

Mechanisms of disease

Genetic associations in large versus small studies: an empirical assessment

John P A Ioannidis, Thomas A Trikalinos, Evangelia E Ntzani, Despina G Contopoulos-Ioannidis

Summary

Background Advances in human genetics could help us to assess prognosis on an individual basis and to optimise the management of complex diseases. However, different studies on the same genetic association sometimes have discrepant results. Our aim was to assess how often large studies arrive at different conclusions than smaller studies, and whether this situation arises more frequently when findings of first published studies disagree with those of subsequent research.

Methods We examined the results of 55 meta-analyses (579 study comparisons) of genetic associations and tested whether the magnitude of the genetic effect differs in large versus smaller studies.

Findings We noted significant between-study heterogeneity in 26 (47%) meta-analyses. The magnitude of the genetic effect differed significantly in large versus smaller studies in ten (18%), 20 (36%), and 21 (38%) meta-analyses with tests of rank correlation, regression on SE, and regression on inverse of variance, respectively. The largest studies generally yielded more conservative results than the complete meta-analyses, which included all studies ($p=0.005$). In 14 (26%) meta-analyses the proposed association was significantly stronger in the first studies than in subsequent research. Only in nine (16%) meta-analyses was the genetic association significant and replicated without hints of heterogeneity or bias. There was little concordance in first versus subsequent discrepancies, and large versus small discrepancies.

Interpretation Genuine heterogeneity and bias could affect the results of genetic association studies. Genetic risk factors for complex diseases should be assessed cautiously and, if possible, using large scale evidence.

Lancet 2003; **361**: 567–71

Departments of Hygiene and Epidemiology (J P A Ioannidis MD, T A Trikalinos MD, E E Ntzani MD, D G Contopoulos-Ioannidis MD) **and Paediatrics** (D G Contopoulos-Ioannidis), **University of Ioannina School of Medicine, Ioannina, Greece; Biomedical Research Institute, Foundation for Research and Technology Hellas, Ioannina** (J P A Ioannidis); **Division of Clinical Care Research, Tufts–New England Medical Center, Boston, MA, USA** (J P A Ioannidis, T A Trikalinos); **and Department of Pediatrics, George Washington University School of Medicine and Health Sciences, Washington, DC, USA** (D G Contopoulos-Ioannidis)

Correspondence to: Dr John P A Ioannidis, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece (e-mail: jioannid@cc.uoi.gr)

Introduction

With more than one million polymorphisms (gene variant sites) in the human genome,¹ the number of genetic association studies of various diseases and clinical outcomes is increasing rapidly.^{2–5} In many instances, the same polymorphisms and disease outcomes are being assessed repeatedly, sometimes with conflicting results.⁶ A systematic approach, such as meta-analysis,⁷ can help to analyse the extensive accumulated information.⁸ Previously, we have used meta-analysis to show that initially proposed genetic associations for a range of medical outcomes are sometimes refuted by subsequent evidence.⁶

Furthermore, results of small studies might differ significantly from the results of larger studies; since limited numbers of samples are available, most studies in genetics are small. Experience from other clinical domains suggests that small studies could yield more favourable outcomes than larger studies.⁹ This observation might suggest either genuine heterogeneity, or a bias against publication of small studies with negative results (publication bias).¹⁰

Several statistical tests have been used in other domains of clinical research to examine whether the magnitude of the recorded effect relates to study size. We used such methods to empirically assess a large sample of genetic association studies and their meta-analyses across diverse medical specialties. We addressed the questions: how often does the magnitude of the genetic effect differ in large versus small studies, and do such discrepancies arise in the same instances in which first results differ from subsequent research?

