Review

The Quality of Meta-Analyses of Genetic Association Studies: A Review With Recommendations

Cosetta Minelli*, John R. Thompson, Keith R. Abrams, Ammarin Thakkinstian, and John Attia

* Correspondence to Dr. Cosetta Minelli, Institute of Genetic Medicine, EURAC Research, Viale Druso 1, 39100 Bolzano, Italy (e-mail: cosetta.minelli@eurac.edu).

Initially submitted December 24, 2008; accepted for publication June 25, 2009.

Although there has been a rapid rise in the publication of meta-analyses of genetic association studies, little is known about their methodological quality. The authors reviewed the quality of 120 randomly selected genetic meta-analyses published between 2005 and 2007. Data extracted included issues of general relevance and other issues specific to genetic epidemiology. Quality was markedly poorer in the 26% of the meta-analyses that accompanied a report on a primary study. Such meta-analyses were predominantly published in specialist journals, and their quality was positively associated with the impact factor of the journal. Among the meta-analyses that did not accompany a primary study, Human Genome Epidemiology reviews tended to score better than the others, although the comparison was limited by relatively small numbers. Comparison of the overall quality with that of genetic meta-analyses published before 2000 showed improvement in both conduct and reporting. However, the quality of the handling of specific genetic issues remains disappointingly low. For a few key general quality issues, the authors compared their findings with findings in other fields of medicine and found that general quality was similar. On the basis of this review, the authors provide practical recommendations for the conduct and reporting of genetic meta-analyses.

epidemiologic methods; genetics; meta-analysis; principal component analysis

Abbreviations: HuGE, Human Genome Epidemiology; HWE, Hardy-Weinberg equilibrium; PCA, principal component analysis.

Meta-analysis—the pooling of results or data across a number of studies—is advocated as a valuable tool in clinical research, not only because it increases the power to detect an association but also because it helps make sense of conflicting results. It is particularly useful in genetic epidemiology, where there has been a massive increase in the number of published primary studies and the proportion of gene-disease associations that have been replicated is disappointingly low (1).

Genuine population diversity may explain part of the lack of replication, but the key reasons are probably methodological, due to small sample sizes and bias (2). Important biases are publication bias and reporting bias, whereby authors only publish a paper if they obtain statistically significant findings or only report those associations which reach statistical significance. One consequence of this is the “first-study effect,” whereby the first study published on a gene-disease association suggests a genetic effect that is not found, or is found with much smaller magnitude, in subsequent studies (3). As a result, new findings are generally treated skeptically until they have been replicated. Meta-analysis plays an important role in assessing that replication and in providing an estimate of the size of the genetic effect.

The importance of evidence synthesis in genetic association research is illustrated by Figure 1, which shows the rapid increase in published meta-analyses of genetic association studies over the last decade. It is less certain whether this increase has been accompanied by an improvement in quality. Previous work evaluating the quality of genetic meta-analyses published up to 2000 showed common flaws in their conduct and reporting (4–6), which may have led to bias and misleading results.

In this paper, we review the methods used in recent meta-analyses of genetic association studies, assess quality, and...
compare that quality with the quality of earlier genetic and nongenetic meta-analyses. Based on the findings of the review, we provide practical recommendations for the conduct and reporting of genetic meta-analyses.

MATERIALS AND METHODS

Review of meta-analyses

Identification of papers. We identified papers published in 2005, 2006, and 2007 by searching the HuGE Reviews Archive (http://www.cdc.gov/genomics/hugenet/reviews_arch.htm), a database of published meta-analyses of genetic association studies maintained by the Human Genome Epidemiology Network. For each year, we randomly selected 40 papers that were published in English and that reported meta-analyses of summary data on gene-disease associations with binary outcomes. We excluded systematic reviews that did not contain a quantitative synthesis of results and meta-analyses based on individual patient data.

Data extraction. Data were extracted using an extraction form (see Web Figure 1, which is posted on the Journal’s Web site (http://aje.oxfordjournals.org/)) with explicit definitions for each field. The topics covered by the form are listed in Table 1 and include issues of general relevance to all meta-analyses and other issues that are only of concern in genetic epidemiology.

For the papers from 2005, data extraction was performed independently by 2 reviewers. Regular meetings were held to reach consensus, discuss the reasons for disagreements, and refine the definitions of the data extraction fields. The final definitions are given in Web Figure 2 (http://aje.oxfordjournals.org/). This iterative process reduced apparent disagreements until we found that remaining discrepancies were due to information being missed by one of the reviewers. For the papers from 2006 and 2007, each article was reviewed by 1 author, and we addressed the problem of overlooked information by using a computer program that searched for a list of keywords in an electronic version of the article (available on request from the authors).

