# A Test for Genetic Association that Incorporates Information about Deviation from Hardy-Weinberg Proportions in Cases

Sanjay Shete, Ph. D.
M. D. Anderson Cancer Center
Houston, Texas
sshete@mdanderson.org

# Hardy-Weinberg Equilibrium

- Hardy-Weinberg equilibrium is one of the most important principles in population genetics.

- Consider a locus with two alleles: A and a

- Let p be the frequency of allele A in the population. q= 1-p the frequency of allele a.

- The H-W Proportion: The frequencies of three possible genotypes AA, Aa (or aA), and aa are $p^2$, 2pq, and $q^2$, respectively.

# Hardy-Weinberg Proportion

- Several programs exist to test whether SNP genotypes are in HWP.
- Pearson’s Chi-square test: compares observed genotype frequencies to expected genotype frequencies under the HWP assumption. Important assumption for this test is large sample size (not the total but in each cells).
- Fisher’s exact test: Accurate but computationally intensive.
- Recently, MCMC methods have been proposed which are quite accurate.

•References: Guo and Thompson (1992) Biometrics, 48:361
Wigginton et al. (2005) AJHG, 76:887-893

# Case-Control Study and HWP

- Case-Controls study design has been a work-horse of association studies.
- Cases are subjects with disease of interest and controls are subjects without the disease.
- HWP is assessed in control subjects as a quality control tool (Graffelman and Camarena 2008; Gomes et al. 1999, Tapper et al. 2005, Hosking et al. 2004).
- Typically, SNPs that are not in HWP in controls are removed for genetic association studies.

# Cases and HWP

- Departure of genotypic frequencies from HWP in cases may provide additional evidence of association between a genetic marker and disease (Feder et al. 1996; Nielsen, Ehm, and Weir 1998; Jiang et al. 2001; Czika and Weir 2004; Wittke-Thompson, Pluzhnikov, and Cox 2005).
- If the SNP is causal or in LD with causal mutation, it is likely to show departure from HWP.
- We used exact test to test for the HWP in cases.

# Linkage Disequilibrium

- Linkage Disequilibrium (LD) is an association (correlation) between the genotypes at two or more loci.

- Disease phenotype and marker genotype(s) association is found due to proximity of putative disease locus and the marker locus.

# Why perform an association study?

- Locate causal variants in the genome.
- Estimate attributable risk due to causal variants.
- To predict clinical outcomes using associated variant → prediction, treatment response

# Case-Control Association study

- Traditionally, regression (GLM) based approaches are used to assess genetic association between SNPs and disease.
- Logarithm of odds is modeled as linear function of predictor variables. A likelihood ratio test can be performed to assess significance of beta coefficient.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$$

# Case-Control Association study

- We propose combining these two test statistics: logistic regression association test and test for departure from Hardy-Weinberg proportions.
- Both tests provide information about association between SNPs and disease.
- These two tests use different aspects of datasets.
- The two tests are statistically correlated.
- Only cases are used in HWP proportion test.
- Both cases and controls are used in logistic regression.

# Tail Strength Measure

- Consider *m* null hypotheses and let the associated p-values be $p_i$, i=1,…,*m*
- The global hypothesis is that all the individual hypothesis hold simultaneously
- Let $p_{(1)}<p_{(2)}<\ldots<p_{(m)}$ be the ordered p-values. Then, the tail strength measure (Taylor and Tibshirani, 2006) is defined as

$$TS(p_1, p_2, \ldots, p_m) = \frac{1}{m}\sum_{i=1}^{m}\left(1 - p_{(i)}\frac{m+1}{i}\right)$$

# Tail Strength Measure

- Note that under null hypothesis (global), each $p_i$ is uniformly distributed so that the ordered p-values follows a beta distribution with mean *i/(m+1).*
- Hence, TS has expected value zero under the null hypothesis.
- TS is closely related to FDR approach to multiple hypothesis testing. Using this property, they also derived asymptotic distribution of TS when m (the number of hypotheses) are large.
- TS is also closely related to area under ROC.

# Tail Strength Measure

- TS calculates the linear combination of the difference between ordered p-value and its expected value.
- It gives more weight to the smaller p-values so that it is more sensitive to deviations in the tail.
- When TS approaches one, it implies smaller p-values than one would expect by chance which indicates evidence against global null hypothesis.
- TS would be more powerful than each of its component tests.

# Hypothesis for HWP and Logistic Regression

- ➢ We are interested in two hypothesis
- $H_{o1}$ = There is no association between SNP and disease
- $H_{02}$ = SNP genotypes are in Hardy-Weinberg proportions

Let $p_1$ be the p-value obtained for testing $H_{o1}$ based on logistic regression (we used likelihood ratio test)

Let $p_2$ be the p-value obtained for testing $H_{o2}$ (based on Exact HWP test)

# Tail Strength Measure

- We can not use the asymptotic distribution as we have only two hypotheses.
- Let $p_{(1)}$ and $p_{(2)}$ be ordered p-values. The tail strength measure is

$$TS(p_1, p_2) = \frac{1}{2}\left(\left(1 - p_{(1)} \times 3\right) + \left(1 - p_{(2)} \times \frac{3}{2}\right)\right)$$

- The range of TS is [-1.25, 1] because $0 < p_{(1)} < p_{(2)} < 1$

# Tail Strength Measure

- Because $p_{(1)}$ and $p_{(2)}$ follow beta distribution, using a bivariate transformation we can derive explicit formula for probability density function of TS

$$f_{TS}(x) = \begin{cases} \frac{8}{27}\left(\frac{5}{2} + 2x\right), & if\ x \in [-1.25, 0.25], \\ \frac{32}{27}(1 - x), & if\ x \in (0.25, 1.00]. \end{cases}$$

- For observed value $TS^*$ , the exact p-value formula is

$$p - value = P(TS > TS^*) = \int_{TS^*}^{1} f_{TS}(x)dx.$$

# Tail Strength Measure

- In TS, p-values are compared to the expected p-value.
- In many situations, median-based estimators are more robust to extreme observations.
- Because we are concerned with small p-values, median-based tail strength measure may be more robust.
- Therefore, we developed a tail strength median measure, TSM.

# Tail Strength Median Measure

- In the TSM. Linear combination of differences between p-values and corresponding median values under null hypothesis are considered.
- The median values of $p_{(1)}$ and $p_{(2)}$ are 1-1/sqrt(2) and 1/sqrt(2), respectively. Therefore, TSM is

$$TSM(p_1, p_2) = \frac{1}{2}\left(\left(1 - p_{(1)} \times \frac{\sqrt{2}}{\sqrt{2}-1}\right) + \left(1 - p_{(2)} \times \sqrt{2}\right)\right)$$

# Tail Strength Median Measure

- The probability density function of TSM is

$$g_{TSM}(x) = \begin{cases} \dfrac{2\sqrt{2}(\sqrt{2}-1)}{\sqrt{2}+1}\left(\sqrt{2}+x\right), & if\ x \in [-\sqrt{2}, 1-\dfrac{1}{\sqrt{2}}], \\ \dfrac{2\sqrt{2}}{\sqrt{2}+1}(1-x), & if\ x \in (1-\dfrac{1}{\sqrt{2}}, 1.00]. \end{cases}$$

- For observed value TSM$^*$ , the exact p-value formula is

$$p-value = P(TSM > TSM^*) = \int_{TSM^*}^{1} g_{TSM}(x)dx.$$

# Tail Strength Median Measure

- Because the joint distribution of $p_{(1)}$ and $p_{(2)}$ is not symmetric, it may be more appropriate to use TSM.
- Compared to TS, TSM assigns more weight to the smaller p-value and less weight to the larger p-value.
- TSM also has similar relationship to FDR, if one uses median values instead of mean values in FDR.

# Permutation Approach for p-value

- The exact p-values of tail strength measure and tail strength median measure are simple and straightforward to compute.
- Deviations of underlying assumptions might lead to either conservative or liberal the p-values based on the explicit formulas.
- Therefore, we devised a permutation based test to assess significance of TS and TSM.

# Permutation Approach for p-value

- For each permutation step, we resample the SNP values by using genotype frequencies of the entire data set (cases and controls) but keep all other covariate values unchanged.
- By re-sampling SNP values from both cases and controls, there is no association between SNP and disease status.
- For each permutation step, we calculate TS and TSM and p-value is defined as proportion of TS or TSM values that are greater than observed TS or TSM.

