STRengthening the REporting of Genetic Association Studies (STREGA) – An Extension of the STROBE Statement

Julian Little1,2, Julian P.T. Higgins3, John P.A. Ioannidis4, David Moher2, France Gagnon5, Erik von Elm1, Muin J. Khoury7, Barbara Cohen8, George Davey-Smith9, Jeremy Grimshaw6, Paul Scheet10, Marta Gwinn11, Robin E. Williamson12, Guang Yong Zou13, Kim Hutchings14, Candice Y. Johnson2, Valerie Tait15, Miriam Wiens16, Jean Golding17, Cornelia van Duijn18, John McLaughlin19, Andrew Paterson20, George Wells21, Isabel Fortier22, Matthew Freedman23, Maja Zecevic24, Richard King25, Claire Infante-Rivard26, Alex Stewart27, Nick Birkett1

1Canada Research Chair in Human Genome Epidemiology; 2Department of Epidemiology and Community Medicine, University of Ottawa, Ontario, Canada; 3MRC Biostatistics Unit, Cambridge, UK; 4Department of Hygiene and Epidemiology, School of Medicine, University of Ioannina, Greece and Center for Genetic Epidemiology and Modeling, Tufts University School of Medicine, Boston, USA; 5CIHR New Investigator and Canada Research Chair in Genetic Epidemiology, University of Toronto, Dalla Lana School of Public Health, Toronto, Ontario, Canada; 6Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland and German Cochrane Centre, Department of Medical Biometry and Medical Informatics, University Medical Centre, Freiburg, Germany; 7National Office of Public Health Genomics, Centers for Disease Control & Prevention, Atlanta, USA; 8Former Senior Editor, Public Library of Science, San Francisco, CA; 9MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine, University of Bristol, UK; 10Canada Research Chair in Health Knowledge Transfer and Uptake, Clinical Epidemiology Program, Ottawa Health Research Institute, Department of Medicine, University of Ottawa; 11University of Texas, MD Anderson Cancer Center, Department of Epidemiology, Houston, Texas; 12Deputy Editor, American Journal of Human Genetics, Boston, USA; 13Department of Epidemiology and Biostatistics, University of Western Ontario, London, Ontario, Canada and Robarts Clinical Trials, Robarts Research Institute, London, Ontario, Canada; 14Editor, Paediatric and Perinatal Epidemiology, Bristol, UK; 15Editor, European Journal of Epidemiology, Rotterdam, Netherlands; 16Cancer Care Ontario, Toronto, and Prosserman Centre for Health Research at the Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada; 17Canada Research Chair in Genetics of Complex Diseases, Hospital for Sick Children (SickKids), Toronto, Canada; 18Director, Cardiovascular Research Methods Centre, University of Ottawa Heart Institute Ottawa, Ontario, Canada; 19Genome Quebec & P3G Observatory, McGill University and Genome Quebec Innovation Center Docteur Penfield, Montréal QC Canada; 20 Dana-Farber Cancer Institute, Boston, MA, USA; 21Senior Editor, Lancet, New York; 22Editor, Genetics in Medicine, Minneapolis, MN; 23Canada Research Chair-James McGill Professor Department of Epidemiology, Biostatistics and Occupational Health Faculty of Medicine, McGill University, Montreal, Canada; 24University of Ottawa Heart Institute, Ottawa, Ontario, Canada

Correspondence to: Dr Julian Little, Department of Epidemiology and Community Medicine, University of Ottawa, 455 Smyth Rd, Ottawa, Ontario K1H 8M5, Canada. E-mail: jllittle@uottawa.ca

Abstract

Making sense of rapidly evolving evidence on genetic associations is crucial to making genuine advances in human genomics and the eventual integration of this information in the practice of medicine and public health. Assessment of the strengths and weaknesses of this evidence, and hence the ability to synthesize it, has been limited by inadequate reporting of results. The STRengthening the REporting of Genetic Association studies (STREGA) initiative builds on the STrengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement and provides additions to 12 of the 22 items on the STROBE checklist. The additions concern population stratification, genotyping errors, modelling haplotype variation, Hardy-Weinberg equilibrium, replication, selection of participants, rationale for choice of genes and variants, treatment effects in studying quantitative traits, statistical methods, relatedness, reporting of descriptive and outcome data, and the volume of data issues that are important to consider in genetic association studies. The STREGA recommendations do not
prescribe or dictate how a genetic association study should be designed but seek to enhance the transparency of its reporting, regardless of choices made during design, conduct, or analysis.

**Key words:** gene-disease associations; genetics; gene-environment interaction; systematic review; meta analysis; reporting recommendations; epidemiology; genome-wide association

**Introduction**

The rapidly evolving evidence on genetic associations is crucial to integrating human genomics into the practice of medicine and public health [1,2]. Genetic factors are likely to affect the occurrence of numerous common diseases, and therefore identifying and characterizing the associated risk (or protection) will be important in improving the understanding of etiology and potentially for developing interventions based on genetic information. The number of publications on the associations between genes and diseases has increased tremendously; with more than 34,000 published articles, the annual number has more than doubled between 2001 and 2008 [3,4]. Articles on genetic associations have been published in about 1500 journals and in several languages.

Despite the many similarities between genetic association studies and “classical” observational epidemiologic studies (that is, cross-sectional, case-control, and cohort) of lifestyle and environmental factors, genetic association studies present several specific challenges including an unprecedented volume of new data [5,6] and the likelihood of very small individual effects. Genes may operate in complex pathways with gene-environment and gene-gene interactions [7]. Moreover, the current evidence base on gene-disease associations is fraught with methodological problems [8-10]. Inadequate reporting of results, even from well-conducted studies, hampers assessment of a study’s strengths and weaknesses, and hence the integration of evidence [11].

