Second Multi-IC Symposium on the Application of Genomic Technologies to Population-Based Studies: Facilitating Collaboration in Genome-Wide Association Studies

Challenges in analysis: computing platforms, spurious associations

Joan E. Bailey-Wilson, Ph.D.
Head, Statistical Genetics Section
Co-Chief, Inherited Disease Research Branch
NHGRI, NIH

A "significant" Population Association may be due to:

- a causal allele - the associated allele IS the disease/trait allele
- an associated allele at a marker locus which is in linkage disequilibrium with the disease/trait allele at the tightly linked disease locus
- association without linkage as a result of other causes such as population stratification or confounding (uninformative about disease causality)
- Type I error (false positive)

Association Due to Linkage Disequilibrium

- In small isolated populations with only a small number of founders, linkage disequilibrium may extend for a long distance around the disease locus.
- **In most populations, linkage disequilibrium can only be detected for extremely tightly linked loci or when the "marker" locus is in fact the disease locus**.
- This is what we hope for when we observe a significant association in a GWAS study.
- But – what about spurious associations – what causes them?? To understand that we need to talk about Hypotheses, hypothesis testing and statistical error rates

Hypotheses in GWAS Studies

- GWAS studies are often said to be "fishing expeditions" without hypotheses.
- This is NOT true!!!
- All statistical tests have a null hypothesis called "H_0:" (that we try to reject) and an alternate hypothesis called "H_a" that is a reasonable alternative if the null hypothesis is rejected.

Hypotheses in GWAS Studies

- The **null hypothesis** for a GWAS is "None of the SNP loci genotyped in these data are associated with the disease of interest."
- The **alternate hypothesis** is "At least 1 of the genotyped SNPs is associated with the disease of interest in these data".
- We calculate **statistics** to measure the **strength** of the association of each SNP (or haplotypes of SNPS) with the disease of interest and significance levels (p-values).
- What do these p-values MEAN????

Significance Levels

- **P-value** – The probability of observing a test statistic as large or larger than the one that **was observed** in your data **if the null hypothesis is true.**
- **As the p-value becomes small, then we feel more comfortable saying that the null hypothesis of "no association" is not true.**
- We pick a **threshold** and when the observed p-value is lower than this threshold, we say the test is **significant** at that p-value and that we have **rejected the null hypothesis.**

Error Rates

- **Significant association** – We generally want there to be only a 5% chance of making an error if we reject the null hypothesis of no association – so we pick a p-value threshold of 5%. This means that if we repeat the same test 100 times, we would say there **was** a significant association 5 times, even though it was not true
- **Type I error** – Saying that the null hypothesis of NO ASSOCIATION is rejected when in fact it is true and there are not any associations of the SNPS with the disease.

Error Rates

- **Type I error rate** – the probability of falsely rejecting the null hypothesis of no association
- Should be 5% if you set a p-value threshold of 5%
- **Type II error** – the probability of FAILING to reject the null hypothesis when it is false – i.e. the probability of NOT saying that a SNP is associated with disease in your sample when in fact it really is associated in the population.
- **Power = 1-Type II Error**

Error Rates

- You want to **MINIMIZE** Type I error rate and **MAXIMIZE** Power so that you don't say a SNP is associated when it is not (spurious association) and you also don't MISS true associations.
- Making the threshold for a significant test very small will make Type I error small **but** this will reduce power.
- We usually pick a **threshold** of 5% (p=0.05) so we only have a 5% risk of a spurious association.

Error Rates

- **BUT there are many things that can increase your Type I error rate and cause spurious associations!**