Methods

Selection criteria and database

We updated a database⁶ of 36 meta-analyses, by including 19 new qualifying meta-analyses and replacing four with more recent ones. We identified meta-analyses by searching Medline (most recent search February, 2002) with the terms genetics/polymorphism(s)/allele(s), mutation(s), and meta-analysis (type of publication) or systematic review. We included meta-analyses if they dealt with non-HLA genetic markers, disease outcomes were discrete (binary), 2×2 tables were available for every study, and data were available for at least three studies published in at least two different years. Most endpoints in genetic association studies are binary. When several meta-analyses addressed the same association, we retained only the most up to date one with the largest sample size to avoid duplication. We excluded family-based studies, because their analysis is different from that of other case-control studies. Definition and selection of the preliminary studies and of the genetic contrasts in the 55 meta-analyses followed the rules previously described.⁶ In three instances, in which the first study had been superseded by an expanded, partly overlapping study by the same investigators, we retained the first study, since the comparison of first versus subsequent evidence was one of our objectives. If the extent of overlap was known, we retained the expanded study but subtracted the first study counts.

GLOSSARY**DERSIMONIAN AND LAIRD**

A random effects method for combining results of different studies on the same association; it assumes that the genetic effect might be genuinely different across the populations of the various studies. The test is used to estimate an average population-specific effect and the dispersion of the population-specific effects.

KENDALL'S TAU RANK CORRELATION

A non-parametric correlation test that has been proposed by Begg and Mazumdar as a way to examine whether the results of a study (the effect size) are related to its variance.

MANTEL-HAENSZEL

A fixed effects method for combining results of different studies on the same association; it assumes that there is a unique true genetic effect, and the various studies differ in their observed findings due to chance alone.

Statistical methods

The odds ratio (OR) was used to measure the strength of the genetic association. OR has advantages over risk ratio¹¹ and is more appropriate for assessment of heterogeneity than absolute measures such as risk difference.¹² We estimated between-study heterogeneity (statistical dissimilarity of study results) for each meta-analysis with the χ^2 -based Q statistic (regarded as significant for $p < 0.10$).¹³

Data were combined with both fixed and random effects models. The MANTEL-HAENSZEL fixed effects model¹⁴ ascribes differences between study results to chance alone. Studies are weighted by the inverse of their within-study variance. Random effects allow variation of OR between different studies. The method of DERSIMONIAN AND LAIRD¹⁵ provides an empirical estimate of the between-study variance that is added to the within-study variance of each study. Random effects give wider CIs when between-study heterogeneity exists, otherwise fixed and random effects estimates are similar. Random effects are provided in the results, unless otherwise stated.

There is no absolute definition of large studies. Use of an arbitrary cutoff of 1000 individuals¹⁶ is problematic, since studies of more than 1000 patients are uncommon in genetic epidemiology. Therefore, to identify discrepancies in large versus small studies, we chose to consider all studies on the same genetic association relative to each other. We used methods that test whether large studies give systematically different results from smaller ones.^{17,18} Study magnitude is expressed by the SE of the effect, variance (squared SE), or study weight (inverse of variance); a large study has smaller SE. We used three tests previously proposed in meta-analyses of randomised trials: the KENDALL'S TAU RANK CORRELATION coefficient between the natural logarithm of the OR ($\ln[\text{OR}]$) and its variance;¹⁹ a linear regression of $\ln(\text{OR})$ on its SE, which might be more powerful than Kendall's tau,^{18,20} and a linear regression of $\ln(\text{OR})$ on the inverse of its variance. Least squares regressions were weighted by the inverse of the variance. Regression slopes provide evidence for bias or heterogeneity. These tests might be underpowered when few studies are included^{19,21} and a p value of less than 0.10 is judged to be significant. For weighted regression analyses, the SE of the slope is corrected by dividing by the square root mean square residual.²²

For each meta-analysis, we identified the first published study.⁵ When there was no clear first study, we considered all the possible first studies published in different journals in the same year. We assessed whether the OR in first versus subsequent studies differed beyond chance ($p < 0.05$) with a z statistic—ie, the difference of the natural logarithms of the two ORs divided by the SE of this difference.¹⁶

We investigated whether discrepancies of large versus small studies coexisted with discrepancies of first versus subsequent studies, and whether such discrepancies coexisted with significant between-study heterogeneity.

Analyses used SPSS (version 10.0; SPSS, Chicago, USA), and Meta-Analyst (J Lau, Boston, MA). p values are two-tailed.