Although several commentaries on the conduct, appraisal and/or reporting of genetic association studies have so far been published [12-39], their recommendations differ. For example, some papers suggest that replication of findings should be part of the publication [12,13,16,17,23,26,34,36] whereas others consider this suggestion unnecessary or even unreasonable [21,40-44]. In many publications, the guidance has focused on genetic association studies of specific diseases [14,15,17,19,22,23,25,26,31-38] or the design and conduct of genetic association studies [13-15,19,20,22,23,25,30-32,35,36] rather than on the quality of the reporting.

Despite increasing recognition of these problems, the quality of reporting genetic association studies needs to be improved [45-49]. For example, an assessment of a random sample of 315 genetic association studies published from 2001 to 2003 found that most studies provided some qualitative descriptions of the study participants (for example, origin and enrolment criteria), but reporting of quantitative descriptors such as age and sex was variable [49]. In addition, completeness of reporting of methods that allow readers to assess potential biases (for example, number of exclusions or number of samples that could not be genotyped) varied [49]. Only some studies described methods to validate genotyping or mentioned whether research staff were blinded to outcome. The same problems persisted in a smaller sample of studies published in 2006 [49]. Lack of transparency and incomplete reporting have raised concerns in a range of health research fields [11,50-53] and poor reporting has been associated with biased estimates of effects in clinical intervention studies [54].

The main goal of this article is to propose and justify a set of guiding principles for reporting results of genetic association studies. The epidemiology community has recently developed the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) Statement for cross-sectional, case-control, and cohort studies [55,56]. Given the relevance of general epidemiologic principles for genetic association studies, we propose recommendations in an extension of the STROBE Statement called the STrengthening the REpo rting of Genetic Association studies (STREGA) Statement. The recommendations of the STROBE Statement have a strong foundation because they are based on empirical evidence on the reporting of observational studies, and they involved extensive consultations in the epidemiologic research community [56]. We have sought to identify gaps
and areas of controversy in the evidence regarding potential biases in genetic association studies. With the recommendations, we have indicated available empirical or theoretical work that has demonstrated or suggested that a methodological feature of a study can influence the direction or magnitude of the association observed. We acknowledge that for many items, no such evidence exists. The intended audience for the reporting guideline is broad and includes epidemiologists, geneticists, statisticians, clinician scientists, and laboratory-based investigators who undertake genetic association studies. In addition, it includes "users" of such studies who wish to understand the basic premise, design, and limitations of genetic association studies in order to interpret the results. The field of genetic associations is evolving very rapidly with the advent of genome-wide association investigations, high-throughput platforms assessing genetic variability beyond common single nucleotide polymorphisms (SNPs) (for example, copy number variants, rare variants), and eventually routine full sequencing of samples from large populations. Our recommendations are not intended to support or oppose the choice of any particular study design or method. Instead, they are intended to maximize the transparency, quality and completeness of reporting of what was done and found in a particular study.

Methods
A multidisciplinary group developed the STREGA Statement by using literature review, workshop presentations and discussion, and iterative electronic correspondence after the workshop. Thirty-three of 74 invitees participated in the STREGA workshop in Ottawa, Ontario, Canada, in June, 2006. Participants included epidemiologists, geneticists, statisticians, journal editors and graduate students.

Before the workshop, an electronic search was performed to identify existing reporting guidance for genetic association studies. Workshop participants were also asked to identify any additional guidance. They prepared brief presentations on existing reporting guidelines, empirical evidence on reporting of genetic association studies, the development of the STROBE Statement, and several key areas for discussion that were identified on the basis of consultations before the workshop. These areas included the selection and participation of study participants, rationale for choice of genes and variants investigated, genotyping errors, methods for inferring haplotypes, population stratification, assessment of Hardy-Weinberg equilibrium (HWE), multiple testing, reporting of quantitative (continuous) outcomes, selectively reporting study results, joint effects and inference of causation in single studies. Additional resources to inform workshop participants were the HuGENet handbook [57,58], examples of data extraction forms from systematic reviews or meta-analyses, articles on guideline development [59,60] and the checklists developed for STROBE. To harmonize our recommendations for genetic association studies with those for observational epidemiologic studies, we communicated with the STROBE group during the development process and sought their comments on the STREGA draft documents. We also provided comments on the developing STROBE Statement and its associated explanation and elaboration document [56].

Results
In Table 1, we present the STREGA recommendations, an extension to the STROBE checklist [55] for genetic association studies. The resulting STREGA checklist provides additions to 12 of the 22 items on the STROBE checklist. During the workshop and subsequent consultations, we identified five main areas of special interest that are specific to, or especially relevant in, genetic association studies: genotyping errors, population stratification, modelling haplotype variation, HWE and replication. We elaborate on each of these areas, starting each section with the corresponding STREGA recommendation, followed by a brief outline of the issue and an explanation for the recommendations. Complementary information on these areas and the rationale for additional STREGA recommendations relating to selection of participants, choice of genes and variants selected, treatment effects in studying quantitative traits, statistical methods, relatedness, reporting of descriptive and outcome data, and issues of data volume, are presented in Table 2.

