Int. J. Epidemiol. Advance Access published March 15, 2006

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology
doi:10.1093/ije/dyl003

Commentary: Grading the credibility of molecular evidence for complex diseases

John P A Ioannidis

Accepted 11 August 2005

Keywords Epidemiology, evidence, grading, bias, false discovery, complex diseases, Bayesian credibility

Dissecting the aetiology of complex diseases has been a great challenge for biomedical research, including epidemiology. Several thinkers,[1–4] including Buchanan *et al*.[5] recently, have focused on the unquestionable difficulties of this ambitious enterprise and the great obstacles encountered in the way. Some of them have ended up with a futility outlook. Over more than a decade, the debate has ranged wild on whether epidemiology has reached its limits,[6] is either dead or in a vegetative state, should call it a day, and whether 'it is time for scientists to re-think the quest' and realize that 'base metal cannot be turned to gold'.[5]

In this commentary, I will not argue whether the challenges for attaining evidence are formidable in the current, molecular era. I will certainly not propose that the 'hunger of the paying public for easy answers and promises'[5] should be superficially satisfied. However, I will argue that not only the public, but scientists also, have been starving until now for some tractable knowledge. I will also argue that despite all shortcomings, molecular evidence, with molecular epidemiology as a centre piece, does have a future. I will try to demonstrate that reasonable progress can be achieved; that progress will require scientific humility and the realization that many postulated research findings have been false; that false discoveries will continue to be very common; and that we need to adapt from the concept of solid knowledge taken for granted to the concept of tentative information that should be replicated and scrutinized. Finally, I will propose a hierarchy for grading the credibility of molecular evidence in complex diseases that emerges under the current circumstances.

Scientific prehistory (stone age until approximately early 21st century?)

Critics of molecular evidence contrast the difficulties that arise in trying to understand complex diseases against a glorious past of biomedicine. As Buchanan *et al*.[5] claim, the failure of current research is in 'stark contrast to prior decades of success in which both epidemiology and human genetics uncovered major causal risk factors . . . and lent [their] fields . . . a well-earned authority.' I greatly respect past successes of these fields, even more so, since they were achieved with limited means and crude methods. But epidemiology and genetics are not dead or finished disciplines—they are just starting now.

Biomedicine until the mid-20th century had been to a large extent a compilation of unfounded beliefs and often dangerous practices, variously infiltrated by vague dogmatic theories derived from the equally or even more immature and overconfident physicochemical sciences. Who would like to defend an era before the mere existence of randomized trials and when even demonstrating major risk factors such as smoking for lung cancer met with enormous resistance from the establishment,[7] while academic stupidities circulated at large? As we get closer to our times, we may have more resistance in admitting the limitations of our professions and scientific disciplines. Nevertheless, we have increasing empirical evidence that the performance of biomedical research has not been that spectacular to date—in any front, be it basic science, preclinical science and epidemiology, or clinical research. At a minimum, there is certainly much room for improvement. Among 101 publications in major basic science journals between 1979 and 1983 that made clear promises for resulting in major clinical applications, only one materialized in a widespread clinical application in the subsequent 20–25 years.[8] The translation of basic science in less prestigious journals and among the vast majority of basic research that does not even imagine being translated is unknown, but it is likely to be even less efficient.[9,10] Much of the 'basic' science research to date is confined to 'focused' observations recycled within esoteric circles of similarly oriented sub-sub-specialists without any material consequence whatsoever.

When it comes to epidemiological investigation and clinical research, the situation is not necessarily better. There is increasing concern that the quality of both epidemiological investigations[11] and clinical research, including clinical trials,[12–14] has been quite poor until now. Exceptions of brilliant and well-designed studies only reinforce the rule. It is not surprising then that the findings of epidemiological and clinical research are very often refuted. In an evaluation[15] of the 45 most-cited (over 1000 citations each) articles finding significant effects for interventions published between 1990 and 2003, five out of six non-randomized studies had already been either clearly refuted by subsequent research or found to have proposed seemingly exaggerated effects by 2004. Even

Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece E-mail: jioannid@cc.uoi.gr

among randomized trials, a quarter had already been refuted or found to be exaggerated.