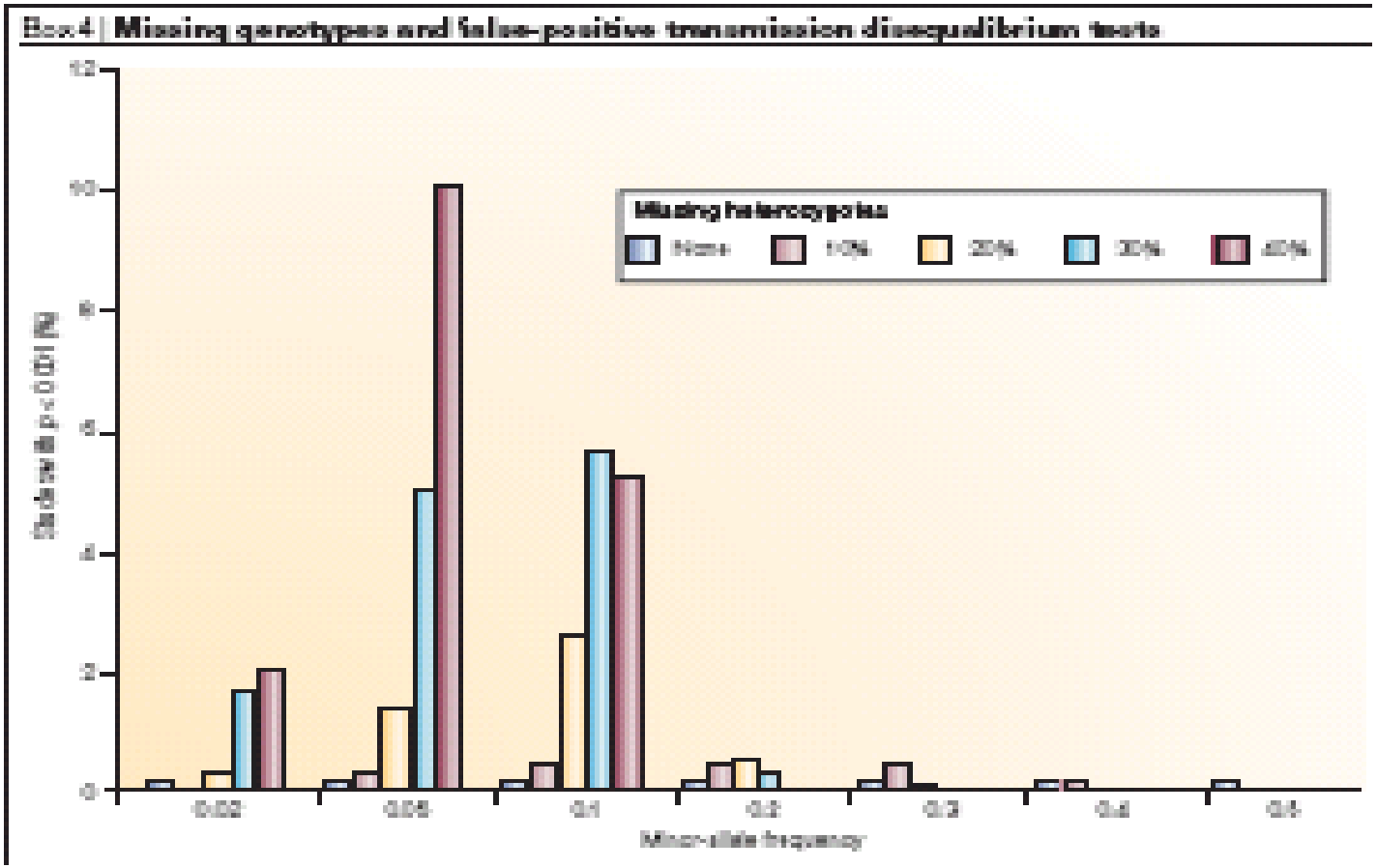
Types of Association Study Designs Can Affect Rate of Spurious Associations

- **Case-Control association tests**
  - Powerful but susceptible to increased Type I errors due to population stratification
  - Can use genomic control methods to detect and adjust for stratification [e.g. Pritchard & Donnelly, *Theoretical Population Biology* **60**, 227–237,2001]
  - Match cases and controls by ethnicity
- **Family based association tests** such as the Transmission Disequilibrium Test (TDT), Haplotype Relative Risk (HRR) method, Pedigree Disequilibrium Test (PDT), and Family-Based Association Test (FBAT)
  - Requires more genotyping for equivalent power but not susceptible increased rates of spurious association due to population stratification

Other Causes of Spurious Associations

- Genotyping errors
  - **Systematic genotyping errors** that preferentially classify heterozygotes as homozygotes or that cause heterozygote genotypes to fail more frequently can increase the rate of spurious associations
  - Test for Hardy-Weinberg equilibrium to detect this and drop markers where the H-W p-value <0.01
  - Drop markers that have genotyping call rates less than 95% (or 99%) since they are more likely to have systematic errors!
  - Repeat genotyping of highly significant SNPs to obtain complete and accurate genotypes

