A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration

Mohammad Ali Mansournia ,[1,2] Gary S Collins,[3,4] Rasmus Oestergaard Nielsen ,[5,6] Maryam Nazemipour,[7] Nicholas P Jewell,[8,9] Douglas G Altman,[3] Michael J Campbell[10]

For numbered affiliations see end of article.

Correspondence to
Professor Mohammad Ali Mansournia, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, 14155-6446 Tehran, Iran; mansournia_ma@yahoo.com and Dr Maryam Nazemipour, Psychosocial Health Research Institute, Iran University of Medical Sciences, 14665-354 Tehran, Iran; nazemipour.m@iums.ac.ir

Douglas G Altman's deceased date: June 3, 2018

Accepted 7 January 2021
Published Online First 29 January 2021

ABSTRACT

Misuse of statistics in medical and sports science research is common and may lead to detrimental consequences to healthcare. Many authors, editors and peer reviewers of medical papers will not have expert knowledge of statistics or may be unconvinced about the importance of applying correct statistics in medical research. Although there are guidelines on reporting statistics in medical papers, a checklist on the more general and commonly seen aspects of statistics to assess when peer-reviewing an article is needed. In this article, we propose a CHecklist for statistical Assessment of Medical Papers (CHAMP) comprising 30 items related to the design and conduct, data analysis, reporting and presentation, and interpretation of a research paper. While CHAMP is primarily aimed at editors and peer reviewers during the statistical assessment of a medical paper, we believe it will serve as a useful reference to improve authors' and readers' practice in their use of statistics in medical research. We strongly encourage editors and peer reviewers to consult CHAMP when assessing manuscripts for potential publication. Authors also may apply CHAMP to ensure the validity of their statistical approach and reporting of medical research, and readers may consider using CHAMP to enhance their statistical assessment of a paper.

The misuse of statistics by implementing flawed methodology in medical and sports science research can lead to unreliable or even incorrect conclusions. The consequences of flawed methodology can have undesirable consequences on public health, patient management and athlete performance.[1] Unfortunately, errors in the study design, statistical analysis, reporting and interpretation of results are common in medical journals[2 3] and raise questions regarding the quality of medical papers.[4]

Sound methodology has been prioritised in the past decades, especially in high-impact factor journals. This is illustrated by the inclusion of more statistical editors and other methodologists (eg, epidemiologists) in the review process. In addition, stakeholders in research have been encouraged to intensify their investments in statistical, epidemiological and methodological education, such as training authors and reviewers, providing online resources, developing (and extending) guidelines and including methods content in regular scientific meetings.[5] There has also been a stronger emphasis on adherence to reporting guidelines (eg, CONsolidated Standards Of Reporting Trials, STrengthening the Reporting of OBservational studies in Epidemiology, STAndards for the Reporting of Diagnostic accuracy studies, REporting recommandations for tumor MARKer prognostic studies, and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis).[6–10]

Still, many medical and sports science journals do not involve statistical experts in the review process. This is unfortunate because the existence of basic statistical errors is more likely when authors, editors and referees do not have sufficient knowledge of statistics and, worse, are unconvinced about the importance of correct statistics in medical research. Rarely do clinical journals systematically assess the use of statistics in submitted papers.[11 12] Thus, even after a paper is published in a scientific journal, it is necessary to read the content with some caution and pay careful attention to whether the statistical design and analysis were appropriate and the conclusions justified. Studies published in high-ranked journals are not immune from methodological or statistical flaws, which were not identified during the peer review process. Although some journals attempt to mitigate against such issues by using statisticians in the review process (as statistical reviewers or statistical editors), guidelines to assess methodological or statistical content in scientific papers would be useful when expert statistical reviewers are unavailable.[5 13 14]