# Simulation Study: Model 1

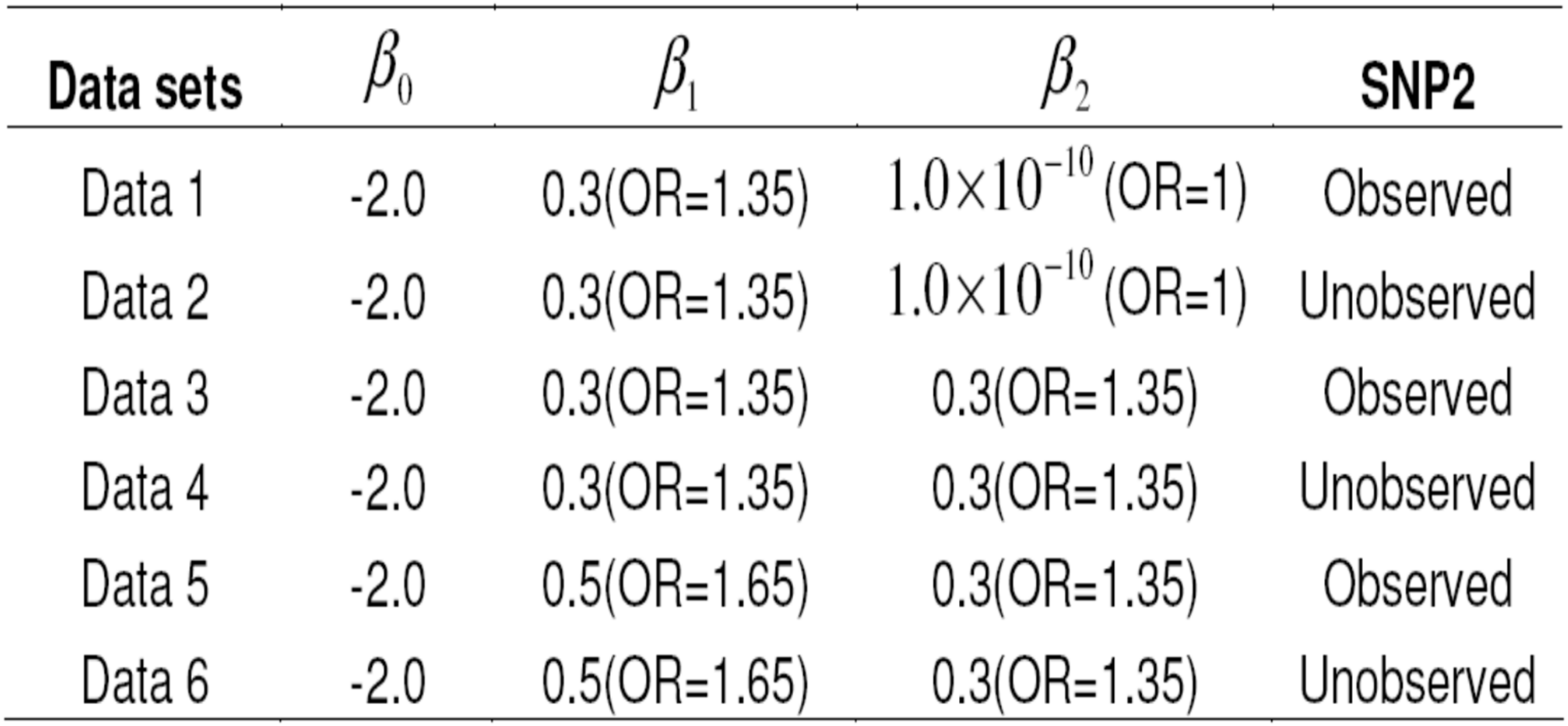
- The two SNPs, $X_1$ and $X_{2,}$ were simulated under the assumption of HWP in the population.
- Minor allele frequencies for $X_1$ and $X_{2,}$ were 10% and 40%, respectively.
- Given the values of SNPs $X_1$ and $X_2$ (coded as additive model), the disease status was simulated using

$$\text{Logit }(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

# Simulation Study: Model 1

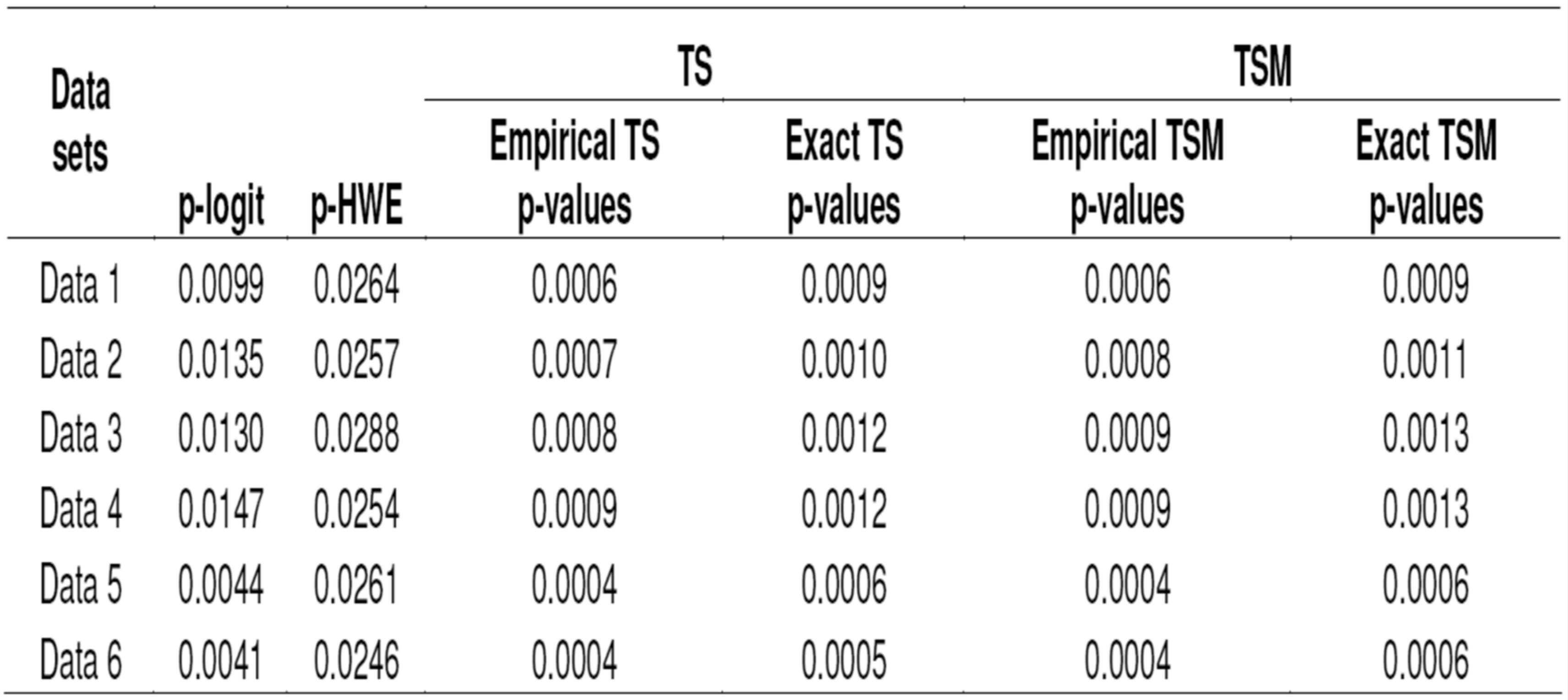
- We simulated 500 cases and 500 controls.
- Even though, we simulated SNPs with HWP in the original population, in case population genotypes may not be in HWP.
- We simulated six scenarios: Different odds ratios for SNPs and whether or not second SNP is observed.

