Biostatistics (2012), **13**, 2, *pp.* 195–203 **Invited Editorial**
doi:10.1093/biostatistics/kxr055
Advance Access publication on January 27, 2012

Learning from our GWAS mistakes: from experimental design to scientific method

CHRISTOPHE G. LAMBERT*
Golden Helix Inc., PO Box 10633, Bozeman, MT 59719, USA
lambert@goldenhelix.com

LAURA J. BLACK
College of Business, Montana State University, PO Box 173040, Bozeman, MT 59717-3040, USA and Greer Black Company, PO Box 3607, Bozeman, MT 59772-3607, USA

SUMMARY

Many public and private genome-wide association studies that we have analyzed include flaws in design, with avoidable confounding appearing as a norm rather than the exception. Rather than recognizing flawed research design and addressing that, a category of quality-control statistical methods has arisen to treat only the symptoms. Reflecting more deeply, we examine elements of current genomic research in light of the traditional scientific method and find that hypotheses are often detached from data collection, experimental design, and causal theories. Association studies independent of causal theories, along with multiple testing errors, too often drive health care and public policy decisions. In an era of large-scale biological research, we ask questions about the role of statistical analyses in advancing coherent theories of diseases and their mechanisms. We advocate for reinterpretation of the scientific method in the context of large-scale data analysis opportunities and for renewed appreciation of falsifiable hypotheses, so that we can learn more from our best mistakes.

Keywords: Association studies; Bioinformatics; Experimental design; GWAS; Scientific method.

1. INTRODUCTION

Often deeper investigation into what appears as an isolated occurrence leads to appreciation of a broader problem. Our starting point for this discussion was observing major batch effects in all but 2 of 30 public and private genome-wide association studies (GWAS) that the first author analyzed over the past several years. Problematically, the dependent variable, case–control status, was correlated with the order of sample collection and/or the order in which samples were batch-processed on genotyping instruments. An example of this can be found in the seminal Wellcome Trust GWAS Study (Wellcome Trust Case Control Consortium [WTCCC], 2007). Members of the consortium independently collected and extracted case and control DNA at different sites; then control samples were genotyped on a series of 96-well plates, while case samples were genotyped on another series of plates, sometimes including more than one disease on a given plate—but cases and controls were not block randomized. Because experimental conditions

*To whom correspondence should be addressed.

vary over time and site, with the unit of the plate providing the largest source of variability from one run to the next, the confounding of case–control status with experimental order was severe. Lambert (2010) contrasted "unfiltered" single nucleotide polymorphism (SNP) and copy number variant (CNV) Manhattan plots for the WTCCC study and a block-randomized Alzheimer's GWAS, finding the former riddled with false positives significant at 10^{-20} (SNP GWAS) and 10^{-200} (CNV GWAS) levels and the latter containing no such artifacts. Early mistakes like this are expected. Our concern is that we detect and learn from them. Are we poised to make similar mistakes with high-throughput sequencing, where hundreds or thousands of spurious associations may lead investigators to questionable conclusions? We cannot answer honestly without examining the philosophy and premises that underlie our experimental methods and the culture in which we execute them.

Our GWAS concerns promise relevance for observational studies in general, which remain pivotal to biological sciences. Here we first describe recurring symptoms of a problem emerging in some published GWAS. Next we recall elements of the traditional scientific method to raise questions about how association studies of large observational data sets can advance biological understanding, particularly by embracing and exploring falsifiable hypotheses. Rather than offer answers, we seek to foster substantive conversations, explorations, and yes, more questions about how our field can persevere in the highest-quality research design and implementation that genomic and similar molecular research can provide. The public, medical research practitioners, funding organizations, and we ourselves need these discussions to achieve society's shared goals for health and well-being.

2. SYMPTOMS OF A BROADER PROBLEM

A common symptom of the problems we investigated is confounding of the outcome of interest with experimental order. One of us raised the issue in a blog post (Lambert, 2010), and Leek *and others* (2010) further elucidated the problem. Alarmingly, confounding can occur many times over by, for example, collecting DNA for cases from sites distinct from those providing DNA for controls, running the cases and controls on different genotyping platforms and on separate sets of plates, and then calculating statistical associations between case–control statuses on these questionably measured genotypes. As an illustration, we analyzed a trio study genotyped by the Broad Institute in which plating placed the fathers' DNA samples on one set of plates, the mothers' samples on another, and the children's samples on a third set. As a result, real genetic differences among family members became confounded with errors in genotyping from plate to plate, making results problematic for publication. We have also seen instances in which DNA is extracted from blood in controls and buccal or saliva cells in disease cases, and then hypotheses comparing cases and controls are confounded with performance characteristics of genotyping on these different DNA sources.