While guidelines on how to report statistics in medical papers exist,[15 16] we propose a general checklist to judge the statistical aspects of a manuscript during peer review. While it is impossible to cover everything, we believe it would be useful to have a basic checklist for assessing the statistical methods used more broadly within medical and sports science research papers. Based on an extensive revision of a previous checklist,[17] we describe CHecklist for statistical Assessment of Medical Papers (CHAMP; table 1) comprising 30 items in the design, analysis, reporting and interpretation stages to aid the peer review of a submitted paper.[18]

DEVELOPMENT AND EXPLANATION OF THE 30-ITEM CHECKLIST

The 30 items in the checklist were selected based on a previous BMJ checklist,[17] an extensive literature review and the authors' collective

Linked
► https://doi.org/10.1136/bjsports-2020-103651
Check for updates

To cite: Mansournia MA, Collins GS, Nielsen RO, *et al*. *Br J Sports Med* 2021;**55**:1009–1017.

Table 1 CHecklist for statistical Assessment of Medical Papers

Design and conduct			
1. Clear description of the goal of research, study objective(s), study design and study population	Yes	Unclear	No
2. Clear description of outcomes, exposures/treatments and covariates, and their measurement methods	Yes	Unclear	No
3. Validity of the study design	Yes	Unclear	No
4. Clear statement and justification of sample size	Yes	Unclear	No
5. Clear declaration of design violations and acceptability of the design violations	Yes	Unclear	No
6. Consistency between the paper and its previously published protocol	Yes	Unclear	No
Data analysis			
7. Correct and complete description of statistical methods	Yes	Unclear	No
8. Valid statistical methods used and assumptions outlined	Yes	Unclear	No
9. Appropriate assessment of treatment effect or interaction between treatment and another covariate	Yes	Unclear	No
10. Correct use of correlation and associational statistical testing	Yes	Unclear	No
11. Appropriate handling of continuous predictors	Yes	Unclear	No
12. CIs do not include impossible values	Yes	Unclear	No
13. Appropriate comparison of baseline characteristics between the study arms in randomised trials	Yes	Unclear	No
14. Correct assessment and adjustment of confounding	Yes	Unclear	No
15. Avoiding model extrapolation not supported by data	Yes	Unclear	No
16. Adequate handling of missing data	Yes	Unclear	No
Reporting and presentation			
17. Adequate and correct description of the data	Yes	Unclear	No
18. Descriptive results provided as occurrence measures with CIs and analytical results provided as association measures and CIs along with p values	Yes	Unclear	No
19. CIs provided for the contrast between groups rather than for each group	Yes	Unclear	No
20. Avoiding selective reporting of analyses and p-hacking	Yes	Unclear	No
21. Appropriate and consistent numerical precisions for effect sizes, test statistics and p values, and reporting the p values rather than their range	Yes	Unclear	No
22. Providing sufficient numerical results that could be included in a subsequent meta-analysis	Yes	Unclear	No
23. Acceptable presentation of figures and tables	Yes	Unclear	No
Interpretation			
24. Interpreting the results based on association measures and 95% CIs along with p values, and correctly interpreting large p values as indecisive results, not evidence of absence of an effect	Yes	Unclear	No
25. Using CIs rather than post hoc power analysis for interpreting the results of studies	Yes	Unclear	No
26. Correctly interpreting occurrence or association measures	Yes	Unclear	No
27. Distinguishing causation from association and correlation	Yes	Unclear	No
28. Results of prespecified analyses are distinguished from the results of exploratory analyses in the interpretation	Yes	Unclear	No
29. Appropriate discussion of the study methodological limitations	Yes	Unclear	No
30. Drawing only conclusions supported by the statistical analysis and no generalisation of the results to subjects outside the target population	Yes	Unclear	No

experience in reviewing the statistical content of numerous papers submitted to a variety of medical journals. The first author produced a checklist draft, the coauthors suggested the addition or removal of items, and all authors approved the final version. Other colleagues provided extensive comments on the paper and are listed in Acknowledgements. Our checklist is not intended to, nor can it, cover all aspects of medical statistics. Our focus is rather on key issues that generally arise in clinical research studies. Therefore, only common statistical issues encountered during the review of research manuscripts were included in CHAMP. Using our checklist requires some primary knowledge of statistics; however, we provide a brief explanation for each item and cite the relevant references for further details. The first six items relate to the design and conduct of research, items 7–16 address data analysis, items 17–23 concern reporting and presentation, and items 24–30 pertain to interpretation.