These data pertain to highly-visible research published in the best, most competitive journals. Less visible research sometimes has been so bad that refutation is not even an option. A scientific claim can be refuted only when it is coherent enough to allow refutation. Much research in the past has been performed with bad methods, for the wrong reasons, under heavy conflicts of interest, or with combinations of these flaws. Even randomized clinical trials suffer from publication bias,[16] time lag bias,[17] and extensive selective reporting.[18] As Doug Altman[19] pointed out several years ago, poor medical research is a scandal.

Are 'macroscopic' risk factors defendable?

Defenders of the 'macroscopic' epidemiology of the pre-molecular era might think of several reasons why to prefer traditional risk factors over mysterious molecular risk factors. Postulated risk factors like obesity, consumption of fruits and vegetables, or consumption of coffee are readily tangible. One can communicate to every lay person that obesity is bad, eating fruit and vegetables is good, and drinking coffee is bad, or good, or does not matter. Anyone can understand what obesity, a peach, or a cup of coffee is. Conversely, most of the genes and gene variants unearthed by molecular research (even when they have a name that is not created simply by a cryptic string of vowels, consonants, and numbers) remain a mystery to the layperson and sometimes also to the scientist-expert. The other postulated problem with most candidate molecular risk factors is that usually one cannot modify them—to date at least. One can advise people to lose weight, eat cucumbers, and drink black coffee in moderation. One cannot tell them to change a single nucleotide polymorphism (SNP) that they already carry in their genome. Last, one might argue that if 200 molecular factors act cumulatively, additively, synergistically, or competitively along with 200 multifarious environmental exposures to generate obesity, why not just focus on the final visible, tangible macroscopic risk factor, i.e. plain body fat, rather than try to chase 400 molecular ones?

The reason why we have to go after these 400 risk factors is that we really have no other good choice. We have tried our best with macroscopic risk factors and this is what we got. A few years ago, the obesity epidemic was estimated to account for an extra 300 000 deaths in the USA,[20] while recently it was claimed that there is hardly any independent effect on mortality, and probably more deaths are caused from underweight than from overweight.[21] Some studies had found that fruit and vegetables reduce the breast cancer risk 10-fold,[22] while recent large cohorts find no effect at all.[23] Coffee consumption was one of the first major 'risk factors' to be proposed and then retracted as a causative factor for pancreatic cancer.[24,25] Buchanan *et al.*[5] provide an excellent list of irreproducible findings of the past. Furthermore, even in the instances where reproducibility has been relatively good, the premise that these factors can be easily modified is spuriously optimistic. Finally, I do not believe that the ability of tabloids to accurately name or describe a risk factor should be taken as scientific proof for its importance. I think it simply means that many low-yield 'macroscopics' should be left to tabloids and epidemiology should proceed to the molecular era.

If macroscopic risk factors are indeed the result of the interplay of hundreds of other more proximal risk factors then it should not be surprising that our results with traditional risk factors have been so unstable to replicate. Large heterogeneity is only to be expected. Results obtained with such risk factors may be difficult, if not impossible, to generalize to other patients and populations. Traditional risk factors are very crude composites of very heterogeneous risk quanta each of which may have a small impact on the macroscopic risk factor and on the final outcomes of interest. Obesity may be possible to describe eventually as a composite of 3000 different sub-types, and fruit consumption may be an extremely complicated biological phenomenon to which we may be doing gross injustice by simply measuring rations consumed per day.

Is epidemiology a match for biology and aetiology?

Molecular 'progress' may be difficult, if common complex diseases are the result of chance or if these diseases largely have a highly 'private' genetic epidemiology. For example, if each case of a lethal disease is caused by a new mutation and each new case represents still a new genetic variant that has not been previously encountered, we can keep cataloguing this knowledge but this will not be of prognostic value for the general population. Even then however, we might gain some interesting scientific knowledge (as opposed to practical, usable knowledge) from the mere evolving catalogue. Chance and private genetic epidemiology may have their share indeed for the aetiology of complex diseases. In fact one might consider 'chance' as the private environmental epidemiology, equivalent of and working with or against the private genetic epidemiology.