Genotyping Errors
Recommendation for reporting of methods (Table 1, item 8(b)): Describe laboratory methods, including source and storage of DNA, genotyping methods and platforms (including the allele calling algorithm used, and its version), error rates and call rates. State the laboratory/centre where genotyping was done. Describe comparability of laboratory methods if there is more than one group. Specify whether genotypes were assigned using all of the data from the study simultaneously.
or in smaller batches.

Recommendation for reporting of results (Table 1, item 13(a)): Report numbers of individuals in whom genotyping was attempted and numbers of individuals in whom genotyping was successful.

Genotyping errors can occur as a result of the DNA sequence flanking the marker of interest, poor quality or quantity of the DNA extracted from biological samples, biochemical artefacts, poor equipment precision or equipment failure, or human error in sample handling, conduct of the array or handling the data obtained from the array [61]. A commentary published in 2005 on the possible causes and consequences of genotyping errors observed that an increasing number of researchers were aware of the problem, but that the effects of such errors had largely been neglected [61]. The magnitude of genotyping errors has been reported to vary between 0.5% and 30% [61-64]. In high-throughput centres, an error rate of 0.5% per genotype has been observed for blind duplicates that were run on the same gel [64]. This lower error rate reflects an explicit choice of markers for which genotyping rates have been found to be highly repeatable and whose individual polymerase chain reactions (PCR) have been optimized. Non-differential genotyping errors, that is, those that do not differ systematically according to outcome status, will usually bias associations towards the null [65-66], just as for other non-differential errors. The most marked bias occurs when genotyping sensitivity is poor and genotype prevalence is high (>85%) or, as the corollary, when genotyping specificity is poor and genotype prevalence is low (<15%) [65]. When measurement of the environmental exposure has substantial error, genotyping errors of the order of 3% can lead to substantial under-estimation of the magnitude of an interaction effect [67]. When there are systematic differences in genotyping according to outcome status (differential error), bias in any direction may occur. Unblinded assessment may lead to differential misclassification. For genome-wide association studies of SNPs, differential misclassification between comparison groups (for example, cases and controls) can occur because of differences in DNA storage, collection or processing protocols, even when the genotyping itself meets the highest possible standards [68]. In this situation, using samples blinded to comparison group to determine the parameters for allele calling could still lead to differential misclassification. To minimize such differential misclassification, it would be necessary to calibrate the software separately for each group. This is one of the reasons for our recommendation to specify whether genotypes were assigned using all of the data from the study simultaneously or in smaller batches.

Population Stratification

Recommendation for reporting of methods (Table 1, item 12(h)): Describe any methods used to assess or address population stratification.

Population stratification is the presence within a population of subgroups among which allele (or genotype; or haplotype) frequencies and disease risks differ. When the groups compared in the study differ in their proportions of the population subgroups, an association between the genotype and the disease being investigated may reflect the genotype being an indicator identifying a population subgroup rather than a causal variant. In this situation, population subgroup is a confounder because it is associated with both genotype frequency and disease risk. The potential implications of population stratification for the validity of genetic association studies have been debated [69-83]. Modelling the possible effect of population stratification (when no effort has been made to address it) suggests that the effect is likely to be small in most situations [75,76,78-80]. Meta-analyses of 43 gene-disease associations comprising 697 individual studies showed consistent associations across groups of different ethnic origin [80], and thus provide evidence against a large effect of population stratification, hidden or otherwise. However, as studies of association and interaction typically address moderate or small effects and hence require large sample sizes, a small bias arising from population stratification may be important [81]. Study design (case-family control studies) and statistical methods [84] have been proposed to address population stratification, but so far few studies have used these suggestions [49]. Most of the early genome-wide association studies used family-based designs or such methods as genomic control and principal components analysis [85,86] to control for stratification. These approaches are particularly appropriate for addressing bias when the identified genetic effects are very small (odds ratio <1.20), as has been the situation in many recent genome-wide association studies [85,87-105]. In view of the debate about the potential implications of population stratification for the validity of genetic association studies, we recommend transparent reporting of the methods used, or stating that none was used, to address this potential problem.
This reporting will enable empirical evidence to accrue about the effects of population stratification and methods to address it.

**Modelling Haplotype Variation**

**Recommendation for reporting of methods (Table 1, item 12(g)): Describe any methods used for inferring genotypes or haplotypes.**

A haplotype is a combination of specific alleles at neighbouring genes that tends to be inherited together. There has been considerable interest in modelling haplotype variation within candidate genes. Typically, the number of haplotypes observed within a gene is much smaller than the theoretical number of all possible haplotypes [106,107]. Motivation for utilizing haplotypes comes, in large part, from the fact that multiple SNPs may “tag” an untyped variant more effectively than a single typed variant. The subset of SNPs used in such an approach is called “haplotype tagging” SNPs. Implicitly, an aim of haplotype tagging is to reduce the number of SNPs that have to be genotyped, while maintaining statistical power to detect an association with the phenotype. Maps of human genetic variation are becoming more complete, and large scale genotypic analysis is becoming increasingly feasible. In consequence, it is possible that modelling haplotype variation will become more focussed on rare causal variants, because these may not be included in the genotyping platforms.

In most current large-scale genetic association studies, data are collected as unphased multilocus genotypes (that is, which alleles are aligned together on particular segments of chromosome is unknown). It is common in such studies to use statistical methods to estimate haplotypes [108-111], and their accuracy and efficiency have been discussed [112-116]. Some methods attempt to make use of a concept called haplotype “blocks” [117,118], but the results of these methods are sensitive to the specific definitions of the “blocks” [119,120]. Reporting of the methods used to infer individual haplotypes and population haplotype frequencies, along with their associated uncertainties should enhance our understanding of the possible effects of different methods of modelling haplotype variation on study results as well as enabling comparison and syntheses of results from different studies.