Indicators of quality. Although our study was designed to evaluate individual quality components, we also used scores as convenient summaries of the main features of the papers. Some aspects of the meta-analysis of genetic association studies are controversial, but there are some very basic indicators of good practice and good reporting that should always be present. We created 2 quality scores based on

Table 1. Main Topics Covered by the Data Extraction Form Used in a Review of 120 Meta-Analyses of Genetic Association Studies Published Between 2005 and 2007

| Topics General to All Meta-Analyses | Topics Specific to Genetic Associations |
|------------------------------------|----------------------------------------|
| Search strategy                    | Genetic model                          |
| Inclusion/exclusion criteria        | Consideration of polymorphism prevalence |
| No. and size of the meta-analyses   | Handling of Hardy-Weinberg equilibrium  |
| Main outcome measure               | Use of biomarkers                      |
| Use of graphical displays          | Use of family studies                  |
| Assessment of heterogeneity        |                                        |
| Fixed- or random-effects models     |                                        |
| Subgroup analyses                  |                                        |
| Reporting of study characteristics  |                                        |
| Consideration of publication bias   |                                        |
such indicators: a “general quality score” that depends on aspects relevant to any meta-analysis and a “genetic quality score” derived from aspects specific to genetics. Twenty-eight quality indicators were identified—19 for general quality and 9 for genetic quality—covering both “positive” and “negative” aspects of the conduct and reporting of genetic meta-analyses. We calculated the 2 quality scores by summing positive factors and subtracting negative factors and scaling the results to lie between 0 and 100, using the theoretical maximum and minimum of the sum.

We then performed a principal component analysis (PCA) (7), with the aim of assessing whether 1) indicators would tend to aggregate within the same papers, forming 2 separate clusters for good and poor quality, and 2) indicators subjectively classified as positive and negative would consistently appear in the right cluster of the PCA analysis. The analysis was performed for both general and genetic indicators. We generated new scores by PCA and compared them with the subjective scores, both formally (correlation coefficient) and graphically by plotting one against the other.

The PCA results strongly supported the definition of positive and negative factors of the general score; the correlation between the subjective and PCA-derived scores was very high (Pearson correlation coefficient = 0.98; \( P < 0.0001 \)). Of the 9 quality indicators for the genetic score, the PCA results supported all but 1 factor, “consideration of biomarkers,” which was subjectively classified as positive but was clustered with the other 2 negative factors in the PCA. After removal of this factor, the correlation between subjective and PCA-derived scores increased from 0.86 (\( P < 0.0001 \)) to 0.96 (\( P < 0.0001 \)). The indicator was therefore removed from the genetic score. The final items included in both general and genetic scores are listed in Table 2. The results of the PCA, including the scatterplots of subjective versus PCA-derived scores and the component plots, are shown in Web Figure 3 (http://aje.oxfordjournals.org/). The component plots show how each individual quality indicator contributed to the discrimination between good and poor quality. For the general score (Web Figure 3, part a), 3 positive factors (duplicate eligibility checking and/or data extraction; forest plot of study-specific results; statistical methods section in paper) and 2 negative factors (search strategy not described; inclusion/exclusion criteria unclear) seemed to capture most of the variability in quality. Corresponding items for the genetic score (Web Figure 3,
part c) were 2 positive factors (consideration of impact of genotyping errors; information on linkage disequilibrium) and 2 negative factors (no data on allele/genotype prevalence; no assessment of Hardy-Weinberg equilibrium (HWE)).

**Subgroup analyses.** We investigated whether the quality of the meta-analyses might be influenced by characteristics of the published articles, by grouping the papers on the basis of whether the meta-analysis: 1) accompanied a primary study; 2) was a Human Genome Epidemiology (HuGE) review; 3) appeared in a general medicine, genetics, or specialty journal; or 4) appeared in a journal with a high impact factor.

**Comparisons with previous work**

We searched the literature to identify previous reports of the quality of published meta-analyses in order to look for evidence of methodological improvement over time and for differences between genetic and nongenetic meta-analyses.

We considered all papers included in MEDLINE (US National Library of Medicine) that had been published between 1966 and August 2008, using the following search strategy: (meta-analy* or systematic review*) and (quality or evaluation or assessment or survey or appraisal or methodology*), with the search field limited to the title of the article. Reference lists of all relevant papers were scanned for further potential studies. We limited inclusion to reviews published in English that had assessed more than 10 meta-analyses. Since different criteria have been used to evaluate the quality of meta-analyses, producing problems of comparability, we included only those reviews which provided information on at least 2 of the following 5 items: 1) reporting of the search strategy; 2) reporting of the inclusion criteria; 3) reporting of the pooling methods; 4) evaluation of statistical heterogeneity; and 5) evaluation of publication bias.