However, probably for most common and important complex diseases, there is a sizeable component of their aetiology that can be ascribed to measurable and reproducible risk factors, be it genetic or environmental, or both. Evolving and improving multivariate predictive models may never reach a coefficient of determination, R^2, of 1.00. However, it is reasonable to expect that we can improve these coefficients of determination for most complex diseases by starting to incorporate molecular factors. Examining how much we can improve them for each disease is actually a very interesting scientific question on its own. Current predictive knowledge is probably largely over-fit, biased, unchallenged, and non-validated. Most claims about strong predictors and large R^2 for many diseases are exaggerated and would not stand the test of large-scale independent validation in unbiased assessments.[26] The threat for poor validation and exaggerated expectations for diagnostic accuracy and prognostic performance of multivariate models may be even greater in the molecular era.[27–30] We may need to take a step back and acknowledge that we know less than we think we do. For many diseases, we may have to acknowledge that we know practically nothing that can stand much testing; anything we learn from now on would be a plus, even if biology and aetiology prove to be private to some extent.

A flood of candidate risk factors

Molecular medicine and molecular epidemiology are characterized by a rapid, exponential increase in the number of putative risk factors that can be measured. A couple of decades ago, the usual practice had been to focus entirely on a single candidate risk factor and at most consider a handful of other measured variables as adjustments to account for confounding. Ten years ago, measurements in the most complicated studies could master only a few dozens of candidate risk factors. Single experiments can currently measure tens and hundreds of thousands of candidate risk factors at the same time. This is typically exemplified by the expanding applications of genomics, proteomics, and metabolomics. It translates to over a 100-fold increase per decade in the candidate risk factors targeted in a typical study. If the trend continues, studies with billions of candidate risk factors may become possible before my generation passes away!

Even if the numbers of candidates reach a plateau soon, the number of potential candidates is already formidable. It is unlikely that many of the probed candidates are really important. Even if chance and private epidemiology do not account for a big proportion of the aetiology of common complex diseases, there is probably a finite number of molecular risk factors that are operating for each disease.

The smaller the average effect size for each of these risk factors, the larger their likely number. As shown in Table 1, we have evidence from studies to date that effect sizes in molecular epidemiology[31–38] may cover the same range as the effect sizes described in pre-molecular medicine and epidemiology. Small effect sizes are not peculiar to the molecular era. In fact, the overwhelming proportion of epidemiological associations ascertained or proposed in the past have been small effects. This applies also to experimental clinical designs. The typical examples are medical treatments: most established medical treatments to date have very modest relative risk reductions, in the range of 10–40%, i.e. small or very small effects. However, the demonstration of such effect sizes has been left to randomized experiments, when it comes to treatment efficacy. It has been questioned whether epidemiological studies can function as well in this range.[39,40] There is thus some understandable agony as epidemiological methods are relied upon to try to capture and validate small or very small effects.

Regardless of the agony, since small and very small effects were being pursued already by macroscopic epidemiology, the challenge is not new. Moreover, I see no reason why someone should worry because epidemiologists *can* measure now a lot of variables. We had good reason to complain in the past that we could only measure a few variables; that our measurements were crude; and that most of the essential players remained unmeasured and unapproachable, collectively buried under residual confounding. Now we can measure a lot of things and the ability to measure keeps improving. This is good news.

There are also other issues to consider this exponential improvement of measurement agility. The availability of tons of complex data may lead to greater temptation to perform data dredging and selectively report the most promising, but biased, results.[30,41] This is certainly a threat and it is probably happening extensively currently as new methodologies appear and investigators are probing into their capabilities. Conversely, the opposite trend may prevail when these scientific disciplines mature. When tons of data are available, the conscious, unconscious, and subconscious need of investigators to data dredge for 'significant' results may diminish. A whole-genome association study of 500 000 SNPs is likely to yield many thousands of putative risk factors and even with a couple of rounds of replication experiments, several dozen candidates may survive for further testing. A mature scientist is faced with the problem of still having too many risk factors on his plate to pursue. Why data dredge then?

