Progressive Statistics for Studies in Sports Medicine and Exercise Science

WILLIAM G. HOPKINS[1], STEPHEN W. MARSHALL[2], ALAN M. BATTERHAM[3], JURI HANIN[4]

[1]Institute of Sport and Recreation Research, AUT University, Auckland, NZ; [2]Departments of Epidemiology, Orthopedics, and Exercise & Sport Science, University of North Carolina at Chapel Hill, Chapel Hill, NC; [3]School of Health and Social Care, University of Teesside, Middlesbrough, UK; [4]KIHU-Research Institute for Olympic Sports, Jyvaskyla, Finland.

ABSTRACT

HOPKINS, W. G., S. W. MARSHALL, A. M. BATTERHAM, and J. HANIN. Progressive Statistics for Studies in Sports Medicine and Exercise Science. *Med. Sci. Sports Exerc.*, Vol. 41, No. 1, pp. 3–12, 2009. Statistical guidelines and expert statements are now available to assist in the analysis and reporting of studies in some biomedical disciplines. We present here a more progressive resource for sample-based studies, meta-analyses and case studies in sports medicine and exercise science. We offer forthright advice on the following controversial or novel issues: using precision of estimation for inferences about population effects in preference to null-hypothesis testing, which is inadequate for assessing clinical or practical importance; justifying sample size via acceptable precision or confidence for clinical decisions rather than via adequate power for statistical significance; showing standard deviations rather than standard errors of the mean, to better communicate magnitude of differences in means and non-uniformity of error; avoiding purely non-parametric analyses, which cannot provide inferences about magnitude and are unnecessary; using regression statistics in validity studies, in preference to the impractical and biased limits of agreement; making greater use of qualitative methods to enrich sample-based quantitative projects; and seeking ethics approval for public access to the depersonalized raw data of a study, to address the need for more scrutiny of research and better meta-analyses. Advice on less contentious issues includes: using covariates in linear models to adjust for confounders, to account for individual differences, and to identify potential mechanisms of an effect; using log transformation to deal with non-uniformity of effects and error; identifying and deleting outliers; presenting descriptive, effect and inferential statistics in appropriate formats; and contending with bias arising from problems with sampling, assignment, blinding, measurement error, and researchers' prejudices. This article should advance the field by stimulating debate, promoting innovative approaches, and serving as a useful checklist for authors, reviewers and editors. **Key Words:** ANALYSIS, CASE, DESIGN, INFERENCE, QUALITATIVE, QUANTITATIVE, SAMPLE

In response to the widespread misuse of statistics in research, several biomedical organizations have published statistical guidelines in their journals, including the International Committee of Medical Journal Editors (www.icmje.org), the American Psychological Association (2), and the American Physiological Society (8). Expert groups have also produced statements about how to publish reports of various kinds of medical research (Table 1). Some medical journals now include links to these statements as part of their instructions to authors.

In this article we provide our view of best practice for the use of statistics in sports medicine and the exercise sciences. The article is similar to those referenced in Table 1 but includes more practical and original material. It should achieve three useful outcomes. First, it should stimulate interest and debate about constructive change in the use of statistics in our disciplines. Secondly, it should help legitimize the innovative or controversial approaches that we and others sometimes have difficulty including in publications. Finally, it should serve as a statistical checklist for researchers, reviewers and editors at the various stages of the research process. Not surprisingly, some of the

TABLE 1. Recent statements of best practice for reporting various kinds of biomedical research.

Interventions (experiments)	CONSORT: Consolidated Standards of Reporting Trials (1,22). See consort-statement.org for statements, explanations and extensions to abstracts and to studies involving equivalence or non-inferiority, clustered randomization, harmful outcomes, non-randomized designs, and various kinds of intervention.
Observational (non-experimental) studies	STROBE: Strengthening the Reporting of Observational Studies in Epidemiology (27,28). See strobe-statement.org for statements and explanations, and see HuGeNet.ca for extension to gene-association studies.
Diagnostic tests	STARD: Standards for Reporting Diagnostic Accuracy (5,6).
Meta-analyses	QUOROM: Quality of Reporting of Meta-analyses (21). MOOSE: Meta-analysis of Observational Studies in Epidemiology (25). See also the Cochrane Handbook (at cochrane.org) and guidelines for meta-analysis of diagnostic tests (19) and of gene-association studies (at HuGeNet.ca).

Address for correspondence: Will G. Hopkins, Ph.D., FACSM, Institute of Sport and Recreation Research, AUT University, Akoranga Drive, Private Bag 92006, Auckland 0627, New Zealand; E-mail: will@clear.net.nz.
Submitted for publication July 2008.
Accepted for publication September 2008.

0195-9131/09/4101-0003/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE_

DOI: 10.1249/MSS.0b013e31818cb278

reviewers of this article disagreed with some of our advice, so we emphasize here that the article represents neither a general consensus amongst experts nor editorial policy for this journal. Indeed, some of our innovations may take decades to become mainstream.

Most of this article is devoted to advice on the various kinds of sample-based studies that comprise the bulk of research in our disciplines. Table 2 and the accompanying notes deal with issues common to all such studies, arranged in the order that the issues arise in a manuscript. This table applies not only to the usual studies of samples of individuals but also to meta-analyses (in which the sample consists of various studies) and quantitative non-clinical case studies (in which the sample consists of repeated observations on one subject). Table 3, which should be used in conjunction with Table 2, deals with additional advice specific to each kind of sample-based study and with clinical and qualitative single-case studies. The sample-based studies in this table are arranged in the approximate descending order of quality of evidence they provide for causality in the relationship between a predictor and dependent variable, followed by the various kinds of methods studies, meta-analyses, and the single-case studies. For more on causality and other issues in choice of design for a study, see Reference (14).

TABLE 2. Generic statistical advice for sample-based studies.

ABSTRACT

- State why you studied the effect(s).
- State the design, including any randomizing and blinding.
- Characterize the subjects who contributed to the estimate of the effect(s) (*final* sample size, sex, skill, status…).
- Ensure all numbers are either in numeric or graphical form in the Results section of the manuscript.
- Show magnitudes and confidence intervals or limits of the most important effect(s). Avoid P values. [Note 1]
- Make a probabilistic statement about clinical, practical, or mechanistic importance of the effect(s).
- The conclusion must not be simply a restatement of results.