Information on common patterns of genetic variation revealed by the International Haplotype Map (HapMap) Project [107] can be applied in the analysis of genome-wide association studies to infer genotypic variation at markers not typed directly in these studies [121,122]. Essentially, these methods perform haplotype-based tests but make use of information on variation in a set of reference samples (for example, HapMap) to guide the specific tests of association, collapsing a potentially large number of haplotypes into two classes (the allelic variation) at each marker. It is expected that these techniques will increase power in individual studies, and will aid in combining data across studies, and even across differing genotyping platforms. If imputation procedures have been used, it is useful to know the method, accuracy thresholds for acceptable imputation, how imputed genotypes were handled or weighted in the analysis, and whether any associations based on imputed genotypes were also verified on the basis of direct genotyping at a subsequent stage.

**Hardy-Weinberg Equilibrium**

**Recommendation for reporting of methods (Table 1, item 12(g)): State whether Hardy-Weinberg equilibrium was considered and, if so, how.**

Hardy-Weinberg equilibrium has become widely accepted as an underlying model in population genetics after Hardy [123] and Weinberg [124] proposed the concept that genotype frequencies at a genetic locus are stable within one generation of random mating; the assumption of HWE is equivalent to the independence of two alleles at a locus. Views differ on whether testing for departure from HWE is a useful method to detect errors or peculiarities in the data set, and also the method of testing [125]. In particular, it has been suggested that deviation from HWE may be a sign of genotyping errors [126-128]. Testing for departure from HWE has a role in detecting gross errors of genotyping in large-scale genotyping projects such as identifying SNPs for which the clustering algorithms used to call genotypes have broken down [85,129]. However, the statistical power to detect less important errors of genotyping by testing for departure from HWE is low [150] and, in hypothetical data, the presence of HWE was generally not altered by the introduction of genotyping errors [131]. Furthermore, the assumptions underlying HWE, including random mating, lack of selection according to genotype, and absence of mutation or gene flow, are rarely met in human populations [132,133]. In five of 42 gene-disease associations assessed in meta-analyses of almost 600 studies, the results of studies that violated HWE significantly differed from results of studies that conformed to the
model [134]. Moreover, the study suggested that exclusion of HWE-violating studies may result in loss of the statistical significance of some postulated gene-disease associations and that adjustment for the magnitude of deviation from the model may also have the same consequence for some other gene-disease associations. Given the differing views about the value of testing for departure from HWE and about the test methods, transparent reporting of whether such testing was done and, if so, the method used, is important for allowing the empirical evidence to accrue.

For massive-testing platforms, such as genome-wide association studies, it might be expected that many false-positive violations of HWE would occur if a lenient P value threshold were set. There is no consensus on the appropriate P value threshold for HWE-related quality control in this setting. So, we recommend that investigators state which threshold they have used, if any, to exclude specific polymorphisms from further consideration. For SNPs with low minor allele frequencies, substantially more significant results than expected by chance have been observed, and the distribution of alleles at these loci has often been found to show departure from HWE.

For genome-wide association studies, another approach that has been used to detect errors or peculiarities in the data set (due to population stratification, genotyping error, HWE deviations or other reasons) has been to construct quantile-quantile (Q/Q) plots whereby observed association statistics or calculated P values for each SNP are ranked in order from smallest to largest and plotted against the expected null distribution [129,130]. The shape of the curve can lend insight into whether or not systematic biases are present.

**Replication**

**Recommendation: State if the study is the first report of a genetic association, a replication effort, or both. (Table 1, item 3)**

Articles that present and synthesize data from several studies in a single report are becoming more common. In particular, many genome-wide association analyses describe several different study populations, sometimes with different study designs and genotyping platforms, and in various stages of discovery and replication [129,130]. When data from several studies are presented in a single original report, each of the constituent studies and the composite results should be fully described. For example, a discussion of sample size and the reason for arriving at that size would include clear differentiation between the initial group (those that were typed with the full set of SNPs) and those that were included in the replication phase only (typed with a reduced set of SNPs) [129,130]. Describing the methods and results in sufficient detail would require substantial space in print, but options for publishing additional information on the study online make this possible.

**Discussion**

The choices made for study design, conduct and data analysis potentially influence the magnitude and direction of results of genetic association studies. However, the empirical evidence on these effects is insufficient. Transparency of reporting is thus essential for developing a better evidence base (Table 2). Transparent reporting helps address gaps in empirical evidence [45], such as the effects of incomplete participation and genotyping errors. It will also help assess the impact of currently controversial issues such as population stratification, methods of inferring haplotypes, departure from HWE and multiple testing on effect estimates under different study conditions.

The STREGA Statement proposes a minimum checklist of items for reporting genetic association studies. The statement has several strengths. First, it is based on existing guidance on reporting observational studies (STROBE). Second, it was developed from discussions of an interdisciplinary group that included epidemiologists, geneticists, statisticians, journal editors, and graduate students, thus reflecting a broad collaborative approach in terminology accessible to scientists from diverse disciplines. Finally, it explicitly describes the rationale for the decisions (Table 2) and has a clear plan for dissemination and evaluation.