Compare this full plate against the epidemiologist of old who took a decade to collect the data on one or two candidate risk factors and then, starved of statistical significance, would be willing to go wild on data fishing to come up with some respectable, but probably spurious, third-level sub-group interaction effect to promote his/her career. In all, the data manipulation is likely to be less if 500 000 candidates are screened all at once rather than if 500 000 candidates were to appear on the screen one-by-one over 1 000 000 years. We need to get to the bottom of it, so let us get there as soon as possible.

Table 1 Effect sizes in the pre-molecular era and in the molecular era[a]

Effect sizes	Putative frequency	Typical examples of postulated risk factors	
		Pre-molecular era	Molecular era
Large (RR > 5)	Rare	Smoking and lung cancer	APOE and Alzheimer's disease[31]
			BRCA1 and breast cancer[32]
Moderate (RR 2–5)	Uncommon	Moderate obesity and cholesterol gallstones	NOD2 and Crohn's disease[33]
			HLA shared epitopes and rheumatoid arthritis[34]
Small (RR 1.2–2)	Common	Racial descent and hypertension	FcγRIIa and SLE[35]
			GSTM1 and bladder cancer[36]
Very small (RR 1–1.2)	Unclear frequency[a]	Passive smoking and lung cancer	GSTM1 and lung cancer[37]
			MTHFR and ischaemic stroke[38]

RR: relative risk.

[a] Presented examples reflect current state of knowledge and are subject to possible refutation in the future; for small and very small effect sizes, it is uncertain whether these risk factors are true, even when evidence is based on large sample sizes from several studies.

Solid knowledge vs tentative information

This new perspective requires a paradigm shift. Until now, scientists were eager to find, discuss, publish, and disseminate information that they felt was true. Refutations of such information created agony, debate, conflicts, and public upheavals. We should now recognize that most of the biomedical information that is likely to be found, discussed, published, and disseminated will be false. Instead of solid knowledge, we should get used to the notion of tentative information. Any single study in the molecular era, no matter how well-designed, well-conducted, well-analysed, and well-presented is probably more likely to be refuted rather than validated. This does not mean we should discredit these data. We should just accept them for what they are: tentative information, some of which, a small portion maybe, may eventually reach higher levels of credibility, while much will be refuted.

Given this perspective, one could consider a grading of credibility of biomedical information in the molecular era. It can be proven[42] that in the absence of overt biases, large effect sizes are probably more credible than smaller effect sizes. However, a key modifying parameter would be the ability to replicate molecular findings again and again in diverse studies. Table 2 shows the range of credibility of biomedical information according to the observed effect size and the extent of replication of the finding.

A very small effect size (relative risk < 1.2), even if found formally statistically significant in a sizeable study, is very unlikely to be true in a research environment of massive discovery testing, where thousands and millions of such findings are likely to pop up by chance. Even with extensive replication across many studies, it would still be more likely for this finding to be false rather than true, unless it is corroborated not only by several similar studies but also by other independent lines of evidence.

Small and moderate effect sizes may reach a credibility of 50% or higher, if such extensive replication is available, while for very large effects credibility may reach up to 90 or 95% at times. Nevertheless, no candidate risk factor can be assumed for certain to be 100% true. The vast majority of proposed risk factors will continue to be at the low end of the credibility range. We need to accept that at any time point, 90% or more of our tentative information base included in our journals, web sites, textbooks (or whatever other forms of information-archiving succeed or replace these forms) will be false.

Table 2 Typical credibility of research findings according to effect size and extent of replication

Effect size (relative risk)	Replication	Typical credibility (%)
Large (>5)	None	10–60
	Limited	30–80
	Extensive	70–95
Moderate (2–5)	None	5–20
	Limited	10–40
	Extensive	50–90
Small (1.2–2)	None	<5
	Limited	2–20
	Extensive	10–70
Very small (1–1.2)	None	<1
	Limited	1–5
	Extensive	2–30