ITEMS 1–6: DESIGN AND CONDUCT

Item 1: clear description of the goal of research, study objective(s), study design and study population

The research goal, study objectives, study design, and study and target populations must be clearly described so that the editors of journals and readers can judge the internal and external validity (generalisability) of the study.

Being explicit about the goal of research is a prerequisite for good science regardless of the scientific discipline. For such clarification, a threefold classification of the research goal may be used: (1) to describe; (2) to predict, which is equivalent to identifying 'who' is at greater risk of experiencing the outcome; or (3) to draw a causal inference, which attempts to explain 'why' the outcome occurs (eg, investigating causal effects).[5 19]

The study objective refers to the rationale behind the study and points to the specific scientific question being addressed. For example, the objective of the heated water-based exercise (HEx) trial, a randomised controlled trial (RCT), was to evaluate the effect of heated water-based exercise training on

24-hour ambulatory blood pressure (BP) levels among patients with resistant hypertension.[20] The study objective is usually provided in the introduction after the rationale has been established.

The study design refers to the type of the study, which is explained in the Methods section.[21] Examples of common study designs include RCTs and observational studies such as cohort, case–control or cross-sectional studies.[22] The study design should be described in detail. In particular, the randomisation procedure in RCTs, follow-up time for cohort studies, control selection for case–control designs and sampling procedure for cross-sectional studies should be adequately explained.[6 7] As a general principle, the study design must be explained sufficiently so that another investigator would be able to repeat the study exactly.

The study population refers to the source population from which data are collected, whereas the target population refers to the population to whom we are going to generalise the study results; the relationship between these two populations may be characterised using inclusion and exclusion criteria and is crucial for assessing generalisability. Returning to the HEx trial, the study population was restricted to persons whose ages were between 40 and 65 years with resistant hypertension for more than 5 years.[20] For both trials and observational studies, it is important to know what proportion of the source population is studied and what proportion of the intended data set is used in the analysis data set. For example, the source population may include all patients admitted to a hospital with a certain condition over a certain period of time. However, the analysis data set may only be 50% of this, for various reasons such as patients refusing consent, measurements not taken and patients dropping out. In the HEx trial, for instance, they screened 125 patients with hypertension to find 32 who met the inclusion criteria with resistant hypertension. This has some bearing on the generalisability of the study and to whom heated water-based exercise training can be given and how likely it is to be relevant to practitioners.

Item 2: clear description of outcomes, exposures/treatments and covariates, and their measurement methods

All variables considered for statistical analysis should be stated clearly in the paper, including outcomes, exposures/treatments, predictors and potential confounders/mediators/effect-measure modifiers (see Box 1). The measurement method and timing of measurement for each of these variables should also be specified. If the goal of the research is to draw a causal inference (explain 'why' the outcome occurs) via observational studies, authors should present their causal assumptions in a causal diagram.[23–25] To exemplify this concept, in a cohort study evaluating the effect of physical activity on functional performance and knee pain in patients with osteoarthritis,[26] physical activity (exposure) was measured using the Physical Activity Scale for the Elderly, and functional performance and self-reported knee pain (outcomes) were measured by the Timed 20-Metre Walk Test and the Western Ontario and McMaster Universities Osteoarthritis Index, respectively. Depressive symptoms were considered a potential confounder and measured using the Center for Epidemiologic Studies Depression Scale. All variables were measured at baseline and in three annual follow-up visits, and a causal method along with a causal diagram representing the study population was used to estimate the effect of interest.[26]