Percentage of spurious associations at varying rates of missing heterozygote data in 1000 parent-child trios, AFTER removing SNPs with call rates < 90% and SNPs with H-W p-value <0.01



Hirschhorn & Daly, Nat Rev Genet 6:95-108, 2005

Genome Wide Association Studies Multiple Testing and Error Rates

- Genotype from 100K to 550K (or more) SNP markers
- Perform association test for each marker with the disease or trait
- **For each test, 5% chance of a spurious association if you use a p-value threshold of 5%**
- **5% X 550,000 = 27,500 spurious associations in a GWAS if you use p=0.05 as the threshold**
- Multiple testing is a **BIG** problem!!!

Adjusting for Multiple Testing

- **Use a smaller p-value for the significance threshold to control the family wide error rate** – i.e. the probability of ANY false positive in your whole study.
- $p=1 \times 10^{-5}$ results in 5 false positives in one GWAS of 550,000 tests
- **$p=1 \times 10^{-7}$ results in a 5.5% chance of a Type I error when performing 550,000 tests**
- P-value needs to be smaller if performing more tests (haplotypes, GXG or GXE interactions)

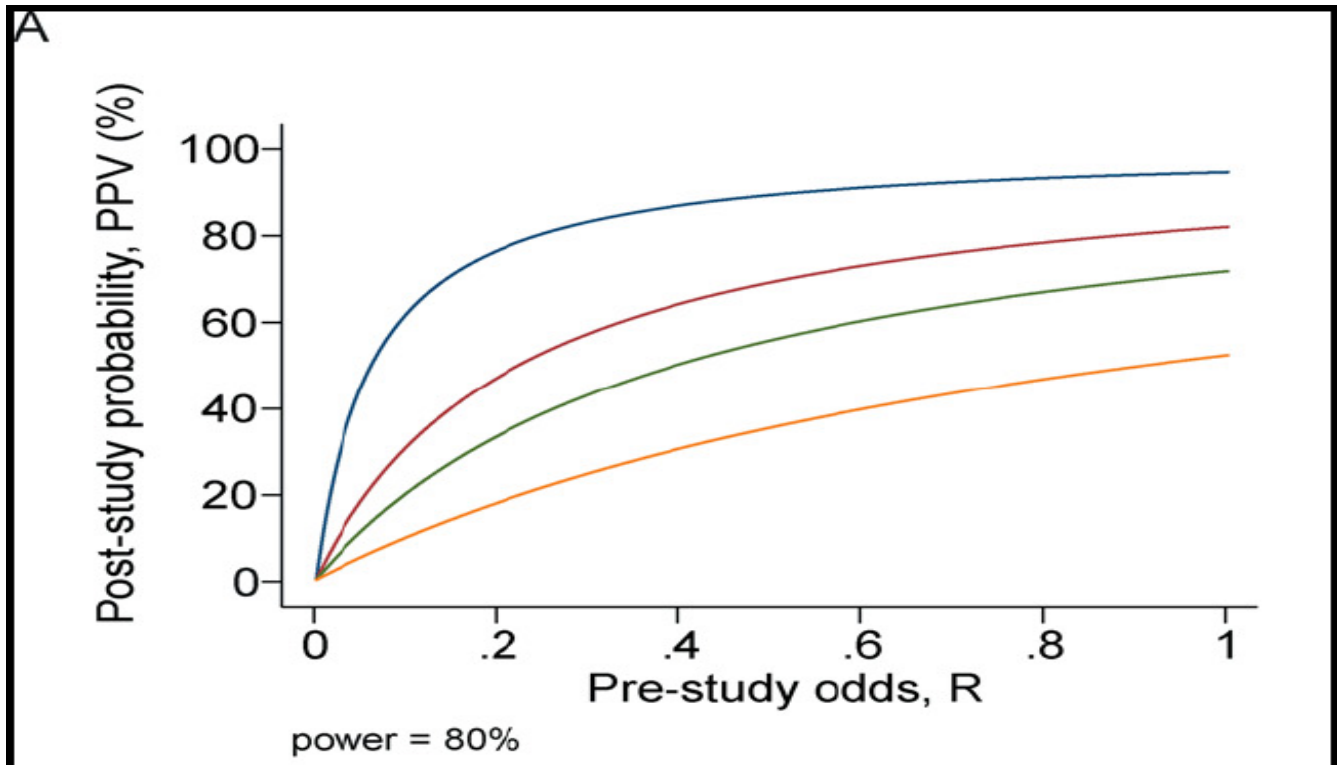
Adjusting for Multiple Testing

- **But** setting a lower p-value threshold means that **larger sample sizes are needed** to have enough power to detect SNPs with real associations, particularly if the size of the effect of the SNP on risk for disease is small.
- Most studies will be underpowered, making it very likely that **many significant associations will be Type I errors or spurious associations**.

Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

- It can be proven that most claimed research findings are false.
- The **probability that a research finding is indeed true** depends on the **prior probability** of it being true (before doing the study), the statistical **power** of the study, and the level of statistical significance (**p-value**)
- **Bias** can further reduce the probability that a research finding is true. It can entail manipulation in the analysis or reporting of findings. Selective or distorted reporting is a typical form of such bias.
- The post-study probability that a significant research finding is true is the **positive predictive value, PPV**.

PPV for any one study decreases if many studies are performed



Blue=1, Red=5, Green=10, Orange=50 studies

Ioannidis JPA (2005). PLoS Med 2(8): e124.

Corollaries

- The **smaller the studies** conducted in a scientific field, the less likely the research findings are to be true.
- The **smaller the effect sizes** in a scientific field, the less likely the research findings are to be true.
- The **greater the number and the lesser the selection of tested relationships** in a scientific field, the less likely the research findings are to be true.
- The **greater the flexibility in designs, definitions, outcomes, and analytical modes** in a scientific field, the less likely the research findings are to be true.
- Too large and too highly significant effects may actually be more likely to be signs of large bias in most fields of modern research.

Ioannidis JPA (2005). PLoS Med 2(8): e124.

Moonesinghe R, Khoury MJ, Janssens ACJW (2007) Most published research findings are false—But a little replication goes a long way. PLoS Med 4(2): e28. doi:10.1371/journal.pmed.0040028

- **Replication**—the performance of another study statistically confirming the same hypothesis—is the cornerstone of science and replication of findings is very important before any causal inference can be drawn.
- **PPVs of research findings increase when more studies have statistically significant results.**

