European Journal of Human Genetics (2015) 23, 863–869

www.nature.com/ejhg

ARTICLE

Homogeneous case subgroups increase power in genetic association studies

Matthew Traylor*[,1], Hugh Markus[1] and Cathryn M Lewis[2,3]

Genome-wide association studies of clinically defined cases against controls have transformed our understanding of the genetic causes of many diseases. However, there are limitations to the simple clinical definitions used in these studies, and GWAS analyses are beginning to explore more refined phenotypes in subgroups of the existing data sets. These analyses are often performed *ad hoc* without considering the power requirements to justify such analyses. Here we derive expressions for the relative power of such subgroup analyses and determine the genotypic relative risks (GRRs) required to achieve equivalent power to a full analysis for relevant scenarios. We show that only modest increases in GRRs may be required to offset the reduction in power from analysing fewer cases, implying that analyses of more genetically homogenous case subgroups may have the potential to identify further associations. We find that, for lower genotypic relative risks in the full sample, subgroup analyses of more homogeneous cases have relatively more power than for higher index genotypic relative risks and that this effect is stronger for rare as opposed to common variants. As GWA studies are likely to have now identified the majority of SNPs with stronger effects, these results strongly advocate a renewed effort to identify phenotypically homogeneous disease groups, in which power to detect genetic variants with small effects will be greater. These results suggest that analysis of case subsets could be a powerful strategy to uncover some of the hidden heritability for common complex disorders, particularly in identifying rarer variants of modest effect.
European Journal of Human Genetics (2015) **23,** 863–869; doi:10.1038/ejhg.2014.194; published online 1 October 2014

INTRODUCTION

In recent years, genome-wide association studies of clinically defined case phenotypes against controls have transformed our understanding of the common genetic causes of many diseases. Hundreds of common genetic variants have been identified that confer small but significant proportions of disease risk.[1–7] The use of clinical phenotypes to define case sets has simplified the collection of sets of diseased cases and has enabled easy interpretation of the impact of disease loci. However, there are limitations to such a simple approach to phenotyping,[8–10] particularly in the presence of heterogeneity.[11]

First, such phenotype definitions depend on adequate diagnostic sensitivity and specificity, which is challenging in some diseases. For example, in Alzheimer's Disease, where a number of common variants have been shown to confer risk of the disease,[2] the majority of cases are diagnosed based on clinical criteria (eg, DSM-IV criteria). Post-mortem data show that clinical diagnoses are imperfect, with specificity and sensitivity of $<80\%$.[12] This leads to underestimation of the effects of associated SNPs, and worse could lead to false positive results, where associations in reality are with diseases misdiagnosed as Alzheimer's Disease. Second, such clinical diagnoses ignore underlying heterogeneity in disease pathogenesis where case subtyping might be more appropriate. Examples of diseases with genetically distinct subgroups include ischaemic stroke, where at least three distinct pathologies (cardioembolic, small vessel and large vessel) lead to stroke events;[13,14] migraine, where cases with or without aura have distinct genetic susceptibility factors;[15] and rheumatoid arthritis, where anti-citrullinated peptide antibody-negative individuals show distinct genetic associations, particularly in the HLA region.[16,17] Third, heterogeneity in genetic susceptibility to disease may exist. For example, cases with later disease onset have more exposure to environmental risk factors, and therefore under a liability threshold model will have a weaker genetic susceptibility.[18,19] Similarly, individuals with type 2 diabetes with higher body mass index may have decreased genetic susceptibility to the disease.[18,20]

Analysis of subgroups of cases in GWAS data may therefore be valuable to identify further associations. Such analyses have been performed,[13,15,20,21] but this has generally been carried out without consideration of the relative power of these analyses, and the conditions under which such analyses are advantageous are not well understood. To resolve this, we seek to answer two questions. First, what increase in genotypic relative risk in a disease subgroup is required to achieve equivalent power to a full analysis? Second, what is the relationship between power in a full analysis and a subgroup analysis, and how is this affected by the size of the genetic effect and its allele frequency? We first derive formulae for the relative power of subgroup analyses to a full analysis and use these to study the power of subgroup analyses for scenarios relevant to GWAS. We then derive expressions for the genotypic relative risk required in a subgroup analysis to achieve equivalent power to a full sample and evaluate this relationship for plausible scenarios. Finally, by interrogating the power relationship between a full and subgroup analysis for fixed proportional increases in genotypic relative risk for the subgroup, we show that subgroup analyses are advantageous in identifying genetic variants with increasingly smaller effects.