The STREGA recommendations are available at www.strega-statement.org. We welcome comments, which will be used to refine future versions of the recommendations. We note that little is known about the most effective ways to apply reporting guidelines in practice, and that therefore it has been suggested that editors and authors collect, analyze, and report their experiences in using such guidelines [135]. We consider that the STREGA recommendations can be used by authors, peer reviewers and editors to improve the reporting of genetic association studies. We invite journals to endorse STREGA, for example by including STREGA and its Web address in their Instructions for Authors and by advising authors and peer reviewers to use the checklist as a guide. It has been suggested that
reporting guidelines are most helpful if authors keep the general content of the guideline items in mind as they write their initial drafts, then refer to the details of individual items as they critically appraise what they have written during the revision process [135]. We emphasize that the STREGA reporting guidelines should not be used for screening submitted manuscripts to determine the quality or validity of the study being reported. Adherence to the recommendations may make some manuscripts longer, and this may be seen as a drawback in an era of limited space in a print journal. However, the ability to post information on the Web should alleviate this concern. The place in which supplementary information is presented can be decided by authors and editors of the individual journal.

We hope that the recommendations stimulate transparent and improved reporting of genetic association studies. In turn, better reporting of original studies would facilitate the synthesis of available research results and the further development of study methods in genetic epidemiology with the ultimate goal of improving the understanding of the role of genetic factors in the cause of diseases.

**Grant support**

By the Institutes of Genetics and of Nutrition, Metabolism and Diabetes, Canadian Institutes of Health Research; Genome Canada; Biotechnology, Genomics and Population Health Branch, Public Health Agency of Canada; Affymetrix; DNA Genotek; TrialStat!; and GeneSens. The funders had no role in the decision to submit the article or in its preparation.

**Acknowledgements**

The authors thank Kyle Vogan and Allen Wilcox for their participation in the workshop and for their comments; Michele Cargill (Affymetrix Inc) and Aaron del Duca (DNA Genotek) for their participation in the workshop as observers; and the Public Population Project in Genomics (P3G), hosted by the University of Montreal and supported by Genome Canada and Genome Quebec. This article was made possible thanks to input and discussion by the P3G International Working Group on Epidemiology and Biostatistics, discussion held in Montreal, May 2007. The authors also thank the reviewers for their very thoughtful feedback, and Silvia Visentin, Rob Moriarity, Morgan Macneill and Valery L'Heureux for administrative support. We were unable to contact Barbara Cohen to confirm her involvement in the latest version of this article.