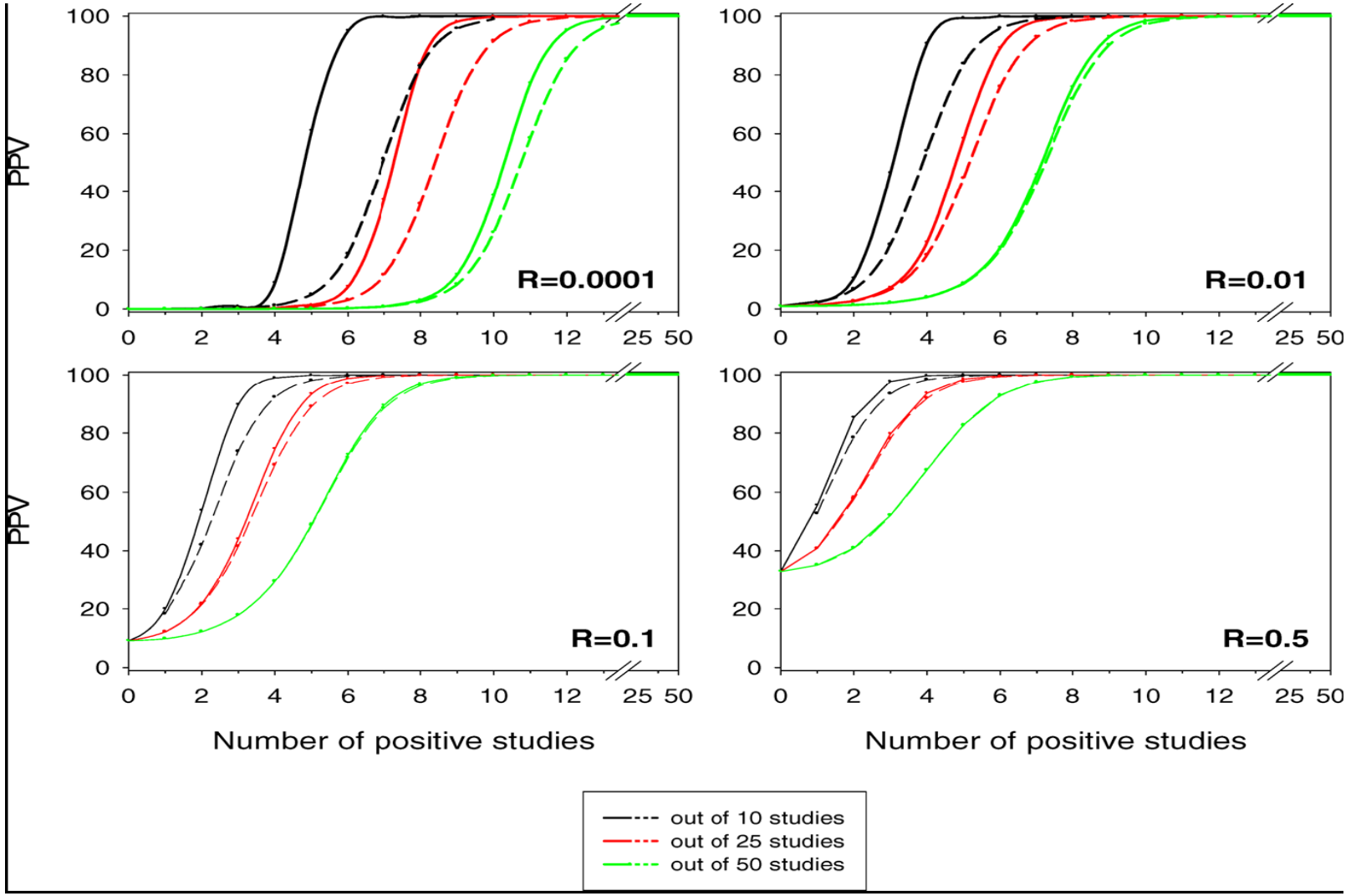


Figure 2. Positive Predictive Value for Research Findings Being True for At Least *r* Positive Studies Out of Ten, 25, and 50 Studies for Pre-Study Odds *R* of 0.0001, 0.01, 0.1, and 0.5 (α =0.05) Dashed lines refer to power of 0.2 and solid lines to power of 0.8.

Moonesinghe R, Khoury MJ, Janssens ACJW (2007). PLoS Med 4(2): e28. doi:10.1371/journal.pmed.0040028

Replication to control spurious associations

- True replication requires that the exact same finding is reexamined in the same way.
- Original and replication studies should be adequately powered
- Bias against reporting negative studies will seriously decrease the chance that a finding is true

How to ensure replication studies are adequately powered?

- Zollner S, Pritchard JK Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data AJHG 80:605-615, 2007
- Difficulties of replication occur because most genuine associations have small to moderate effects on risk of disease
- Therefore, there is generally incomplete power to detect associations in any given study

Zollner S, Pritchard JK Overcoming the Winner's Curse: Estimating Penetrance Parameters from Case-Control Data

AJHG 80:605-615, 2007

- The odds ratio (a measure of risk of disease due to the associated risk allele) is almost always overestimated in the initial association study
- Using these overestimates to plan the sample size needed for the replication study results in lower power than expected
- Zollner & Pritchard developed a method for estimating frequency and genotypic penetrances at the associated locus that do not overestimate the effect of the risk allele on disease risk