INTRODUCTION

- Explain the need for the study.
  - Justify choice of a particular population of subjects.
  - Justify choice of design here, if it is one of the reasons for doing the study.
- State an achievable aim or resolvable question about the magnitude of the effect(s). Avoid hypotheses. [Note 1]

METHODS

Subjects

- Explain the recruitment process and eligibility criteria for acquiring the sample from a population.
  - Justify any stratification aimed at proportions of subjects with certain characteristics in the sample.
- Include permission for public access to depersonalized raw data in your application for ethics approval. [Note 2]

Design

- Describe any pilot study aimed at measurement properties of the variables and feasibility of the design.
- To justify sample size, avoid adequate power for statistical significance. Instead, estimate or reference the smallest important values for the most important effects and use with one or more of the following approaches, taking into account any multiple inferences and quantification of individual differences or responses [Notes 3, 4]:
  - adequate precision for a trivial outcome, smallest expected outcome, or comparison with a published outcome;
  - acceptably low rates of wrong clinical decisions;
  - adequacy of sample size in similar published studies;
  - limited availability of subjects or resources (in which case state the smallest magnitude of effect your study could estimate adequately).
- Detail the timings of all assessments and interventions.
- See also Table 3 for advice on design of specific kinds of study.

Measures

- Justify choice of *dependent* and *predictor variables* in terms of practicality and measurement properties specific to the subjects and conditions of the study. Use variables with the smallest errors.
- Justify choice of potential *moderator variables*: subject characteristics or differences/changes in conditions or protocols that could affect the outcome and that are included in the analysis as predictors to reduce confounding and account for individual differences.
- Justify choice of potential *mediator variables*: measures that could be associated with the dependent variable because of a causal link from a predictor and that are included in an analysis of the mechanism of the effect of the predictor. [Note 5]
- Consider including open-ended interviews or other qualitative methods, which afford serendipity and flexibility in data acquisition.
  - Use in a pilot phase aimed at defining purpose and methods, during data gathering in the project itself, and in a follow-up assessment of the project with stakeholders.

Analysis

- Describe any initial screening for miscodings, for example using stem-and-leaf plots or frequency tables.
- Justify any imputation of missing values and associated adjustment to analyses.
- Describe the model used to derive the effect. [Note 6]
  - Justify inclusion or exclusion of main effects, polynomial terms and interactions in a linear model.
  - Explain the theoretical basis for use of any non-linear model.
  - Provide citations or evidence from simulations that any unusual or innovative data-mining technique you used to derive effects (neural nets, genetic algorithms, decision trees, rule induction) should give trustworthy estimates with your data.
  - Explain how you dealt with repeated measures or other clustering of observations.
- Avoid purely non-parametric analyses. [Note 7]
- If the dependent variable is continuous, indicate whether you dealt with non-uniformity of effects and/or error by transforming the dependent variable, by modeling different errors in a single analysis, and/or by performing and combining separate analyses for independent groups. [Note 8]
- Explain how you identified and dealt with outliers, and give a plausible reason for their presence. [Note 9]
- Indicate how you dealt with the magnitude of the effect of linear continuous predictors or moderators, either as the effect of 2 SD, or as a partial correlation, or by parsing into independent subgroups. [Note 10]
- Indicate how you performed any subsidiary mechanisms analysis with potential mediator variables, either using linear modeling or (for interventions) an analysis of change scores. [Note 5]
- Describe how you performed any sensitivity analysis, in which you investigated quantitatively, either by simulation or by simple calculation, the effect of error of measurement and other potential sources of bias on the magnitude and uncertainty of the effect statistic(s).
- Explain how you made inferences about the true (infinite-sample) value of each effect. [Note 1]
  - Show confidence intervals or limits.
  - Justify a value for the smallest important magnitude, then base the inference on the disposition of the confidence interval relative to substantial magnitudes.
  - For effects with clinical or practical application, make a decision about

utility by estimating chances of benefit and harm.
  - Avoid the traditional approach of statistical significance based on a null-hypothesis test using a P value.
  - Explain any adjustment for multiple inferences. [Note 3]
- Include this statement, when appropriate: measures of centrality and dispersion are mean ± SD.
  - Add the following statement, when appropriate: for variables that were log transformed before modeling, the mean shown is the back-transformed mean of the log transform, and the dispersion is a coefficient of variation (%) or ×/÷ factor SD.
  - The range (minimum-maximum) is sometimes informative, but beware that it is strongly biased by sample size.
  - Avoid medians and other quantiles, except when parsing into subgroups.
  - Never show standard errors of means. [Note 11]
- See also Table 3 for advice on analysis of specific kinds of study.

RESULTS

Subject Characteristics

- Describe the flow of number of subjects from those who were first approached about participation through those who ended up providing data for the effects.
- Show a table of descriptive statistics of variables in important groups of the subjects included in the final analysis, not the subjects you first recruited.
  - For numeric variables, show mean ± SD. [Note 11]
  - For nominal variables, show percent of subjects.
  - Summarize the characteristics of dropouts (subjects lost to follow-up) if they represent a substantial proportion (>10%) of the original sample or if their loss is likely to substantially bias the outcome. Be precise about which groups they were in when they dropped out and why they dropped out.
- See also Table 3 for advice on reporting subject characteristics in specific kinds of study.

Outcome Statistics

- Avoid all exact duplication of data between tables, figures, and text.
- When adjustment for subject characteristics and other potential confounders is substantial, show unadjusted and adjusted outcomes.
- Use standardized differences or changes in means to assess qualitative magnitudes of the differences, but there is generally no need to show the standardized values. [Note 1]
- If the most important effect is unclear, provide a qualitative interpretation of its uncertainty. (For example, it is unlikely to have a small beneficial effect and very unlikely to be moderately beneficial.) State the approximate sample size that would be needed to make it clear.
- See also Table 3 for advice on outcome statistics in specific kinds of study.