If this level of confounding seems a thing of the past, one need to look no further than the recent *Science* paper "Genetic signature of exceptional longevity in humans" (Sebastiani *and others*, 2010), which underwent editorial reevaluation (Alberts, 2010) and subsequent retraction (Sebastiani *and others*, 2011). Interestingly, the editorial concern and subsequent retraction cite a lack of quality control and genotyping errors peculiar to a genotyping platform as the problematic issue, rather than a flawed experimental design. We believe that this study can become a learning path for improving experimental design, if we converse openly among the scientific community.

A "lack of quality control" provides a clue that we are observing a symptom of a broader problem; indeed, a whole class of post-experiment statistical methods has emerged to address confounding. In GWAS, for instance, genotype measurements may be filtered based on call rate for departure from Hardy–Weinberg equilibrium or for having a low minor allele frequency. These methods that remove outliers resulting from flawed experimental design represent a palliative, not a cure. In our view, the GWAS research community has too often accommodated bad experimental design with automated post-experiment

cleanup. We infer this in part because sometimes researchers seem unaware that experimental design can be a source of confounding, resulting in suspect scientific conclusions. While good statistical practice examines every outlier, experimental designs for large-scale hypothesis testing have produced so many outliers that the field has made it standard practice to automate discarding outlying data. In our analyses of GWAS with consistent protocols for sample collection so that avoidable sources of variability are removed, with proper design of the genotyping experiment over plates ("blocking what you can and randomizing what you cannot" [Box *and others*, 1978]), we are learning that post-experiment automated filters are unneeded. Rather, one can examine each signal of association in turn for biological relevance.

3. Reconsidering "experiments"

A first step in increasing clarity in biostatistics experimental design acknowledges the broader continuum in which biostatistical research resides. In the scientific method, an experiment arbitrates between competing models or hypotheses. Unlike with mice and fruit flies, experiments involving humans cannot take the liberty of cloning 2 copies of a person, disabling a gene on one, keeping all other factors constant, and observing the effects. Thus, most studies on humans are observational because it is not practical to fit the system under study into a laboratory setting. In observational studies, therefore, great attention must be given to identifying and accounting for confounding factors.

Even though we often refer to GWAS as "experiments," they are observational studies or quasi-experiments. In this realm, there are 2 useful types of studies. One seeks signals (associations among variables) as inputs for generating hypotheses. The other starts with a hypothesis—or multiple candidate hypotheses (Chamberlin, 1890)—and designs one or more studies that pursue systematic hypothesis falsification. Both kinds of studies using observational data aim to generate and refine cohesive causal theories consistent with, or not contradicted by, numerous observations of the empirical world. In a paper advocating applying Karl Popper's philosophy of science to epidemiology, Buck (1975) acknowledged that some scholars relegated the epidemiologist to the role of data gatherer through association testing. But she invoked the example of John Snow, a "father" of modern epidemiology, who identified the "causal" mechanism of a cholera outbreak in London, as part of the strong deductive-reasoning heritage of the field. Buck characterized many epidemiological studies as "quasi-experiments" but urged epidemiologists to form hypotheses prior to designing collection of observational data and, moreover, to embrace the "exciting process of deduction" since systematic refutation of hypotheses deduced from theory can inform and alter causal understanding.

The obvious benefits of large-scale observational data sets lie in increasing the statistical power and the chances that the findings will be generalizable beyond the sample. In a quasi-experimental context, however, we must consider how to design collection and analysis of observational data to yield the strongest possible findings elucidating mechanisms in studied biological systems. What processes can we establish to increase recognition of noncontrolled variables in experiments? To add nuance to these challenges, we reflect on current practices in data collection and experimental execution.