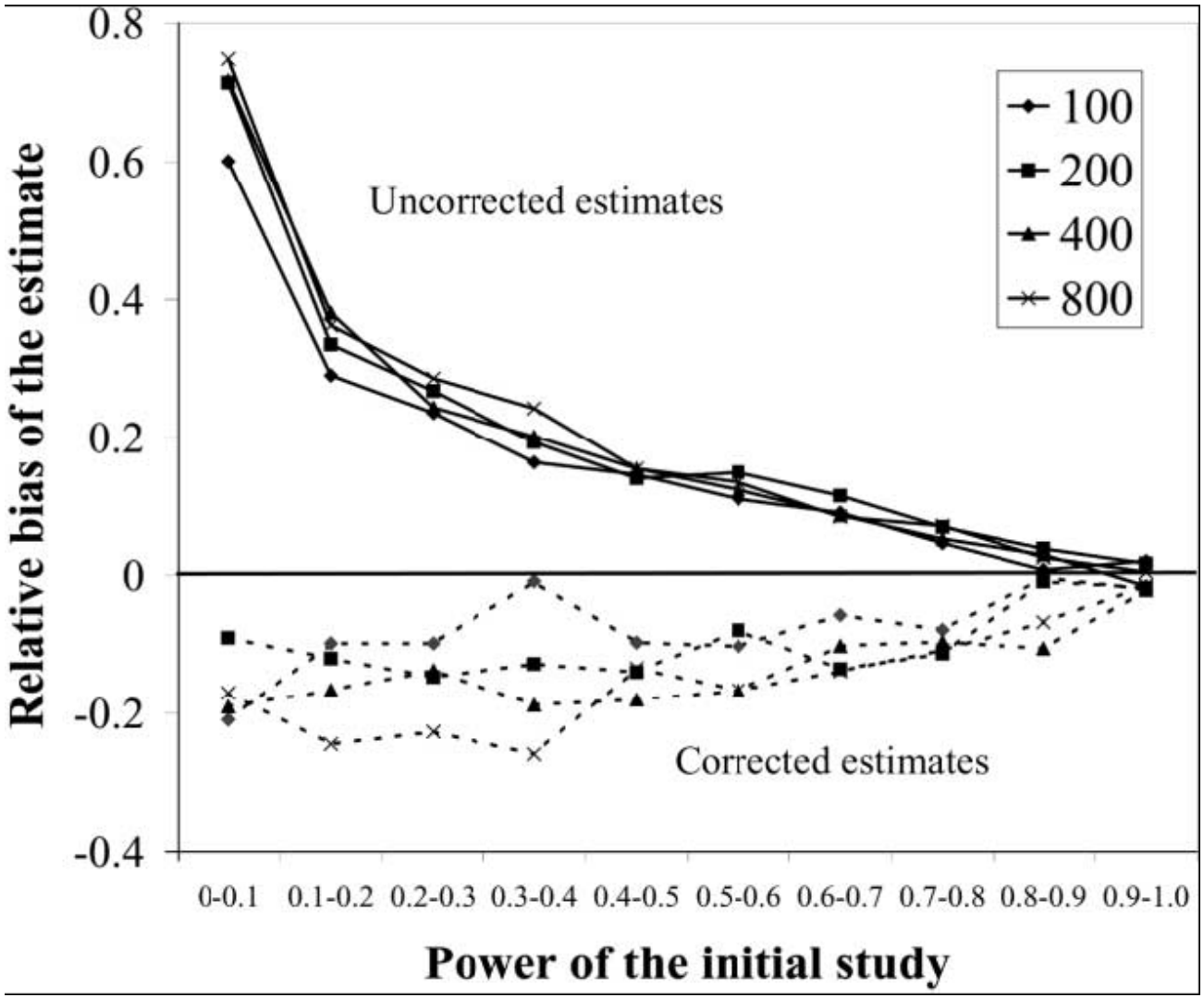


Figure 2. Bias of the uncorrected and corrected estimates of the additive genetic effect. The vertical axis indicates the average relative bias observed in each power category. The solid lines show the bias of estimates of penetrance parameters that were generated without correction for ascertainment, whereas the dashed lines show the bias of estimates generated while correcting for ascertainment.

Zollner S, Pritchard JK AJHG 80:605-615, 2007

Replication Studies

- **Should be adequately powered**
- **Should test the same phenotype**
- **Should test the same associated SNP**, even if other nearby SNPs are also tested
- At least some replication studies should be in the same population
- Should use the same methods as the original report

Computing Issues

- **HUGE** amount of data
- Can take **DAYS** just to download these datasets
- Large amounts of disk space and fast processors with fast I/O speeds are necessary to manipulate these data
- Thousands of files of genotypes for one project!!
- Merging phenotypes with genotypes is a BIG job!

Computing Issues

- Some programs available to help manipulate these files
- PLINK (free)
- BC/SNPmax
- HelixTree

Computing Issues

- Computations for 550,000 tests take much longer than for 1 or even 100 tests as in most candidate gene studies. (5500 times longer than 100 tests)
- Doing a GWAS on one PC will take an unacceptably long time!!
- Need a cluster of fast computers to enable GWAS analyses to finish in a reasonable time.