**RESULTS**

**Review of meta-analyses**

**Selection of studies.** The meta-analysis articles were classified in the HuGE Reviews Archive by area of medicine, as papers on: neoplasms (n = 28); endocrine, nutritional, and metabolic diseases and immunity disorders (n = 25); diseases of the circulatory system (n = 18); diseases of the nervous system and sense organs (n = 16); mental disorders (n = 12); and other topics (n = 21). Of the 120 papers identified, 74% reported results of 1 or more meta-analyses, while in the other 26%, investigators reported results from their own primary study, accompanied by a meta-analysis based on one of their main findings. The median number of primary studies included in a paper was 13 (interquartile range, 8–22), while the median number of separate meta-analyses performed was 2 (interquartile range, 1–3; range, 1–26). Each of these meta-analyses included 2–119 primary studies. The median size of the largest meta-analysis in each of the 120 articles was 11 studies (interquartile range, 8–18). Among the 96 papers that clearly reported the design of the studies included, 5 meta-analyzed family-based studies, another 4 combined family studies with population studies, and the rest analyzed only population-based studies, including case-control, cohort, and cross-sectional studies or a mixture of the 3.

**Indicators of quality in conduct or reporting.** Results for the individual quality indicators are shown in Table 2. Detailed results on all of the other items considered, for the whole sample and by subgroup, are shown in Web Tables 1–6 (http://aje.oxfordjournals.org/). Overall, the papers scored better on general indicators than on genetic indicators, although basic issues such as reporting of inclusion/exclusion criteria and assessment of publication bias were ignored by more than one-fourth of the meta-analyses. In approximately one-third of the papers, investigators had performed duplicate checking of eligibility/data extraction and estimated the magnitude of the between-study heterogeneity along with the test, suggesting thoroughness in the conduct of the meta-analysis. Study quality assessment was performed in only 12 meta-analyses. These analyses used different sets of criteria, usually based on checklists developed for the evaluation of epidemiologic studies, with the addition of items specific to genetics. In only 9 meta-analyses did investigators evaluate the first-study effect.

**Primary study effect.** Whether the meta-analysis was the primary focus of the article or was an accompaniment to the authors’ own primary study had a large impact on quality. We accessed supplementary materials posted on a journal’s or author’s Web site whenever such materials were mentioned. Table 3 shows the factors that varied significantly (P < 0.05) between meta-analyses published with and without an accompanying primary study. The factors all pointed to poorer quality when the meta-analysis was not the primary focus, with strong differences for most aspects of the conduct and reporting of the meta-analysis. The quality was much lower for both the general quality indicators and the genetic quality indicators, with differences of 36% (P < 0.0001) and 35% (P < 0.001), respectively. In more than half of the meta-analyses that accompanied a primary study, the investigators did not even specify what databases they had searched.

**Quality of HuGE reviews.** None of the HuGE reviews accompanied a primary study, so they were compared with the other articles that did not include a primary study. There were 16 (18%) HuGE reviews, 14 published in the American Journal of Epidemiology and 2 in Genetics in Medicine, and 73 comparator articles.

There was no significant difference in general quality scores between HuGE reviews and other meta-analyses. For the genetic score, there was a nonsignificant trend towards better quality in HuGE reviews (16% difference; P = 0.090). This reflected a higher proportion of meta-analyses in which investigators tested for departures from HWE (63% vs. 47%; P = 0.281) and a higher proportion of papers in which investigators considered the possibility of genotyping errors in the primary studies (25% vs. 8%; P = 0.076). Half of the HuGE reviews, as compared with
one-fifth of the others ($P = 0.021$), used pairwise comparisons that do not require the assumption of a genetic model. The tendency toward better reporting of primary study characteristics in HuGE meta-analyses can be seen in several features (Web Tables 3 and 4), particularly for disease definition (81% vs. 51%; $P = 0.029$) and for evaluation of the prevalence of the genetic variant in the population (81% vs. 44%; $P = 0.011$).

Impact factor and journal type. The 31 meta-analyses that accompanied primary studies were generally of poor quality and were predominantly published in the specialist journals (81%). Among meta-analyses in papers without a primary study, no differences in either general or genetic quality were found between the 3 types of journals. Detailed results are shown in Web Tables 5 and 6.

We performed a regression analysis to assess whether the quality scores varied with the (log) impact factor of the journal. The overall quality score among the 31 meta-analyses that accompanied primary studies was positively associated with the impact factor of the journals ($P = 0.01$), with such an association being mainly confined to the general quality indicators. Among the other 89 articles, neither the general quality nor the genetic quality showed any trend with impact factor.

