Sequential genome-wide association studies for pharmacovigilance

Patrick Kelly

5th International Conference on Multiple Comparison Procedures

10 July 2007

Medical and Pharmaceutical Statistics Research Unit

The University of Reading, UK

In collaboration with…

John Whitehead
Yinghui (Echo) Zhou
MPS Research Unit, University of Reading

Nigel Stallard
Warwick Medical School, University of Warwick

Clive Bowman
Research Statistics Unit , GSK

Funded by GSK

Outline

1. Pharmacovigilance and genetics
2. Fixed sample size test
3. Sequential testing
4. An example
5. Summary & issues

Pharmacovigilance & genetics

Adverse events (AEs) may prevent regulatory approval

If AEs are limited to some subgroup

drug can be targeted

adverse event rate reduced

drug may be safely used

Pharmacovigilance & genetics

Can we identify whether patients with AEs differ from those that do not have AEs?

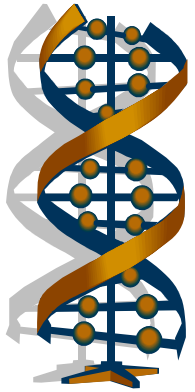
Present

Patient medical history

Future

Large scale genotyping becoming available

e.g. > 500,000 SNP markers across genome



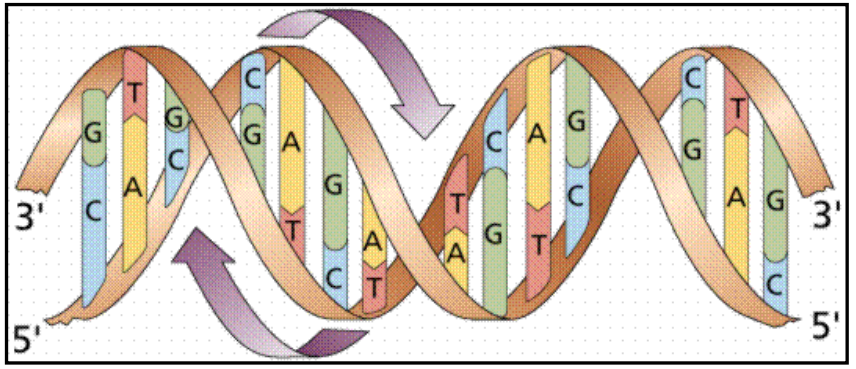
Pharmacovigilance & genetics

SNP (single nucleotide polymorphism)

A single base in the DNA sequence that can be observed to be different between individuals in the population

e.g. **AACGTATGGACCGA**

AACGTATCGACCGA



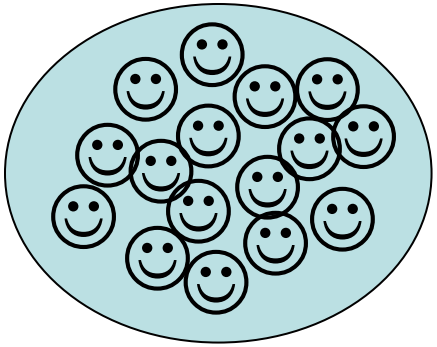
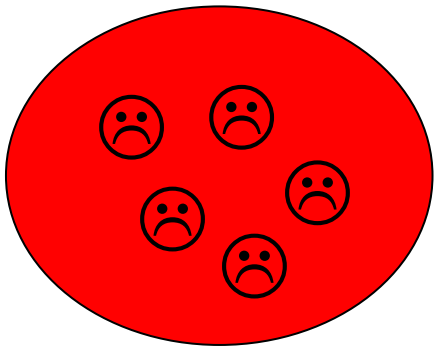
Since chromosomes occur in pairs we have three genotypes

Pharmacovigilance & genetics

Case = patient with AE ☹

Control = patient without AE ☺

Are cases genetically different from controls?