# Simulation Study: Model 1



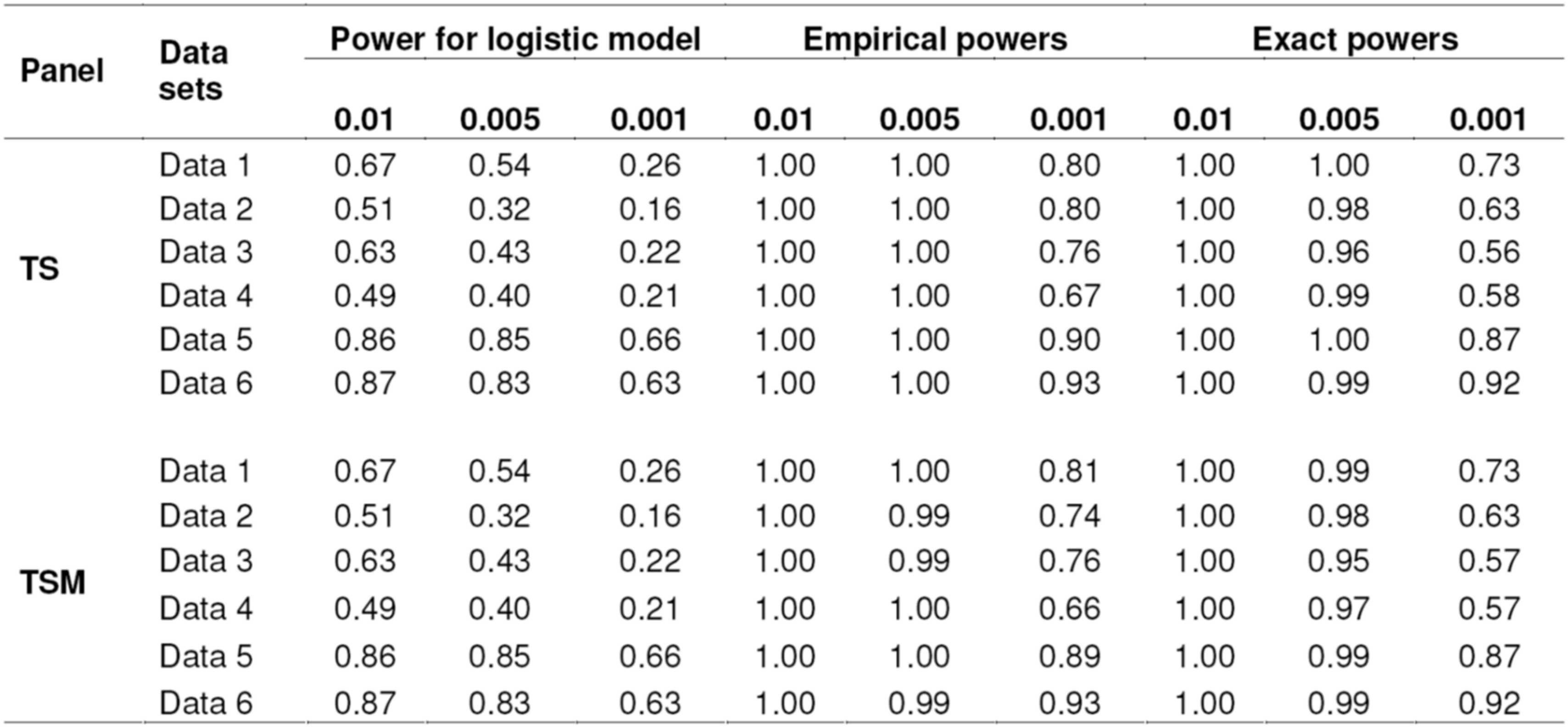
| Data sets | $\beta_0$ | $\beta_1$ | $\beta_2$ | SNP2 |
|---|---|---|---|---|
| Data 1 | -2.0 | 0.3(OR=1.35) | $1.0\times10^{-10}$ (OR=1) | Observed |
| Data 2 | -2.0 | 0.3(OR=1.35) | $1.0\times10^{-10}$ (OR=1) | Unobserved |
| Data 3 | -2.0 | 0.3(OR=1.35) | 0.3(OR=1.35) | Observed |
| Data 4 | -2.0 | 0.3(OR=1.35) | 0.3(OR=1.35) | Unobserved |
| Data 5 | -2.0 | 0.5(OR=1.65) | 0.3(OR=1.35) | Observed |
| Data 6 | -2.0 | 0.5(OR=1.65) | 0.3(OR=1.35) | Unobserved |

# Results (Model 1):Average p-values based on 100 replicates



| Data sets | p-logit | p-HWE | TS | | TSM | |
|---|---|---|---|---|---|---|
| | | | Empirical TS p-values | Exact TS p-values | Empirical TSM p-values | Exact TSM p-values |
| Data 1 | 0.0099 | 0.0264 | 0.0006 | 0.0009 | 0.0006 | 0.0009 |
| Data 2 | 0.0135 | 0.0257 | 0.0007 | 0.0010 | 0.0008 | 0.0011 |
| Data 3 | 0.0130 | 0.0288 | 0.0008 | 0.0012 | 0.0009 | 0.0013 |
| Data 4 | 0.0147 | 0.0254 | 0.0009 | 0.0012 | 0.0009 | 0.0013 |
| Data 5 | 0.0044 | 0.0261 | 0.0004 | 0.0006 | 0.0004 | 0.0006 |
| Data 6 | 0.0041 | 0.0246 | 0.0004 | 0.0005 | 0.0004 | 0.0006 |

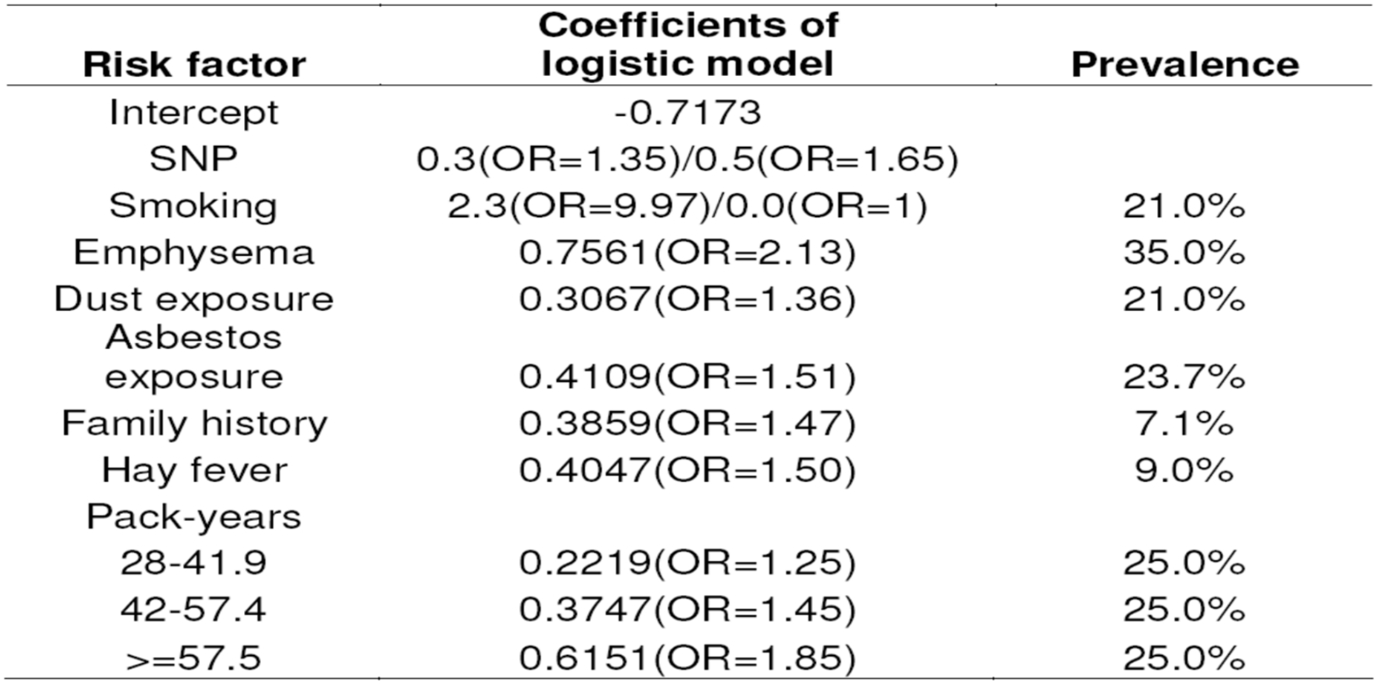
# Results (Model 1):Power comparison based on 100 replicates