4. Reviewing data collection and experimental execution in the GWAS context

While an observational study should perhaps be initiated with hypotheses in mind, in current practice, collecting large amounts of biological data on populations of humans or other organisms often commences with an assumption that many hypotheses will be generated "after" data collection. These data sets can yield fruitful exploration, yet we must recognize the risks. When a data analyst using a data set collected by others (for possibly unrelated research questions) forms a hypothesis, he may be unaware of sources of variability in the data. For efficiency's sake, there may be divisions of labor in data collection processes arising from geographic dispersion or quantity of work. From our own experiences, we know that

variations in data collection protocols are common, and locating documentation is difficult. Even analysts who make every effort to understand sources of variability may find little with which to work.

A widespread practice of "borrowing controls" from separate data collection efforts, often for economic reasons, can prove risky, as demonstrated by the Sebastiani *and others* (2010, 2011) paper and subsequent retraction. An important way to retain the large economic savings and statistical power benefits of combining data sets from disparate studies is to block randomize by phenotype and strive for equal proportions of cases and controls in each study. Problems with past studies may be mitigated with filters, imputation, and resequencing, and certainly many valid findings have emerged despite some experimental flaws. But since appropriate design incurs little extra cost, we can hardly justify making similar mistakes going forward, especially as we enter an era dominated by large-scale experimentation with next-generation sequencing.

Divisions of labor arising from specializing in parts of experimental execution may introduce additional separations between hypotheses and experimental design. The trio experiment mentioned above represents an instance of disconnections between laboratory scientists who plated the biological samples and the research analysts and their scientific questions. Some may view divisions of labor among data collectors, laboratory scientists, and research analysts as a necessary evil. But when we consider the high costs of collecting large amounts of biological data and the higher costs of implementing new or changed health care policies, it seems an inordinate and unacceptable risk "not" to involve a statistician in designing both sample collection and protocols for performing subsequent measurements. The growing use of post-experiment "quality control" efforts may in large part arise from failures to obtain substantive statistician participation early in experimental designs. We advocate for ongoing conversations about what approaches offer the most scientifically rigorous ways to probe the data made available in large observational data sets and what processes can ensure coherent experimental design and execution across roles, given gaps in time, space, and research purposes. To explore further, we must acknowledge that association studies using large-scale observational data sets use "hypotheses" in ways different from many other experimental undertakings.

5. "HYPOTHESES" IN THE ASSOCIATION STUDY CONTEXT

Collecting samples without specifying preliminary hypotheses leads to a frame of mind in which genetic assays are run without an experimental design that accounts explicitly for the hypotheses to be tested—in other words, not taking steps to randomize and so prevent biases in data. As with many large biological data collection efforts, GWAS are premised on an overarching hypothesis that a disease has a genetic cause, rather than many more granular hypotheses designed to challenge a causal theory of how a disease manifests. A GWAS might test a million null hypotheses that case–control status and a given genomic SNP are statistically independent over the population sample. The experiment to falsify these null hypotheses is an automated search for indications of association between genetic variants and disease status.

In the traditional scientific method, a hypothesis is a proposed explanation for a phenomenon, but scientists use the hypothesis to deduce additional predicted effects which, if observed, can corroborate the hypothesis but not prove it and, if not observed as predicted, can falsify the hypothesis. A single observation can falsify a traditional scientific hypothesis.

Such a hypothesis might describe a force or property common to all members of a population. A hypothesis of this sort must make universal claims in order to produce repeatable experiments; then a hypothesis tested by experimenting on a member of the population could be refuted for the entire population. If we abstract a hypothesis-forming process this way—"I have observed property X in a number of instances of A; I hypothesize that all other instances of A will have property X"—then the underlying premise is "All A's are effectively identical." If all A's are effectively identical, then we should accept that a corroborated or refuted hypothesis can inform us about the universe of A's. If A's are billiard

balls, and we are looking at properties of how they transfer momentum on a billiard table, we would expect all instances of these A's to have the same properties. If, instead, A's are people, and we seek to identify universal rules about behaviors or effects of environment on disease status applicable to all people, the premise of uniformity can lead us astray (see Meehl, 1990, pp. 200–201, for his discussion of *ceteris paribus*). Once we categorize nonidentical objects together and make cause–effect statements about them, we can speak about population parameters—but in this context a single observation cannot falsify a statistical hypothesis about a population of unlike objects.