Conduct a test  versus 

Pharmacovigilance & genetics

Cases ☹

Rare
Conduct fixed sample size test ✗
Drug withdrawn due to safety concerns

Aim

Conduct a test as each new case occurs
i.e. a sequential approach

Controls ☺

Patients in same trial may become cases ✗
Want a large number
⇒ Select in advance, i.e. fixed in number

Fixed sample test - one SNP

3 genotypes – AA, AG, GG

$k = 0, 1, 2$

Are the genotype probabilities for cases the same as the controls?

$$H_0: p_{☹0} = p_{☺0},\ p_{☹1} = p_{☺1},\ p_{☹2} = p_{☺2}$$

$p_{☹0}, p_{☹1}, p_{☹2}$ genotype probabilities for controls

$p_{☺0}, p_{☺1}, p_{☺2}$ genotype probabilities for cases

Fixed sample test - one SNP

Can test H_0 using a likelihood ratio test statistic, Λ

$$\Lambda = \frac{\prod_{k=0}^{2} \binom{n_{\text{☺}k} + n_{\text{☹}k}}{n_{\text{☺}k} + n_{\text{☹}k}}}{\prod_{k=0}^{2} \binom{n_{\text{☺}k}}{n_{\text{☺}}}^{n_{\text{☺}}} \prod_{k=0}^{2} \binom{n_{\text{☹}k}}{n_{\text{☹}}}^{n_{\text{☹}}}}$$

Controls $n_{\text{☺}0}$, $n_{\text{☺}1}$, $n_{\text{☺}2}$ $\quad n_{\text{☺}} = n_{\text{☺}0} + n_{\text{☺}1} + n_{\text{☺}2}$

Cases $n_{\text{☹}0}$, $n_{\text{☹}1}$, $n_{\text{☹}2}$ $\quad n_{\text{☹}} = n_{\text{☹}0} + n_{\text{☹}1} + n_{\text{☹}2}$

Under H_0 , $-2 \log \Lambda \sim \chi^2_2$

Fixed sample test – many SNPs

Suppose there are L SNPs

Are the genotype probabilities for cases the same as the controls for **all L SNPs**?

H_0: $p_{☹01} = p_{☺01}, p_{☹11} = p_{☺11}, p_{☹21} = p_{☺21}$; (SNP 1)

$p_{☹02} = p_{☺02}, p_{☹12} = p_{☺12}, p_{☹22} = p_{☺22}$; (SNP 2)

...

$p_{☹0L} = p_{☺0L}, p_{☹1L} = p_{☺1L}, p_{☹2L} = p_{☺2L}$ (SNP L)

Fixed sample test – many SNPs

Likelihood ratio test statistic, Λ, becomes

$$\Lambda = \prod_{j=1}^{L} \left[\frac{\prod_{k=0}^{2} \binom{n_{\text{☺}kj} + n_{\text{☹}kj}}{n_{\text{☺}kj}}^{n_{\text{☺}kj} + n_{\text{☹}kj}}}{\prod_{k=0}^{2} \binom{n_{\text{☺}kj}}{n_{\text{☺}}}^{n_{\text{☺}}} \prod_{k=0}^{2} \binom{n_{\text{☹}kj}}{n_{\text{☹}}}^{n_{\text{☹}kj}}} \right]$$

$n_{\text{☺}0j}$, $n_{\text{☺}1j}$, $n_{\text{☺}2j}$ for controls at SNP j

$n_{\text{☹}0j}$, $n_{\text{☹}1j}$, $n_{\text{☹}2j}$ for cases at SNP j

Fixed sample test – many SNPs

If asymptotics hold and SNPs are independent then

under H_0, $-2 \log \Lambda \sim \chi^2_{2L}$

However!