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

A list of the meta-analyses, genetic association studies, and preliminary studies included in our investigation can be found at <http://image.thelancet.com/extras/02art8007webappendix.pdf>. Across the 55 meta-analyses (table), we had 579 study comparisons (mean 10.4, median 8 per meta-analysis). Total sample size (alleles or patients) per meta-analysis exceeded 1000 in all but six meta-analyses. 45 study comparisons (in 24 meta-analyses) had sample sizes of greater than 1000. 27 meta-analyses (49%) showed a significant genetic association (random effects). Of these, only 13 showed a 50% increase in genetic risk and none showed a doubling in genetic risk for the genetically unfavourable group. Only five meta-analyses had at least one study with a sample size of greater than 1000 that reached significance on its own; in two of these, there were additional studies with sample size of greater than 1000 but without significance. For 26 meta-analyses (47%), there was significant between-study heterogeneity (table). When we excluded the first study or studies, 22 meta-analyses (40%) showed a significant association by random effects and 17 of the 50 meta-analyses with two or more remaining studies (34%) still had significant heterogeneity.

We noted significant differences in the genetic effects of large versus small studies for ten (18%), 20 (36%), and 21 (38%) meta-analyses based on the three tests of rank correlation, regression on the SE, and regression on the inverse of the variance, respectively (see webtable 1 at <http://image.thelancet.com/extras/02art8007webtable1.pdf>). The largest studies in every specialty provided more conservative findings than those suggested by the overall meta-analysis in 38 cases, whereas the opposite was recorded in 17 meta-analyses ($p = 0.005$). When restricted to the 27 meta-analyses with overall significant genetic associations, in 19 instances the largest study had more conservative findings than the overall meta-analysis ($p = 0.034$).

The results of the first studies differed significantly from the results of subsequent research in 14 (26%) meta-analyses, on the basis of both fixed and random effects (see webtable 2 at <http://image.thelancet.com/extras/02art8007webtable2.pdf>). In all 14 instances, the proposed genetic association was stronger in first than subsequent studies. Another four meta-analyses showed a significant difference between first and subsequent studies by fixed effects only.

Of 26 meta-analyses with significant heterogeneity, 25 had a large versus small discrepancy with at least one of the three tests, or a first versus subsequent discrepancy (figure). In seven meta-analyses, there was no significant heterogeneity, but there was a large versus small discrepancy or a first versus subsequent discrepancy. Finally, in 22 (40%) meta-analyses, there was no evidence of either heterogeneity or bias. In eleven of these 22 cases, the meta-analysis had shown an overall significant genetic association (random effects); in nine of those the association was still significant when the first studies were excluded.