### Table 3.

| Quality Indicator                                      | Papers With M-A Only ($n = 89$) | Papers With M-A + Primary Study ($n = 31$) | $P$ Value$^a$ |
|--------------------------------------------------------|---------------------------------|------------------------------------------|--------------|
| General indicators of quality                          |                                 |                                          |              |
| Search strategy                                        |                                 |                                          |              |
| Databases listed                                       | 100                             | 45                                       | <0.001       |
| End date stated                                        | 93                              | 35                                       | <0.001       |
| Search terms listed                                    | 93                              | 42                                       | <0.001       |
| Inclusion/exclusion criteria reported                  | 92                              | 29                                       | <0.001       |
| Duplicate eligibility assessment                       | 17                              | 0                                        | 0.011        |
| Duplicate data extraction                              | 36                              | 3                                        | <0.001       |
| Use of random-effects models                           | 87                              | 48                                       | <0.001       |
| Heterogeneity test                                     | 99                              | 77                                       | <0.001       |
| Measure of size of heterogeneity                       | 36                              | 16                                       | 0.044        |
| Primary study sizes reported                           | 81                              | 52                                       | 0.004        |
| Primary study disease definitions reported             | 56                              | 6                                        | <0.001       |
| Primary study ethnicity/location reported              | 88                              | 42                                       | <0.001       |
| Graphical evaluation of publication bias               | 55                              | 23                                       | 0.002        |
| Statistical test of publication bias                   | 70                              | 26                                       | <0.001       |
| Cumulative meta-analysis                               | 20                              | 3                                        | 0.024        |
| Study influence assessment                             | 25                              | 6                                        | 0.036        |
| Genetic indicators of quality                          |                                 |                                          |              |
| Testing for Hardy-Weinberg equilibrium                 | 49                              | 16                                       | 0.001        |
| Allele/genotype counts reported                        | 64                              | 42                                       | 0.036        |
| Data on prevalence of polymorphism(s)                  | 51                              | 13                                       | <0.001       |
| Quality score$^b$                                      |                                 |                                          |              |
| General score                                          | 66.1 (10.6)                     | 41.8 (14.5)                              | <0.0001      |
| Genetic score                                          | 34.0 (18.0)                     | 22.2 (11.0)                              | <0.001       |
| Overall score                                          | 56.6 (10.2)                     | 36.0 (12.0)                              | <0.0001      |

Abbreviations: M-A, meta-analysis; SD, standard deviation.

$^a$ $P$ values were based on a 2-tailed Fisher’s exact test for quality indicators and the Mann-Whitney U test for quality scores.

$^b$ Quality scores were based on positive and negative indicators of good practice and good reporting. The general quality score depended on aspects relevant to any meta-analysis, and the genetic quality score was derived from aspects specific to genetics. The 2 quality scores were calculated by summing positive factors and subtracting negative factors and scaling the results to lie between 0 and 100, using the theoretical maximum and minimum of the sum.
Comparison with previous work

Previous reviews. The electronic search yielded 957 “hits,” from which 73 full-text articles were obtained and evaluated for eligibility. Four papers were added after cross-checking of reference lists, and 1 paper was added because it was known to the authors. Fifty papers were subsequently excluded, most because of the low number of meta-analyses considered (≤10) or the reporting of only 1 of the 5 criteria. A flow chart detailing inclusion and reasons for exclusion is presented in Web Figure 4 (http://aje.oxfordjournals.org/). Table 4 presents the results of the remaining 29 articles, reporting on 28 reviews. The reviews considered 16–272 meta-analyses. Only 1 review considered the meta-analysis of genetic association studies, while most of the others focused on randomized clinical trials. The Oxman and Guyatt scale, also known as the Overview Quality Assessment Questionnaire (8), was the most commonly used quality scale (62%), although it was often accompanied by information on additional quality indicators.

Trend in quality over time. By comparing our findings on recently published meta-analyses with those of Attia et al. (4) on 37 meta-analyses published before 2000, we can see improvement over time in both reporting and conduct. Reporting of the search methods increased from 65% in older meta-analyses to 86% in recent meta-analyses (P = 0.008), and the same was found for reporting of inclusion criteria, which increased from 49% to 74% (P = 0.005). Improvement in the conduct of meta-analyses is suggested by a substantial increase in the assessment of heterogeneity (from 76% to 93%; P = 0.005) and an extraordinary increase in the evaluation of publication bias (from 19% to 71%; P < 0.001). None of the meta-analyses published before 2000 used a formal statistical test to evaluate the presence of publication bias, as compared with 58% of recent meta-analyses, the majority of which used Egger’s test (introduced in 1997).

Improvement in genetic quality factors is less marked. HWE was tested in individual studies in only 24% of the meta-analyses published before 2000 versus 41% of those included in our review (P = 0.081). No difference was found for the choice of the genetic model, with results being based solely on a per-allele analysis in 24% of the papers versus 21% (P = 0.653). Interestingly, in our review, only 6 of the 25 papers (24%) that used solely a per-allele analysis performed a test for HWE, even though HWE is required for such a per-allele test. Similarly, only 1 of those 25 (4%) provided a reason for choosing an additive genetic model. For papers using per-genotype analyses, the proportion of those assuming a specific genetic model did not change: 33% in older meta-analyses versus 27% in recent ones (P = 0.53). Among these meta-analyses, the proportion of papers in which investigators provided justification for their model choice was even lower for recent meta-analyses than for older ones (25% vs. 67%; P = 0.016).