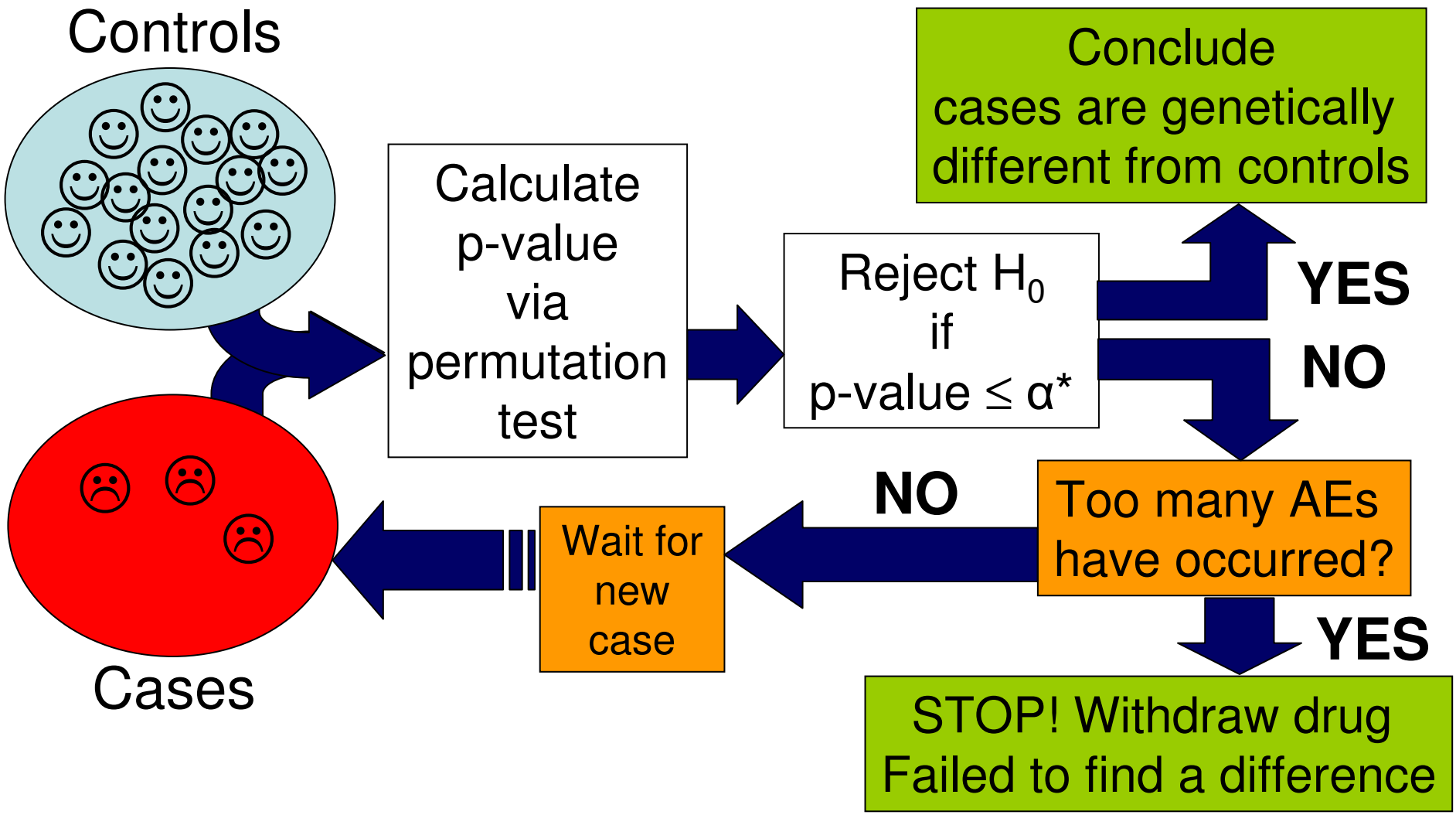
Some genotypes are rare $\Rightarrow$ asymptotics poor

SNPs are not independent (linkage disequilibrium)

Overcome problem by using a **permutation test** to obtain p-value

Correlation structure of genome is preserved

Sequential testing



Sequential testing

We plan up to M tests

Need to account for the multiple testing

Otherwise type I error rate, α, will be inflated

How can we control the type I error rate?

- Adjust the nominal significance level α* accordingly
- Can't use standard sequential methods