Box 1 Glossary

Association: Statistical dependence, referring to any relationship between two variables.
Association measure: A measure of association between two variables, either in absolute or in relative terms. Absolute association measures are differences in occurrence measures e.g., risk difference and rate difference. Relative association measures are ratios of occurrence measures e.g., risk ratio, rate ratio, and odds ratio.
Causal diagram (Causal directed acyclic graph (DAG)): A diagram which includes nodes linked by directed arrows and has two properties: (i) the absence of an arrow between two variables implies the absence of a direct causal effect, and (ii) all shared causes of any pair of variables are included in the graph.
Collider: A variable that is a common effect of two other variables.
Effect-measure modifier: A variable that modifies the effect of the exposure on the outcome.
Confounder: A variable that is on the common cause path (confounding path) of the exposure and outcome.
Confounding: A bias created by a common cause of the exposure and outcome.
Correlation: Any monotonic (either entirely non-increasing or entirely non-decreasing) association.
Data dredging (Data fishing): The misuse of data analysis to find patterns which can be presented as statistically significant.
Design effect (in survey): The ratio of the variance of an estimator from a sampling scheme to the variance of the estimator from simple random sampling with the same sample size.
Effect (Causal effect): In the potential outcome (counterfactual) framework of causation, we say that A has a (causal) effect on B in a population of units if there is at least one unit for which changing A will change B.
Linearity assumption: An assumption underlying regression models imposed by inclusion of quantitative predictors which should be assessed.
Mediator: A variable that is affected by the exposure and also affects the outcome.
Null hypothesis: A hypothesis which is assumed in hypothesis testing, often corresponds to no association between two variables in the population.
Occurrence measure: A measure of disease frequency such as risk (incidence proportion), incidence rate, and prevalence.
Sparse-data bias: A bias arisen as a result of sparse data, leading to inflation of the effect size estimates.

Item 3: validity of the study design

The design should be valid and match the research question without introducing bias in the study results. For example, an editor should be able to assess whether the controls in a case–control study were adequately representative of the source population of the cases. Alternatively, in a clinical trial, it should be clear whether there was one (or more) control groups, and if so, whether patients were randomised to treatment or control, and if so, whether the randomisation method and allocation concealment were appropriate.

Item 4: clear statement and justification of sample size

The manuscript should have a section clearly justifying the sample size.[27] When a sample size calculation is warranted, the sample size section should be described in enough detail

to allow replication of the estimate, along with a clear rationale (supported by references) on choice of values used in the calculation, the outcome for which the calculation is based on, including the minimum clinically important effect size.[28 29] For example, typical sample size calculations aim to ensure that the study contains a sufficiently large precision for estimates of occurrence measures (eg, risk) or association measures (eg, risk ratio),[30 31] or that there is an adequate power to detect genuine effects (eg, true differences) if they exist (statistical tests). Attrition/loss to follow-up/non-response and design effects (eg, due to clustering) should be taken into consideration. Guidance for sample size calculation for prediction model development and validation has been described previously.[32–34]

Item 5: clear declaration of design violations and acceptability of the design violations

Design violations frequently occur in research. Non-response in surveys, censoring (loss to follow-up or competing risks) in prospective studies[35] and non-compliance with the study treatments in RCTs are examples and should be declared explicitly in the paper.[36 37] Given the validity of the design, the acceptability of violations should be assessed. For example, was an observed non-response/censoring proportion too high? What were the reasons for data loss, and is this level acceptable to achieve the scientific goals of the study?

Item 6: consistency between the paper and its previously published protocol

The reviewer should identify inconsistencies with any published protocol (and where relevant, registry information) regarding important features of the study, including sample size, primary/secondary/exploratory outcomes and statistical methods.