What Meehl (1967) called the "strong" form of null hypothesis testing, compatible with Popper's (2002) approach of falsifying a scientific theory, is in line with Fisher's original formulation of a test of significance (Fisher, 1955, 1956; Gigerenzer, 2004; Gill, 1999). Fisher's test specified a null hypothesis which would be falsified if data analysis did not yield a significance level determined by the test statistic appropriate to the assumed or known data distribution. Falsifying hypotheses derived from theory can advance causal understanding in science, and hypothesis testing to "confirm" theories is statistically and scientifically unsound.

Notably, biostatistics design methods have roots in Fisher's agricultural experiments (Fienberg and Tanur, 1966), but some basic principles underlying that work have been overlooked as the methods have migrated to other research realms. Fisher (1956) opposed misuses of null hypothesis testing, and Gigerenzer (2004, p. 589) characterized Fisher's view of null hypothesis testing as "the most primitive type of statistical analyses … used only for problems about which *we have no or very little knowledge*" (italics in original). It is thus appropriate to apply this method when one first examines a disease and has no idea where to look for genetic factors that may be causative, as we do in GWAS. Nevertheless, we can recognize that, without a theoretical context, a hypothesis (whether corroborated or disconfirmed) can do little to enhance understanding of cause and effect. Moreover, without a clear falsifiable stance—one that has implications for the theory—associations do not necessarily contribute deeply to science.

6. Multiple testing revisited

Historically, epidemiology has focused on minimizing Type II error (missing a relationship in the data), often ignoring multiple testing considerations, while traditional statistical study has focused on minimizing Type I error (incorrectly attributing a relationship in data better explained by random chance). When traditional epidemiology met the field of GWAS, a flurry of papers reported findings which eventually became viewed as nonreplicable. As a result, genome-wide thresholds of significance were instituted, and replication in an independent sample became the gold standard for GWAS publication. It would appear the field learned from its mistakes. Consider, though, that when genome-wide significance is not reached, researchers commonly take forward, say, 20–40 nominally significant signals from a GWAS and then run association tests for those signals in a second study, concluding that all the signals with a p-value <0.05 have replicated (no Bonferroni adjustment). Frequently 1 or 2 associations replicate—which is also the number expected by random chance. Then to further "confirm" the signal, the "replicated" signals from the 20–40 signals tested in the second study are combined with the previous study data to compute p-values considered genome-wide significant. This method has been propagated in publications, leading us to wonder if standard practice could become to publish random signals and tell a plausible biological story about the findings.

Few large-scale GWAS have used high-throughput sequencing. Yet there are dramatic success stories of researchers locating the causal variant for a rare disease by sequencing a few affected and unaffected members of a family. What distinguished these is the explicit search for causative variants, working from the hypothesis that the causative variant lies in a protein-coding region and must be present in all affected members of a family and absent in the others. Because such a small number of variants satisfy these and

other filtering criteria, a putative cause can be identified, and predicted effects can be tested, by seeing, for instance, if the mutation causes protein-binding disruption in a subsequent experiment.

7. BUILDING ON ASSOCIATION STUDIES FINDINGS

GWAS research relies on the premise that, if there is a causal link between a genetic variant and a disease, we will see significant association. As noted, however, significant associations do not imply causation. If we can perturb many different parts of a biological system and see a disease, what specifically causes the disease? With association studies, we metaphorically poke different parts of a biological web to assert that a particular strand might cause the disease status. Perhaps, a more aggregated causal explanation is needed, such as if the sum of various components perturb the system's equilibrium by a certain degree—if many strands of the web are stretched beyond a certain extent—the effect is to push the system into a disease state. We also expect that future biostatistics research will require extensive longitudinal pathway measures to elucidate causal relationships in aggregated biological systems.

Because every biological system exists in a varying environment to which it responds, and because delays between sub-part interactions and their effects in observable macro variables can create nonlinearities or tipping points, the interrelationships in biological systems can appear intractable. Although genome research has sometimes been cast as capable of providing unequivocal cause–effect explanations, an association between biological system "case" status and a tiny subset of a biological system such as an SNP may not be as informative as society would hope. Pathway analysis approaches show promise, and we must still guard against publishing more unfalsifiable correlations without causal follow-up. We can also discuss openly analytical approaches that render association more meaningful in the context of interdependent biological systems. In an era of molecular research, we may reconsider the usefulness of the concept of named diseases, classifications that may have been formed long ago based on symptoms of like appearances that may have unlike causes.