Sequential testing

Standard multiple testing methods

Sidak $\alpha^* = 1 - (1 - \alpha)^{1/M}$

e.g. $M = 20$, $\alpha = 0.05 \Rightarrow \alpha^* \approx 0.0026$

Assumes tests are independent

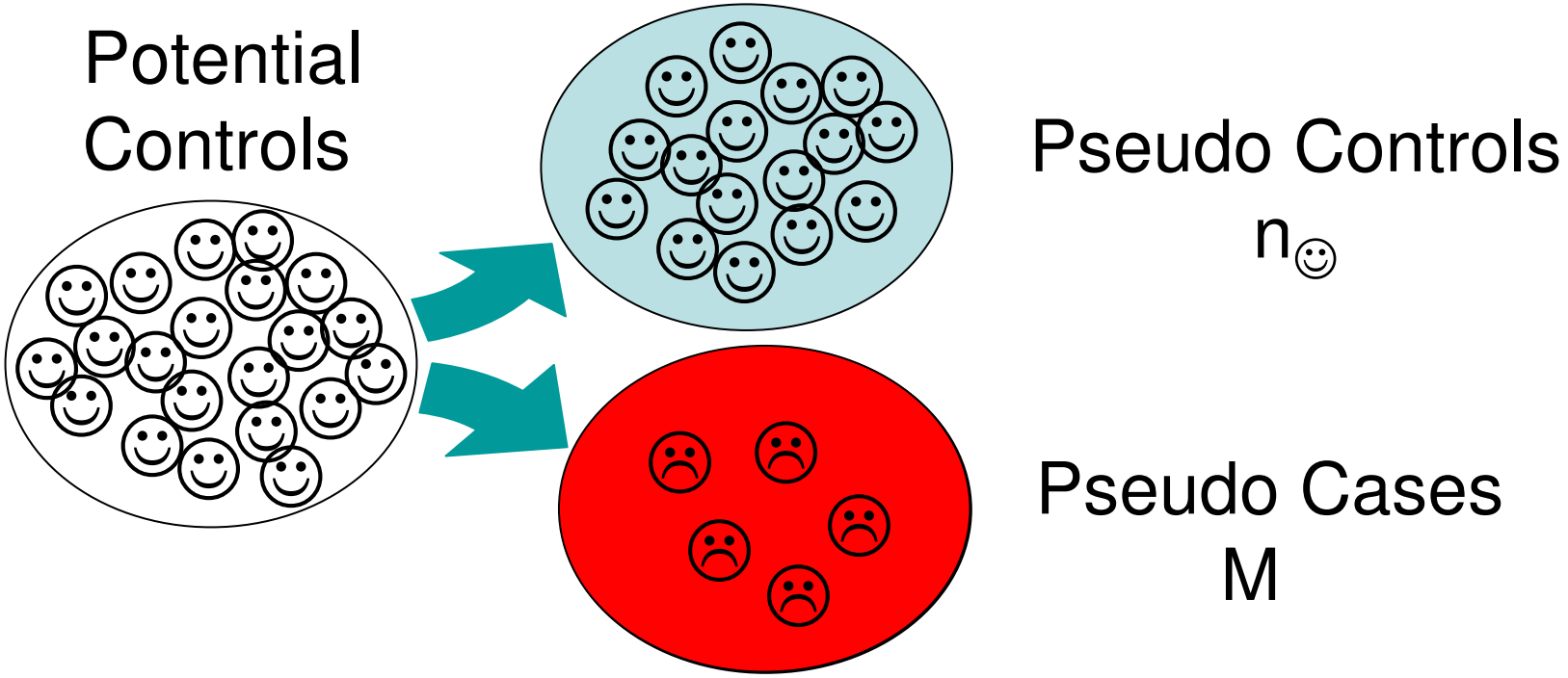
But tests are correlated!

Analysis will be conservative

Sequential testing

Sampling-based approach

Randomly select



Sequential testing

Sampling-based approach

With pseudo data set

conduct sequential procedure

using significance level α^*

Reject H_0?