ITEMS 7–16: DATA ANALYSIS

Item 7: correct and complete description of statistical methods

A separate part in the Methods section of the manuscript should be devoted to the description of the statistical procedures. Both descriptive and analytical statistical methods should be sufficiently described so that the methods can be assessed by a statistical reviewer to judge their suitability and completeness in addressing the study objectives.

Item 8: valid statistical methods used and assumptions outlined

The validity of statistical analyses relies on some assumptions. For example, the independent t-test for the comparison of two means requires three assumptions: independence of the observations, normality and homogeneity of variance.[38] As another example, all expected values for a χ^2 test must be more than 1, and at most 20% of the expected values can be less than 5. These statistical assumptions should be judged as a matter of context or assessed using appropriate methods such as a normal probability plot for checking the normality assumption.[39] In this regard, an alternative statistical test should be applied if some assumptions are clearly violated. It should be noted that some statistical tests are robust against mild-to-moderate violations of some assumptions. For the t-test, lack of normality and lack of homogeneity of variance do not necessarily invalidate the t-test, whereas lack of independence of the outcome variables will imply the results are invalid.[40] It has been demonstrated that the independent t-test can be valid but suboptimal for the ordinal scaled data (eg, a variable with values 0, 1, 2, 3) even with a sample size of 20.[41]

An important but often ignored aspect in practice is that ratio estimates such as the estimated odds ratio (OR), risk ratio and rate ratio are biased away from the null value. This bias is amplified with sparse data known as sparse-data bias.[42] A sign of sparse data is an unrealistically large ratio estimate or confidence limit which is simply an artefact of sparse data. For example, an OR >10 for a non-communicable disease should be considered as a warning sign for sparse-data bias. In the extreme, an empty cell leads to an absurd OR estimate of infinity, known as separation.[43] Special statistical methods such as penalisation or Bayesian methods must be applied to decrease the sparse-data bias.[43 44] Some other important considerations in statistical analysis are (1) accounting for correlation in the analysis of correlated data (eg, variables with repeated measurements in longitudinal studies,[45] cluster randomised trials[46] and complex surveys[47]); (2) accounting for matching in the analysis of matched case–control and cohort data[48–50]; (3) considering ordering of several groups in the analysis; (4) considering censoring in the analysis of survival data; (5) adjusting for baseline values of the outcome in the analysis of randomised clinical trials[28]; (6) correct calculation and interpretation of the population attributable fraction[51 52]; (7) adjusting for overfitting using shrinkage or penalisation methods when developing a prediction model[53 54]; and (8) assessment of similarity and consistency assumptions in network meta-analysis.[55]

Item 9: appropriate assessment of treatment effect or interaction between treatment and another covariate

Appropriate statistical tests should be used for the assessment of treatment effects and potential interactions. Assessment of overlapping treatment group-specific confidence intervals (CIs can be misleading.[56–58] Thus, the comparison of the CIs of the treatment groups should not be used as an alternative to the statistical test of treatment effect. Moreover, comparing p values for the treatment effect at each level of the covariate (eg, men and women) should not be used as an alternative for an interaction test between the treatment and covariate. For example, in the case of observing p value <0.05 in men and p value >0.05 in women, one might incorrectly conclude that gender was an effect modifier.[59] Similarly, we cannot conclude no effect modification if the CIs of the subgroups are overlapping.[60]

Item 10: correct use of correlation and associational statistical testing

The misuse of correlation and associational statistical testing is not uncommon. As an example, correlation should not be used for assessing the agreement between two methods in methods-comparison studies.[61] To see why, two measures of X and Y are perfectly correlated but in poor agreement if X is twice Y. Similarly, we cannot infer that the two methods agree well because the p value is large enough using the statistical testing of the means such as paired t-test. In fact, a high variance of differences indicates poor agreement but also increases the chance that the paired t-test will result in a large p value, and thus the methods will appear to agree.[1]