Is this new? Others may disagree with my view, but I think that 90% or more of our tentative information base is already false anyhow. Regardless of whether I am correct or not about this, the molecular era makes it clear cut that we need to recognize with humility and judiciousness these credibility limits. This has major implications on how research findings are interpreted and eventually used. It also has major implications on why transparent and comprehensive replication becomes so important. We cannot rely on single studies. Single studies are purely hypothesis-generating, and they are important to respect and register, but they will not provide the final answer. We need transparent and complete cataloguing/registration of all studies that are happening in any specific field. Selective reporting and non-publication of results[43] becomes even more unacceptable. Global collaboration and transparency are essential. Evidence is not static, but its credibility needs to be continuously updated and reappraised.[44]

In this direction, consortia of investigators working on the same topic with individual-level information may become the gold standard.[45] These consortia would not hinder individual thinking and brilliant new research ideas, but whatever research findings are produced by individual teams would then be possible to test and validate or refute across the network of other investigators working on the same topic. Such an approach is already being adopted in several fields,[46] and a network of networks has recently been created in human genome epidemiology[45] in an effort to share experiences on how to launch, build, maintain, and expand such networks of investigators. Such networks may also help to keep updating the molecular evidence base of their fields.

Grading of molecular evidence

Given the above considerations, one might consider a grading of molecular evidence, as shown in Table 3. The grading considers five axes: effect size, amount and replication of the evidence, protection from bias, biological credibility, and relevance.

I have already discussed above the first two axes above, effect size and replication of the evidence. One should add some caveats regarding effect sizes. In the presence of bias, the observed effect sizes may actually simply measure nothing else but the average net bias operating in a scientific field. As shown previously,[42] under conditions of bias, larger effect sizes simply mean that more bias is operating—a materialization of the 'too good to be true' situation.[47] Given the massive data being procured, we actually have a better chance of measuring bias empirically in molecular research. Thus the third axis (bias in the evidence) becomes very important. In the presence of demonstrable strong bias, even strong effects that have been observed across several studies may need to be abandoned. Whole fields of scientific inquiry may be dismantled in this way. The problem is that for most currently conducted

Table 3 Proposed grading of credibility in molecular evidence

First axis: Effect size
1.1 Very small or small effect size (relative risk < 2)
1.2 Moderate effect size (relative risk 2–5)
1.3 Large effect size (relative risk > 5)
Second axis: Amount and replication of evidence
2.1 Single or few scattered studies
2.2 Meta-analyses of group data
2.3 Large-scale evidence from inclusive networks
Third axis: Protection from bias
3.1 Clear presence of strong bias in the evidence
3.2 Uncertain about the presence of bias
3.3 Clear strong protection from bias
Fourth axis: Biological credibility
4.1 No functional/biological data or negative data
4.2 Limited or controversial functional/biological data
4.3 Convincing functional/biological data
Fifth axis: Relevance
5.1 No clinical or public health applicability
5.2 Limited clinical or public health applicability
5.3 Considerable clinical/public health applicability

research, we are given very limited or no information by which to judge whether bias has been present or not. Reporting in epidemiological studies is esoteric and elliptical. Hopefully, the efforts to improve transparency in the design, conduct, and reporting of research will remedy the situation.[48,49] Otherwise, it is safe to interpret results as if at least some bias is present, until proven otherwise.

Biological credibility is also important to consider, but we also need more data on how this should be operationalized. It is easy to make *post hoc* claims about biological plausibility.[50] Much of that reasoning may be silly, but there is no validated scale to measure silliness. However, we do start to accumulate data on various molecular associations from other biological/functional avenues of evidence. We need more empirical data in order to understand what exactly they mean, i.e. how much they should change the credibility of some molecular epidemiological findings.[51,52] For example, how credible should associations be when an SNP is located in a non-coding region or a conserved region? How much should the credibility improve when a luciferase experiment demonstrates an effect of this polymorphism on transcription? How much should negative functional data weight on the credibility of an epidemiological association?

Despite some evidence that functional biological data are important,[51,52] it is unavoidable that different experimental conditions may give somewhat different results.[51] The increase or decrease in credibility based on biological and functional data will continue to be a subjective choice. However, we can now start posing questions and dissecting the components of biological plausibility. Given the massive accumulated data, we can start having some empirical evidence on what each piece of biology means. It will not be perfect, but it will be something that we can start measuring. Until now, this speculative part belonged to the Discussion sections of papers and approached poetry more than science. Eventually the likelihood ratios conferred by biological reasoning may become steeper across levels of evidence.