Repeat many times (e.g 10,000) to estimate α

Find α^* which gives required α

Example

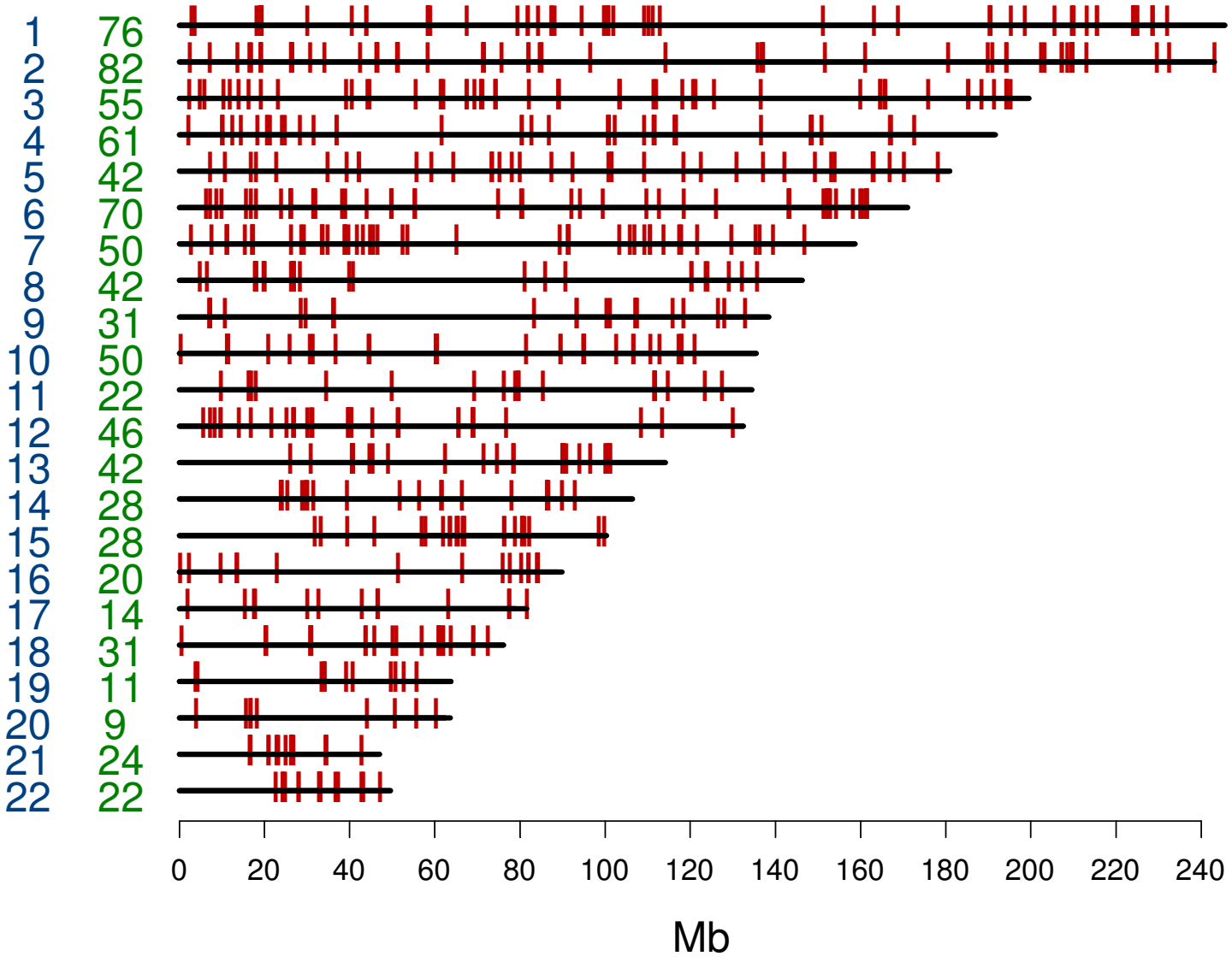