ID	Disease/outcome	Gene (polymorphism), contrast	Sample size* (number of studies)	OR (F) (95% CI)	OR (R) (95% CI)	p Het
1	Myocardial infarction	ACE (insertion/deletion), DD vs DI+II	18 664 (15)	1.20 (1.10–1.31)	1.28 (1.09–1.50)	<0.001
2	Ischaemic heart disease	ACE (insertion/deletion), DD vs DI+II	21 876 (17)	1.16 (1.08–1.25)	1.20 (1.06–1.36)	0.004
3	ICVD	ACE (insertion/deletion), DD vs DI+II	11 394 (6)	1.18 (1.01–1.37)	1.21 (0.98–1.50)	0.187
4	Poor clozapine response	HTR2A (102T/C), CC vs CT+TT	733 (6)	1.64 (1.18–2.28)	1.65 (1.19–2.29)	0.514
5	Poor clozapine response	HTR2A (H452Y), YY vs HY+HH	676 (5)	5.55 (1.15–26.8)	3.37 (0.97–11.6)	0.941
6	Vascular disease	MTHFR (677C/T), TT vs CC	6947 (23)	1.08 (0.95–1.22)	1.12 (0.92–1.37)	<0.001
7	Lung cancer	CYP2D6 (deficient oxidation), poor metabolisers vs others	5162 (14)	0.67 (0.53–0.84)	0.63 (0.42–0.93)	0.013
8	Dementia in Down's syndrome	APOE (ε2/ε3/ε4), allele ε2 vs ε3+ε4	1130 (9)	0.41 (0.22–0.78)	0.45 (0.17–1.17)	0.044
9	Schizophrenia	DRD3 (Bal1), 11+22 vs 12	5121 (25)	1.07 (0.95–1.20)	1.08 (0.95–1.23)	0.204
10	Bipolar affective disorder	MAOA (Fnu4HI), allele 1 vs 2	962 (3)	1.30 (0.96–1.75)	1.42 (0.82–2.46)	0.052
11	Bipolar affective disorder	MAOA (CA), allele 122 vs others	1932 (7)	0.95 (0.75–1.21)	0.95 (0.69–1.31)	0.115
12	Bipolar affective disorder	TH (tetranucleotide repeat), allele 1 vs others	2901 (8)	0.92 (0.78–1.08)	0.97 (0.74–1.26)	0.019
13	Unipolar affective disorder	TH (tetranucleotide repeat), allele 1 vs others	1128 (3)	1.17 (0.91–1.51)	1.17 (0.91–1.51)	0.370
14	NIDDM	KCNJ11/KIR6.2-BIR (E23K), KK vs EK+EE	888 (4)	1.94 (1.30–2.88)	1.93 (1.29–2.87)	0.777
15	Lung cancer	GSTM1 (gene deletion), null/null vs others	9724 (21)	1.17 (1.07–1.27)	1.18 (1.04–1.34)	0.007
16	Lung cancer	CYP1A1 (4889A/G), GG vs AA+AG	2392 (6)	1.45 (0.80–2.62)	2.07 (0.71–6.08)	0.106
17	Lung cancer	CYP1A1 (MspI), +/- vs others	4263 (12)	1.21 (0.87–1.70)	1.26 (0.84–1.89)	0.294
18	Myocardial infarction	SERPINE1/PAI-1 promoter (4G/5G), 4G/4G vs 5G/5G	3381 (9)	1.32 (1.15–1.52)	1.55 (1.16–2.08)	0.001
19	Parkinson's disease	CYP2D6 (1934G-A), allele 4 vs others	7029 (14)	1.17 (1.03–1.32)	1.18 (1.00–1.40)	0.094
20	Essential hypertension	AGT (M235T), allele T235 vs M235	4698 (6)	1.22 (1.06–1.42)	1.44 (1.04–2.00)	0.002
21	Cancer	HRAS/HRAS1 (rare alleles) rare vs common alleles	8542 (24)	1.91 (1.62–2.27)	1.84 (1.54–2.21)	0.364
22	Left ventricular hypertrophy	ACE (insertion/deletion), allele D vs I	8186 (12)	1.08 (0.97–1.21)	1.13 (0.95–1.33)	0.085
23	Bladder cancer	NAT2 (slow acetylation alleles), slow/slow vs others	5836 (20)	1.38 (1.23–1.55)	1.43 (1.20–1.71)	0.010
24	ICVD	APOE (ε2/ε3/ε4), allele ε4 vs others	3632 (9)	1.71 (1.38–2.11)	1.69 (1.37–2.09)	0.449
25	Non-syndromic cleft lip	TGFA (TaqI), allele 2 vs 1	5272 (9)	1.50 (1.23–1.82)	1.58 (1.13–2.21)	0.012
26	Alcoholism	DRD2 (TaqIA), allele A1 vs A2	3826 (15)	1.50 (1.26–1.79)	1.60 (1.19–2.15)	0.002
27	Ischaemic stroke	ACE (insertion/deletion), DD vs DI+II	2160 (6)	1.42 (1.16–1.74)	1.58 (1.11–2.25)	0.022