Item 11: appropriate handling of continuous predictors

Reviewers should be wary of studies that have dichotomised or categorised continuous variables–this should be generally avoided.[62] Bias, inefficiency and residual confounding may also result from dichotomising/categorising a continuous variable and using it as a categorical variable in a model. Continuous variables should be retained as continuous and their functional

form be examined, as a linearity assumption may not be correct. Approaches for handling continuous predictors include fractional polynomials or regression splines.[62–65]

Item 12: CIs do not include impossible values

A valid CI should exclude impossible values. For instance, a simple Wald CI for a proportion ($P \pm 1.96\sqrt{\frac{P(1-P)}{n}}$) is not valid when p is close to 0 or 1, and may yield negative values outside the possible range for a proportion ($0 \leq p \leq 1$).[66] To remedy such conditions, the Wilson score or Agresti-Coull interval can be applied.[6]

Item 13: appropriate comparison of baseline characteristics between the study arms in randomised trials

In a randomised clinical trial, any baseline characteristic difference between groups should be due to chance (or unreported bias). Reviewers should look out for any statistical testing at the baseline as reporting p values does not make sense.[67] The decision on which baseline characteristics (prognostic factors) are included in any adjustment should be prespecified in the protocol and should be based on the subject-matter knowledge, not on p values. The differences between groups in baseline characteristics should be identified by their size and discussed in terms of potential implications for the interpretation of the results.

Item 14: correct assessment and adjustment of confounding

An important goal of health research is drawing a causal inference. Here, the interest is in the causal effect of an exposure on the outcome. The major source of bias threatening causality studies, including observational studies as well as randomised studies (with small-to-moderate sample size), is confounding.[68–71] Confounding can be controlled in the design phase (eg, through restriction or matching) or analysis phase (eg, using regression models, standardisation or propensity score methods).[72–74] Selection of confounders should be based on a priori causal knowledge, often represented in causal diagrams,[23 75–77] *not* p values (eg, using stepwise approaches). Automated statistical procedures, such as stepwise regression, do not discriminate between confounders and other covariates like mediators or colliders which should not be adjusted for in the analysis. Moreover, stepwise regression is only based on the association between confounders and outcome, and disregards the association between the confounders and exposure. Thus, stepwise procedures should not be used for confounder selection. In practice, many confounders (and exposures and outcomes)[78 79] are time-varying, and the so-called 'causal methods' should be applied for the appropriate adjustment of time-varying confounders.[80 81] Similarly, in studies evaluating the prognostic effect of a new variable, adjustment for existing prognostic factors should be routinely performed, and variable selection of the existing factors is not generally needed.[53]

Item 15: avoiding model extrapolation not supported by data

The goal of interest in many health studies is predicting an outcome from one or more explanatory variables using a regression model. The model is valid only within the range of observed data on the explanatory variables, and we cannot make prediction for people outside the range. This is known as model extrapolation.[82] Suppose we have found a linear relation between body mass index (BMI) and BP based on the following equation in a cohort study:

$$BP = A + B * (BMI)$$

Now the intercept, A, cannot be interpreted because it corresponds to the expected BP value of a person with BMI of zero! The remedy is centring BMI and including the centred variable (BMI−average BMI) in the model so that the new intercept refers to the expected BP value of a person with the average BMI in the population.

As another example, suppose the following linear relation holds in an RCT:

$$BP = A + B * (TRT) + C * (BMI) + D * (TRT * BMI)$$

where TRT denotes treatment (1: intervention, 0: placebo) and TRT^*BMI is the product term (interaction term) between treatment and BMI. In this model, the parameter B cannot be interpreted on its own because it is the mean difference in BP between two treatment groups for a person with BMI of zero. Again, the solution is centring BMI and including centred BMI and the product term between TRT and centred BMI in the model so that B' (coefficient of TRT in the new model) refers to the mean difference in BP of a person with the average BMI in the population.