Numbers

- Insert a space between numbers and units, with the exception of % and °. Examples: 70 $ml.min^{-1}.kg^{-1}$; 90%.
- Insert a hyphen between numbers and units only when grammatically necessary: the test lasted 4 min; it was a 4-min test.
- Ensure that units shown in column or row headers of a table are consistent with the data in the cells of the table.
- Round up numbers to improve clarity.
  - Round up percents, SD, and the "±" version of confidence limits to two significant digits. A third digit is sometimes appropriate to convey adequate accuracy when the first digit is "1"; for example, 12.6% vs 13%. A single digit is often appropriate for small percents (<1%) and some subject characteristics.
  - Match the precision of the mean to the precision of the SD. In these properly presented examples, the true values of the means are the same, but they are rounded differently to match their different SD: 4.567 ± 0.071, 4.57 ± 0.71, 4.6 ± 7.1, 5 ± 71, 0 ± 710, 0 ± 7100.
  - Similarly, match the precision of an effect statistic to that of its confidence limits.
- Express a confidence *interval* using "to" (e.g., the effect was 3.2 units; 90% confidence interval -0.3 to 6.7 units) or express confidence *limits* using "±" (3.2 units; 90% confidence limits ±3.5 units).
  - Drop the wording "90% confidence interval/limits" for subsequent effects, but retain consistent punctuation (e.g., 2.1%; ±3.6%). Note that there is a semicolon or comma before the "±" and no space after it for confidence limits, but there is a space and no other punctuation each side of a "±" denoting an SD. Check your abstract and results sections carefully for consistency of such punctuation.
  - Confidence limits for effects derived from back-transformed logs can be expressed as an exact ×/÷factor by taking the square root of the upper limit divided by the lower limit. Confidence limits of measurement errors and of other SD can be expressed in the same way, but the resulting ×/÷factor becomes less accurate as degrees of freedom fall below 10.
- When effects and confidence limits derived via log transformation are less than ~±25%, show as percent effects; otherwise show as factor effects. Examples: -3%, -14 to 6%; 17%, ±6%; a factor of 0.46, 0.18 to 1.15; a factor of 2.3, ×/÷1.5.
- Do not use P-value inequalities, which oversimplify inferences and complicate or ruin subsequent meta-analysis.
  - Where brevity is required, replace with the ± or ×/÷ form of confidence limits. Example: "active group 4.6 units, control group 3.6 units (P>0.05)" becomes "active group 4.6 units, control group 3.6 units (95% confidence limits ±1.3 units)".
  - If you accede to an editor's demand for P values, use two significant digits for P≥0.10 and one for P<0.10. Examples: P=0.56, P=0.10, P=0.07, P=0.003, P=0.00006 (or 6E-5).

Figures

- Use figures sparingly and only to highlight key outcomes.
- Show a scattergram of individual values or residuals only to highlight the presence and nature of unusual non-linearity or non-uniformity.
  - Most non-uniformity can be summarized non-graphically, succinctly and more informatively with appropriate SD for appropriate subgroups.
  - Do not show a scattergram of individual values that can be summarized by a correlation coefficient. (Exception: validity studies.)
- Use line diagrams for means of repeated measurements. Use bar graphs for single observations of means of groups of different subjects.
- In line diagrams and scattergrams, choose symbols to highlight similarities and differences in groups or treatments.
  - Make the symbols too large rather than too small.
  - Explain the meaning of symbols using a key on the figure rather than in the legend.
  - Place the key sensibly to avoid wasting space.
  - Where possible, label lines directly rather than via a key.
- Use a log scale for variables that required log transformation when the range of values plotted is greater than ~×1.25.
- Show SD of group means to convey a sense of magnitude of effects. [Note 11]
  - For mean change scores, convey magnitude by showing a bar to the side indicating one SD of composite baseline scores.
- In figures summarizing effects, show bars for confidence intervals rather than asterisks for P values.
  - State the level of confidence on the figure or in the legend.
  - Where possible, show the range of trivial effects on the figure using shading or dotted lines. Regions defining small, moderate and large effects can sometimes be shown successfully.

DISCUSSION

- Avoid restating any numeric values exactly, other than to compare your findings with those in the literature.
- Avoid introducing new data.
- Be clear about the population to which your effect statistics apply, but consider their wider applicability.
- Interpret a mechanisms analysis cautiously. [Note 5]
- Assess the possible bias arising from the following sources:
  - confounding by non-representativeness or imbalance in the sampling or assignment of subjects, when the relevant subject characteristics could affect the dependent variable and have not been adjusted for by inclusion in the model;

- random or systematic error in a continuous variable or classification error in a nominal variable; [Note 12]
- choosing the largest or smallest of several effects that have overlapping confidence intervals; [Note 3]
- your prejudices or desire for an outcome, which can lead you to filter data inappropriately and misinterpret effects.

Note 1

Inferences are evidence-based conclusions about the true nature of something. The traditional approach to inferences in research on samples is an assertion about whether the effect is statistically significant or "real", based on a P value. Specifically, when the range of uncertainty in the true value of an effect represented by the 95% confidence interval does not include the zero or null value, P is <0.05, the effect "can't be zero", so the null hypothesis is rejected and the effect is termed significant; otherwise P is >0.05 and the effect is non-significant. A fundamental theoretical dilemma with this approach is the fact that the null hypothesis is always false; indeed, with a large enough sample size all effects are statistically significant. On a more practical level, the failure of this approach to deal adequately with the real-world importance of an effect is evident in the frequent misinterpretation of a non-significant effect as a null or trivial effect, even when it is likely to be substantial. A significant effect that is likely to be trivial is also often misinterpreted as substantial.

A more realistic and intuitive approach to inferences is based on where the confidence interval lies in relation to threshold values for substantial effects rather than the null value (4). If the confidence interval includes values that are substantial in some positive and negative sense, such as beneficial and harmful, you state in plain language that the effect could be substantially positive *and* negative, or more simply that the effect is *unclear*. Any other disposition of the confidence interval relative to the thresholds represents a clear outcome that can be reported as trivial, positive or negative, depending on the observed value of the effect. Such magnitude-based inferences about effects can be made more accurate and informative by qualifying them with probabilities that reflect the uncertainty in the true value: *possibly harmful, very likely substantially positive, unclear but likely to be beneficial,* and so on. The qualitative probabilistic terms can be assigned using the following scale (16): <0.5%, most unlikely, almost certainly not; 0.5-5%, very unlikely; 5-25%, unlikely, probably not; 25-75%, possibly; 75-95%, likely, probably; 95-99.5%, very likely; >99.5%, most likely, almost certainly. Research on the perception of probability could result in small adjustments to this scale.

