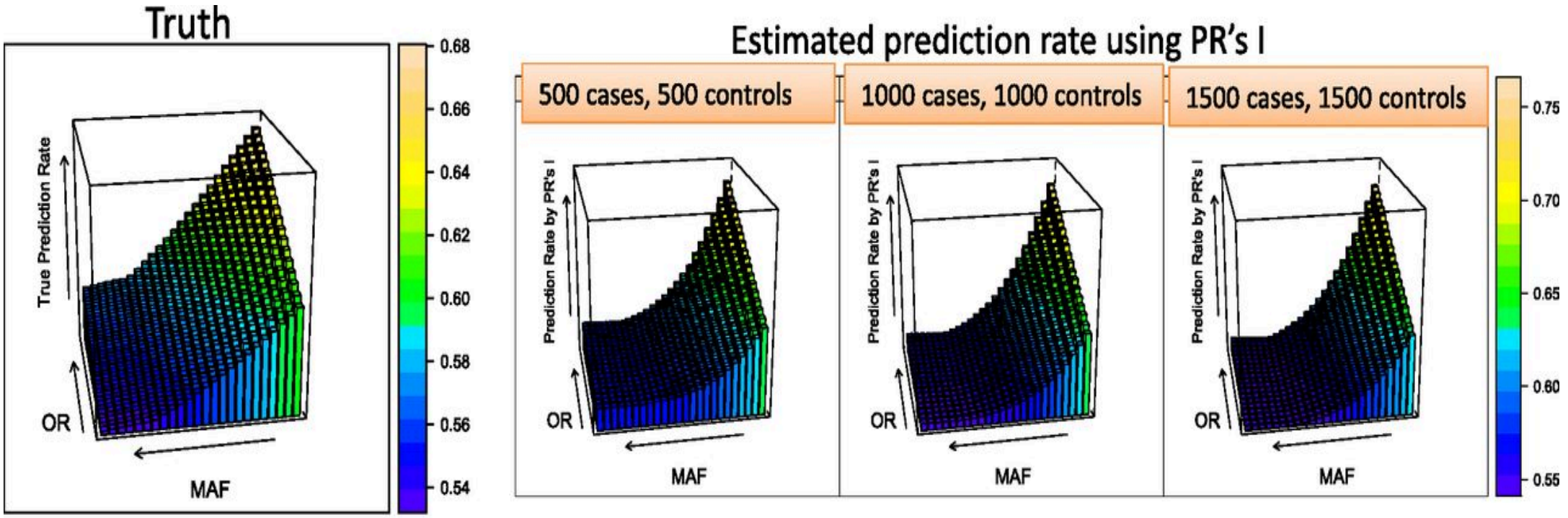


Each vertical bar is one variable set (VS). Its height and color represent the "importance" of a given VS. The taller and lighter (towards yellow) a bar, the more important the VS. In this example, 546 variable sets are considered with different MAF and OR settings. Three settings of sample sizes are considered.





Comparing Significance test with I score



Each vertical bar is one variable set.
546 variable modules are considered.

Applying I score to Real Breast Cancer Data

Table 1. Real breast cancer example: Five genes in the top returned predictive variable set from van't Veer data

Systematic name	Gene name	Marginal *P* value
Contig45347_RC	KIAA1683	0.008
NM_005145	GNG7	0.54
Z34893	ICAP-1A	0.15
NM_006121	KRT1	0.9
NM_004701	CCNB2	0.003

I Score

- Y_i= i^{th} individual
- $\bar{Y}$ = Mean of all Y values
- s = SD of all Y values
- $\bar{Y}_j$ = Mean of all Y values in cell j
- n_j = No of individuals in cell j
- n = Total no of individuals

$$I = \sum_{j=1}^{m_1} \frac{n_j}{n} \frac{\left(\overline{Y}_j - \overline{Y}\right)^2}{s^2/n_j} = \frac{\sum_{j=1}^{m_1} n_j^2 \left(\overline{Y}_j - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2},$$

Conclusion

- In order to apply efficient techniques ,we need to know the underlying distribution
- Real examples are difficult to analyze because of large number of variables
- exploration away from significance-based methodologies and toward prediction-oriented ones is encouraged
- The partition retention method, helps in reducing prediction error from 30% to 8% on a long-studied breast cancer data set.

References

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4653162/pdf/pnas.201518285.pdf
- https://en.wikipedia.org/wiki/Genome-wide_association_study
- https://www.cebm.net/2014/02/likelihood-ratios/
- https://arxiv.org/pdf/1009.5744.pdf