Abacavir : anti-viral treatment for HIV

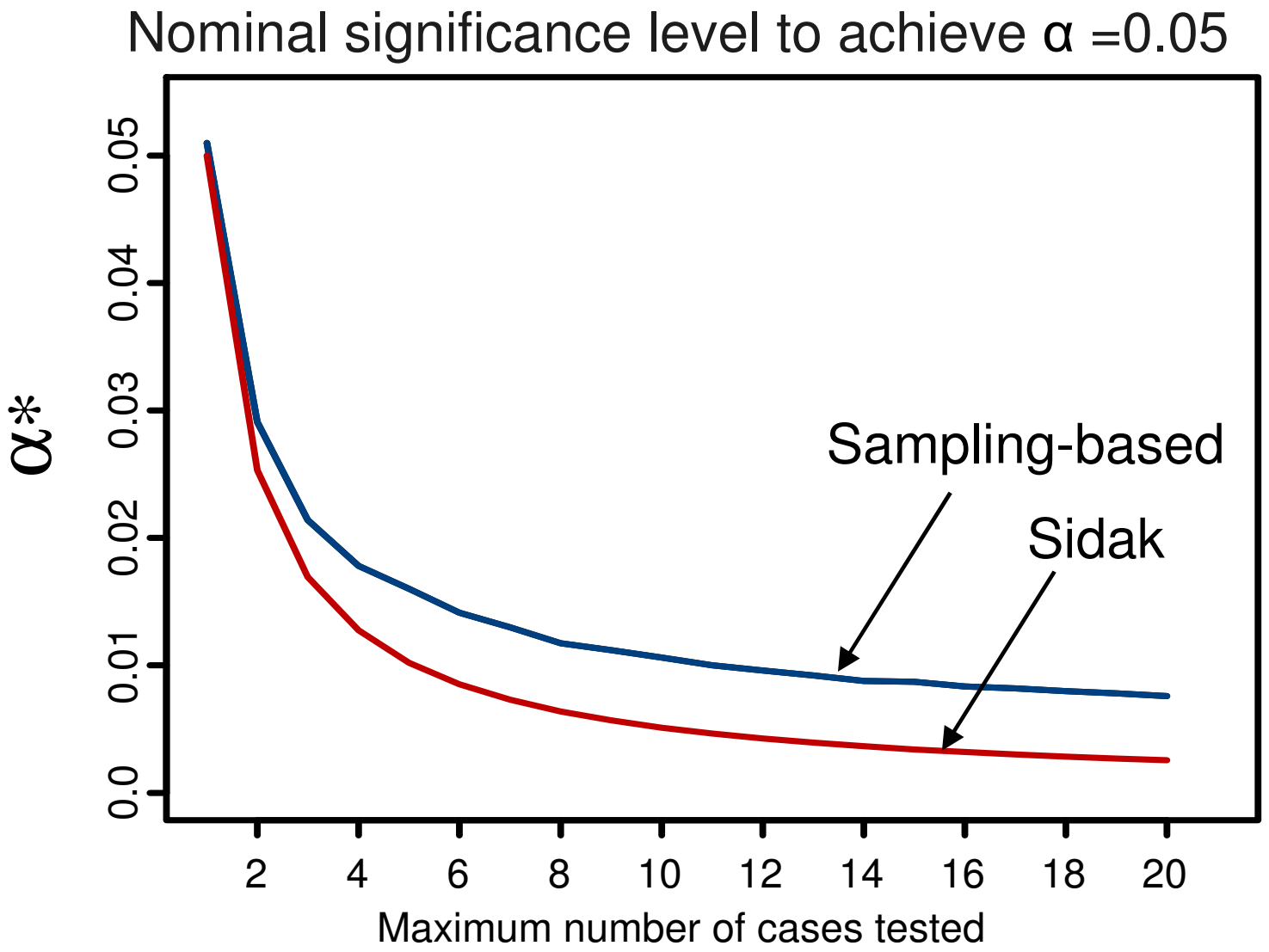
AE: hypersensitivity (5-8% of patients)