[1]Clinical Neurosciences, University of Cambridge, Cambridge, UK; [2]Department of Medical and Molecular Genetics, King's College London, London, UK; [3]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK
*Correspondence: M Traylor, Clinical Neurosciences, University of Cambridge, R3, Box 83, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. Tel: +44 012 2321 7718; E-mail: mt628@medschl.cam.ac.uk
Received 25 April 2014; revised 16 July 2014; accepted 20 August 2014; published online 1 October 2014

MATERIALS AND METHODS

Relative power of subgroup analyses

To better understand the relationship in power between a full and subgroup analysis, formulae for the ratio of non-centrality parameter (*NCP*) from analysis of a subgroup of cases to a full analysis were derived. These formulae were then used to study properties of the *NCP* ratio for appropriate risk allele frequencies (RAF), index genotypic relative risks with respect to the causal variant (λ), and proportional increases in GRRs in the subset analysis (κ).

Expressions for the ratio in power of a subset analysis to a full analysis were first derived using the framework from Yang *et al*[22] as follows. In the context of a case–control study of a complex disease with prevalence K, consider a variant with two alleles (A, a) with frequency p and (1–p). Assuming a multiplicative model of allele effects, the *NCP* of a χ^2 test for association can be expressed as follows:

$$NCP = \frac{2p(1-p)(\lambda-1)^2 v(1-v)N}{(1-K)^2[1+p(\lambda-1)]^2}$$

where N is the total sample size and v denotes the proportion of the overall sample that are cases.[22] *NCP*, used to calculate power analytically, is the value that determines the degree of noncentrality of the chi-squared distribution under the alternative hypothesis. The power of a χ^2 test for association can be found by integrating under the noncentral χ^2 distribution for given *NCP* and degrees of freedom of the test. Although *NCP* does not directly equate to power, it is a suitable proxy measure and is used here to represent power, as in previous studies.[22]

Of interest is the ratio in *NCP* of analysis of a subset of the cases (N_2 individuals, prevalence K_2, proportion of cases v_2, and GRR λ_2) to the power of a full study with N_1 individuals, prevalence K_1, proportion of cases v_1, and GRR λ_1, where N_2 is a subset of N_1 (retaining all controls) in which the variant has a stronger effect.

ie, $N_1 > N_2$ and $\lambda_1 < \lambda_2$.The ratio of *NCP* between two such analyses can be expressed as:

$$\frac{NCP_2}{NCP_1} = \frac{\frac{2p_2(1-p_2)(\lambda_2-1)^2 v_2(1-v_2)N_2}{(1-K_2)^2[1+p_2(\lambda_2-1)]^2}}{\frac{2p_1(1-p_1)(\lambda_1-1)^2 v_1(1-v_1)N_1}{(1-K_1)^2[1+p_1(\lambda_1-1)]^2}}$$
$$= \left(\frac{v_2(1-v_2)}{v_1(1-v_1)}\right)\left(\frac{N_2}{N_1}\right)\left(\frac{2p_2(1-p_2)(\lambda_2-1)^2(1-K_1)^2\,[1+p_1(\lambda_1-1)]^2}{2p_1(1-p_1)(\lambda_1-1)^2(1-K_2)^2\,[1+p_2(\lambda_2-1)]^2}\right)$$

p_1 and p_2 are defined in the union of cases and controls in each situation. In the context of GWAS, where genetic effects are small, and particularly when number of controls exceeds that of cases, it can be assumed that $2p_1(1-p_1) \cong 2p_2(1-p_2)$ in the two scenarios as changes in allele frequency between the full and subgroup analysis will be minimal.

Thus we obtain the following expression (formula (1)) for the ratio in power of a subgroup analysis to power of a full analysis:

$$\frac{NCP_2}{NCP_1} \cong \varphi\left(\frac{(\lambda_2-1)^2\,[1+p(\lambda_1-1)]^2}{(\lambda_1-1)^2\,[1+p(\lambda_2-1)]^2}\right) \quad (1)$$

where

$$\varphi = \left(\frac{v_2(1-v_2)}{v_1(1-v_1)}\right)\left(\frac{N_2}{N_1}\right)\left(\frac{(1-K_1)^2}{(1-K_2)^2}\right)$$

To evaluate the derived formula, we calculated *NCP* ratios for a case–control study of 2000 cases and 2000 controls compared with a subgroup of 1000 cases and 2000 controls for a RAF = 0.25 and for index GRR (λ_1) of 1.1, 1.2, and 1.3, each for five subgroup GRR (λ_2) using formula (1). We then compared these with the equivalent *NCP* ratios from Genetic Power Calculator, calculating each *NCP* separately and determining the appropriate ratio, to confirm that our estimates matched those obtained from an alternative approach.[23]

Genotypic relative risk required for equivalent power in the subgroup analysis

Having derived expressions for relative power of subgroup analyses, we then derived an expression for the GRR (λ_2) required in the subgroup to achieve equivalent power as the full analysis. From equation (1) above, we set the full expression for *NCP* ratio equal to one and solve for λ_2:

$$\varphi\left(\frac{(\lambda_2-1)^2[1+p(\lambda_1-1)]^2}{(\lambda_1-1)^2[1+p(\lambda_2-1)]^2}\right) = 1$$