The last axis to consider is the relevance of a molecular research finding. Seen at face value, most research findings have no major clinical or public health relevance. We should be ready to acknowledge this fact. Identifying a risk factor that accounts for 1–5% of the risk of a disease is a venerable target, and most venerable targets are likely to be of this sort. However, this does not mean that screening for this risk factor should automatically be introduced to the general population.[53,54] It is unavoidable that most molecular discoveries will continue to have limited relevance for public health and clinical medicine.

This humility should not be embarrassing. Nor should the public be fed with easy answers. A cursory look through newspapers around the globe would suggest that the final total cure for cancer is discovered many times per week. Scientists should teach themselves and the public that scientific progress is making smaller steps, and these steps should be respected for what they teach us. In particular, we need to learn from our mistakes, biases, and misconceptions. We may never turn base metal to gold, but we still have a lot of fascinating things to learn from the basest of metals in molecular epidemiology.

References

[1] Skrabanek P. Has risk-factor epidemiology outlived its usefulness? *Am J Epidemiol* 1993;**138:**1016–17.

[2] Davey Smith G. Reflections on the limitations to epidemiology. *J Clin Epidemiol* 2001;**54:**325–31.

[3] Davey Smith G, Ebrahim S. Epidemiology—is it time to call it a day? *Int J Epidemiol* 2001;**30:**1–11.

[4] Le Fanu J. *The Rise and Fall of Modern Medicine*. New York: Little Brown, 1999.

[5] Buchanan AV, Weiss KM, Fullerton SM. Dissecting complex disease: the quest for the Philosopher's Stone? *Int J Epidemiol* doi:10.1093/ije/dyl001.

[6] Taubes G. Epidemiology faces its limits. *Science* 1995;**269:**164–69.

[7] Doll R. Fifty years of research on tobacco. *J Epidemiol Biostat* 2000; **5:**321–29.

[8] Contopoulos-Ioannidis DG, Ntzani E, Ioannidis JP. Translation of highly promising basic science research into clinical applications. *Am J Med* 2003;**114:**477–84.

[9] Crowley WF Jr. Translation of basic research into useful treatments: how often does it occur? *Am J Med* 2003;**114:**503–05.

[10] Ioannidis JP. Materializing research promises: opportunities, priorities and conflicts in translational medicine. *J Transl Med* 2004;**2:**5.

[11] Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, Kalish LA *et al*. Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ* 2004;**329:**883.

[12] Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;**323:**42–46.

[13] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273:** 408–12.

[14] Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C *et al*. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;**287:**2973–82.

[15] Ioannidis JPA. Contradicted and initially stronger effects in highly-cited clinical research. *JAMA* 2005;**294:**218–28.

[16] Dickersin K, Min YI. Publication bias: the problem that won't go away. *Ann N Y Acad Sci* 1993;**703:**135–46.

[17] Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* 1998;**279:**281–86.

[18] Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004; **291:**2457–65.

[19] Altman DG. The scandal of poor medical research. *BMJ* 1994;**308:** 283–84.

[20] Allison DB, Fontaine KR, Manson JE, Stevens J, VanItallie TB. Annual deaths attributable to obesity in the United States. *JAMA* 1999;**282:**1530–38.

[21] Flegal KM, Graubard BI, Williamson DF, Gail MH. Excess deaths associated with underweight, overweight, and obesity. *JAMA* 2005; **293**:1861–67.

[22] Katsouyanni K, Trichopoulos D, Boyle P, Xirouchaki E, Trichopoulou A, Lisseos B *et al*. Diet and breast cancer: a case-control study in Greece. *Int J Cancer* 1986;**38:**815–20.

[23] van Gils CH, Peeters PH, Bueno-de-Mesquita HB, Boshuizen HC, Lahmann PH, Clavel-Chapelon F *et al*. Consumption of vegetables and fruits and risk of breast cancer. *JAMA* 2005;**293**:183–93.

[24] MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and cancer of the pancreas. *N Engl J Med* 1981;**304:**630–33.

[25] Hsieh CC, MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and pancreatic cancer (Chapter 2). *N Engl J Med* 1986;**315:** 587–89.