Use of thresholds for moderate and large effects allows even more informative inferential assertions about magnitude, such as *probably moderately positive, possibly associated with small increase in risk, almost certainly large gain,* and so on. As yet, only a few effect statistics have generally accepted magnitude thresholds for this purpose. Thresholds of 0.1, 0.3 and 0.5 for small, moderate and large correlation coefficients suggested by Cohen (7) can be augmented with 0.7 and 0.9 for very large and extremely large; these translate approximately into 0.20, 0.60, 1.20, 2.0 and 4.0 for standardized differences in means (the mean difference divided by the between-subject SD) and into risk differences of 10%, 30%, 50%, 70% and 90% (see newstats.org/effectmag.html). The latter applied to chances of a medal provide thresholds for change in an athlete's competition time or distance of 0.3, 0.9, 1.6, 2.5 and 4.0 of the within-athlete variation between competitions (17 and WGH, unpublished observations). Magnitude thresholds for risk, hazard and odds ratios require more research, but a risk ratio as low as 1.1 for a factor affecting incidence or prevalence of a condition should be important for the affected population group, even when the condition is rare. Thresholds have been suggested for some diagnostic statistics (20), but more research is needed on these and on thresholds for the more usual measures of validity and reliability.

An appropriate default level of confidence for the confidence interval is 90%, because it implies quite reasonably that an outcome is clear if the true value is very unlikely to be substantial in a positive and/or negative sense. Use of 90% rather than 95% has also been advocated as a way of discouraging readers from reinterpreting the outcome as significant or non-significant at the 5% level (24). In any case, a symmetrical confidence interval of whatever level is appropriate for making only non-clinical or mechanistic inferences. An inference or decision about clinical or practical utility should be based on probabilities of harm and benefit that reflect the greater importance of avoiding use of a harmful effect than failing to use a beneficial effect. Suggested default probabilities for declaring an effect clinically beneficial are <0.5% (most unlikely) for harm and >25% (possible) for benefit (16). A clinically unclear effect is therefore possibly beneficial (>25%) with an unacceptable risk of harm (>0.5%). These probabilities correspond to a ratio of ~60 for odds of benefit to odds of harm, a suggested default when sample sizes are sub- or supra-optimal (16). Note that even when an effect is unclear, you can often make a useful probabilistic statement about how big or small it could be, and your findings should contribute to a meta-analysis.

Magnitude-based inferences as outlined above represent a subset of the kinds of inference that are possible using so-called Bayesian statistics, in which the researcher combines the study outcome with uncertainty in the effect prior to the study to get the posterior (updated) uncertainty in the effect. A qualitative version of this approach is an implicit and important part of the Discussion section of most studies, but in our view specification of the prior uncertainty is too subjective to apply the approach quantitatively. Researchers may also have difficulty accessing and using the computational procedures. On the other hand, confidence limits and probabilities related to threshold magnitudes can be derived readily via a spreadsheet (16) by making the same assumptions about sampling distributions that statistical packages use to derive P values. Bootstrapping, in which a sampling distribution for an effect is derived by resampling from the original sample thousands of times, also provides a robust approach to computing confidence limits and magnitude-based probabilities when data or modeling are too complex to derive a sampling distribution analytically.

Note 2

Public Access to depersonalized data, when feasible, serves the needs of the wider community by allowing more thorough scrutiny of data than that afforded by peer review and by leading to better meta-analyses. Make this statement in your initial application for ethics approval, and state that the data will be available indefinitely at a website or on request without compromising the subjects' privacy.

Note 3

Multiple Inferences. Any conclusive inference about an effect could be wrong, and the more effects you investigate, the greater the chance of making an error. If you test multiple hypotheses, there is inflation of the Type I error rate: an increase in the chance that a null effect will turn up statistically significant. The usual remedy of making the tests more conservative is not appropriate for the most important pre-planned effect, it is seldom applied consistently to all other effects reported in a paper, and it creates problems for meta-analysts and other readers who want to assess effects in isolation. We therefore concur with others (e.g., 23) who advise against adjusting the Type I error rate or confidence level of confidence intervals for multiple effects.

For several important clinical or practical effects, you should constrain the increase in the chances of making clinical errors. Overall chances of benefit and harm for several interdependent effects can be estimated properly by bootstrapping, but a more practical and conservative approach is to assume the effects are independent and to estimate errors approximately by addition. The sum of the chances of harm of all the effects that separately are clinically useful should not exceed 0.5% (or your chosen maximum rate for Type 1 clinical errors; Note 4); otherwise you should declare fewer effects useful and acknowledge that your study is underpowered. Your study is also underpowered if the sum of chances of benefit of all effects that separately are not clinically useful exceeds 25% (or your chosen Type 2 clinical error rate). When your sample size is small, reduce the chance that the study will be underpowered by designing and analyzing it for fewer effects.

A problem with inferences about several effects with overlapping confidence intervals is misidentification of the largest (or smallest) and upward (or downward) bias in its magnitude. In simulations the bias is of the order of the average standard error of the outcome statistic, which is approximately one-third the width of the average 90% confidence interval (WGH, unpublished observations). Acknowledge such bias when your aim is to quantify the largest or smallest of several effects.

Note 4

Sample Sizes that give acceptable precision with 90% confidence limits are similar to those based on a Type 1 clinical error of 0.5% (the chance of using an effect that is harmful) and a Type 2 clinical error of 25% (the chance of not using an effect that is beneficial). The sample sizes are approximately one-third those based on the traditional approach of an 80% chance of statistical significance at the 5% level when the true effect has the smallest important value. Until hypothesis testing loses respectability, you should include the traditional and new approaches in applications for ethical approval and funding.

Whatever approach you use, sample size needs to be quadrupled to adequately estimate individual differences or responses and effects of covariates on the main effect. Larger samples are also needed to keep clinical error rates for clinical or practical decisions acceptable when there is more than one important effect in a study (Note 3). See Reference (12) for a spreadsheet and details of these and many other sample-size issues.

Note 5

Mechanisms. In a mechanisms analysis, you determine the extent to which a putative mechanism variable mediates an effect through being in a causal chain linking the predictor to the dependent variable of the effect. For an effect derived from a linear model, the contribution of the mechanism (or mediator) variable is represented by the reduction in the effect when the variable is included in the model as another predictor. Any such reduction is a necessary but not sufficient condition for the variable to contribute to the mechanism of the effect, because a causal role can be established definitively only in a separate controlled trial designed for that purpose.

For interventions, you can also examine a plot of change scores of the dependent variable vs those of potential mediators, but beware that a relationship will not be obvious in the scattergram if individual responses are small relative to measurement error. Mechanism variables are particularly useful in unblinded interventions, because evidence of a mechanism that cannot arise from expectation (placebo or nocebo) effects is also evidence that at least part of the effect of the intervention is not due to such effects.