If we make the substitutions, $\gamma = \lambda_1 - 1$ and $\delta = \lambda_2 - 1$, we have:

$$\varphi\left(\frac{(\delta)^2\,[1+p(\gamma)]^2}{(\gamma)^2\,[1+p(\delta)]^2}\right) = 1.$$

This simplifies to a quadratic equation for δ:

$$\delta^2\left(\varphi\left(\frac{1+2\,p\gamma+p^2\gamma^2}{\gamma^2}\right)-p^2\right) - 2p\delta - 1 = 0$$

Using the quadratic formula to solve for δ gives:

$$\delta = 2p\sqrt{(-2p)^2 + 4\left(\varphi\left(\frac{1+2\,p\gamma+p^2\gamma^2}{\gamma^2}\right)-p^2\right)}$$

Discarding the implausible negative solution and expressing in terms of GRR in the subgroup analysis λ_2, we obtain the following expression (formula (2)) for the GRR in the case subgroup λ_2 required to obtain equivalent power in a subgroup analysis:

$$\lambda_2 = \left(2p - 1 + \sqrt{(-2p)^2 + 4\left(\varphi\left(\frac{1+2\,p\gamma+p^2\gamma^2}{\gamma^2}\right)-p^2\right)}\right) \quad (2)$$

We then used the above formula to calculate the GRR (λ_2) required in the subgroup analysis to achieve equivalent power to the full analysis for four scenarios in which a genetic variant was assumed to have a given GRR (1.05, 1.1, 1.2, or 1.3) in the full sample, and for different proportions of discarded cases in the subgroup analysis for three risk allele frequencies (RAF = 0.01, 0.25, 0.5), assuming an equal number of cases and controls in the full sample.

Relative power for fixed proportional increase in odds ratio

Additionally, we sought to study how the *NCP* ratio between the full and subgroup analyses was affected by index GRR (λ_1) in the full sample. We generated results using formula (1) for intervals of GRR ($1.05 \le \lambda_1 \le 1.35$) in the full sample and for a fixed proportional increase in GRR in the subgroup, $(\lambda_2) = \kappa \times (\lambda_1)$, where $\kappa = 1.05$, 1.10, 1.20, 1.30, using three minor allele frequencies (RAF = 0.01, 0.25, 0.5).

RESULTS

Relative power of subgroup analyses

We identified properties of the relative power of a subgroup analysis, using formula (1). We set disease prevalence at 1% throughout, but we note that the results were almost completely insensitive to this value. We used three index GRRs ($\lambda_1 = 1.1$, 1.2, 1.3) and, for simplicity, assumed that λ_2 is a product of λ_1, that is,

$$\lambda_2 = \kappa\lambda_1, \kappa > 1$$

Index GRRs were in the range of those found in previous GWAS studies[1,2,5,6,13] and were chosen in order to represent SNPs that might be identified in future studies. Results were analysed by κ, where $\lambda_2 = \kappa \times \lambda_1$, rather than specific GRRs to enable comparison across different index GRRs and different proportions of discarded cases. κ values (range = 1.00–1.10) were chosen to reflect modest changes in GRR in case subsets and are similar to those observed in our previous analysis of the effect of age-at-onset in ischaemic stroke.[19] The *NCP* ratio was calculated for proportional increases in GRR (κ) at three minor allele frequencies, assuming either 25 or 50% of cases were discarded (Figure 1).

As expected, increasing values of κ increased the relative power (*NCP* ratio) of the subset analysis. When 50% of cases were discarded,