| Panel | Data sets | Power for logistic model | | | Empirical powers | | | Exact powers | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.01 | 0.005 | 0.001 | 0.01 | 0.005 | 0.001 | 0.01 | 0.005 | 0.001 |
| TS | Data 1 | 0.67 | 0.54 | 0.26 | 1.00 | 1.00 | 0.80 | 1.00 | 1.00 | 0.73 |
| | Data 2 | 0.51 | 0.32 | 0.16 | 1.00 | 1.00 | 0.80 | 1.00 | 0.98 | 0.63 |
| | Data 3 | 0.63 | 0.43 | 0.22 | 1.00 | 1.00 | 0.76 | 1.00 | 0.96 | 0.56 |
| | Data 4 | 0.49 | 0.40 | 0.21 | 1.00 | 1.00 | 0.67 | 1.00 | 0.99 | 0.58 |
| | Data 5 | 0.86 | 0.85 | 0.66 | 1.00 | 1.00 | 0.90 | 1.00 | 1.00 | 0.87 |
| | Data 6 | 0.87 | 0.83 | 0.63 | 1.00 | 1.00 | 0.93 | 1.00 | 0.99 | 0.92 |
| TSM | Data 1 | 0.67 | 0.54 | 0.26 | 1.00 | 1.00 | 0.81 | 1.00 | 0.99 | 0.73 |
| | Data 2 | 0.51 | 0.32 | 0.16 | 1.00 | 0.99 | 0.74 | 1.00 | 0.98 | 0.63 |
| | Data 3 | 0.63 | 0.43 | 0.22 | 1.00 | 0.99 | 0.76 | 1.00 | 0.95 | 0.57 |
| | Data 4 | 0.49 | 0.40 | 0.21 | 1.00 | 1.00 | 0.66 | 1.00 | 0.97 | 0.57 |
| | Data 5 | 0.86 | 0.85 | 0.66 | 1.00 | 1.00 | 0.89 | 1.00 | 0.99 | 0.87 |
| | Data 6 | 0.87 | 0.83 | 0.63 | 1.00 | 0.99 | 0.93 | 1.00 | 0.99 | 0.92 |

# Simulation Study: Model 2

- We simulated data from a lung cancer model (Spitz et al. 2007).

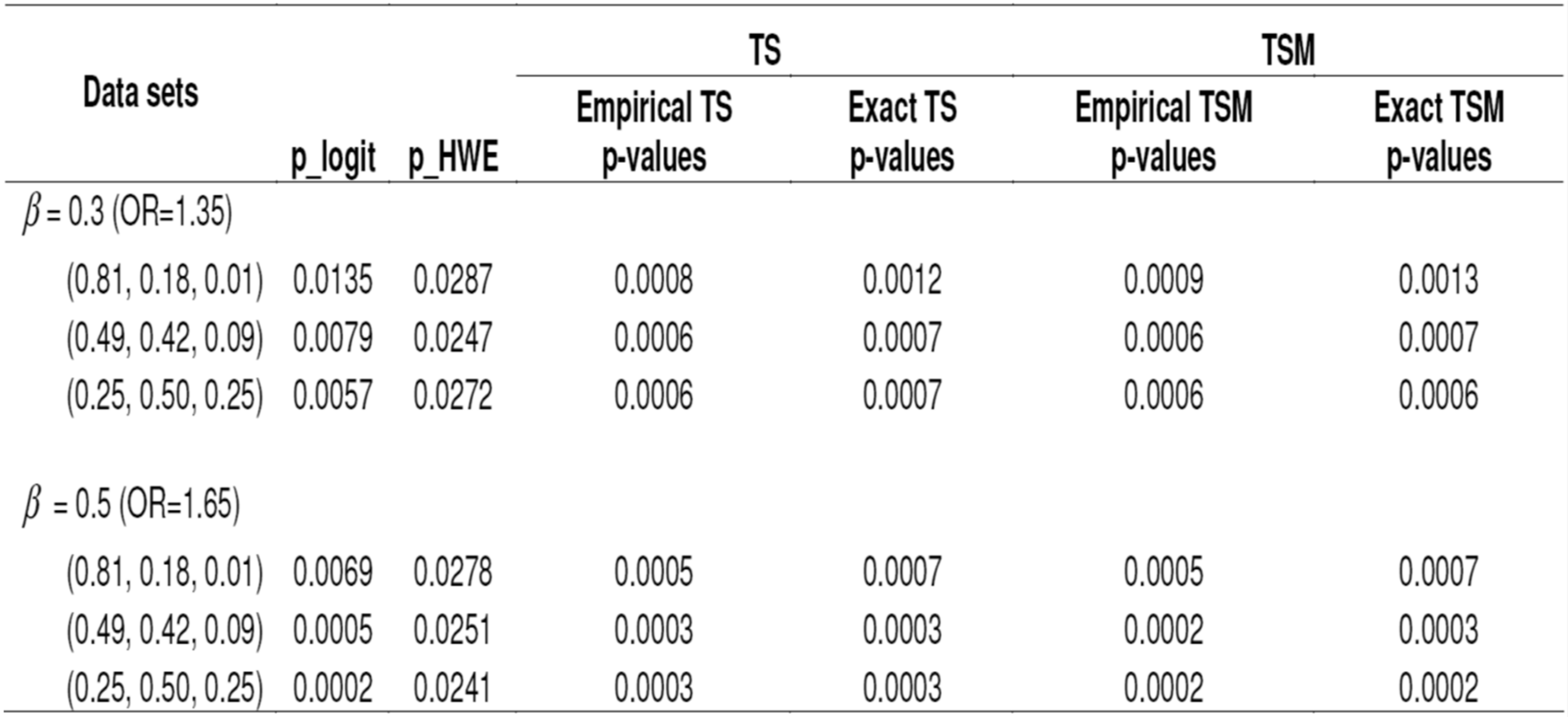


| **Risk factor** | **Coefficients of logistic model** | **Prevalence** |
|---|---|---|
| Intercept | -0.7173 | |
| SNP | 0.3(OR=1.35)/0.5(OR=1.65) | |
| Smoking | 2.3(OR=9.97)/0.0(OR=1) | 21.0% |
| Emphysema | 0.7561(OR=2.13) | 35.0% |
| Dust exposure | 0.3067(OR=1.36) | 21.0% |
| Asbestos exposure | 0.4109(OR=1.51) | 23.7% |
| Family history | 0.3859(OR=1.47) | 7.1% |
| Hay fever | 0.4047(OR=1.50) | 9.0% |
| Pack-years | | |
| 28-41.9 | 0.2219(OR=1.25) | 25.0% |
| 42-57.4 | 0.3747(OR=1.45) | 25.0% |
| >=57.5 | 0.6151(OR=1.85) | 25.0% |