[26] Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19:**453–73.

[27] Ransohoff DF. Lessons from controversy: ovarian cancer screening and serum proteomics. *J Natl Cancer Inst* 2005;**97:**315–19.

[28] Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 2004;**4:**309–14.

[29] Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005;**5:**142–49.

[30] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005; **365:**488–92.

[31] Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R *et al*. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 1997;**278:**1349–56.

[32] Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL *et al*. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* 2003;**72:**1117–30.

[33] Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. *Am J Gastroenterol* 2004;**99:**2393–404.

[34] Ioannidis JP, Tarassi K, Papadopoulos IA, Voulgari PV, Boki KA, Papasteriades CA *et al*. Shared epitopes and rheumatoid arthritis: disease associations in Greece and meta-analysis of Mediterranean European populations. *Semin Arthritis Rheum* 2002;**31:**361–70.

[35] Karassa FB, Trikalinos TA, Ioannidis JP; FcgammaRIIa-SLE Meta-Analysis Investigators. Role of the Fcgamma receptor IIa polymorphism in susceptibility to systemic lupus erythematosus and lupus nephritis: a meta-analysis. *Arthritis Rheum* 2002;**46:**1563–71.

[36] Engel LS, Taioli E, Pfeiffer R, Garcia-Closas M, Marcus PM, Lan Q *et al*. Pooled analysis and meta-analysis of glutathione *S*-transferase M1 and bladder cancer: a HuGE review. *Am J Epidemiol* 2002;**156:**95–109.

[37] Benhamou S, Lee WJ, Alexandrie AK, Boffetta P, Bouchardy C, Butkiewicz D *et al*. Meta- and pooled analyses of the effects of glutathione *S*-transferase M1 polymorphisms and smoking on lung cancer risk. *Carcinogenesis* 2002;**23:**1343–50.

[38] Cronin S, Furie KL, Kelly PJ. Dose-related association of MTHFR 677T allele with risk of ischemic stroke. Evidence from a cumulative meta-analysis. *Stroke* 2005;**36:**1581–87.

[39] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342:**1887–92.

[40] Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG *et al*. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286:**821–30.

[41] Ioannidis JP. Microarrays and molecular research: noise discovery? *Lancet* 2005;**365:**454–55.

[42] Ioannidis JPA. Why most published research findings are false? *PLoS Med* 2005;**2:**e124.

[43] Chalmers I. Underreporting research is scientific misconduct. *JAMA* 1990;**263:**1405–08.

[44] Ioannidis J, Lau J. Evolution of treatment effects over time: empirical insight from recursive cumulative metaanalyses. *Proc Natl Acad Sci USA* 2001;**98:**831–36.

[45] Ioannidis JPA, Bernstein J, Boffetta P, Danesh J, Dolan S, Hartge P *et al*. A network of investigator networks in human genome epidemiology. *Am J Epidemiol* 2005;**162:**302–04.

[46] Ioannidis JP, Rosenberg PS, Goedert JJ, O'Brien TR; International Meta-analysis of HIV Host Genetics. Commentary: meta-analysis of individual participants' data in genetic epidemiology. *Am J Epidemiol* 2002;**156:**204–10.

[47] Higgins JP, Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int J Epidemiol* 2002;**31:**96–104.

[48] Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D *et al*. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001;**134:**663–94.

[49] Standards of Reporting for Observational Studies in Epidemiology (STROBE), ESF Workshop, convened by Matthias Egger, University of Bristol, 2004.

[50] Ioannidis JPA. Genetic associations: false or true? *Trends Mol Med* 2003;**9:**135–38.

[51] Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies. *Nat Rev Genet* 2004;**5:**589–97.

[52] Zhu Y, Spitz MR, Amos CI, Lin J, Schabath MB, Wu X. An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. *Cancer Res* 2004;**64:**2251–57.

[53] Haga SB, Khoury MJ, Burke W. Genomic profiling to promote a healthy lifestyle: not ready for prime time. *Nat Genet* 2003;**34:**347–50.

[54] Khoury MJ, McCabe LL, McCabe ER. Population screening in the age of genomic medicine. *N Engl J Med* 2003;**348:**50–58.