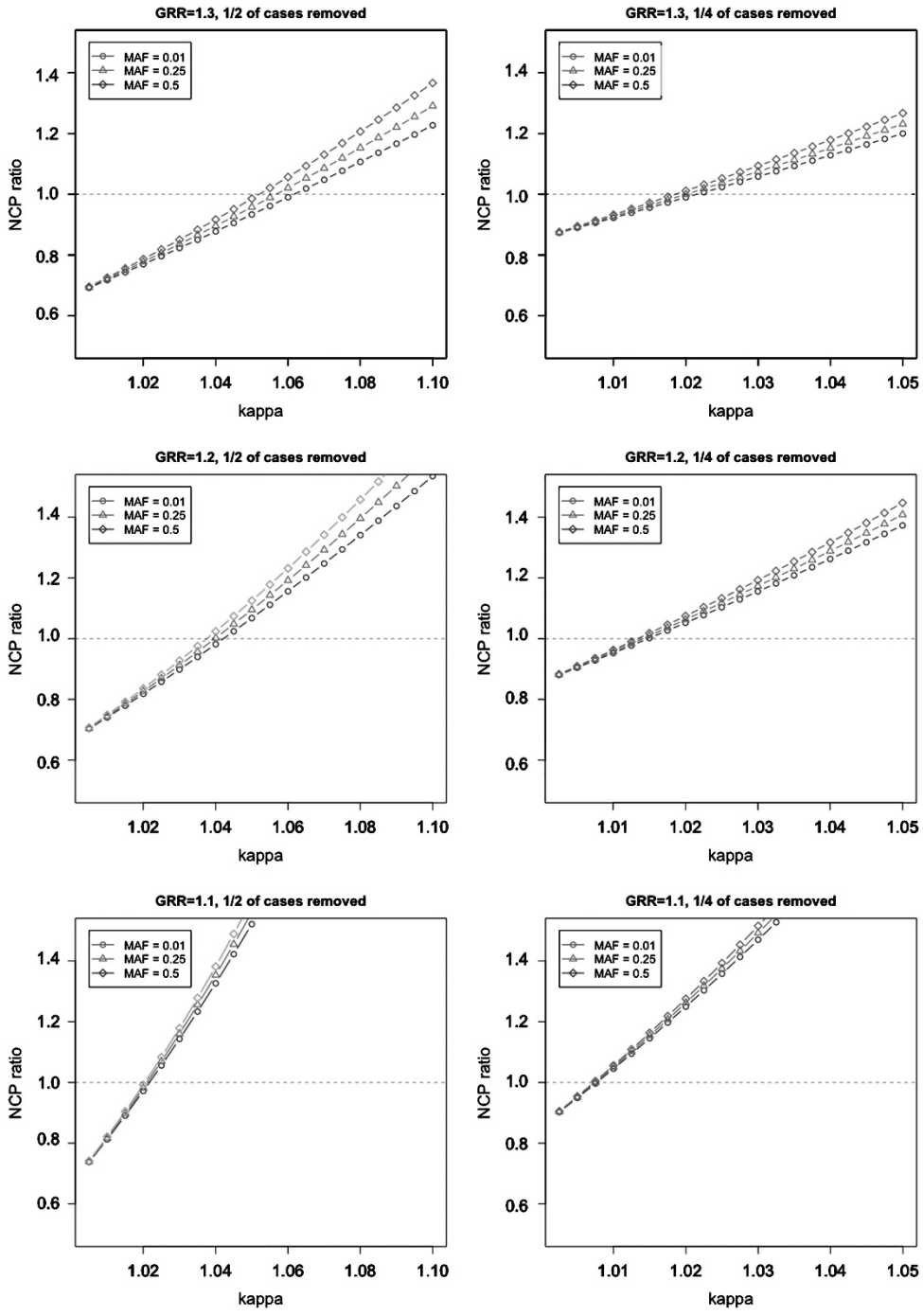

Figure 1 Ratio in power between analyses of all cases and case-subset, for different minor allele frequency and subset size. MAF, minor allele frequency; kappa (κ), proportional increase in GRR. Horizontal line at NCP ratio = 1 denotes the kappa (κ) value for which power in the subset analysis exceeds that of the full analysis.

the threshold of κ at which relative study power became greater in the subset analysis was higher than when 25% of cases were discarded. This indicates that, as expected, the required proportional increase in GRR is correlated with the proportion of cases discarded: a higher proportion of discarded, and therefore smaller retained case sample size, requires a higher GRR to achieve the same power (*NCP*).

We compared the *NCP* ratios from our formula with those calculated from comparing two sets of results generated from Genetic Power Calculator.[23] The concordance between the results was nearly exact ($r = 0.999$, $r^2 = 0.999$), showing that our simple formulae reproduce the results obtained when performing the calculations using alternative approaches.

Genotypic relative risk required for equivalent power in subgroup analysis

We calculated the λ_2 value required to achieve equivalent power in the subgroup for intervals of proportions of cases discarded in the subgroup using formula (2) (Figure 2). The results showed that