# Simulation Study: Model 2

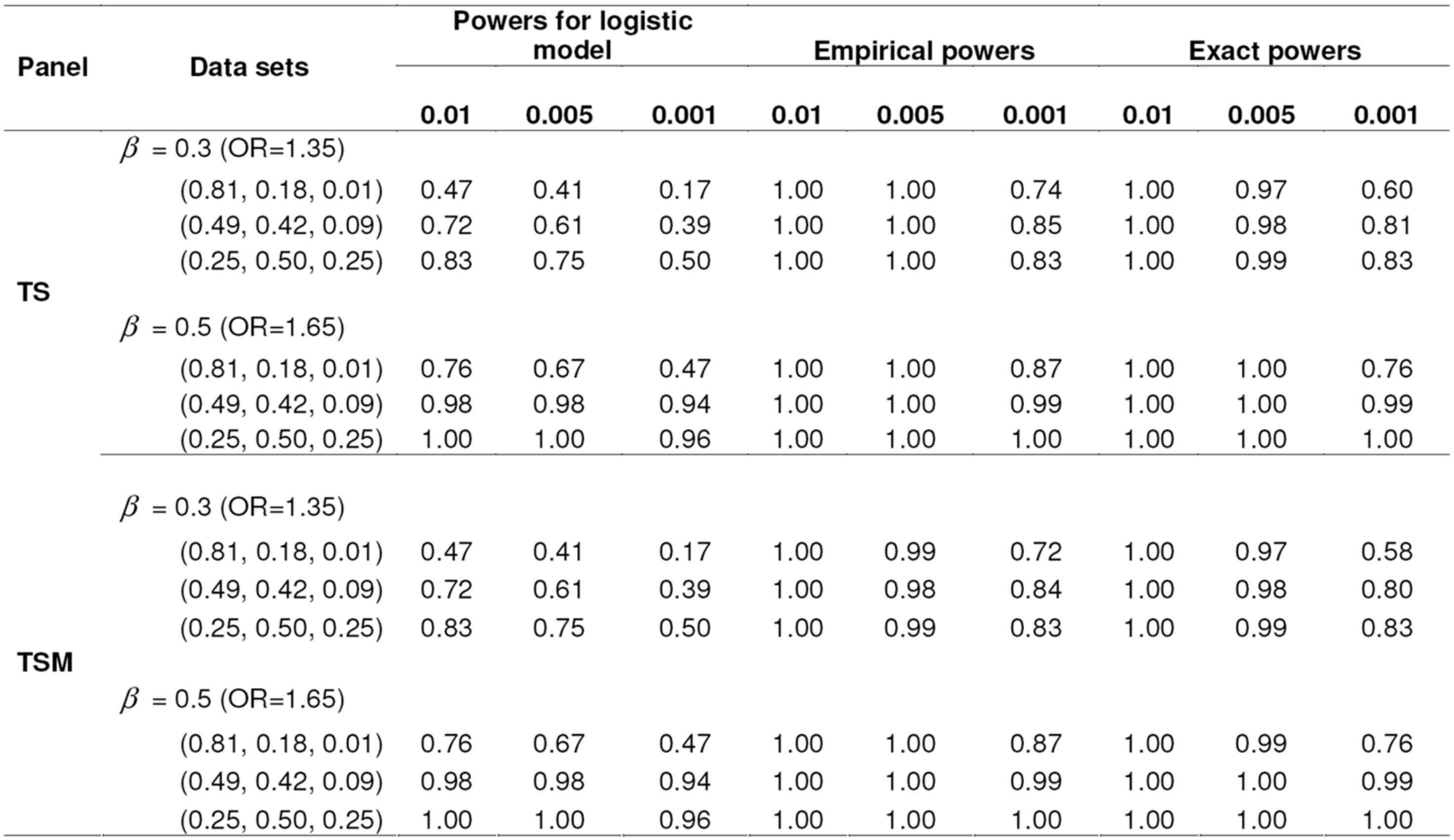
- We simulated two models: general lung cancer model and lung cancer model for current smokers
- All the odds ratios used for simulation are from Spitz et al. paper.
- For current smoking model, cigarette smoking odds ratio was 1.0.
- Prevalence of risk factors was used from various published papers.
- 500 cases and 500 controls were simulated.

# Results (General): Average p-values based on 100 replicates



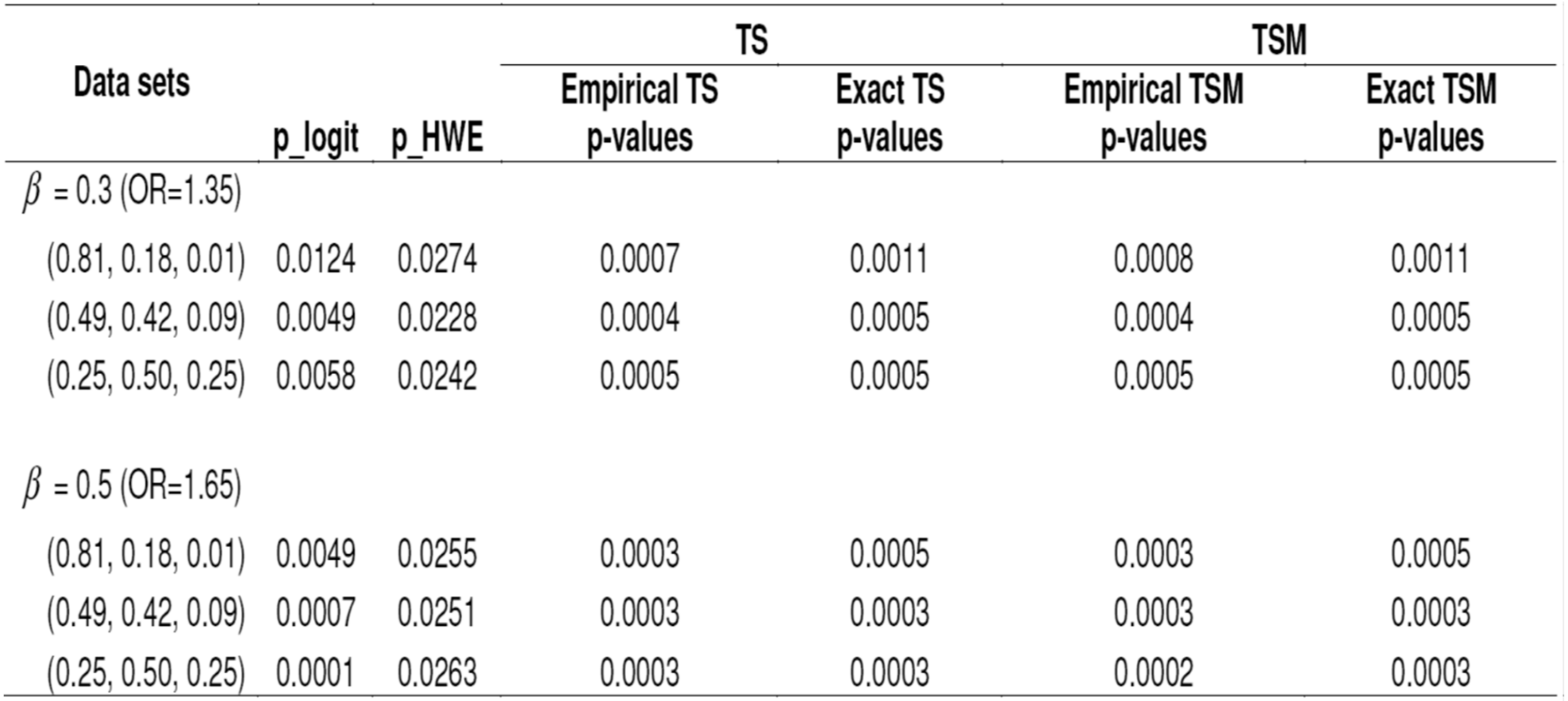
| Data sets | | | TS | | TSM | |
|---|---|---|---|---|---|---|
| | p_logit | p_HWE | Empirical TS p-values | Exact TS p-values | Empirical TSM p-values | Exact TSM p-values |
| $\beta$ = 0.3 (OR=1.35) | | | | | | |
| (0.81, 0.18, 0.01) | 0.0135 | 0.0287 | 0.0008 | 0.0012 | 0.0009 | 0.0013 |
| (0.49, 0.42, 0.09) | 0.0079 | 0.0247 | 0.0006 | 0.0007 | 0.0006 | 0.0007 |
| (0.25, 0.50, 0.25) | 0.0057 | 0.0272 | 0.0006 | 0.0007 | 0.0006 | 0.0006 |
| $\beta$ = 0.5 (OR=1.65) | | | | | | |
| (0.81, 0.18, 0.01) | 0.0069 | 0.0278 | 0.0005 | 0.0007 | 0.0005 | 0.0007 |
| (0.49, 0.42, 0.09) | 0.0005 | 0.0251 | 0.0003 | 0.0003 | 0.0002 | 0.0003 |
| (0.25, 0.50, 0.25) | 0.0002 | 0.0241 | 0.0003 | 0.0003 | 0.0002 | 0.0002 |