28	Diabetic nephropathy	ACE (insertion/deletion), II vs ID+DD	5393 (20)	0.73 (0.63–0.83)	0.68 (0.55–0.84)	0.006
29	Neural tube defects	MTHFR (677C/T), TT vs CT+CC	3880 (13)	1.68 (1.38–2.05)	1.67 (1.26–2.23)	0.105
30	Neural tube defects	MTHFR (677C/T) mother, TT vs CT+CC	1955 (8)	1.95 (1.44–2.65)	1.98 (1.46–2.68)	0.844
31	Neural tube defects	MTHFR (677C/T) father, TT vs CT+CC	950 (5)	1.09 (0.62–1.93)	1.15 (0.65–2.05)	0.568
32	Ischaemic heart disease	APOE (ε2/ε3/ε4), ε4/ε3+ε4/ε2+ε4/ε4 vs ε3/ε3	8 962 (9)	1.27 (1.14–1.42)	1.39 (1.11–1.73)	<0.001
33	Ischaemic heart disease	LPL (D9N), ND vs DD	2 022 (3)	1.37 (0.81–2.33)	1.36 (0.80–2.32)	0.861
34	Ischaemic heart disease	LPL (N291S), SN vs NN	13 115 (4)	1.15 (0.91–1.46)	1.15 (0.91–1.46)	0.996
35	Ischaemic heart disease	LPL (S447X), XS vs SS	4067 (5)	0.84 (0.70–1.00)	0.84 (0.70–1.00)	0.969
36	Alcoholic liver disease	CYP2E1/CYP2E1 (RsaI), allele c2 vs others	4178 (9)	1.54 (1.04–2.30)	1.41 (0.78–2.55)	0.119
37	Myocardial infarction	FGB/FGB promoter (455G/A), AA vs GG	1561 (3)	0.68 (0.47–0.99)	0.68 (0.47–1.00)	0.839
38	Myocardial infarction	F5 (1691G/A), AA+AG vs GG	5937 (12)	1.29 (1.03–1.61)	1.32 (1.03–1.70)	0.336
39	Myocardial infarction	F2 (20210G/A), AA+AG vs GG	5637 (7)	1.11 (0.79–1.56)	1.20 (0.80–1.80)	0.281
40	Ulcerative colitis	IL1RN (86-BP DUP), carriers of ε2 vs others	2835 (8)	1.23 (1.04–1.45)	1.18 (0.88–1.57)	0.011
41	CAD	ITGB3 (L33P), A2A2 vs A1A2+A1A1	17 315 (31)	1.07 (1.00–1.14)	1.10 (0.99–1.21)	0.004
42	Fractures	COL1A1 (2046G/T), ss+Ss vs SS	3580 (13)	1.39 (1.17–1.65)	1.43 (1.13–1.81)	0.088
43	Bipolar disorder	DRD3 (Bal1), allele 1 vs allele 2	3392 (9)	1.01 (0.87–1.16)	1.01 (0.87–1.16)	0.545
44	Parkinson's disease	MAPT (allele A0), allele A0 vs others	2090 (5)	1.53 (1.23–1.91)	1.52 (1.22–1.90)	0.445
45	Bulimia	HTR2A (1438G/A), allele A vs G	1126 (3)	1.32 (1.04–1.69)	1.33 (1.04–1.69)	0.427
46	Anorexia nervosa	HTR2A (1438G/A), allele A vs G	3698 (7)	1.36 (1.19–1.57)	1.42 (1.05–1.93)	<0.001
47	Bladder cancer	GSTM1 (gene deletion), null/null vs others	4724 (15)	1.50 (1.32–1.71)	1.54 (1.27–1.86)	0.037
48	SLE nephritis	FCGR2A (R131H), RR vs RH+HH	2801 (24)	1.12 (0.93–1.35)	1.11 (0.88–1.41)	0.123
49	SLE	FCGR2A (R131H), RR vs RH+HH	4708 (21)	1.29 (1.12–1.48)	1.29 (1.10–1.52)	0.218
50	Parkinson's disease	COMT (V158M), MM+MV vs VV	964 (3)	1.31 (0.84–2.04)	1.37 (0.68–2.76)	0.097
51	Recurrent early pregnancy loss	MTHFR (677C/T), TT vs CT+CC	1097 (6)	1.37 (0.95–1.98)	1.31 (0.78–2.20)	0.146
52	Alzheimer's disease	MAPT (extended haplotypes), H1H1 vs H1H2+H2H2	3377 (7)	1.03 (0.89–1.19)	1.03 (0.89–1.20)	0.377
53	Alzheimer's disease	LPR/LPR exon3 (766 C/T), CC vs CT+TT	4097 (8)	1.37 (1.18–1.58)	1.37 (1.08–1.73)	0.021
54	IgA nephropathy	ACE (insertion/deletion), DD vs DI+II	1774 (7)	1.26 (1.02–1.55)	1.26 (0.99–1.61)	0.294
55	Heparin-induced thrombocytopenia	FCGR2A (R131H), RR vs RH+HH	1939 (6)	1.02 (0.81–1.29)	0.81 (0.49–1.34)	0.023