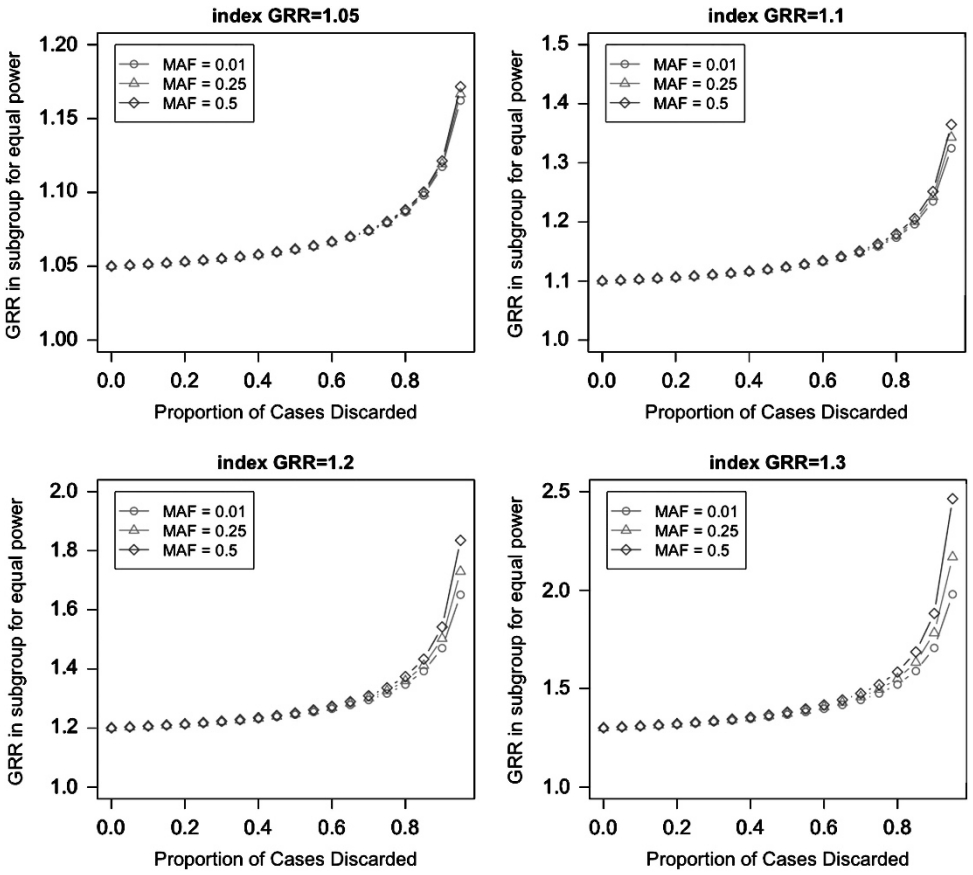

Figure 2 Genotypic relative risk in subgroup required to achieve equivalent power as a full analysis for intervals of proportions of discarded cases and three minor allele frequencies. MAF, minor allele frequency.

Table 1 λ_2 values required in subgroup analysis to achieve equivalent power as full sample for given index genotypic relative risk λ_1 and proportions of discarded cases assuming RAF = 0.10

	λ_2 *for given proportions of discarded cases*		
λ_1	*25%*	*50%*	*75%*
1.05	1.054	1.061	1.079
1.10	1.108	1.123	1.160
1.20	1.217	1.247	1.326
1.30	1.326	1.373	1.496

λ_1, genotypic relative risk in full sample; λ_2, genotypic relative risk in subgroup.

relatively small increases in GRR in the subgroup were required to achieve equivalent power as a full analysis. This was particularly notable for lower index GRRs. For example, for a GRR of 1.05 in the full sample, a GRR of 1.079 in 25% of the cases achieves equivalent power (Table 1). Similarly, for a GRR of 1.10, a GRR of 1.16 in 25% of cases achieves equivalent power. This clearly shows that if stronger genetic effects exist in subgroups of data sets, a large proportion of cases can be discarded without loss of power.

The results also showed that the relative power of a subset analysis is greatest for rare genetic variants at fixed values of λ_1 and κ. This result was consistent across all scenarios studied (Figures 1 and 2). For example, for a genetic variant with index GRR $\lambda_1 = 1.3$ and RAF = 0.01, analysis of a subset of 25% of cases has more power if the proportional increase in the GRR is $\kappa > 1.14$ ($\lambda_2 > 1.48$). The proportionate increase in index GRR required for equal power increases with RAF: for the same scenario, but assuming a variant with a RAF = 0.5, the analysis has more power for $\kappa > 1.17$ ($\lambda_2 > 1.52$). These results show that subgroup analyses have comparatively more power to identify rare as opposed to common variants.

Relative power for fixed proportional increase in genotypic relative risk

Finally, we interrogated the relationship of the *NCP* ratio for different index GRRs (λ_1) in the full sample, fixing κ values in each case. For four κ values (1.05, 1.1, 1.2, 1.3) and across three minor allele frequencies (0.01, 0.25, 0.5), the *NCP* ratio monotonically increased as index GRRs decreased (Figure 3). This effect was particularly strong for index GRR < 1.15, below which the curve increased dramatically. For example, for a variant with RAF = 0.25 and a GRR of 1.3 in the full sample, if the variant has an effect 1.05 times stronger in half the cases, then analysis of this subgroup does not have as much study power than the full analysis (*NCP* ratio = 0.96). Conversely, for the same variant with same proportional increase in half the cases, but an index GRR = 1.05, then analysis of the subgroup has considerably more study power (*NCP* ratio = 2.73). This effect clearly shows that, for smaller genetic effects, the proportional increase in power for analysis of a homogeneous subgroup of the cases increases greatly and emphasizes that homogeneous disease subgroups in which genetic effects may be larger are better suited for detection of small genetic effects.