# Results (General): Power comparison based on 100 replicates



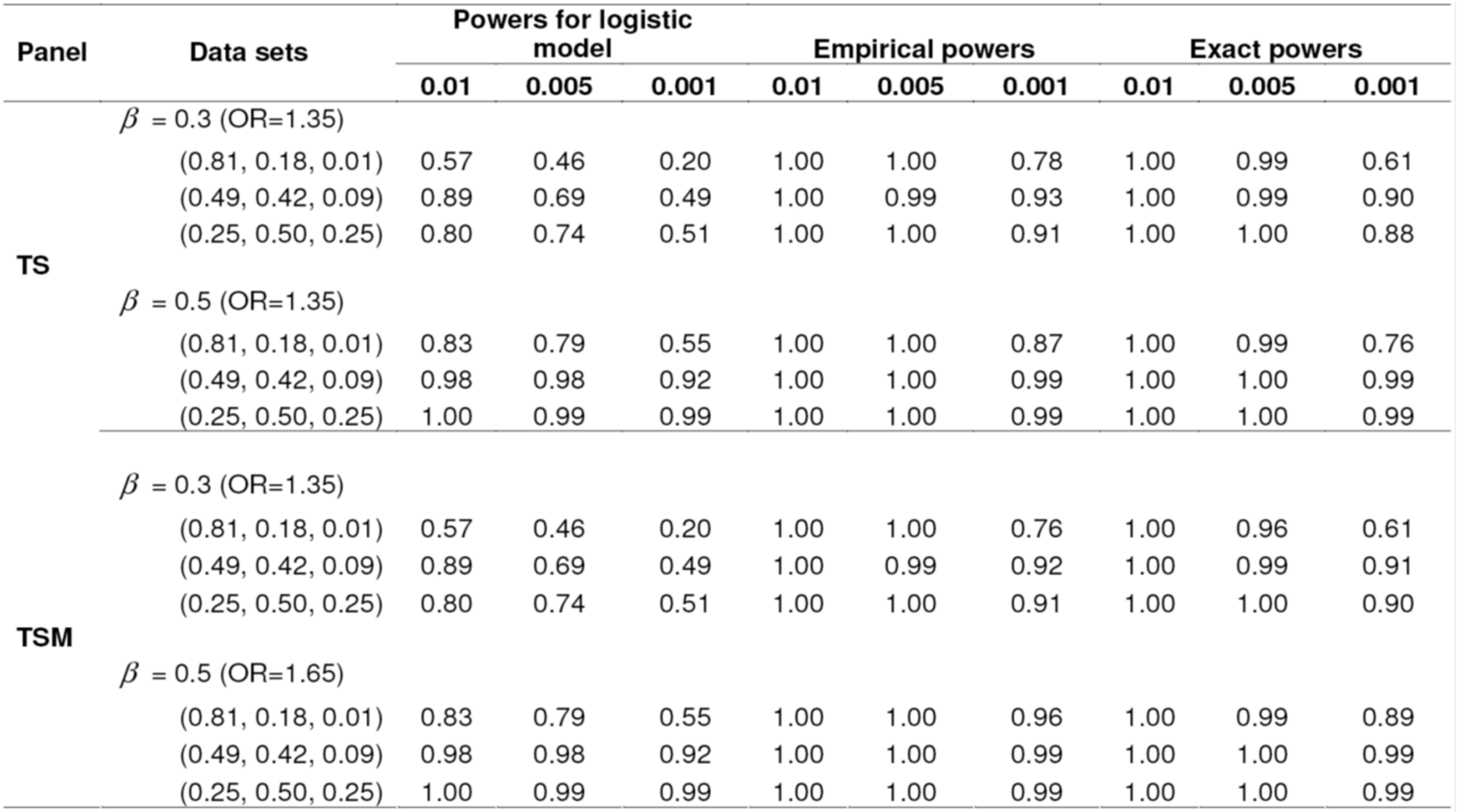
| Panel | Data sets | Powers for logistic model | | | Empirical powers | | | Exact powers | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.01 | 0.005 | 0.001 | 0.01 | 0.005 | 0.001 | 0.01 | 0.005 | 0.001 |
| TS | $\beta$ = 0.3 (OR=1.35) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.47 | 0.41 | 0.17 | 1.00 | 1.00 | 0.74 | 1.00 | 0.97 | 0.60 |
| | (0.49, 0.42, 0.09) | 0.72 | 0.61 | 0.39 | 1.00 | 1.00 | 0.85 | 1.00 | 0.98 | 0.81 |
| | (0.25, 0.50, 0.25) | 0.83 | 0.75 | 0.50 | 1.00 | 1.00 | 0.83 | 1.00 | 0.99 | 0.83 |
| | $\beta$ = 0.5 (OR=1.65) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.76 | 0.67 | 0.47 | 1.00 | 1.00 | 0.87 | 1.00 | 1.00 | 0.76 |
| | (0.49, 0.42, 0.09) | 0.98 | 0.98 | 0.94 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 |
| | (0.25, 0.50, 0.25) | 1.00 | 1.00 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| TSM | $\beta$ = 0.3 (OR=1.35) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.47 | 0.41 | 0.17 | 1.00 | 0.99 | 0.72 | 1.00 | 0.97 | 0.58 |
| | (0.49, 0.42, 0.09) | 0.72 | 0.61 | 0.39 | 1.00 | 0.98 | 0.84 | 1.00 | 0.98 | 0.80 |
| | (0.25, 0.50, 0.25) | 0.83 | 0.75 | 0.50 | 1.00 | 0.99 | 0.83 | 1.00 | 0.99 | 0.83 |
| | $\beta$ = 0.5 (OR=1.65) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.76 | 0.67 | 0.47 | 1.00 | 1.00 | 0.87 | 1.00 | 0.99 | 0.76 |
| | (0.49, 0.42, 0.09) | 0.98 | 0.98 | 0.94 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 |
| | (0.25, 0.50, 0.25) | 1.00 | 1.00 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

# Results (Smokers): Average p-values based on 100 replicates



| Data sets | p_logit | p_HWE | TS<br>Empirical TS p-values | TS<br>Exact TS p-values | TSM<br>Empirical TSM p-values | TSM<br>Exact TSM p-values |
|---|---|---|---|---|---|---|
| $\beta$ = 0.3 (OR=1.35) | | | | | | |
| (0.81, 0.18, 0.01) | 0.0124 | 0.0274 | 0.0007 | 0.0011 | 0.0008 | 0.0011 |
| (0.49, 0.42, 0.09) | 0.0049 | 0.0228 | 0.0004 | 0.0005 | 0.0004 | 0.0005 |
| (0.25, 0.50, 0.25) | 0.0058 | 0.0242 | 0.0005 | 0.0005 | 0.0005 | 0.0005 |
| $\beta$ = 0.5 (OR=1.65) | | | | | | |
| (0.81, 0.18, 0.01) | 0.0049 | 0.0255 | 0.0003 | 0.0005 | 0.0003 | 0.0005 |
| (0.49, 0.42, 0.09) | 0.0007 | 0.0251 | 0.0003 | 0.0003 | 0.0003 | 0.0003 |
| (0.25, 0.50, 0.25) | 0.0001 | 0.0263 | 0.0003 | 0.0003 | 0.0002 | 0.0003 |