Comparison with quality in other fields. The results shown in Table 4 suggest that the quality of meta-analysis across different fields of research varies widely. When comparing meta-analyses published before 2002 with the more recent ones, a time trend of improvement is noticeable for all items, apart from the assessment of publication bias. This important issue seems to be consistently overlooked in meta-analysis across fields and over time, with the exception of the papers included in our review. No clear difference is evident between reviews on meta-analyses of randomized clinical trials and reviews which include meta-analyses of other study designs. The quality of Cochrane systematic reviews and meta-analyses is reported to be higher, with the exception of the 2002 paper by Shea et al. (54), where Cochrane papers were found to score no better, and for some quality items even worse, than others (Table 4). The authors suggested that this finding was due to low quality of reporting rather than conduct, and they pointed out how the Cochrane Collaboration had taken steps to improve the quality of its reviews in the period since their study (54).

DISCUSSION

The general quality of current genetic meta-analyses is very variable, although on average it is similar to that observed in other fields of medicine, and there is evidence of an improvement since Attia et al.’s earlier review (4). Recognition of potential problems, such as publication bias, is not always accompanied by appropriate action. The quality of the handling of specifically genetic factors is disappointingly low and does not seem to be improving. HuGE reviews scored better than the other analyses in this respect, but the comparison is hard to interpret because of the relatively small numbers.

Quality was markedly poorer in the 26% of the meta-analyses that accompanied the report of a primary study. This may reflect partly the expertise of the authors and the time available for performing the meta-analysis and partly the amount of space available in a given journal. Investigators sometimes add a meta-analysis in order to put their results into context. Although this is, in principle, a way of strengthening the evidence, it can lead to a “quick-and-dirty” meta-analysis. A poorly conducted or poorly reported meta-analysis is of little scientific use and may mislead. If space is a problem, details should be posted on a Web site.

The use of quality scores has limitations, in the same way as shown for the quality assessment of primary studies included in systematic reviews and meta-analyses (9). Our interest was to evaluate individual quality components, and we used scores only as convenient summaries for comparing quality across different types of meta-analyses. We created 2 quality scores, one with general quality indicators and the other with quality indicators specific to genetics, based on our subjective judgment. It is interesting that when a more objective approach to the assessment of quality was applied using PCA, the correspondence of the PCA-derived scores with our subjective scores was very high. The only exception was an item that we had selected as a genetic indicator of good quality—the consideration of biomarkers or intermediate phenotypes in the paper—which proved not to be a good marker of quality in the PCA. When this item was deleted from the genetic score, the correlation between “subjective” and “objective” scores was 96%. The PCA results suggest that the quality indicators used in this paper,
when taken as a set, can discriminate well between meta-analyses of good and poor quality, even if individually they have differing relevance and impact on quality. The results of the PCA also suggest that fewer indicators might be equally able to discriminate between good and poor quality for both the general and the genetic aspects of meta-analysis.

**Recommendations**

**Stating the rationale for the meta-analysis.** We found that the rationale for the meta-analysis was often not stated clearly. A meta-analysis might be exploratory or it might, for instance, be intended to replicate or support a result from a primary study or to evaluate sources of heterogeneity. When a finding from a primary study is conclusive, the aim of a meta-analysis may be to see whether the same association holds in other populations; when it is not conclusive, the aim might be to see whether it becomes so when combined with evidence from other studies. In the former case, the meta-analysis would not include the primary study, while in the latter it would. However, in our review, all 31 meta-analyses which accompanied a primary study included that study in the meta-analysis.

**Identification and selection of studies.** Although the majority of the authors in our review described their search strategy, few of them did so in a way that allowed reproducibility. Details on the search strategy could be made available as Web material if space were limited. The numbers of papers identified at each stage should be reported, possibly using the flow diagram suggested in the QUOROM statement (10). In our review, fewer than one-fourth of the meta-analyses provided this information.

There is no consensus on whether combining family- and population-based studies within a single meta-analysis is appropriate, the main concern being that gene-gene and gene-environment interactions might play different roles in the 2 types of studies. However, a number of authors have argued in favor of their combination (11–13), and this is recommended in the HuGE Review Handbook (14) when the available evidence is limited. Family studies based on the transmission disequilibrium test can be combined with population-based studies that use a per-allele approach, provided that the assumptions of additivity and HWE hold. When family- and population-based studies are combined, it is important to perform a sensitivity analysis.

Study identification is time-consuming, which may explain why only 13% of the papers in our review included a duplicate reading of the titles and abstracts. However, this is an important step that protects against selection bias. The selected articles should be read and information extracted onto a predesigned data extraction form, again in duplicate. In the meta-analyses included in our review, duplicate data extraction was performed in 28% of all papers and 50% of the HuGE reviews.