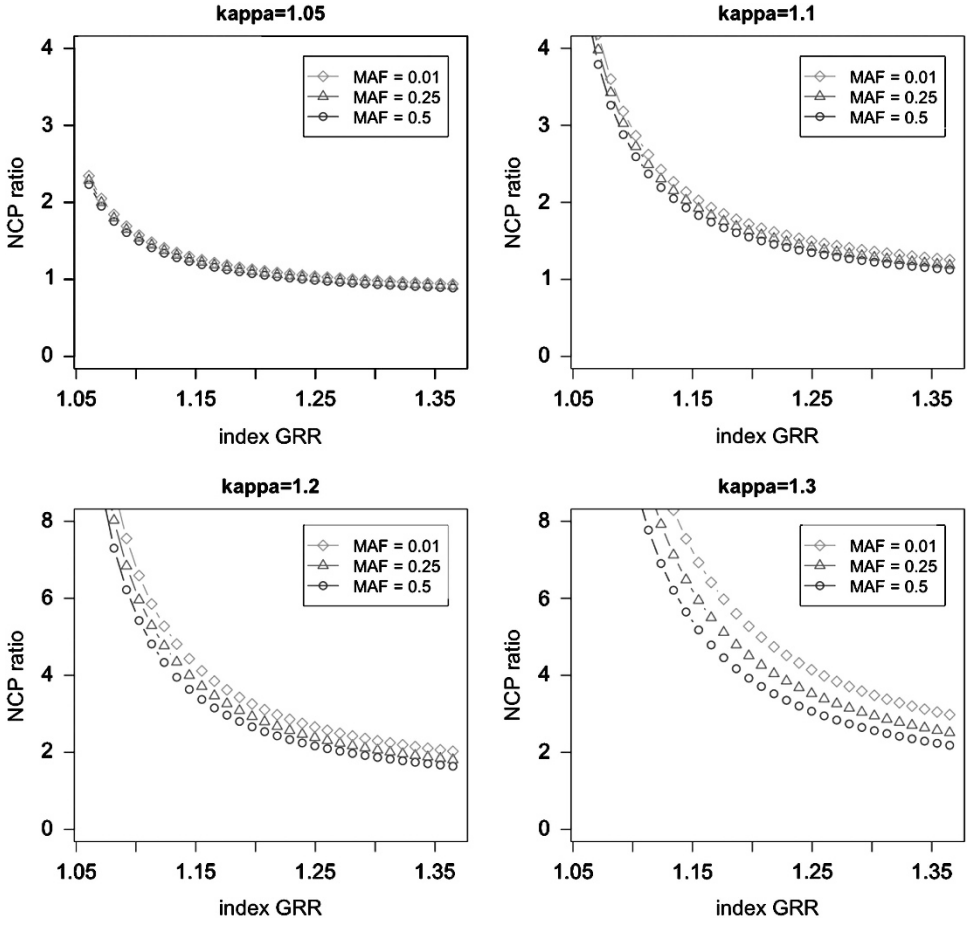

Figure 3 Ratio in power between an analysis considering all cases to an analysis considering a subset of cases for different minor allele frequency with fixed proportional increase in genotypic relative risk (kappa (κ)) in the subset. MAF, minor allele frequency; kappa (κ), proportional increase in genotypic relative risk.

All statistical analysis was performed using the *R* statistical software. All formulae and scripts used to generate plots are available from https://sites.google.com/site/mtraylor263/software/case-subgroup-power-analysis.

DISCUSSION

We have developed a framework that elucidates the power relationship between a full GWAS analysis, and analysis of a subgroup of the cases in which genetic effects were stronger, while retaining all controls. We derived an expression for the ratio in power between a subgroup analysis and a full analysis and used this to study the power properties of subgroup analyses. A simplifying assumption regarding the frequency of genetic variants in the two analyses enabled the broad properties of the power ratio to be studied. This assumption is valid for GWAS, particularly where controls normally exceed cases by at least twofold. This enabled identification of two important results.

