Commentary

Reflections on Publication Criteria for Genetic Association Studies

Colin B. Begg

Memorial Sloan-Kettering Cancer Center, New York, New York

In a recent editorial, the editors outlined stringent new criteria for prioritizing studies for publication (1). This initiative was prompted by the broad perception that the literature is rife with associations of genetic variants with cancer, most of which are not reproduced, and thus considered to be false positives (2). The goal of improving the validity of published results is laudable, and I applaud the editors' initiative to move in the direction of articulating prioritization criteria. However, I have some reservations about the proposed strategies. The editors' statement touched on many themes, and I will focus on the ones that give me cause for concern.

The editors would give priority to publishing associations involving variants with biological plausibility, such as variants with documented functional significance or those with other prior evidence for targeting the variant. However, it is not clear that our knowledge-base is sufficient for an effective sorting of candidate variants in this way. Bioinformatics computational tools are advocated, but these tools are very much at the development stage and their validity is largely untested in the context of identifying cancer susceptibility. Recently, the journal published a study of the SIFT tool, which predicts functional significance on the basis of evolutionary conservation (3). In this study, the tool predicted that 36% of a panel of DNA repair variants are likely to be damaging to protein function versus 76% of a panel of variants of known functional significance. Another study went further and associated the SIFT predictions with reported odds ratios of DNA repair variants from case-control studies, finding a significant trend (4). However, the authors' scatter plot shows wide variability. Both these studies show a high potential for falsely excluding from study variants that could have cancer relevance.

A second theme of the editors is that genes cannot be considered in isolation, that they function within pathways, and that their actions in any case are likely to involve interactions with environmental factors. The fact that genes act in concert does not negate the value of studying them in isolation. The progress of science is incremental. Epidemiology as a field is at its best when it identifies isolated facts with reasonable confidence in carefully designed studies, thereby stimulating further mechanistic investigation. An epidemiologic study that tries to do too much will very likely fail.

The preceding theme is mirrored in the guidelines for statistical analysis, where studies of multiple genes and multilevel interactions are mandated. The editors seem to want analyses that will be directed at providing comprehensive biological modeling of the results, as opposed to the more conventional analyses that address relatively simple hypotheses. This is perhaps the most bothersome recommendation, because it is likely to direct authors away from methods that endeavor to ''let the data speak for themselves'' toward a much more exploratory style. The editors also recommend that studies be accompanied by array or proteomic evaluations of etiologic heterogeneity, and by implication that these analyses should lead to corresponding subgroup evaluations of the genetic variants. Although the search for etiologic heterogeneity is certainly of value, this combination of strategies is likely to encourage data-dredging, and may well promote the kinds of false alarms that the editors seek to avert.

Finally, the editors express a clear preference for confirmatory investigations as opposed to original studies identifying new variants. This preference begs the question, Where will the original findings be published that will lead to the confirmatory studies that the journal seeks? Whereas confirmation is a critical feature of the scientific method, identification of novel associations is the most rewarding stage in the process of establishing cancer risk. These findings do frequently arise in small studies, to be sure, and the majority may indeed be false positives, perhaps to some extent because the studies are typically not as methodologically rigorous as the larger confirmatory studies. But false positives are the inevitable product of the search for associations in the vastness of the human genome, and replication is the tool to sort the wheat from the chaff (5). Good study design, recognition of the limitations of exploratory investigations, and reporting of the results with an appropriate level of skepticism about the conclusiveness of the findings are the strategies to deal with this phenomenon, not discrimination against studies merely because they are small or produce unexpected or novel results.

Despite the many false positives, and the insidious promotion of unsubstantiated findings to the lay press, I believe that the methodologic tradition in epidemiology is strong. At its best, this tradition encompasses concern for careful study design, a respect for the power of data over theorizing (i.e., a tradition of empiricism), an analytic style that favors attention to relatively simple hypotheses, and a commitment to publish results of completed studies regardless of the conclusions. Recently, in some quarters, this tradition has been derided as ''black box'' research, but its merits have been strongly defended also (6). Indeed, most of what we know about cancer risk has emerged serendipitously in empirical investigations. These important aspects of the scientific method are not mentioned in the new publication guidelines, but they deserve prominence. Instead, the guidelines require the use of new methods before they have been shown to be effective in the study of cancer risk (e.g., bioinformatics techniques and ideas for incorporating pathway information into the statistical analyses of risk), they would discard powerful simple end points (e.g., single gene associations) in favor of complex ones that will be difficult for referees to critique and for other investigators to reproduce, and they are likely to encourage an analytic style that will lead to data dredging (the push for the exploration of interactions and the creation of subgroups for analysis based on array

Cancer Epidemiol Biomarkers Prev 2005;14(6):1364–5

techniques). Although I am enthusiastic about further development of all the exciting new techniques mentioned, they should not be allowed to displace the established methodologic tradition. We must not throw out the baby with the bathwater. I suggest that the editors open up these pages to a period of debate on the criteria and their application and reconsider them in the light of that debate before implementing them in practice.

Acknowledgment

I thank Bruce Armstrong for careful review and valuable advice on this article.

References

1. Rebbeck TR, Martinez ME, Sellers TA, et al. Genetic variation and cancer: improving the environment for publication of association studies. Cancer Epidemiol Biomarkers Prev 2004;13:1985–6.
2. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet 2001;29:306–9.
3. Savas S, Kim DY, Ahmad MF, et al. Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. Cancer Epidemiol Biomarkers Prev 2004;13:801–7.
4. Zhu Y, Spitz MR, Amos CI, et al. An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. Cancer Res 2004;64:2251–7.
5. Thomas DC, Clayton D. Betting odds and genetic associations. J Natl Cancer Inst 2004;96:421–3.
6. Greenland S, Gago-Dominguez M, Casteleo JE. The value of risk-factor (''black-box'') epidemiology. Epidemiology 2004;15:529–35.

Reflections on Publication Criteria for Genetic Association Studies

Colin B. Begg

Cancer Epidemiol Biomarkers Prev 2005;14:1364-1365.