**Choice of the genetic model.** A key decision in any meta-analysis is which genetic model to adopt. In our review, 1 in 5 papers used a per-allele analysis. This requires an additive model and HWE (15). However, these assumptions are often ignored. In only 1 (4%) of the papers that relied on a per-allele analysis did investigators provide a reason, and in only one-fourth did they test for HWE. For the per-genotype approach, a genetic model was assumed in approximately one-fourth of the papers reviewed. The assumption of a genetic model increases statistical power, but there must be a priori knowledge to support the choice. In practice, there is often only poor-quality information on the genetic model, especially when the allele frequency is low. Among meta-analyses that assumed a genetic model, in only one-fourth of the papers did the authors provide justification for their choice. Assuming a “wrong” genetic model is a potential source of bias (16). When the underlying genetic model is unknown, it is better to use pairwise comparisons of the 3 genotypes, and the loss of power may be limited through the use of a bivariate meta-analysis (17). Alternatively, data can be utilized more efficiently using a “genetic model-free” approach, which estimates the genetic model from the data (16, 17).

**Assessment of heterogeneity.** Assessing the presence of between-study heterogeneity is a crucial step in any meta-analysis. In the vast majority of the meta-analyses reviewed, investigators performed a statistical test for heterogeneity, although only one-third of them estimated its magnitude. Testing for heterogeneity alone is unsatisfactory, not only because it does not provide evidence on the extent of the problem but also because of the low power of the tests. The magnitude of heterogeneity can be directly measured by the between-study variance. In the meta-analyses reviewed, heterogeneity was mainly quantified using $I^2$, a measure proposed by Higgins et al. (18) and defined as the percentage of total variation in study estimates explained by heterogeneity rather than sampling error. However, as Higgins and Thompson (19) have pointed out, $I^2$ better describes the impact of heterogeneity on the meta-analysis than the magnitude of heterogeneity.

If the studies are heterogeneous in a way not anticipated in the hypothesis, the first priority should be to investigate the causes. Perhaps heterogeneity is due to a few aberrant studies, or perhaps it is due to geographic or methodological factors. This can be evaluated by means of meta-regression or by using subgroup analyses accompanied by formal testing, although statistical significance has to be interpreted in the light of the low power of interaction tests (20). In the review, subgroup analyses were performed in 70% of the meta-analyses, but only 10% formally tested for interaction. Meta-regression based on study-level characteristics was used in 18% of the papers, while patient-level characteristics were used in 13%. Without individual patient data, the use of patient-level characteristics in meta-regression should be discouraged, as it is difficult to interpret (21) and has very little power to explain heterogeneity (22).

Ideally, study quality should be investigated as a cause of heterogeneity. Our review shows that this is rarely done, perhaps because there is no consensus on the best quality scoring model. Although many authors have proposed checklists (1, 23–26), no synthesis of this work has been carried out.