First, it was shown that, as GWAS sample sizes increase, and the detectable genetic effects of SNP variants become smaller, the power of a subset analysis in which variants have stronger effects becomes proportionally greater than a full analysis. Calculating *NCP* ratios for ratios of GRR (κ) in the full and subset analyses supported this observation: at lower index GRR, *NCP* ratios increased dramatically. This clearly shows that, when attempting to identify genetic variants with smaller effects, improvements in power become more substantial for subgroup analyses, particularly for index GRR < 1.15. In a recent GWAS meta-analyses of rheumatoid arthritis,[24] schizophrenia,[6] multiple sclerosis,[25] Alzheimer's disease,[2] coronary artery disease,[3] and breast cancer,[26] only 39 of the 339 associated variants showed overall odds ratios > 1.15, suggesting that the majority of variants with effects greater than this threshold have now been identified. Our results show that analysing homogeneous disease subgroups forms a powerful strategy to identify further variants with effects in this range, as stronger effects may be present. Second, we also showed that the relative increase in power for a subset analysis is consistently greater for rare as opposed to common variants, although this effect was more modest. These results imply that searching for rare variants will particularly be aided by subtyping of disease cases into genetically distinct groups. Importantly, both of these results also hold for a lower case:control ratio (Supplementary Figure S1). Several methods can be used to identify genetically distinct disease subgroups in which a stratified GWAS analysis could be performed. Genetic correlations between disease subgroups can be calculated using the GREML methods,[27] which use linear mixed models to obtain estimates of the genetic correlation between the groups. This approach showed shared genetic susceptibility to psychiatric disorders[28] and that Tourette syndrome and obsessive-compulsive disorder are genetically distinct.[29] The approach can easily be adapted to interrogate disease subgroups, where a low genetic correlation in a well-powered sample

would imply distinct genetic architecture. Similarly, genetic risk profile scoring, in which the cumulative effect of genome-wide SNPs is used to test for differences between sets of cases and controls, can be used for the same purpose.[30] This has been used in analyses comparing multiple sclerosis with amyotrophic lateral sclerosis[31] and Parkinson's disease with Alzheimer's disease.[32] Genome-wide genetic correlations between diseases can be calculated in combination with these methods using the framework created by Dudbridge.[33] Finally, polygenic rare variant analysis approaches can be used to identify disease groups that have a polygenic contribution from rare variants and can be used to identify disease subgroups.[34] These methods will be valuable for identifying the genetically distinct groups that would benefit from further association analysis.

A compelling example of the benefits of subgroup analyses can be found in ischaemic stroke, where subtyping of cases based of clinical and radiological criteria has enabled identification of the first common variants associated with the disease.[13,14,35–37] Importantly, in the largest meta-analysis to date, all associations were with ischaemic stroke subtypes, and these showed much stronger association than in an analysis with all ischaemic stroke ($\kappa = 1.24$, 1.23, 1.19, and 1.11 for rs2107595 (*HDAC9*), rs6843082 (*PITX2*), rs879324 (*ZFHX3*), and rs2383207 (*9p21*), respectively).[13] Further to this, analyses of early onset cases suggest even stronger associations with young onset cases at these loci.[19] This example clearly shows that with careful subtyping, genetic studies can provide new information on heterogeneous diseases, such as stroke.

Several other considerations should be made when interpreting our conclusions. First, our results are expressed in terms of genotypic relative risk. For rarer diseases, these values are almost equivalent to odds ratios, the preferred measure in most GWAS studies. This approach may therefore be more intuitive to researchers. However, an alternative approach would have been to benchmark our comparisons in terms of the variance explained by a locus. In particular, this may affect comparisons across the allele frequency spectrum. Second, our approach does not take into account the multiple testing correction, which it might be appropriate to make when performing multiple analyses on a single dataset. Factoring this correction into the analyses would have the effect of increasing the GRR required to achieve equivalent power in the subset analysis. Third, it should be noted that, for complex diseases, the underlying genetic architecture is unknown. Therefore the optimal approach to splitting the data into homogeneous groups remains elusive. Indeed, in some diseases, splitting the cases into groups may not prove beneficial. Interrogation of GWAS data sets with GREML methods,[27] polygenic scoring,[30] and polygenic rare variant association methods,[34] as discussed above, may help to shed some light on this.

Many have been critical of GWAS for not identifying a large proportion of disease variance and for only identifying risk variants with small effects.[38,39] However, GWAS have been very successful, particularly in auto-immune and metabolic disorders, where hundreds of associated genetic variants have been identified.[4] Our results strongly advocate a renewed effort to identify genetically distinct disease groups with increased phenotypic homogeneity within existing data sets, in which power to detect genetic variants with small effects will be greater. This will be particularly important as the focus of genetic studies turns from common to rare variation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Matthew Traylor is funded by a Stroke Association Project Grant (TSA 2013/01). We acknowledge support from the National Institutes of Health Research Biomedical Research Centre at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