8. FALSIFYING HYPOTHESES AND VALUING MISTAKES

For over a century, periodically authors call for renewing the scientific method and warn against seeking confirmation of an explanation of reality, rather than trying to falsify it (see Chamberlin, 1890; Ioannidis, 2005; Platt, 1964, for notable examples of the genre). Calls for clarity and rigor in experimental design remain as relevant as ever, especially since public and private entities invest increasing resources in large-scale biological research. While our purpose here is to raise questions rather than presume that we know answers, we nevertheless suggest some actions that can strengthen our field.

We can undertake to learn methods that help us recognize what we do not know. Scientists take as given the methodologies described in their particular literature and practiced by leaders in their field. The critical mass of methods appearing in the literature matters because, as Kuhn (1962) noted, an accumulation of apparent consensus can discourage diffusion of other valuable approaches because they are viewed as inconsistent with or simply different from the dominant paradigm. A biologist trained in biology alone may find herself unprepared for large-scale experimentation unless she also undertakes systematic study of statistical methods. Let us talk openly about interdisciplinary collaborative approaches that lead us to unlearn previously unquestioned assumptions as well as create new knowledge.

We can refrain from overstating in publication "what science tells us." Both contemporary statistical methods and the traditional scientific method agree that nothing can be accepted as certain. We can articulate clearly limits to generalizing findings from sampled populations. When conducting research to inform public policies, we can advocate for complementary causal mechanism studies even as we recognize that sometimes policy and health practice decisions must be made urgently in light of associations.

A key may lie in editorial boards and reviewers for journals such as this since publication is a major gateway to scientific reputation. Almost all association-based studies assert that the identified signal makes sense in terms of either previously published associations or causal studies of biological mechanisms related to the signal. If journals were to insist that association studies also suggest possible experiments that could falsify a putative theory of causation based on association, the quality and durability of association studies could increase.

Mainstream publications also provide a point of leverage. Since many people do not understand that association is not equal to causation, we can direct public attention to findings and theories that have withstood extensive attempts at falsification. A survey of medical findings reported in newsprint (Bartlett *and others*, 2002) found that observational studies were much more likely than randomized clinical trials to be reported in the press, despite that randomized studies are proportionally overrepresented in research press releases and are less subject to biases common in observational studies. Aware that media favor reporting observational studies, we can diligently report findings using the recommendations of the STROBE guidelines (von Elm *and others*, 2007).

Management scientist and operations researcher Ackoff (2006) said he was often asked regarding his advocacy of systems thinking, "If this way of thinking is as good as you say it is, why don't more organizations use it?" Ackoff answered that the challenge exists in adopting any transforming idea: we are afraid of making mistakes, errors of commission. In a world that punishes mistakes and regularly disregards often more devastating errors of omission, publishing correlations may provide a temptingly reassuring path. Correlations are not easily falsified; nonreproducibility may be explained in many ways, including by population differences. Professionally, potential reward and drastically reduced risk can result from publishing correlations. Applying well-published but perhaps flawed methods of experimental design also presents a path of safety; if one uses a method published in *Nature* or *Science*, reviewers may be less inclined to question one's approach. Consistent with Kuhn, researchers (Young *and others*, 2008) suggest that a pattern of "following the leader" is further amplified by path dependency resulting when "early decisions by a few influential individuals as to the importance of an area of investigation consolidate … the trajectory" of research (p. 1420).

The problem with avoiding mistakes, Ackoff noted, is that we learn, that is, update our mental models of reality, only "through" mistakes, through falsification, not from what confirms our expectations. As long as our culture sets professional impact and personal security at odds with the scientific method's requirement of falsifying hypotheses, its use will be hindered. The scientific progress achieved despite the underlying cultural conflict is amazing—and yet imagine what may emerge if this obstacle were removed. A real lever for scientific progress may lie in how we treat our children, one another, and foremost ourselves with respect to our mistakes.