Note 6

Linear Models. An effect statistic is derived from a model (equation) linking a dependent (the "Y" variable) to a predictor and usually other predictors (the "X" variables or covariates). The model is linear if the dependent can be expressed as a sum of terms, each term being a coefficient times a predictor or a product of predictors (interactions, including polynomials), plus one or more terms for random errors. The effect statistic is the predictor's coefficient or some derived form of it. It follows from the additive nature of such models that the value of the effect statistic is formally equivalent to the value expected when the other predictors in the model are held constant. Linear models therefore automatically provide adjustment for potential confounders and estimates of the effect of potential mechanism variables. A variable that covaries with a predictor and dependent variable is a confounder if it causes some of the covariance and is a mechanism if it mediates it. The reduction of an effect when such a variable is included in a linear model is the contribution of the variable to the effect, and the remaining effect is independent of (adjusted for) the variable.

The usual models are linear and include: regression, ANOVA, general linear and mixed for a continuous dependent; logistic regression, Poisson regression, negative binomial regression and generalized linear modeling for

events (a dichotomous or count dependent); and proportional-hazards regression for a time-to-event dependent. Special linear models include factor analysis and structural equation modeling.

For repeated measures or other clustering of observations of a continuous dependent variable, avoid the problem of interdependence of observations by using within-subject modeling, in which you combine each subject's repeated measurements into a single measure (unit of analysis) for subsequent modeling; alternatively, account for the interdependence using the more powerful approach of mixed (multilevel or hierarchical) modeling, in which you estimate different random effects or errors within and between clusters. Avoid repeated-measures ANOVA, which sometimes fails to account properly for different errors. For clustered event-type dependents (proportions or counts), use generalized estimation equations.

Note 7

Non-parametric Analysis. A requirement for deriving inferential statistics with the family of general linear models is normality of the sampling distribution of the outcome statistic. Although there is no test that data meet this requirement, the central-limit theorem ensures that the sampling distribution is close enough to normal for accurate inferences, even when sample sizes are small (~10) and especially after a transformation that reduces any marked skewness in the dependent variable or non-uniformity of error. Testing for normality of the dependent variable and any related decision to use purely non-parametric analyses (which are based on rank transformation and do not use linear or other parametric models) are therefore misguided. Such analyses lack power for small sample sizes, do not permit adjustment for covariates, and do not permit inferences about magnitude. Rank transformation followed by parametric analysis can be appropriate (Note 8), and ironically, the distribution of a rank-transformed variable is grossly non-normal.

Note 8

Non-uniformity of effect or error in linear models can produce incorrect estimates and confidence limits. Check for non-uniformity by comparing standard deviations of the dependent variable in different subgroups or by examining plots of the dependent variable or its residuals for differences in scatter (heteroscedasticity) with different predicted values and/or different values of the predictors.

Differences in standard deviations or errors between groups can be taken into account for simple comparisons of means by using the unequal-variances t statistic. With more complex models use mixed modeling to allow for and estimate different standard deviations in different groups or with different treatments. For a simpler robust approach with independent subgroups, perform separate analyses then compare the outcomes using a spreadsheet (15).

Transformation of the dependent variable is another approach to reducing non-uniformity, especially when there are differences in scatter for different predicted values. For many dependent variables, effects and errors are uniform when expressed as factors or percents; log transformation converts these to uniform additive effects, which can be modeled linearly then expressed as factors or percents after back transformation. Always use log transformation for such variables, even when a narrow range in the dependent variable effectively eliminates non-uniformity.

Rank transformation eliminates non-uniformity for most dependent variables and models, but it results in loss of precision with a small sample size and should therefore be used as a last resort. To perform the analysis, sort all observations by the value of the dependent variable, assign each observation a rank (consecutive integer), then use the rank as the dependent variable in a liner model. Such analyses are often referred to incorrectly as non-parametric.

Use the transformed variable, not the raw variable, to gauge magnitudes of correlations and of standardized differences or changes in means. Back-transform the mean effect to a mean in raw units and its confidence limits to percents or factors (for log transformation) or to raw units at the mean of the transformed variable or at an appropriate value of the raw variable (for all other transformations). When analysis of a transformed variable produces impossible values for an effect or a confidence limit (e.g., a negative rank with the rank transformation), the assumption of normality of the sampling distribution of the effect is violated and the analysis is therefore untrustworthy. Appropriate use of bootstrapping avoids this problem.

Note 9

Outliers for a continuous dependent variable represent a kind of non-uniformity that appears on a plot of residuals vs predicteds as individual points with much larger residuals than other points. To delete the outliers in an objective fashion, set a threshold by first standardizing the residuals (dividing by their standard deviation). The resulting residuals are t statistics, and with the assumption of normality, a threshold for values that would occur rarely (<5% of the time is a good default) depends on sample size. Approximate sample sizes and thresholds for the absolute value of t are: <~50, >3.5; ~500, >4.0; ~5000, >4.5; ~50,000, >5.0. Some packages identify outliers more accurately using statistics that account for the lower frequency of large residuals further away from the mean predicted value of the dependent.

Note 10

Effect of Continuous Predictors. The use of two standard deviations (SD) to gauge the effect of a continuous predictor ensures congruence between Cohen's threshold magnitudes for correlations and standardized differences (Note 1). Two SD of a normally distributed predictor also corresponds approximately to the mean separation of lower and upper tertiles (2.2 SD). The SD is ideally the variation in the predictor after adjustment for other predictors; the effect of 2 SD in a correlational study is then equivalent to, and can be replaced by, the partial correlation (the square root of the fraction of variance explained by the predictor after adjustment for all other predictors).

A grossly skewed predictor can produce incorrect estimates or confidence limits, so it should be transformed to reduce skewness. Log transformation is often suitable for skewed predictors that have only positive values; as simple linear predictors their effects are then expressed per factor

or percent change of their original units. Alternatively, a skewed predictor can be parsed into quantiles (usually 2-5 subgroups with equal numbers of observations) and included in the model as a nominal variable or as an ordinal variable (a numeric variable with integer values). Parsing is also appropriate for a predictor that is likely to have a non-linear effect not easily or realistically modeled as a polynomial.