# Results (Smokers): Power comparison based on 100 replicates

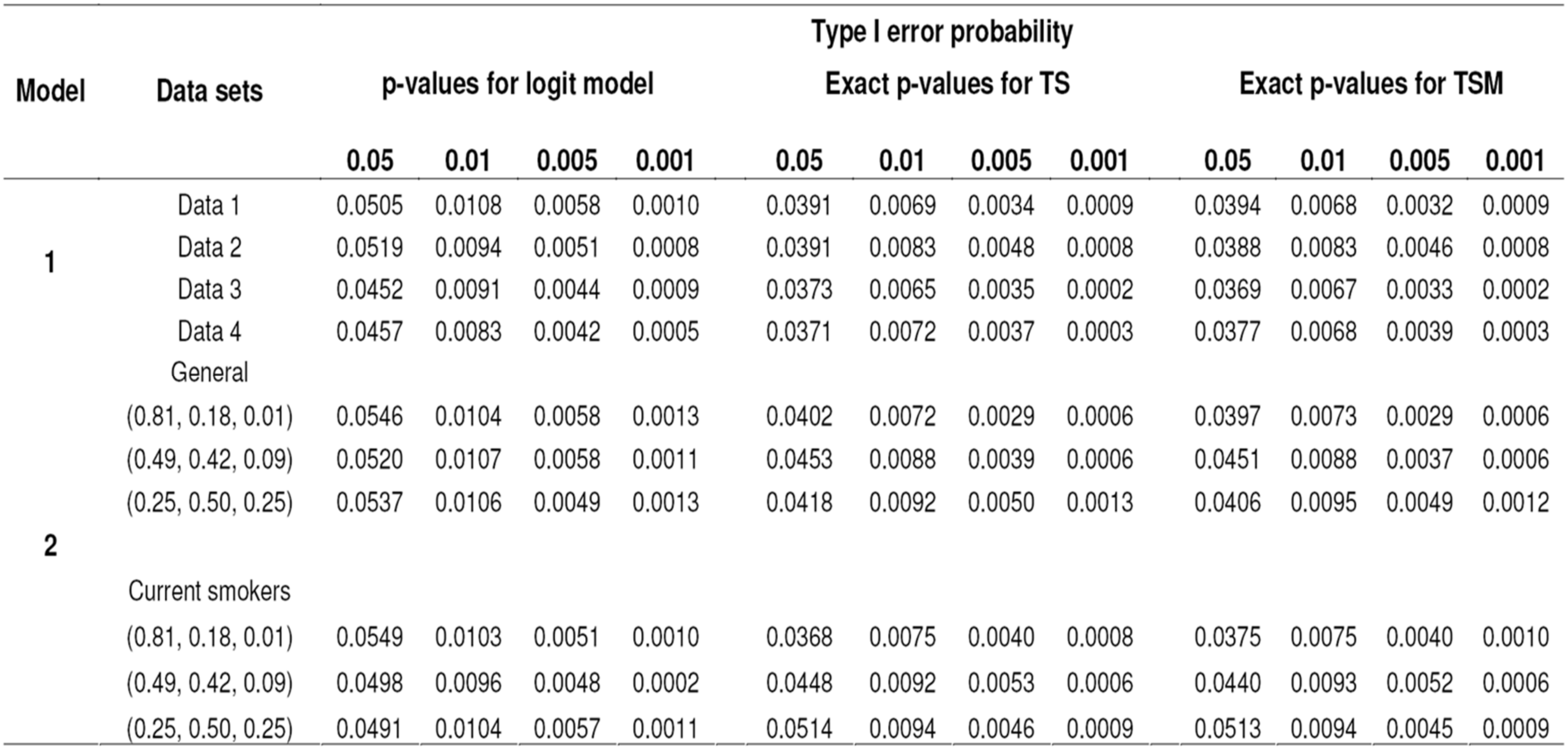


| Panel | Data sets | Powers for logistic model | | | Empirical powers | | | Exact powers | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.01 | 0.005 | 0.001 | 0.01 | 0.005 | 0.001 | 0.01 | 0.005 | 0.001 |
| TS | $\beta$ = 0.3 (OR=1.35) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.57 | 0.46 | 0.20 | 1.00 | 1.00 | 0.78 | 1.00 | 0.99 | 0.61 |
| | (0.49, 0.42, 0.09) | 0.89 | 0.69 | 0.49 | 1.00 | 0.99 | 0.93 | 1.00 | 0.99 | 0.90 |
| | (0.25, 0.50, 0.25) | 0.80 | 0.74 | 0.51 | 1.00 | 1.00 | 0.91 | 1.00 | 1.00 | 0.88 |
| | $\beta$ = 0.5 (OR=1.35) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.83 | 0.79 | 0.55 | 1.00 | 1.00 | 0.87 | 1.00 | 0.99 | 0.76 |
| | (0.49, 0.42, 0.09) | 0.98 | 0.98 | 0.92 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 |
| | (0.25, 0.50, 0.25) | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 |
| TSM | $\beta$ = 0.3 (OR=1.35) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.57 | 0.46 | 0.20 | 1.00 | 1.00 | 0.76 | 1.00 | 0.96 | 0.61 |
| | (0.49, 0.42, 0.09) | 0.89 | 0.69 | 0.49 | 1.00 | 0.99 | 0.92 | 1.00 | 0.99 | 0.91 |
| | (0.25, 0.50, 0.25) | 0.80 | 0.74 | 0.51 | 1.00 | 1.00 | 0.91 | 1.00 | 1.00 | 0.90 |
| | $\beta$ = 0.5 (OR=1.65) | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.83 | 0.79 | 0.55 | 1.00 | 1.00 | 0.96 | 1.00 | 0.99 | 0.89 |
| | (0.49, 0.42, 0.09) | 0.98 | 0.98 | 0.92 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 |
| | (0.25, 0.50, 0.25) | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 |

# Simulation Study: Type 1 error

- We simulated data sets under the null hypothesis that SNP is not associated with the disease.
- The simulation model for this was identical to that for models 1 and 2, except that the beta coefficient for the SNP was zero (OR=1).
- We simulated 10,000 replicates, each with 500 cases and 500 controls.

# Results (type 1 error): Average p-values based on 100 replicates



| Model | Data sets | Type I error probability | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | p-values for logit model | | | | Exact p-values for TS | | | | Exact p-values for TSM | | | |
| | | 0.05 | 0.01 | 0.005 | 0.001 | 0.05 | 0.01 | 0.005 | 0.001 | 0.05 | 0.01 | 0.005 | 0.001 |
| 1 | Data 1 | 0.0505 | 0.0108 | 0.0058 | 0.0010 | 0.0391 | 0.0069 | 0.0034 | 0.0009 | 0.0394 | 0.0068 | 0.0032 | 0.0009 |
| | Data 2 | 0.0519 | 0.0094 | 0.0051 | 0.0008 | 0.0391 | 0.0083 | 0.0048 | 0.0008 | 0.0388 | 0.0083 | 0.0046 | 0.0008 |
| | Data 3 | 0.0452 | 0.0091 | 0.0044 | 0.0009 | 0.0373 | 0.0065 | 0.0035 | 0.0002 | 0.0369 | 0.0067 | 0.0033 | 0.0002 |
| | Data 4 | 0.0457 | 0.0083 | 0.0042 | 0.0005 | 0.0371 | 0.0072 | 0.0037 | 0.0003 | 0.0377 | 0.0068 | 0.0039 | 0.0003 |
| 2 | General | | | | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.0546 | 0.0104 | 0.0058 | 0.0013 | 0.0402 | 0.0072 | 0.0029 | 0.0006 | 0.0397 | 0.0073 | 0.0029 | 0.0006 |
| | (0.49, 0.42, 0.09) | 0.0520 | 0.0107 | 0.0058 | 0.0011 | 0.0453 | 0.0088 | 0.0039 | 0.0006 | 0.0451 | 0.0088 | 0.0037 | 0.0006 |
| | (0.25, 0.50, 0.25) | 0.0537 | 0.0106 | 0.0049 | 0.0013 | 0.0418 | 0.0092 | 0.0050 | 0.0013 | 0.0406 | 0.0095 | 0.0049 | 0.0012 |
| | Current smokers | | | | | | | | | | | | |
| | (0.81, 0.18, 0.01) | 0.0549 | 0.0103 | 0.0051 | 0.0010 | 0.0368 | 0.0075 | 0.0040 | 0.0008 | 0.0375 | 0.0075 | 0.0040 | 0.0010 |
| | (0.49, 0.42, 0.09) | 0.0498 | 0.0096 | 0.0048 | 0.0002 | 0.0448 | 0.0092 | 0.0053 | 0.0006 | 0.0440 | 0.0093 | 0.0052 | 0.0006 |
| | (0.25, 0.50, 0.25) | 0.0491 | 0.0104 | 0.0057 | 0.0011 | 0.0514 | 0.0094 | 0.0046 | 0.0009 | 0.0513 | 0.0094 | 0.0045 | 0.0009 |