Persistent inference of multiple falsifiable hypotheses challenging existing theories provides one of the fastest ways to probe the unknown. Popper wrote that bold or daring scientific conjectures set apart great science from the rest, and he called a conjecture "daring if and only if it takes a great risk of being false—if matters could be otherwise, and seem at the time to be otherwise" (Popper, 1985, pp. 118–119). Bold claims that are just as boldly disproved by (even our own) scientific experiments merit scientific praise and publication—and this is what society wants of science, to probe and provide causal explanations on issues of widespread concern. We can contribute by rethinking assumptions and practices underlying statistical analyses of observational data sets to increase the rigor of experimental design, execution, and interpretation in efforts to falsify, systematically, our best explanations of what we think we know.

Acknowledgements

The authors thank the editors and 2 anonymous reviewers for very helpful suggestions. *Conflict of Interest:* None declared.

REFERENCES

ACKOFF, R. L. (2006). Why few organizations adopt systems thinking. *Systems Research and Behavioral Science* **23**, 705–708.

ALBERTS, B. (2010). Editorial expression of concern. *Science* **330**, 912.

BARTLETT, C., STERNE, J. AND EGGER, M. (2002). What is newsworthy? Longitudinal study of the reporting of medical research in two British newspapers. *British Medical Journal* **325**, 4.

BOX, G. E. P., HUNTER, W. G. AND HUNTER, J. S. (1978). *Statistics for Experimenters*. New York: John Wiley & Sons, Inc.

BUCK, C. (1975). Popper's philosophy for epidemiologists. *International Journal of Epidemiology* **4**, 10.

CHAMBERLIN, T. C. (1890). The method of multiple working hypotheses. *Science (old series)* **15**, 92–96; reprinted 1965 **148**, 754–759.

FIENBERG, S. E. AND TANUR, J. M. (1966). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review* **64**, 237–253.

FISHER, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society (B)* **17**, 9.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh, UK: Oliver & Boyd.

GIGERENZER, G. (2004). Mindless statistics. *The Journal of Socio-Economics* **33**, 20.

GILL, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly* **52**, 28.

IOANNIDIS, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine* **2**, e124.

KUHN, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

LAMBERT, C. G. (2010). *Stop Ignoring Experimental Design (or My Head Will Explode)*. http://blog.goldenhelix.com/?p=322.

LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GERMAN, D., BAGGERLY, K. AND IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**, 7.

MEEHL, P. E. (1967). Theory testing in psychology and physics: a methodological paradox. *Philosophy of Science* **34**, 103–115.

MEEHL, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports* **66**, 195–244.

PLATT, J. R. (1964). Strong inference. *Science* **146**, 7.

POPPER, K. (1985). The problem of demarcation, Chapter 8 (1974). In: Miller, D. (editor), *Popper Selections*. Princeton, NJ: Princeton University Press, pp. 118–119.

POPPER, K. (2002). *Conjectures and Refutations*. London: Routledge Classics.

SEBASTIANI, P., SOLOVIEFF, N., PUCA, A., HARTLEY, S. W., MELISTA, E., ANDERSEN, S., DWORKIS, D. A., WILK, J. B., MYERS, R. H., STEINBERG, M. H. *and others* (2010). Genetic signatures of exceptional longevity in humans. *Science* **330**, 912.

SEBASTIANI, P., SOLOVIEFF, N., PUCA, A., HARTLEY, S. W., MELISTA, E., ANDERSEN, S., DWORKIS, D. A., WILK, J. B., MYERS, R. H., STEINBERG, M. H. *and others* (2011). Retraction. *Science* **333**, 404.

VON ELM, E., ALTMAN, D. G., EGGER, M., POCOCK, S. J., GØTZSCHE, P. C., VANDENBROUCKE, J. P. AND STROBE INITIATIVE (2007). Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *British Medical Journal* **335**, 806–808.

WELLCOME TRUST CASE CONTROL CONSORTIUM (WTCCC) (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 18.

YOUNG, N. S., IOANNIDIS, J. P. A. AND AL-UBAYDLI, O. (2008). Why current publication practices may distort science. *PLoS Medicine* **5**, e201.

[*Received April 1, 2011; revised December 14, 2011; accepted for publication December 16, 2011*]