---

**Table 1. STREGA reporting recommendations, extended from STROBE Statement.**

<table>
<thead>
<tr>
<th>Item</th>
<th>STROBE Guideline</th>
<th>Extension for Genetic Association Studies (STREGA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title and Abstract</td>
<td>1 Indicate the study’s design with a commonly used term in the title or the abstract.</td>
<td></td>
</tr>
</tbody>
</table>
(b) Provide in the abstract an informative and balanced summary of what was done and what was found. |
| Introduction | 2 Explain the scientific background and rationale for the investigation being reported. |
| Objectives | 3 State specific objectives, including any pre-specified hypotheses. |  
State if the study is the first report of a genetic association, a replication effort, or both. |
| Methods | 4 Present key elements of study design early in the paper. |
| Setting | 5 Describe the setting, locations and relevant dates, including periods of recruitment, exposure, follow-up, and data collection. |
| Participants | 6 (a) Cohort study – Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up. |  
Case-control study – Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls. |  
Give information on the criteria and methods for selection of subsets of participants from a larger study, when relevant. |
<table>
<thead>
<tr>
<th>Item</th>
<th>Item number</th>
<th>STROBE Guideline</th>
<th>Extension for Genetic Association Studies (STREGA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional study</td>
<td></td>
<td>Give the eligibility criteria, and the sources and methods of selection of participants.</td>
<td></td>
</tr>
<tr>
<td>(b) Cohort study</td>
<td></td>
<td>For matched studies, give matching criteria and number of exposed and unexposed.</td>
<td></td>
</tr>
<tr>
<td>Case-control study</td>
<td></td>
<td>For matched studies, give matching criteria and the number of controls per case.</td>
<td></td>
</tr>
<tr>
<td>Variables</td>
<td>7</td>
<td>(a) Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.</td>
<td>(b) Clearly define genetic exposures (genetic variants) using a widely-used nomenclature system. Identify variables likely to be associated with population stratification (confounding by ethnic origin).</td>
</tr>
<tr>
<td>Data sources measurement</td>
<td>6*</td>
<td>(a) For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.</td>
<td>(b) Describe laboratory methods, including source and storage of DNA, genotyping methods and platforms (including the allele calling algorithm used, and its version), error rates and call rates. State the laboratory/centre where genotyping was done. Describe comparability of laboratory methods if there is more than one group. Specify whether genotypes were assigned using all of the data from the study simultaneously or in smaller batches.</td>
</tr>
<tr>
<td>Bias</td>
<td>9</td>
<td>(a) Describe any efforts to address potential sources of bias.</td>
<td>(b) For quantitative outcome variables, specify if any investigation of potential bias resulting from pharmacotherapy was undertaken. If relevant, describe the nature and magnitude of the potential bias, and explain what approach was used to deal with this.</td>
</tr>
<tr>
<td>Study size</td>
<td>10</td>
<td>Explain how the study size was arrived at.</td>
<td>If applicable, describe how effects of treatment were dealt with.</td>
</tr>
<tr>
<td>Quantitative variables</td>
<td>11</td>
<td>Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why.</td>
<td></td>
</tr>
<tr>
<td>Statistical methods</td>
<td>12</td>
<td>(a) Describe all statistical methods, including those used to control for confounding.</td>
<td>State software version used and options (or settings) chosen.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b) Describe any methods used to examine subgroups and interactions.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c) Explain how missing data were addressed.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(d) Cohort study – If applicable, explain how loss to follow-up was addressed.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Case-control study – If applicable, explain how matching of cases and controls was addressed.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cross-sectional study – If applicable, describe analytical methods taking account of sampling strategy.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(e) Describe any sensitivity analyses.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(f) State whether Hardy-Weinberg equilibrium was considered and, if so, how.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(g) Describe any methods used for inferring genotypes or haplotypes.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(h) Describe any methods used to assess or address population stratification.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(i) Describe any methods used to address multiple comparisons or to control risk of false positive findings.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(j) Describe any methods used to address and correct for relatedness among subjects</td>
<td></td>
</tr>
<tr>
<td>Item</td>
<td>Item number</td>
<td>STROBE Guideline</td>
<td>Extension for Genetic Association studies (STREGA)</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>------------------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>Results</td>
<td>13*</td>
<td>(a) Report the numbers of individuals at each stage of the study – e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed.</td>
<td>Report numbers of individuals in whom genotyping was attempted and numbers of individuals in whom genotyping was successful.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b) Give reasons for non-participation at each stage.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c) Consider use of a flow diagram.</td>
<td></td>
</tr>
<tr>
<td>Descriptive data</td>
<td>14*</td>
<td>(a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders.</td>
<td>Consider giving information by genotype.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b) Indicate the number of participants with missing data for each variable of interest.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c) Cohort study – Summarize follow-up time, e.g., average and total amount.</td>
<td></td>
</tr>
<tr>
<td>Outcome data</td>
<td>15*</td>
<td>Cohort study: Report numbers of outcome events or summary measures over time.</td>
<td>Report outcomes (phenotypes) for each genotype category over time.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Case-control study – Report numbers in each exposure category, or summary measures of exposure.</td>
<td>Report numbers in each genotype category.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cross-sectional study – Report numbers of outcome events or summary measures.</td>
<td>Report outcomes (phenotypes) for each genotype category.</td>
</tr>
<tr>
<td>Main results</td>
<td>16</td>
<td>(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence intervals). Make clear which confounders were adjusted for and why they were included.</td>
<td>(d) Report results of any adjustments for multiple comparisons.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(b) Report category boundaries when continuous variables were categorized.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period.</td>
<td></td>
</tr>
<tr>
<td>Other analyses</td>
<td>17</td>
<td>(a) Report other analyses done – e.g., analyses of subgroups and interactions, and sensitivity analyses.</td>
<td>(b) If numerous genetic exposures (genetic variants) were examined, summarize results from all analyses undertaken.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(c) If detailed results are available elsewhere, state how they can be accessed.</td>
</tr>
<tr>
<td>Discussion</td>
<td>18</td>
<td>Summarize key results with reference to study objectives.</td>
<td></td>
</tr>
<tr>
<td>Limitations</td>
<td>19</td>
<td>Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias.</td>
<td></td>
</tr>
<tr>
<td>Interpretation</td>
<td>20</td>
<td>Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence.</td>
<td></td>
</tr>
<tr>
<td>Generalizability</td>
<td>21</td>
<td>Discuss the generalizability (external validity) of the study results.</td>
<td></td>
</tr>
<tr>
<td>Other Information</td>
<td>22</td>
<td>Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based.</td>
<td></td>
</tr>
</tbody>
</table>

STREGA = STrengthening the REporting of Genetic Association studies; STROBE = STrengthening the Reporting of Observational Studies in Epidemiology.

* Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.
### Table 2. Rationale for Inclusion of Topics in the STREGA Recommendations.