*ie, number of individuals or alleles. For ID 7, 15, 23, and 47 some studies inferred genotype from phenotype. ACE=angiotensin converting enzyme.

APOE=apolipoprotein E. CAD=coronary artery disease; COL1A1=collagen 1a1; COMT=catechol-O-methyltransferase; CVD=cerebrovascular disease; CYP=cytochrome p450. DRD2/DRD3=dopamine receptor D2/D3. F=fixed effects. F2/5=factor II/V. FCGR2A=low-affinity receptor of the Fc domain of immunoglobulin G. FGB=fibrinogen β-chain; GSTM1=glutathione S-transferase M1. Het=heterogeneity. HTR2A=5-hydroxytryptamine receptor 2A. ICVD=ischaemic cerebrovascular disease. IL1RN=interleukin 1 receptor antagonist. ITGB3=platelet glycoprotein receptor IIIa. KIR/BIR=K+ inwardly rectifier channel/beta cell inward rectifier. LPL=lipoprotein lipase. LPR=low-density lipoprotein receptor-related protein. MAOA=monoamine oxidase A. MAPT=microtubule-associated (tau) protein. MTHFR=methylenetetrahydrofolate reductase. NAT-2=N-acetyltransferase 2. NIDDM=non-insulin dependent diabetes mellitus. PAI-1=plasminogen activator inhibitor 1. R=random effects. SLE=systemic lupus erythematosus. TGFA=transforming growth factor A. TH=tyrosine hydroxylase.

Meta-analyses included in our study

There was little concordance between the presence of discrepancies between large and small studies and between first and subsequent studies (see webtable 3 at <http://image.thelancet.com/extras/02art8007webtable3.pdf>). In 17 of 25 meta-analyses in which there was evidence for a large versus small discrepancy

by at least one test, there was no significant discrepancy between the first and subsequent studies. In six of the 14 instances in which we recorded a discrepancy in first versus subsequent studies, there was no large versus small discrepancy with any of the three tests.

Meta-analysis ID	Sample size	Significant heterogeneity	Tests					Significant association (RE)	Significant association (FE)
			First vs subsequent (RE)	First vs subsequent (FE)	Rank correlation	SE regression	1/Var regression		
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									
33									
34									
35									
36									
37									
38									
39									
40									
41									
42									
43									
44									
45									
46									
47									
48									
49									
50									
51									
52									
53									
54									
55									

Heterogeneity and bias

Meta-analysis ID numbers correspond to those in table 1. Total sample size (participants or alleles): 0–1000 (white), 1001–5000 (grey), and >5000 (black). Heterogeneity and tests of bias and heterogeneity are black if significant. Also shown are presence of significant association in the overall meta-analysis by random effects (RE) and by fixed effects (FE): black=significance both in the overall meta-analysis and when the first studies are excluded; grey=significance in the overall meta-analysis that is lost when the first studies are excluded. Var=variance.

Whenever the Kendall’s tau rank correlation was significant, results of both the regression tests were also significant, with two exceptions. The two regression methods showed a high degree of concordance: they were both significant in 17 instances, only one of the two was significant in seven instances, and they were both non-significant in 31 instances.