Note 11

SEM vs SD. The standard error of the mean (SEM = SD/√(group sample size)) is the sampling variation in a group mean, which is the expected typical variation in the mean from sample to sample. Some researchers argue that, as such, this measure communicates uncertainty in the mean and is therefore preferable to the SD. A related widespread belief is that non-overlap of SEM bars on a graph indicates a difference that is statistically significant at the 5% level. Even if statistical significance was the preferred approach to inferences, this belief is justified only when the SEM in the two groups are equal, and for comparisons of changes in means, only when the SEM are for means of change scores. Standard error bars on a time-series graph of means of repeated measurements thus convey a false impression of significance or non-significance, and therefore, to avoid confusion, SEM should not be shown for any data. In any case, researchers are interested not in the uncertainty in a single mean but in the uncertainty of an effect involving means, usually a simple comparison of two means. Confidence intervals or related inferential statistics are used to report uncertainty in such effects, making the SEM redundant and inferior.

The above represents compelling arguments for not using the SEM, but there are even more compelling arguments for using the SD. First, it helps to assess non-uniformity, which manifests as different SD in different groups. Secondly, it can signpost the likely need for log transformation, when the SD of a variable that can have only positive values is of magnitude similar to or greater than the mean. Finally and most importantly, the SD communicates the magnitude of differences or changes between means, which by default should be assessed relative to the usual between-subject SD (Note 1). The manner in which the SEM depends on sample size makes it unsuitable for any of these applications, whereas the SD is practically unbiased for sample sizes ~10 or more (9).

Note 12

Error-related Bias. Random error or random misclassification in a variable attenuates effects involving the variable and widens the confidence interval. (Exception: random error in a continuous dependent variable does not attenuate effects of predictors on means of the variable.) After adjustment of the variable for any systematic difference from a criterion in a validity study with subjects similar to those in your study, it follows from statistical first principles that the correction for attenuation of an effect derived directly from the variable's coefficient in a linear model is $1/v^2$, where v is the validity correlation coefficient; the correction for a correlation with the variable is 1/v. In this context, a useful estimate for the upper bound of v is the square root of the short-term reliability correlation.

When one variable in an effect has *systematic* error or misclassification that is substantially correlated with the value of the other variable, the effect will be biased up or down, depending on the correlation. Example: a spurious beneficial effect of physical activity on health could arise from healthier people exaggerating their self-reported activity.

Substantial random or systematic error of measurement in a covariate used to adjust for confounding results in partial or unpredictable adjustment respectively and thereby renders untrustworthy any claim about the presence or absence of the effect after adjustment. This problem applies also to a mechanisms analysis involving such a covariate.

TABLE 3. Additional statistical advice for specific sample-based and single-case designs.

INTERVENTIONS

Design

- Justify any choice between pre-post vs post-only and between parallel-group vs crossover designs. Avoid single-group (uncontrolled) designs if possible. See Reference (3) for more.
- Investigate more than one experimental treatment only when sample size is adequate for multiple comparisons. [Note 4]
- Explain any randomization of subjects to treatment groups or treatment sequences, any stratification (balancing of numbers in subject-characteristic subgroups), and any minimization of differences of means of subject characteristics in treatment groups. State whether/how randomization to groups or sequences was concealed from researchers.
- Detail any blinding of subjects and researchers.
- Detail the timing and nature of assessments and interventions.

Analysis

- Indicate how you included, excluded or adjusted for subjects who showed substantial non-compliance with protocols or treatments or who were lost to follow-up.
- In a parallel-groups trial, estimate and adjust for the potential confounding effect of any substantial differences in mean characteristics between groups.
  - In pre-post trials in particular, estimate and adjust for the effect of baseline score of the dependent variable on the treatment effect. Such adjustment eliminates any effect of regression to the mean, whereby a difference between groups at baseline arising from error of measurement produces an artifactual treatment effect.

Subject Characteristics

- For continuous dependent and mediator variables, show mean and SD in the subject-characteristics table only at baseline.

Outcome Statistics: Continuous Dependents

(For event-type dependents, see the section below on prospective cohort studies.)

- Baseline means and SD that appear in text or a table can be duplicated in a line diagram summarizing means and SD at all assay times
- Show means and SD of change scores in each group.
- Show the unadjusted and any relevant adjusted differences between the mean changes in treatment and control (or exposed and unexposed) groups, with confidence limits.
- Show the standard error of measurement derived from repeated baseline tests and/or pre-post change scores in a control group.
- Include an analysis for individual responses derived from the SD of the change scores. In post-only crossovers this analysis requires an as-

sumption about, or separate estimation of, error of measurement over the time between treatments.

Discussion

- If there was lack or failure of blinding, estimate bias due to placebo and nocebo effects (outcomes better and worse than no treatment arising purely from expectation with the experimental and control treatments respectively).

COHORT STUDIES

Design

- Describe the methods of follow-up.

Analysis

- Indicate how you included, excluded or adjusted for subjects who showed substantial non-compliance with protocols or treatments or who were lost to follow-up.
- Adjust effects for any substantial difference between groups at baseline.

Outcome Statistics: Event Dependents

(For continuous dependents, see the section above on interventions.)

- When the outcome is assessed at fixed time points, show percentage of subjects in each group who experienced the event at each point.
- When subjects experience multiple events, show raw or factor means and SD of counts per subject.
- When the outcome is time to event, display survival curves for the treatment or exposure groups.
- Show effects as the risk, odds or hazard ratios adjusted for relevant subject characteristics.
  - Present them also in a clinically meaningful way by making any appropriate assumptions about incidence, prevalence, or exposure to convert the ratios to risks (proportions affected) and risk difference between groups or for different values of predictors, along with confidence limits for the risk ratio and/or risk difference (18).
  - Adjusted mean time to event and its ratio or difference between groups is a clinically useful way to present some outcomes.

Discussion

- Take into account the fact that confounding can bias the risk ratio by as much as ×/÷2.0-3.0 in most cohort and case-control studies (26).

CASE-CONTROL STUDIES

Design

- Explain how you tried to choose controls from the same population giving rise to the cases.
- Justify the case:control ratio. (Note that >5 controls per case or >5 cases per control give no useful increase in precision.)
- *Case-crossovers*: describe how you defined the time windows for assessing case and control periods.

Outcome Statistics

- Present risk-factor outcomes in a clinically meaningful way by converting the odds ratio (which is a hazard ratio with incidence density sampling) to a risk ratio and/or risk difference between control and exposed subjects in an equivalent cohort study over a realistic period (18).

Discussion

- See the Discussion point on confounding in cohort studies.
- Estimate bias due to under-matching, over-matching or other mismatching of controls.

CROSS-SECTIONAL STUDIES

Outcome Statistics

- Show simple unadjusted effects and effects adjusted for all other predictors in the model.