<table>
<thead>
<tr>
<th>Specific issues in Genetic Association Studies</th>
<th>Rationale for inclusion in STREGA Recommendations</th>
<th>Item(s) in STREGA</th>
<th>Specific suggestions for reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main areas of special interest</strong> (See also main text)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotyping errors (misclassification of exposure)</td>
<td>Non-differential genotyping errors will usually bias associations towards the null. When there are systematic differences in genotyping according to outcome status (differential error), bias in any direction may occur.</td>
<td>8(b): Describe laboratory methods, including source and storage of DNA, genotyping methods and platforms (including the allele calling algorithm used, and its version), error rates and call rates. State the laboratory centre where genotyping was done. Describe comparability of laboratory methods if there is more than one group. Specify whether genotypes were assigned using all of the data from the study simultaneously or in smaller batches. 13(a): Report numbers of individuals in whom genotyping was attempted and numbers of individuals in whom genotyping was successful.</td>
<td>Factors affecting the potential extent of misclassification (information bias) of genotype include the types and quality of samples, timing of collection, and the method used for genotyping. When high throughput platforms are used, it is important to report not only the platform used but also the allele calling algorithm and its version. Different calling algorithms have different strengths and weaknesses (59 and supplementary information 60). For example, some of the currently used algorithms are notably less accurate in assigning genotypes to single nucleotide polymorphisms with low minor allele frequencies (&lt;0.10) than to single nucleotide polymorphisms with higher minor allele frequencies (12). Algorithms are continually being improved. Reporting the allele calling algorithm and its version will help readers to interpret reported results, and it is critical for reproducing the results of the study given the same intermediate output files summarizing intensity of hybridization. For some high throughput platforms, the user may choose to assign genotypes using all of the data from the study simultaneously, or in smaller batches, such as by plate (61) and supplementary information (62). This choice can affect both the overall call rate and the robustness of the calls. For case-control studies, whether genotyping was done blind to case-control status should be reported, along with the reason for this decision.</td>
</tr>
<tr>
<td>Population stratification (confounding by ethnic origin)</td>
<td>When study sub-populations differ in allele (or genotype) frequencies and disease risks, then confounding will occur if these sub-populations are unevenly distributed across exposure groups (or between cases and controls).</td>
<td>12(b): Describe any methods used to assess or address population stratification.</td>
<td>In view of the debate about the potential implications of population stratification for the validity of genetic association studies, transparent reporting of the methods used, or stating that none was used, to address this potential problem is important for allowing the empirical evidence to accrue. Ethnicity information should be presented (see for example Winker 63), as should genetic markers or other variables likely to be associated with population stratification. Details of case-family control designs should be provided if they are used. As several methods of adjusting for population stratification have been proposed 64, explicit documentation of the methods is needed.</td>
</tr>
<tr>
<td>Modelling haplotype variation</td>
<td>In designs considered in this article, haplotypes have to be inferred because of lack of available family information. There are diverse methods for inferring haplotypes.</td>
<td>12(g): Describe any methods used for inferring haplotypes.</td>
<td>When discrete “windows” are used to summarize haplotypes, variation in the definition of these may complicate comparisons across studies, as results may be sensitive to choice of windows. Related “imputation” strategies are also in use 65, 66, 67. It is important to give details on haplotype inference and, when possible, uncertainty. Additional considerations for reporting include the strategy for dealing with rare haplotypes, window size and construction (if used) and choice of software.</td>
</tr>
<tr>
<td>Hardy-Weinberg equilibrium (HWE)</td>
<td>Departure from Hardy-Weinberg equilibrium may indicate errors or peculiarities in the data. Empirical assessments have found that 20%-69% of genetic associations were reported with some indication about conformity with Hardy-Weinberg equilibrium, and that among some of these, there were limitations or errors in its assessment 68.</td>
<td>12(f): State whether Hardy-Weinberg equilibrium was considered and, if so, how.</td>
<td>Any statistical tests or measures should be described, as should any procedure to allow for deviations from Hardy-Weinberg equilibrium in evaluating genetic associations 69.</td>
</tr>
<tr>
<td>Replication</td>
<td>Publications that present and synthesize data from several studies in a single report are becoming more common.</td>
<td>3: State if the study is the first report of a genetic association, a replication effort, or both.</td>
<td>The selected criteria for claiming successful replication should also be explicitly documented.</td>
</tr>
</tbody>
</table>

**Additional issues**

| Selection of participants | Selection bias may occur if (i) genetic associations are investigated in one or more subsets of participants (sub-samples) from a particular study; or (ii) there is differential non-participation in groups being compared; or (iii) there are differential genotyping call rates in groups being compared. | 6(a): Give information on the criteria and methods for selection of subsets of participants from a larger study, when relevant. 13(a): Report numbers of individuals in whom genotyping was attempted and numbers of individuals in whom genotyping was successful. | Inclusion and exclusion criteria, sources and methods of selection of sub-samples should be specified, stating whether these were based on a priori or post hoc considerations. |
### Rationale for choice of genes and variants investigated

Without an explicit rationale, it is difficult to judge the potential for selective reporting of study results. There is strong empirical evidence from randomised controlled trials that reporting of trial outcomes is frequently incomplete and bias in favour of statistically significant findings. Some evidence is also available in pharmacoepidemiology.

#### 7
- Clearly define genetic exposures (genetic variants) using a widely-used nomenclature system. Identify variables likely to be associated with population stratification (confounding by ethnic origin).

#### 8
- For quantitative outcome variables, specify if any investigation of potential bias resulting from pharmacotherapy was undertaken. If relevant, describe the nature and magnitude of the potential bias, and explain what approach was used to deal with this.
- If applicable, describe how effects of treatment were dealt with.

### Treatment effects in studies of quantitative traits

A study of a quantitative variable may be compromised when the trait is subjected to the effects of a treatment for example, the study of a lipid-related trait for which several individuals are taking lipid-lowering medication. Without appropriate correction, this can lead to bias in estimating the effect and loss of power.

#### 9
- Several methods of adjusting for treatment effects have been proposed. As the approach to deal with treatment effects may have an important impact on both the power of the study and the interpretation of the results, explicit documentation of the selected strategy is needed.

### Statistical methods

**Analysis methods** should be transparent and replicable, and genetic association studies are often performed using specialised software.

#### 10
- **State software version used and options (or settings) chosen.**

### Relatedness

The methods of analysis used in family-based studies are different from those used in studies that are based on unrelated cases and controls. Moreover, even in the studies that are based on apparently unrelated cases and controls, some individuals may have some connection and may be (distant) relatives, and this is particularly common in small, isolated populations, for example, Iceland. This may need to be probed with appropriate methods and adjusted for in the analysis of the data.

#### 11
- **Describe any methods used to address or correct for relatedness among subjects.**

### Reporting of descriptive and outcome data

The synthesis of findings across studies depends on the availability of sufficiently detailed data.

#### 12
- **Consider giving information by genotype.**
- **Cohort study - Report outcomes (phenotypes) for each genotype category over time.**
- **Case-control study - Report numbers in each genotype category.**
- **Cross-sectional study - Report outcomes (phenotypes) for each genotype category.**

### Volume of data

The key problem is of possible false-positive results and selective reporting of these. Type 1 errors are particularly relevant to the conduct of genome-wide association studies. A large search among hundreds of thousands of genetic variants can be expected by chance alone to find thousands of false positive results (odds ratios significantly different from 1.0).