Significant between-study heterogeneity was related to large versus small discrepancies (OR 3.19 [p=0.12] for Kendall’s tau, 16.37 [p<0.001] for regression on the SE, 19.50 [p<0.001] for regression on the inverse variance) and first versus subsequent discrepancies (OR 3.91 [p=0.043]).

Discussion

Our findings indicated that only 16% of genetic associations identified were subsequently replicated with formal statistical significance, without heterogeneity or bias. In more than half the genetic associations we assessed, we noted either a differential magnitude of effect in large versus smaller studies, or differences in first versus subsequent results. Typically, large studies and subsequent research suggested weak associations or no association at all, compared with strong associations proposed by smaller studies and first research. The strong relation between lack of fit to simple fixed effects models and significance in the statistical tests could indicate bias in genetic association studies. However, these tests might not always distinguish bias from genuine heterogeneity.

Concerns have been voiced with respect to the replication of genetic associations.^{6,23} In a non-quantitative review,²³ only six of 166 putative genetic associations that had been studied several times were consistently replicated with p values of less than 0.05 in 75% or more of the studies. However, another 91 had p values of less than 0.05 in at least two different studies. Claims of significance should be interpreted cautiously, and effect sizes should be compared across studies. Several examples of non-replication with individual-level data meta-analyses for specific polymorphisms have been recorded.^{24,25} Inflated estimates might be a problem in the wider sphere of genetics. Results of simulation studies suggest that upward estimation bias could also affect genomewide scans²⁶ and that penetrance might also be overestimated.²⁷

A structured approach is needed to deal with bias and heterogeneity in this setting. First, overall heterogeneity should be assessed. Second, results of first versus subsequent studies should be routinely assessed to ascertain whether they concur and whether evidence exists for an evolution of the summary effect over time.^{28,29} Third, tests that check for a differential magnitude of effect in large versus small studies should be done. When no evidence of heterogeneity or bias exists, either the results of all studies agree, or inadequate numbers of studies have been completed. When hints of heterogeneity and bias are present, several explanations are possible.

Publication bias¹⁰ or time lag bias³⁰ can occur. Retrospectively, results of small, non-significant, unpublished studies are difficult to locate. Statistical approaches to correct for missing studies³¹ are precarious. In view of the rapid pace of genetics research, publication bias could be prominent. International registries where all investigators contribute their data and international meta-analyses could diminish the threat of publication bias.³²

Misclassification bias from errors in case-control assignment or genotyping errors should also be considered.^{33,34} Confounding by ethnic origin, age, or other variables can also cause bias. Genuine heterogeneity could be due to variation in frequency of alleles, variable effects of linkage disequilibrium for other important genetic markers, variable disease expression, or differential disease susceptibility across the studied populations. In the presence of large biological and environmental variability, genetic effects can differ across different populations or even among generations within a population. However, with highly context-dependent associations, estimates of genetic risk are hard to extrapolate for use in

clinical decision-making. Variables reflecting genuine heterogeneity may or may not be related to study timing and study size. Unfortunately, their in-depth assessment requires detailed information that is rarely collected and reported. A consortium-based approach could also offer advantages in this respect. Since most associations refer to small odds ratios (1.20–1.50), single studies with several hundred participants are greatly underpowered. Several thousands of patients are necessary to address these genetic risk factors, and more than 10 000 individuals need to be studied for adequately powered analyses in the presence of genuine genetic heterogeneity. The resources needed to obtain large-scale evidence should be carefully considered in each case and balanced against feasibility, and the pragmatic usefulness of the sought information. For very heterogeneous and context-dependent genetic variables, even very large-scale studies might not yield valid answers.

Post-hoc explanations of variability can usually only generate hypotheses. To disentangle bias from genuine heterogeneity may often be difficult, since there is no gold-standard test for discrimination. Furthermore, results of any statistical test might indicate significant results by chance alone. *p* values obtained from regression diagnostics can depend on the regression model used. Meta-regressions with random-effects modelling might lead to higher *p* values. The rank correlation could have lower type 1 and higher type 2 error than regression diagnostics. Assessment of the strength of heterogeneity or bias can be useful, as can focusing on the absolute magnitude of postulated genetic effects.