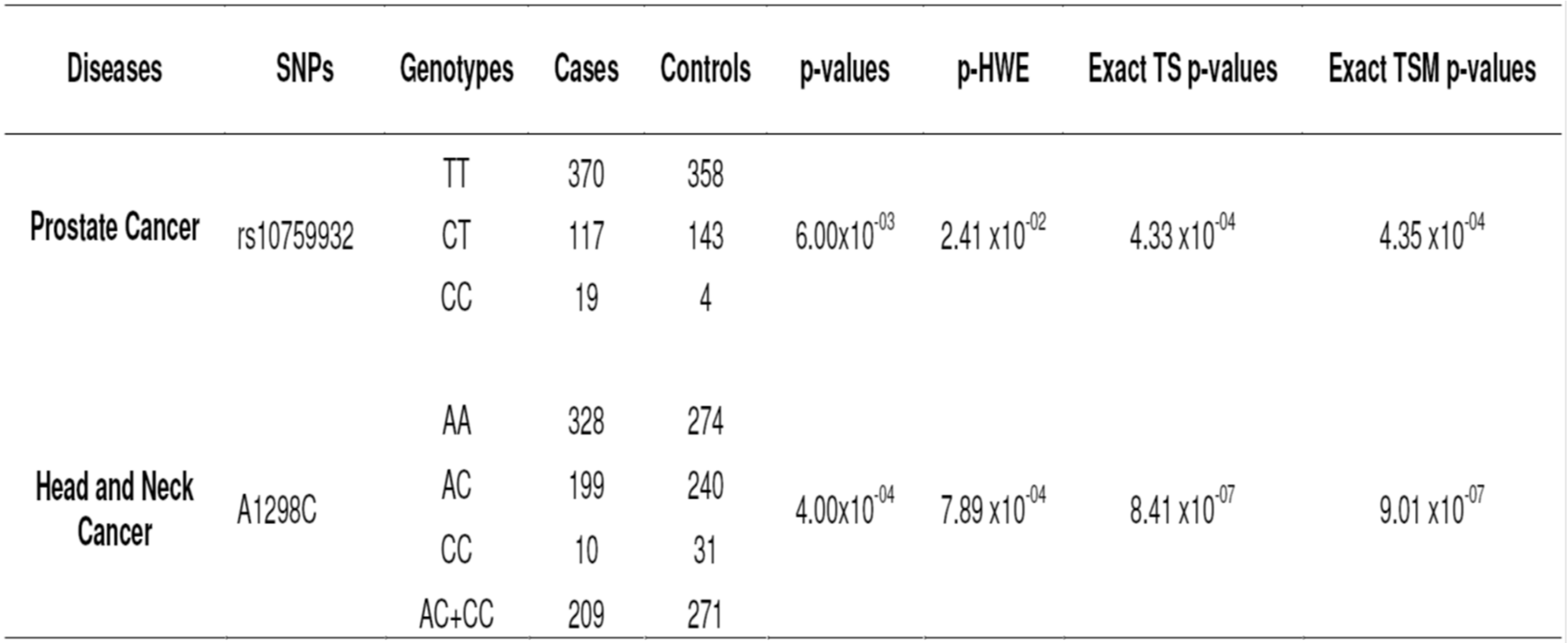
# Real Data Application

- We applied the approach to two real data sets.
- Prostate Cancer (Cheng et al. 2007): These authors investigated role of toll-like receptor 4 in prostate cancer.
- Sample size is 506 cases and 506 controls. We used the SNP, rs10759932, which is most significantly associated risk factor with the disease.
- P-value from regression approach was used from paper. We calculated exact HWP p-value and TS and TSM p-values.

# Real Data Application

- Head and Neck Cancer (Neumann et al. 2005): We investigated role of methylenetetrahydrofolate reductase (MTHFR) 1298AC/CC genotypes with H and N cancer.
- Sample size is 537 cases and 545 controls. We used the SNP, MTHFR A2198C, which is most significantly associated protective factor with the disease.
- P-value from regression approach was used from paper. We calculated exact HWP p-value and TS and TSM p-values.

# Real Data Application



| Diseases | SNPs | Genotypes | Cases | Controls | p-values | p-HWE | Exact TS p-values | Exact TSM p-values |
|---|---|---|---|---|---|---|---|---|
| **Prostate Cancer** | rs10759932 | TT | 370 | 358 | $6.00x10^{-03}$ | $2.41\ x10^{-02}$ | $4.33\ x10^{-04}$ | $4.35\ x10^{-04}$ |
| | | CT | 117 | 143 | | | | |
| | | CC | 19 | 4 | | | | |
| **Head and Neck Cancer** | A1298C | AA | 328 | 274 | $4.00x10^{-04}$ | $7.89\ x10^{-04}$ | $8.41\ x10^{-07}$ | $9.01\ x10^{-07}$ |
| | | AC | 199 | 240 | | | | |
| | | CC | 10 | 31 | | | | |
| | | AC+CC | 209 | 271 | | | | |

# Discussion

- We proposed an approach to assess genetic association that can include information about deviation of genotypic frequencies from the expected Hardy-Weinberg proportions in the case population.
- The proposed method is more powerful than the traditional approach.
- The two measure TS and TSM perform approximately similar.

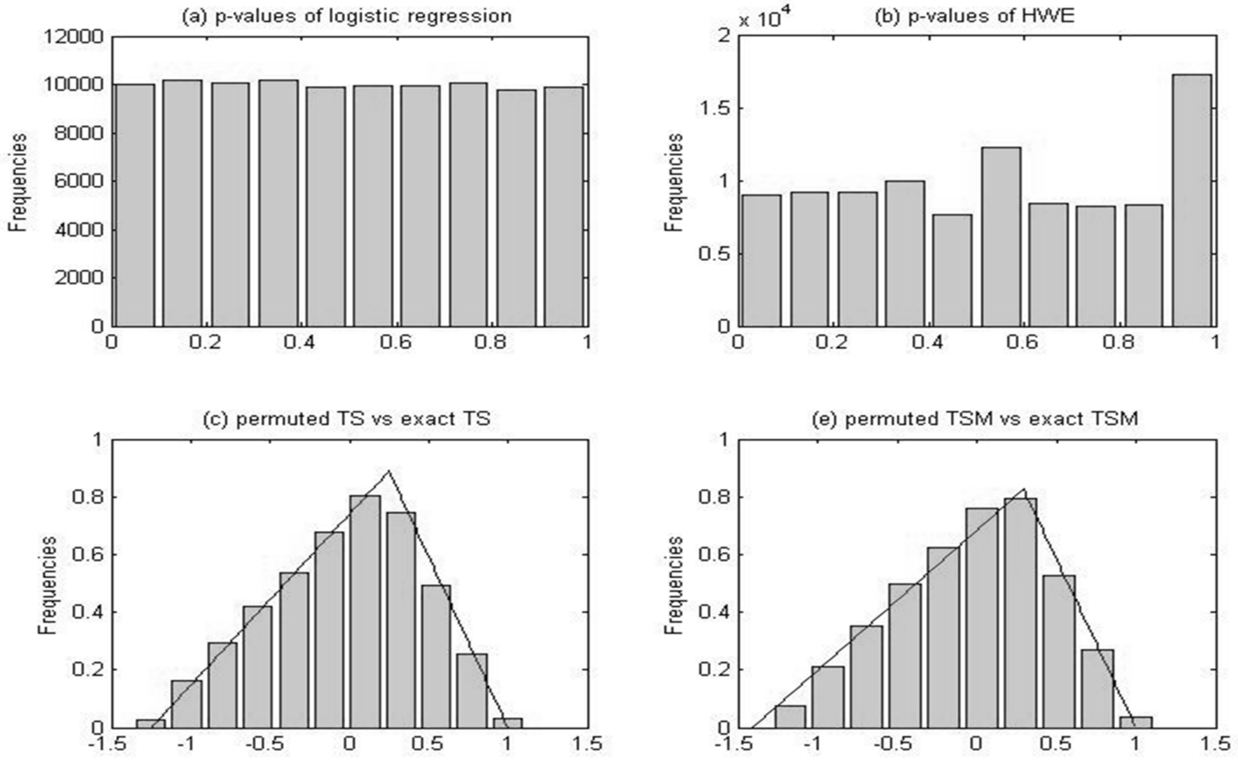
# Discussion

- The genotypes in cases may NOT be in the Hardy-Weinberg proportions because of several reasons such as penetrance of SNP, allele frequency etc.
- The test is not applicable in such situations.
- The genotypes in controls subjects may also NOT be in the Hardy-Weinberg proportions (which could also indicate association).
- Our test can include HWP deviations in controls too.
- Genomewide significance may have very high significance and may not need this approach.

# Permutation Test

- To examine performance of permutation test, we picked one replicate.
- OR = 1.65 and genotype frequencies (.49, .42, .09).
- Permutated logistic p-values, permuted HWP test p-values, permuted TS or TSM Vs exact TS or TSM (from formula).
- P-values are approximately uniformly distributed for logistic and HWP test(?)
- TS and TSM distributions are quite accurate.

# Permutation Test


(a) p-values of logistic regression
Frequencies
(b) p-values of HWE
x 10^4
Frequencies
(c) permuted TS vs exact TS
Frequencies
(e) permuted TSM vs exact TSM
Frequencies