**Use of random effects.** When unanticipated heterogeneity cannot be explained, a decision needs to be made on
| Authors and Year (Ref. No.) | Field of Research (Type of Primary Studies) | No. of M-A’s | Publication Date(s) of M-A’s | Search Methods Reported | Inclusion Criteria Reported | Pooling Methods Reported | Statistical Heterogeneity Assessed | Publication Bias Assessed |
|-----------------------------|---------------------------------------------|--------------|------------------------------|-------------------------|---------------------------|---------------------------|----------------------------------|--------------------------|
| Current review              | Genetic epidemiology (genetic association studies) | 120          | 2005–2007                    | 86                      | 74                        | 97                        | 93                               | 71                       |
| Attia et al., 2003 (4)      | Genetic epidemiology (genetic association studies) | 37           | 1991–2000                    | 65                      | 49                        | 97                        | 76                               | 19                       |
| De Vito et al., 2007 (37)a  | Vaccinology (all study designs)               | 121          | 1991–2007 (February)         | 87                      | 79                        | 79                        | 67                               | 26                       |
| Junhua et al., 2007 (38)    | Traditional Chinese medicine (NS)             | 36           | 1978–2006                    | 78                      | 81                        | 67                        | NR                               | NR                       |
| Gerber et al., 2007 (39)b   | Any field (mostly RCTs)                      | 272          | 1993–2002                    | 85                      | NR                        | 100                       | 85                               | 21                       |
| Sheik et al., 2007 (40)a    | Maternal medicine (NS)                        | 39           | Cochrane 2001–2006           | 97                      | 100                       | NR                        | NR                               | NR                       |
| Boluyt et al., 2006 (41)    | Asthma (randomized and quasi-randomized trials) | 14           | Cochrane 2000–2006           | 100                     | 100                       | 100                       | 74                               | NR                       |
| Collier et al., 2006 (42)a  | Dermatology (mostly RCTs)                    | 28           | Cochrane 1999–2004           | 100                     | 96                        | 71                        | NR                               | NR                       |
| Flores-Mir et al., 2006 (43)a | Orthodontics (NS)                        | 16           | 2000–2004                    | 88                      | 100                       | NR                        | NR                               | NR                       |
| Golder et al., 2006 (44)a   | Adverse effects (all study designs)          | 256          | 1994–2005                    | 77                      | NR                        | NR                        | 88                               | NR                       |
| Jorgensen et al., 2006 (45) | Any field (mostly RCTs)                      | 24           | Cochrane 1996–2003           | 96                      | 100                       | 100                       | NR                               | NR                       |
| Shea et al., 2006 (46)a     | Musculoskeletal diseases                     | 57           | Cochrane Up to 2002          | 88                      | 100                       | 95                        | NR                               | NR                       |
| Shea et al., 2006 (47)a     | Any field (mostly RCTs)                      | 53           | Cochrane original            | 81                      | 98                        | 89                        | NR                               | NR                       |
| Delaney et al., 2005 and 2007 (48, 49) | Critical care medicine (NS) | 47       | Cochrane 1994–2003           | 100                     | 98                        | 81                        | NR                               | NR                       |
| Dixon et al., 2005 (50)     | General surgical literature (NS)             | 51           | 1997–2002                    | 67                      | 70                        | 67                        | NR                               | NR                       |
| Lawson et al., 2005 (51)    | Conventional medicine and complementary/alternative medicine (RCTs) | 105 | Conventional Up to 1999 | 49 | 74 | 77 | NR | 15 |
| Palma and Delgado-Rodriguez, 2005 (52) | Cardiovascular (all study designs) | 225 | 1990–2002 | NRc | NRc | NRc | 83 | 11 |
| Moher et al., 2002 (53)     | Pediatric alternative and conventional medicine (NS) | 66           | NS–2001                      | 52                      | 64                        | 41                        | 38                               | 17                       |
| Shea et al., 2002 (54)      | Any field (RCTs)                             | 52           | Cochrane 1993–1996           | 31                      | 74                        | 98                        | 29                               | 8                        |
| Bhandari et al., 2001 (55)  | Orthopedic surgery (RCTs and observational studies) | 40           | 1984–1999                    | 83                      | 78                        | 70                        | NR                               | NR                       |
| Choi et al., 2001 (56)      | Anesthesia (NS)                              | 82           | 1989–1999                    | 73                      | 81                        | 82                        | 35                               | 5                        |
whether to continue with the analysis using a random-effects model (27) or to refrain from pooling. Forcing a random-effects meta-analysis in the presence of large heterogeneity can produce an average estimate that is meaningless. An extreme example of this is genetic "flip-flop" (28, 29), where results from primary studies are significant but point in opposite directions. A meta-analysis which combines such studies might misleadingly suggest a lack of genetic effect. There are no clear guidelines on how much heterogeneity is allowable, but as a rough guide, we would suggest only using a random-effects model if the standard deviation of the between-study variation is less than 25% of the pooled effect size—for instance, a log odds ratio.

**Evaluation of HWE.** HWE should be investigated in the individual studies, since deviation from HWE may reflect methodological problems, such as genotyping error, population stratification, or selection bias (30). Because of low statistical power, it is advisable to measure the magnitude of the deviation from HWE as well as its significance. In our review, HWE was assessed in 41% of the recent papers, and the assessment was limited to statistical testing, apart from 3 meta-analyses in which the magnitude of the deviation was used to adjust the final pooled estimate of the genetic effect. We suggest that lack of HWE should be treated as a reason for further investigation of a primary study rather than as grounds for its exclusion (4, 31).

**Evaluation of publication bias.** Although publication bias was more frequently considered in our review than in most others, in only half of the papers did investigators report using a funnel plot, and slightly more than half reported using a statistical test. Graphical assessment of the possible presence of publication bias is simple and useful. However, judgment based only on visual inspection of funnel plots tends to be inaccurate, as suggested by empirical evaluation (32), and funnel plots are best used in conjunction with a test. A number of tests for publication bias have been proposed, and the choice between them depends on characteristics of the meta-analysis (33). If publication bias is suspected, it may be sensible to concentrate the analysis on the larger studies or to model the dependence on sample size (34). Publication might be faster for studies with positive findings (time-lag bias) (35), and the use of cumulative and recursive meta-analysis can help detect time trends in effect estimates (36). In our review, cumulative meta-analysis was performed in only 16% of the papers.

**Reporting of study-specific data.** Generally, with the exception of HuGE reviews, the reporting of study characteristics needs to be improved. Neither genotype nor allele counts were reported in nearly half of the meta-analyses. Additional material can easily be placed on a Web page, and journal editors and reviewers should encourage this. Other information that should be reported includes the prevalence of the risk allele, linkage disequilibrium in the region, genotyping error rates, and any haplotypes or biomarkers that have been investigated. With the exception of the prevalence of the polymorphisms, which was reported in the majority of HuGE reviews and one-third of the others, such data were rarely given.
Conclusions

Although the general quality of genetic meta-analyses is similar to that observed in other fields of medicine and shows improvement over time, the quality of the handling of specifically genetic factors is disappointingly low and does not seem to be improving. This is perhaps not surprising, given the lack of consensus in the theoretical literature on the best methods to use. The tendency towards better quality of HuGE reviews suggests that the HuGE Review Handbook (14) has positively influenced the conduct and reporting of such meta-analyses, but there is still a long way to go. Meta-analysis features very highly in the hierarchy of evidence, so it is incumbent on investigators performing meta-analyses to be as methodologically rigorous as possible. Development of formal, detailed consensus guidelines, similar to those of QUOROM (10), would be helpful.