Our results suggest that most associations have small or modest effect sizes. Although we included a large sample of genetic associations, these were selected from existing meta-analyses. For many genetic associations there have been only one or two small genetic studies. However, there is no strong reason to believe that bias or heterogeneity would be more or less likely to occur on such occasions. For many genetic associations, truly large studies with thousands of participants might never be done. Therefore, information with respect to genetic associations should be thoroughly examined and cautiously scrutinised before its integration into clinical practice.

Contributors

J P A Ioannidis had the original idea for the project and wrote the first draft of the research protocol. All authors discussed and formulated the final research design. D G Contopoulos-Ioannidis and E E Ntzani extracted data. T A Trikalinos and J P A Ioannidis did statistical analyses, which were discussed and interpreted by all authors. J P A Ioannidis drafted the final report, which was edited and revised by all authors.

Conflict of interest statement

None declared.

Acknowledgments

The investigation was funded by a PENED grant from the General Secretariat for Research and Technology, Greece, and the European Union.

References

- The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; **409**: 928–33.
- Risch N, Merikangas K. The future of genetic studies on complex human diseases. *Science* 1996; **273**: 1516–17.
- Gambaro G, Anglani F, D'Angelo A. Association studies of genetic polymorphisms and complex disease. *Lancet* 2000; **355**: 308–11.
- Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; **405**: 847–56.
- Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev* 2001; **2**: 91–99.
- Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29**: 306–09.
- Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet* 1998; **351**: 123–27.
- Khoury MJ, Little J. Human genome epidemiologic reviews: the beginning of something HuGE. *Am J Epidemiol* 2000; **151**: 2–3.
- Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA* 1998; **279**: 1089–93.
- Dickersin K, Min YI. Publication bias: the problem that won't go away. *Ann N Y Acad Sci* 1993; **703**: 135–46.
- Feinstein AR. Clinical epidemiology: the architecture of clinical research. Philadelphia: WB Saunders, 1985.
- Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J Clin Epidemiol* 2001; **54**: 1046–55.
- Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med* 1997; **127**: 820–26.
- Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959; **22**: 719–48.
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986; **7**: 177–88.
- Cappelleri JC, Ioannidis JP, Schmid CH, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? *JAMA* 1996; **276**: 1332–38.
- Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, 1997; **315**: 629–34.
- Sterne JAC, Egger M, Davey Smith G. Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, 2001; **323**: 101–05.
- Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; **50**: 1088–101.
- Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 2000; **53**: 1119–29.
- Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Stat Med* 2001; **20**: 641–54.
- Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev* 1987; **9**: 1–30.
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002; **4**: 45–61.
- Heo M, Leibel RL, Boyer BB, et al. Pooling analysis of genetic data: the association of leptin receptor (LEPR) polymorphisms with variables related to human adiposity. *Genetics* 2001; **159**: 1163–78.
- Ioannidis JP, Rosenberg J, Goedert J, et al. Effects of CCR5-Delta32, CCR2-64I, and SDF-1 3'A alleles on HIV-1 disease progression: an international meta-analysis of individual patient data. *Ann Intern Med* 2001; **135**: 782–95.
- Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 2001; **69**: 1357–69.
- Begg CB. On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst* 2002; **94**: 1221–26.
- Ioannidis JP, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proc Natl Acad Sci USA*; 2001; **98**: 831–36.
- Lau J, Antman EM, Jimenez-Silva J, et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992; **327**: 248–54.
- Ioannidis JPA. Effect of statistical significance of results on the time to completion and publication of randomized efficacy studies. *JAMA* 1998; **279**: 281–86.
- Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; **56**: 455–63.
- Ioannidis JPA, Rosenberg PS, Goedert JJ, O'Brien TR. Meta-analysis of individual participant's data in genetic epidemiology. *Am J Epidemiol* 2002; **156**: 204–10.
- Bogardus ST Jr, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. *JAMA* 1999; **281**: 1919–26.
- Kelsey JL, Whitmore AS, Evans AS, Thompson WD, eds. Methods in observational epidemiology, 2nd edn. New York: Oxford University Press, 1996.