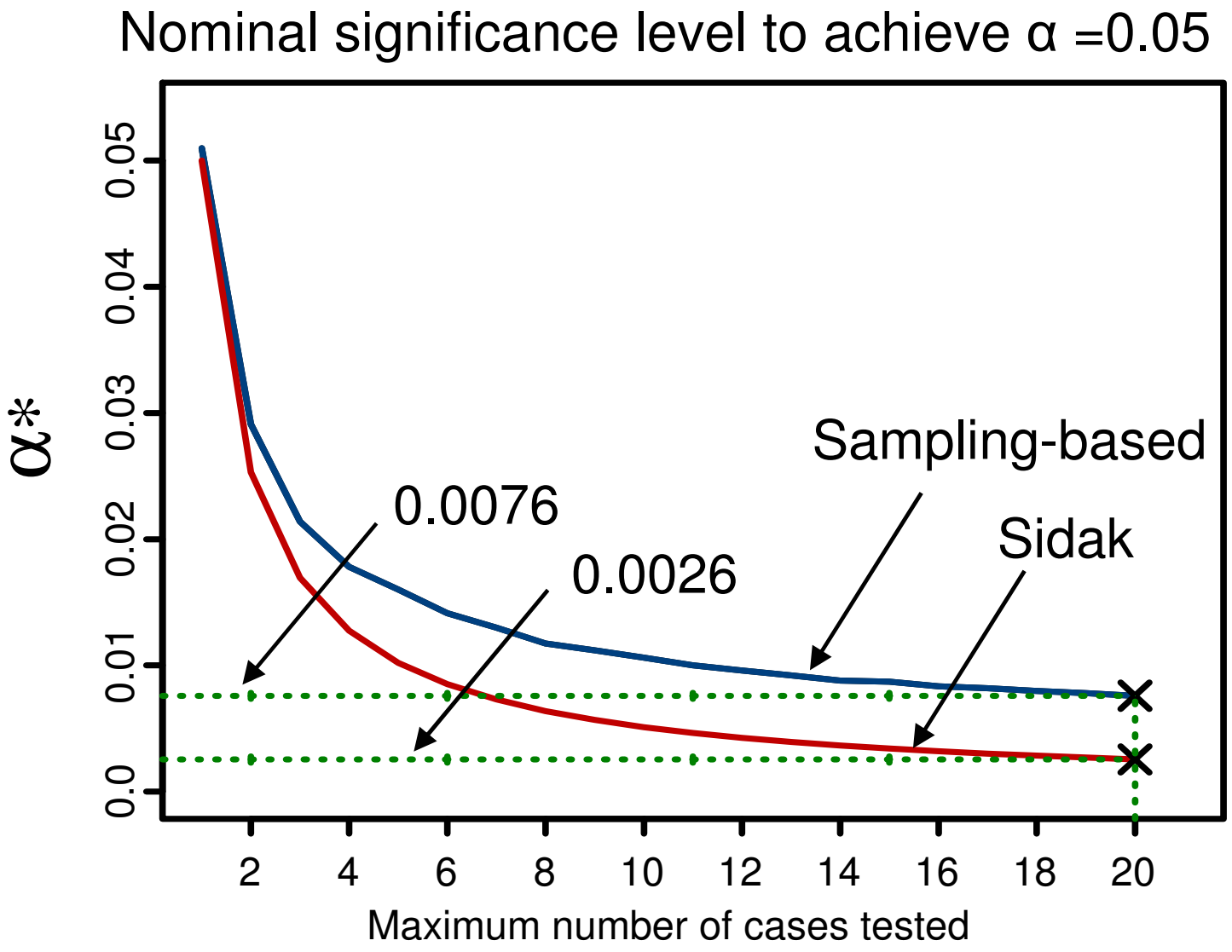
523 patients with AE (pool of cases)