ACKNOWLEDGMENTS

Author affiliations: Institute of Genetic Medicine, EURAC Research, Bolzano, Italy (Cosetta Minelli); Department of Health Sciences, School of Medicine, University of Leicester, Leicester, United Kingdom (John R. Thompson, Keith R. Abrams); Section for Clinical Epidemiology and Biostatistics, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand (Ammarin Thakkinstian); Centre for Clinical Epidemiology and Biostatistics, University of Newcastle, Newcastle, Australia (John Attia); and Hunter Medical Research Institute and Department of General Medicine, John Hunter Hospital, Newcastle, Australia (John Attia).

Drs. Cosetta Minelli and John R. Thompson are joint first authors.

Conflict of interest: none declared.

REFERENCES

1. Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet. 2003;361(9360):865–872.
2. Ioannidis JP, Ntzani EE, Trikalinos TA, et al. Replication validity of genetic association studies. Nat Genet. 2001;29(3):306–309.
3. Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG, et al. Establishment of genetic associations for complex diseases is independent of early study findings. Eur J Hum Genet. 2004;12(9):762–769.
4. Attia J, Thakkinstian A, D’Este C. Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. J Clin Epidemiol. 2003;56(4):297–303.
5. Munafo MR, Flint J. Meta-analysis of genetic association studies. Trends Genet. 2004;20(9):439–444.
6. Salanti G, Sanderson S, Higgins JP. Obstacles and opportunities in meta-analysis of genetic association studies. Genet Med. 2005;7(1):13–20.
7. DeVellis RF. Scale Development: Theory and Applications. 2nd ed. London, United Kingdom: Sage Publications; 2003.
8. Oxman AD, Guyatt GH, Singer J, et al. Agreement among reviewers of review articles. J Clin Epidemiol. 1991;44(1):91–98.
9. Herbison P, Hay-Smith J, Gillespie WJ. Adjustment of meta-analyses on the basis of quality scores should be abandoned. J Clin Epidemiol. 2006;59(12):1249–1256.
10. Moher D, Cook DJ, Eastwood S, et al. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet. 1999;354(9193):1896–1900.
11. Mitchell LE. Relationship between case-control studies and the transmission/disequilibrium test. Genet Epidemiol. 2000;19(3):193–201.
12. Lohmueller KE, Pearce CL, Pike M, et al. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet. 2003;33(2):177–182.
13. Kazeme GR, Farrall M. Integrating case-control and TDT studies. Ann Hum Genet. 2005;69(3):329–335.
14. Little J, Higgins IPT, eds. The HuGENet™ HuGE Review Handbook, Version 1.0. Ottawa, Ontario, Canada: Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa; 2006. (http://www.hugenet.ca). (Accessed December 3, 2008).
15. Schaid DJ, Jacobsen SJ. Biased tests of association: comparisons of allele frequencies when departing from Hardy-Weinberg proportions. Am J Epidemiol. 1999;149(8):706–711.
16. Minelli C, Thompson JR, Abrams KR, et al. The choice of a genetic model in the meta-analysis of molecular association studies. Int J Epidemiol. 2005;34(6):1319–1328.
17. Minelli C, Thompson JR, Abrams KR, et al. Bayesian implementation of a genetic model-free approach to the meta-analysis of genetic association studies. Stat Med. 2005;24(24):3845–3861.
18. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–560.
19. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–1558.
20. Brookes ST, Whitely E, Egger M, et al. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol. 2004;57(3):229–236.
21. Berlin JA, Santanna J, Schmid CH, et al. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med. 2002;21(3):371–387.
22. Lambert PC, Sutton AJ, Abrams KR, et al. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. J Clin Epidemiol. 2002;55(1):86–94.
23. Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. Am J Epidemiol. 2002;156(4):300–310.
24. Thakkinstian A, McEvoy M, Minelli C, et al. Systematic review and meta-analysis of the association between β2-adrenoceptor polymorphisms and asthma: a HuGE review. Am J Epidemiol. 2005;162(3):201–211.
25. Bogardus ST Jr, Conca J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. JAMA. 1999;281(20):1919–1926.
26. Li D, He L. Association study of the G-protein signaling 4 (RGS4) and proline dehydrogenase (PRODH) genes with schizophrenia: a meta-analysis. Eur J Hum Genet. 2006;14(10):1130–1135.