STRUCTURAL EQUATION MODELING

Analysis

- Specify the measurement and structural models using a path diagram.
- Explain the estimation method and the strategy for assessing goodness of fit.
- Demonstrate that all parameters were estimable.

MEASUREMENT STUDIES: VALIDITY

Design

- Justify the cost-effectiveness of the criterion measure, citing studies of its superiority and measurement error.

Analysis

- Use linear or non-linear regression to estimate a calibration equation, a standard error of the estimate, and a validity correlation coefficient.
  - For criterion and practical measures in the same metric, use the calibration equation to estimate bias in the practical measure over its range.
  - Do not calculate limits of agreement or present a Bland-Altman plot. [Note 13]

MEASUREMENT STUDIES: DIAGNOSTIC TESTS

Design

- Document the diagnostic accuracy of the method or combination of methods used as the reference standard.

Analysis

- Calculate the following diagnostic measures, all of which can be useful: the validity correlation coefficient, the kappa coefficient, sensitivity, specificity, positive and negative predictive values, positive and negative diagnostic likelihood ratios, and diagnostic odds ratio.
- For a continuous measure, calculate area under the ROC curve and give the above diagnostic measures for an appropriate threshold.

MEASUREMENT STUDIES: RELIABILITY

Design

- Justify number of trials, raters, items of equipment and/or subjects needed to estimate the various within and between standard deviations.
- Justify times between trials to establish effects due to familiarization (habituation), practice, learning, potentiation, and/or fatigue.

Analysis

- Assess habituation and other order-dependent effects in simple reliability studies by deriving statistics for consecutive pairs of measurements.
- The reliability statistics are the change in the mean between measurements, the standard error of measurement (typical error), and the appropriate intraclass correlation coefficient (or the practically equivalent test-retest Pearson correlation).
  - Do not abbreviate standard error of measurement as SEM, which is confused with standard error of the mean.
  - Avoid limits of agreement. [Note 13]
- With several levels of repeated measurement (e.g., repeated sets, different raters for the same subjects) use judicious averaging or preferably mixed modeling to estimate different errors as random effects.

MEASUREMENT STUDIES: FACTOR STRUCTURE

Design

- Describe any pilot work with experts and subjects to develop or modify wording in any exploratory factor analysis.

Analysis

- Specify the analytic approach (principal components or principal axes), the criteria used to extract factors, the rotation method and factor-loading cutoffs for selection of variables for each factor, and the communalities to justify exclusion of items from the instrument.
- For confirmatory factor analysis use an appropriate structural equation model.
- For each factor calculate the square root of Cronbach's alpha, which is an upper bound for the validity correlation.

META-ANALYSES

Design

- Describe the search strategy and inclusion criteria for identifying relevant studies.
- Explain why you excluded specific studies that other researchers might consider worthy of inclusion.

Analysis
- Explain how you reduced study-estimates to a common metric.
  - Conversion to factor effects (followed by log transformation) is often appropriate for means of continuous variables.
  - Avoid standardization (dividing each estimate by the between-subject SD) until *after* the analysis, using an appropriate between-subject composite SD derived from some or all studies.
  - Hazard ratios are often best for event outcomes.
- Explain derivation of the weighting factor (inverse of the sampling variance, or adjusted sample size if sufficient authors do not provide sufficient inferential information).
- Avoid fixed-effect meta-analysis. State how you performed a random-effect analysis to allow for real differences between study-estimates. With sufficient studies, adjust for study characteristics by including them as fixed effects, and account for any clustering of study-estimates by including extra random effects.
- Use a plot of standard error or 1/√(sample size) vs study-estimate or preferably the t statistic of the solution of each random effect to explore the possibility of publication bias and to identify outlier study-estimates.
- To gauge the effect of 2 SD of predictors [Note 10] representing mean *subject* characteristics, use an appropriate mean of the between-subject SD from selected or all studies, not the SD of the study means.

Study Characteristics
- Show a table of study characteristics, study-estimates, inferential information (provided by authors) and confidence limits (computed by you, when necessary).
  - If the table is too large for publication, make it available at a website or on request.
  - A one-dimensional plot of effects and confidence intervals ("forest plot") represents unnecessary duplication of data in the above table.
- Show a scatterplot of study-estimates with confidence limits to emphasize a relationship with a study characteristic.

SINGLE-CASE STUDIES: QUANTITATIVE NON-CLINICAL

Design
- Regard these as sample-based studies aimed at an inference about the value of an effect statistic in the population of repeated observations on a single subject.
- Justify the choice of design by identifying the closest sample-based design.
- Take into account within-subject error when estimating "sample size" (number of repeated observations).
- State the smallest important effect, which should be the same as for a usual sample-based study.

Analysis
- Account for trends in consecutive observations with appropriate predictors.
  - Check for any remaining autocorrelation, which will appear as a trend in the scatter of a plot of residuals vs time or measurement number.
  - Use an advanced modeling procedure that allows for autocorrelation only if there is a trend that modeling can't remove.
- Make it clear that the inferences apply only to your subject and possibly only to a certain time of year or state.
- Perform separate single-subject analyses when there is more than one case. With an adequate sample of cases, use the usual sample-based repeated-measures analyses.

SINGLE-CASE STUDIES: CLINICAL

Case Description
- For a difficult differential diagnosis, justify the use of appropriate tests by reporting their predictive power, preferably as positive and negative diagnostic likelihood ratios.

Discussion
- Where possible, use a quantitative Bayesian (sequential probabilistic) approach to estimate the likelihoods of contending diagnoses.

SINGLE-CASE STUDIES: QUALITATIVE

Methods
- State and justify your ideological paradigm (e.g., grounded theory).
- Describe your methods for gathering the information, including any attempt to demonstrate congruence of data and concepts by triangulation (use of different methods).
- Describe your formal approach to organizing the information (e.g., dimensions of form, content or quality, magnitude or intensity, context, and time (10)).
- Describe how you reached saturation, when ongoing data collection and analysis generated no new categories or concepts.
- Describe how you solicited feedback from respondents, peers and experts to address trustworthiness of the outcome.
- Analyze a sufficiently large sample of cases or assessments of an individual by coding the characteristics and outcomes of each case (assessment) into variables and by following the advice for the appropriate sample-based study. [Note 14]

Results and Discussion
- Address the likelihood of alternative interpretations or outcomes.
- To generalize beyond a single case or assessment, consider how differences in subject or case characteristics could have affected the outcome.