1 Franke A, McGovern DP, Barrett JC *et al*: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2010; **42**: 1118–1125.
2 Lambert JC, Ibrahim-Verbaas CA, Harold D *et al*: Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 2013; **45**: 1452–1458.
3 Deloukas P, Kanoni S, Willenborg C *et al*: Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet* 2013; **45**: 25–33.
4 Visscher PM, Brown MA, McCarthy MI, Yang J: Five years of GWAS discovery. *Am J Hum Genet* 2012; **90**: 7–24.
5 Morris AP, Voight BF, Teslovich TM *et al*: Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012; **44**: 981–990.
6 Ripke S, O'Dushlaine C, Chambert K *et al*: Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013; **45**: 1150–1159.
7 Ehret GB, Munroe PB, Rice KM *et al*: Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* 2011; **478**: 103–109.
8 Plomin R, Haworth CM, Davis OS: Common disorders are quantitative traits. *Nat Rev Genet* 2009; **10**: 872–878.
9 Mitchell KJ: What is complex about complex disorders? *Genome Biol* 2012; **13**: 237.
10 Girirajan S, Eichler EE: Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet* 2010; **19**: R176–R187.
11 Manchia M, Cullis J, Turecki G, Rouleau GA, Uher R, Alda M: The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. *PLoS One* 2013; **8**: e76295.
12 Beach TG, Monsell SE, Phillips LE, Kukull W: Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005-2010. *J Neuropathol Exp Neurol* 2012; **71**: 266–273.
13 Traylor M, Farrall M, Holliday EG *et al*: Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol* 2012; **11**: 951–962.
14 Bellenguez C, Bevan S, Gschwendtner A *et al*: Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. *Nat Genet* 2012; **44**: 328–333.
15 Anttila V, Winsvold BS, Gormley P *et al*: Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nat Genet* 2013; **45**: 912–917.
16 Padyukov L, Seielstad M, Ong RT *et al*: A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Ann Rheum Dis* 2011; **70**: 259–265.
17 Ohmura K, Terao C, Maruya E *et al*: Anti-citrullinated peptide antibody-negative RA is a genetically distinct subset: a definitive study using only bone-erosive ACPA-negative rheumatoid arthritis. *Rheumatology (Oxford)* 2010; **49**: 2298–2304.
18 Zaitlen N, Lindstrom S, Pasaniuc B *et al*: Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet* 2012; **8**: e1003032.
19 Traylor M, Bevan S, Rothwell PM *et al*: Using phenotypic heterogeneity to increase the power of genome-wide association studies: application to age at onset of ischaemic stroke subphenotypes. *Genet Epidemiol* 2013; **37**: 495–503.
20 Perry JR, Voight BF, Yengo L *et al*: Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genet* 2012; **8**: e1002741.
21 Li Y, Sheu CC, Ye Y *et al*: Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol* 2010; **11**: 321–330.
22 Yang J, Wray NR, Visscher PM: Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. *Genet Epidemiol* 2010; **34**: 254–257.
23 Purcell S, Cherny SS, Sham PC: Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; **19**: 149–150.
24 Okada Y, Wu D, Trynka G *et al*: Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 2014; **506**: 376–381.
25 Beecham AH, Patsopoulos NA, Xifara DK *et al*: Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* 2013; **45**: 1353–1360.
26 Michailidou K, Hall P, Gonzalez-Neira A *et al*: Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 2013; **45**: 353–361, 361e351-352.
27 Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR: Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 2012; **28**: 2540–2542.

28 Lee SH, Ripke S, Neale BM *et al*: Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet* 2013; **45**: 984–994.
29 Davis LK, Yu D, Keenan CL *et al*: Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet* 2013; **9**: e1003864.
30 Purcell SM, Wray NR, Stone JL *et al*: Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009; **460**: 748–752.
31 Goris A, van Setten J, Diekstra F *et al*: No evidence for shared genetic basis of common variants in multiple sclerosis and amyotrophic lateral sclerosis. *Hum Mol Genet* 2014; **23**: 1916–1922.
32 Moskvina V, Harold D, Russo G *et al*: Analysis of genome-wide association studies of Alzheimer disease and of Parkinson disease to determine if these 2 diseases share a common genetic risk. *JAMA Neurol* 2013; **70**: 1268–1276.
33 Dudbridge F: Power and predictive accuracy of polygenic risk scores. *PLoS Genet* 2013; **9**: e1003348.
34 Chan Y, Lim ET, Sandholm N *et al*: An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. *Am J Hum Genet* 2014; **94**: 437–452.
35 Holliday EG, Maguire JM, Evans TJ *et al*: Common variants at 6p21.1 are associated with large artery atherosclerotic stroke. *Nat Genet* 2012; **44**: 1147–1151.
36 Gudbjartsson DF, Holm H, Gretarsdottir S *et al*: A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat Genet* 2009; **41**: 876–878.
37 Gretarsdottir S, Thorleifsson G, Manolescu A *et al*: Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Ann Neurol* 2008; **64**: 402–409.
38 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
39 McClellan J, King MC: Genetic heterogeneity in human disease. *Cell* 2010; **141**: 210–217.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)