#### 13
- **Describe any methods used to address multiple comparisons or to control risk of false positive findings.**
- **Report results of any adjustments for multiple comparisons.**
- **If numerous genetic exposures (genetic variants) were examined, summarise results from all analyses undertaken.**
- **If detailed results are available elsewhere, state how they can be accessed.**

Genome-wide association studies collect information on a very large number of genetic variants concomitantly. Initiatives to make the entire database transparent and available online may supply a definitive solution to the problem of selective reporting.

Availability of raw data may help interested investigators reproduce the published analyses and also pursue additional analyses. A potential drawback of public data availability is that investigators using the data second-hand may not be aware of limitations or other problems that were originally encountered, unless those are also transparently reported. In this regard, collaboration of the data users with the original investigators may be beneficial. Issues of consent and confidentiality may also complicate what data can be shared, and how. It would be useful for published reports to specify not only what data can be accessed and where, but also briefly mention the procedure. For articles that have used publicly available data, it would be useful to clarify whether the original investigators were also involved and if so, how.

The volume of data analyzed should also be considered in the interpretation of findings.

Examples of methods of summarizing results include giving distribution of P values (frequentist statistics), distribution of effect sizes and specifying false discovery rates.
In order to encourage dissemination of the STREGA Statement, this article has also been published by *Annals of Internal Medicine*, *European Journal of Clinical Investigation*, *European Journal of Epidemiology*, *Genetic Epidemiology*, *Human Genetics*, *Journal of Clinical Epidemiology*, and *PLoS Medicine.* The authors jointly hold the copyright of this article. Please note that M. J. Khoury and M. Gwinn are US Federal Government employees and therefore copyright for these authors is in the public domain.

### References

107) International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype...
116) van Hylckama Vlieg A, Sandkuijl LA, Rosendaal FR, Bertina RM, Vos HL. Candidate gene approach in association studies: Would the factor V leiden mutation have been found by this approach?. Eur J Hum Genet 2004;12:478-82.
154) Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science 2008;322(5898):44.
Genetic epidemiology faces the challenge of understanding and compiling evidence regarding the contribution of genetic risk in common human diseases. This discipline is closely related to traditional epidemiology, focusing on the genetic determinants of diseases and on the joint effects of genes and non-genetic factors [1]. As Shpilberg and colleagues optimistically stated, ‘The sequencing of the human genome offers the greatest opportunity for epidemiology since John Snow discovered the Broad Street pump’ [2]. The understanding of this challenge, together with a tremendous advance in genotyping technologies and the inclusion of DNA collection in observational studies, has resulted in an ever increasing number of genetic association studies published over the past years. The literature database edited by the CDC [3] reveals that the annual number of reports on genetic association has increased impressively, with more than 34,000 reports published since 2000. More recently, the ability to perform genome-wide analyses is providing tantalizing new clues to disease causation and therapeutic targets. All these accelerated discoveries of susceptibility genes for common complex diseases fuel expectations that sooner or later genomic information will become an integral part of personalized health care and disease prevention, even though this seems unlikely to happen as quickly as some have claimed [4].

The speed of progress in genetic epidemiology research, however, has not been followed by an equally fast and substantial improvement in the methodological aspects of genetic association studies design, the appraisal and/or the reporting. Several efforts in the past 5 years have been made by different research group to share best practices, tools and methods for design, analysis and appraisal of genetic association studies [5-7]. These have been rarely applied, however, and a possible explanation is that the different recommendations sometimes diverge substantially. This failure prompted a multidisciplinary work group, within the large Human Genome Epidemiology Network, to develop a checklist reporting the essential guidelines to Strengthening the Reporting of Genetic Association studies (STREGA), which has been published in several journals simultaneously [8]. This initiative follows the previously issued recommendations for the improvement of observational studies reporting quality (case-control, cohort and cross-sectional) in conventional epidemiology, namely the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) [9]. As Authors acknowledge, despite the fact that traditional epidemiologic studies share many similarities with genetic association studies, there are specific challenges that need to be addressed in the latter, mainly due to the high likelihood of modest effects related to one single genetic variant, and the vast amount of genetic data that require replication. The STREGA statement does not pass judgment on how researchers should carry out the study; rather it advocates the transparency, quality and completeness in reporting how the study was performed and what was found. STREGA reports 12 additional items to the previous 22 of the STROBE checklist, mainly focused on 5 areas that are of special interest in genetic association studies: genotyping errors, population stratification, modelling haplotype variation, Hardy-Weinberg Equilibrium (HWE), and replication. A systematic review published in 2008 [10] evaluating the reporting practices of 315 randomly selected genetic association studies, showed that 93% of them did not report the validation of genotype data by using duplicate samples. This aspect, together with the evidence that more than 50% of these 315 papers did not test for a possible departure of the HWE [10], is a clear hint that a misclassification of the genotype under study might occurred. As for the issue of population stratification, 41.3% of the studies examined by Yesuspira A et al. [10] did not state whether the study participants were drawn from...
the same ethnic population, which might result in residual confounding. Lastly, the vast majority of the published reports do not state whether their study is the first on a specific issue, or whether it is a replication of a previous study. As such, there is a lot of room for improvement in the reporting of genetic association studies, and in this sense STREGA could assist authors by providing essential information, making articles complete and easy to read. In conclusion we invite journals to endorse STREGA by including the checklist and its internet address in the Instructions for Authors, and advising authors and peer reviewers to use the checklist as a guide, which is what the Italian Journal of Public Health has done in the present issue.

References