Note 13

Limits of Agreement. Bland and Altman introduced limits of agreement (defining a reference interval for the difference between measures) and a plot of subjects' difference vs mean scores of the measures (for checking relative bias and non-uniformity) to address what they thought were shortcomings arising from misuse of validity and reliability correlation coefficients in measurement studies. Simple linear regression nevertheless provides superior statistics in validity studies, for the following reasons: the standard error of the estimate and the validity correlation can show that a measure is entirely suitable for clinical assessment of individuals and for sample-based research, yet the measure would not be interchangeable with a criterion according to the limits of agreement; the validity correlation provides a correction for attenuation (see Note 12), but no such correction is available with limits of agreement; the regression equation provides trustworthy estimates of the bias of one measure relative to the other, whereas the Bland-Altman plot shows artifactual bias for measures with substantially different errors (11); regression statistics can be derived in all validity studies, whereas limits of agreement can be derived from difference scores in only a minority of validity studies ("method-comparison" studies, where both measures are in the same units); finally, limits of agreement in a method-comparison study of a new measure with an existing imprecise measure provide no useful information about the validity of the new measure, whereas the regression validity statistics can be combined with published validity regression statistics for the imprecise measure to correctly estimate validity regression statistics for the new measure.

Arguments have also been presented against the use of limits of agreement as a measure of reliability (13). Additionally, data generally contain several sources of random error, which are invariably estimated as variances in linear models then combined and expressed as standard errors of

measurement and/or correlations. Transformation to limits of agreement is of no further clinical or theoretical value.

Note 14

Qualitative Inferences. Some qualitative researchers believe that it is possible to use qualitative methods to generalize from a sample of qualitatively analyzed cases (or assessments of an individual) to a population (or the individual generally). Others do not even recognize the legitimacy of generalizing. In our view, generalizing is a fundamental obligation that is best met quantitatively, even when the sample is a series of qualitative case studies or assessments.

Chris Bolter, Janet Dufek, Doug Curran-Everett, Patria Hume, George Kelley, Ken Quarrie, Chris Schmid, David Streiner and Martyn Standage provided valuable feedback on drafts, as did nine reviewers on the submitted manuscript. The authors have no professional relationship with a for-profit organization that would benefit from this study; publication does not constitute endorsement by ACSM.

No funding was received for this work from any organization, other than salary support for the authors from their respective institutions.

Editor-in-Chief 's note: This article by Hopkins et al. should be considered invited commentary by the authors. The article has undergone peer review by eight other scientists, each an acknowledged expert in experimental design, statistical analysis, data interpretation, and reporting, and the authors have undertaken extensive revision in response to those reviews and my own reviews. The majority of reviewers recommended publication of the article, but there remain several specific aspects of the discussion, on which authors and reviewers strongly disagreed. Therefore, the Associate Editors and I believe that our scientific community has not yet achieved sufficient "consensus" to establish formal editorial policy about appropriate reporting of research design, data analysis, and results. However, we also believe that Dr. Hopkins and his colleagues have presented a thoughtful, provocative framework of "progressive" recommendations, which merit consideration and discussion. Readers are advised that the recommendations remain the authors' opinion and not the journal's editorial policy, and we encourage your feedback.

REFERENCES

1. Altman DG, Schulz KF, Moher D et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med*. 2001;134:663-94.
2. Anonymous. Publication Manual of the American Psychological Association. 5th ed. Washington DC: APA; 2001.
3. Batterham AM, Hopkins WG. A decision tree for controlled trials. *Sportscience*. 2005;9:33-9.
4. Batterham AM, Hopkins WG. Making meaningful inferences about magnitudes. *Int J Sports Physiol Perform*. 2006;1:50-7. *Sportscience*. 2005;9:6-13.
5. Bossuyt PM, Reitsma JB, Bruns DE et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ*. 2003;326:41-4.
6. Bossuyt PM, Reitsma JB, Bruns DE et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49:7-18.
7. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1988. 567 p.
8. Curran-Everett D, Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society. *J Appl Physiol*. 2004;97:457-9.
9. Gurland J, Tripathi RC. A simple approximation for unbiased estimation of the standard deviation. *Am Stat*. 1971;25(4):30-2.
10. Hanin YL. Performance related emotional states in sport: a qualitative analysis. *Forum: Qualitative Social Research*. 2003;4(1):qualitative-research.net/fqs-texte/1-03/1-hanin-e.htm.
11. Hopkins WG. Bias in Bland-Altman but not regression validity analyses. *Sportscience*. 2004;8:42-6.
12. Hopkins WG. Estimating sample size for magnitude-based inferences. *Sportscience*. 2006;10:63-70.
13. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Med*. 2000;30:1-15.
14. Hopkins WG. Research designs: choosing and fine-tuning a design for your study. *Sportscience*. 2008;12:12-21.
15. Hopkins WG. A spreadsheet for combining outcomes from several subject groups. *Sportscience*. 2006;10:51-3.
16. Hopkins WG. A spreadsheet for deriving a confidence interval, mechanistic inference and clinical inference from a p value. *Sportscience*. 2007;11:16-20.
17. Hopkins WG, Hawley JA, Burke LM. Design and analysis of research on sport performance enhancement. *Med Sci Sports Exerc*. 1999;31:472-85.
18. Hopkins WG, Marshall SW, Quarrie KL, Hume PA. Risk factors and risk statistics for sports injuries. *Clin J Sport Med*. 2007;17:208-10.
19. Irwig L, Tosteson ANA, Gatsonis C et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med*. 1994;120:667-76.
20. Jaeschke R, Guyatt G, Sackett DL. Users'guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA*. 1994;271:389-91.
21. Moher D, Cook DJ, Eastwood S. Improving the quality of reports of meta-analyses of randomised controlled trials. *Lancet*. 1999;354:1896-900.
22. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *Ann Intern Med*. 2001;134:657-62.
23. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316:1236-8.
24. Sterne JAC, Smith GD. Sifting the evidence–what's wrong with significance tests. *BMJ*. 2001;322:226-31.
25. Stroup DF, Berlin JA, Morton SC et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA*. 2000;283:2008-12.
26. Taubes G. Epidemiology faces its limits. *Science*. 1995;269:164-9.
27. Vandenbroucke JP, von Elm E, Altman DG et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Ann Intern Med*. 2007;147:W163-W94.
28. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med*. 2007;147:573-7.