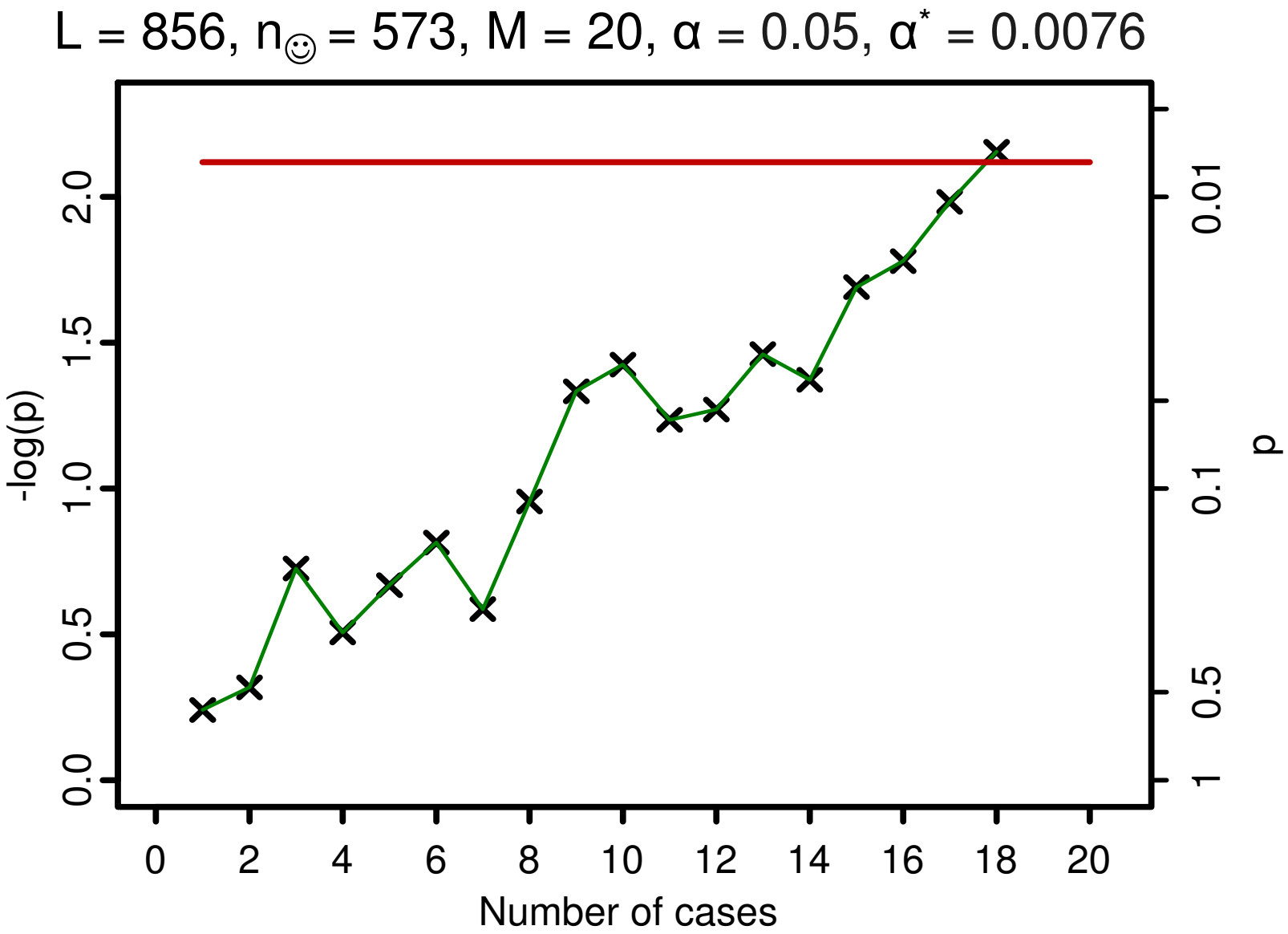
593 patients without AE (pool of controls)

Retrospective application


L = 856 SNP markers
1 76
2 82
3 55
4 61
5 42
6 70
7 50
8 42
9 31
10 50
11 22
12 46
13 42
14 28
15 28
16 20
17 14
18 31
19 11
20 9
21 24
22 22
0 20 40 60 80 100 120 140 160 180 200 220 240
Mb


Nominal significance level to achieve α =0.05
α*
0.0
0.01
0.02
0.03
0.04
0.05
2
4
6
8
10
12
14
16
18
20
Maximum number of cases tested
Sampling-based
Sidak


Nominal significance level to achieve α =0.05
α*
0.0
0.01
0.02
0.03
0.04
0.05
2
4
6
8
10
12
14
16
18
20
Maximum number of cases tested
Sampling-based
Sidak
0.0076
0.0026


L = 856, n☺ = 573, M = 20, α = 0.05, α* = 0.0076
-log(p)
0.0
0.5
1.0
1.5
2.0
p
1
0.5
0.1
0.01
0
2
4
6
8
10
12
14
16
18
20
Number of cases

Summary

Permutations for fixed sample test

- allows for correlation between SNP markers
- obtain a valid global p-value for genome

Sampling-based sequential testing

- adjust nominal significance level to account for repeated testing
- allows for correlation between successive tests
- simple and easy

Summary

Requirements

- must specify the maximum number of tests planned
- must have group of controls in advance
- must have sufficient computing facilities
  - very computer intensive!

Issues

How do you select the controls?

Can we develop a less computer intensive method?

If we reject H_0, how do we identify those patients who should be excluded from receiving the drug?

References

Kelly PJ, Stallard N, Zhou Y, Whitehead J, Bowman C.

Sequential genome-wide association studies for monitoring adverse events in the clinical evaluation of new drugs. *Statistics in Medicine* 2006; 25: 3081-3092

Kelly PJ, Stallard N, Whittaker JC.

Statistical design and analysis of pharmacogenetics trials. *Statistics in Medicine*, 2005; 24: 1495-1508