Item 16: adequate handling of missing data

The methods used for handling missing data should be described and justified in relation to stated assumptions about the missing data (missing completely at random, missing at random and missing not at random), and sensitivity analyses must be done if appropriate. Missing data[83] can introduce selection bias and should be handled using appropriate methods such as multiple imputation[84] and inverse probability weighting.[85] Naïve methods such as complete-case analysis, single imputation using the mean of the observed data, last observation carried forward and the missing indicator method are statistically invalid in general and they can lead to serious bias.[86]

ITEMS 17–23: REPORTING AND PRESENTATION

Item 17: adequate and correct description of the data

The mean and standard deviation (SD) provide a satisfactory summary of data for continuous variables that have a reasonably symmetric distribution. The standard error (SE) is not a sound choice to be used in place of SD.[87] A useful memory aid is to use SD to Describe data and SE to Estimate parameters.[88] Besides, 'mean±SD' is not suitable because it implies the range in which 68% of data are within (not a relevant concept we are looking for), and 'mean (SD)' should be reported instead.[1] In case of having highly skewed quantitative data, median and interquartile range (IQR) are more informative summary statistics for description. It should be noted that the mean/SD ratio of <2 for positive variables is a sign of skewness.[89] Categorical data should be summarised as number and percentage.[90] For cohort data, a summary of follow-up time such as median and IQR should be reported.

Item 18: descriptive results provided as occurrence measures with CIs and analytical results provided as association measures and CIs along with p values

The point estimates of the occurrence measures, for instance, prevalence, risk and incidence rate with 95% CIs, should be reported for descriptive objectives.[90] Alternatively, the point estimates of the association measures, for instance, OR, risk ratio and rate ratio with 95% CIs along with p values, should be reported for analytical objectives as part of the Results section.[91 92]

Item 19: CIs provided for the contrast between groups rather than for each group

For analytical studies like RCTs, the 95% CIs should be given for the contrast between groups rather than for each group.[6] For the BP example mentioned above,[20] the authors reported the mean of BP with 95% CI in each group, but they should have given the mean difference in 24-hour ambulatory BP levels between groups with 95% CI as the aim of the trial was to compare treatment with control.

Item 20: avoiding selective reporting of analyses and p-hacking

All statistical analyses performed should be reported regardless of the results. P-hacking, playing with data to produce the desired p value (upwards as well as downwards), must be avoided.[93–95] This is probably difficult to assess as a reader/reviewer, but usually one would be clued in if there are many more analyses than those stated in the objectives or only statistically significant comparisons are presented when a larger pool of variables were identified in the methods.

Item 21: appropriate and consistent numerical precisions for effect sizes, test statistics and p values, and reporting the p values rather than their range

P values should be reported directly with one or two significant figures even if they are greater than 0.05, for example, p value=0.09 or p value=0.28. One should not focus on 'statistical significance' or dichotomise p values (eg, p<0.05)[96–98] or express them as '0.000' or 'NS'. Nonetheless, spurious precision, too many decimals, in numerical presentation should be avoided.[99 100] For example, typically p values less than 0.001 can be written as <0.001 without harm, and it does not make sense to present percentages with more than one decimal when the sample size is much less than 100.

Item 22: providing sufficient numerical results that could be included in a subsequent meta-analysis

Meta-analyses of randomised trials and observational studies provide high levels of evidence in health research. Providing numerical results in individual studies contributing to subsequent meta-analysis is of special importance. Follow-up score and change score from the baseline are two possible approaches that can be applied to estimate treatment effect in RCTs.[101] While the follow-up score meta-analysis requires after-intervention mean and SD in two groups of intervention and control, the mean and SD of differences from the baseline are prerequisite for performing change-score meta-analysis. However, authors often only report mean and SD before and after intervention. The mean of the difference in each group can be calculated from the difference of the means, but calculating the SD of differences needs a guessed group-specific correlation between baseline and follow-up scores besides before-intervention and after-intervention SDs.

Item 23: acceptable presentation of the figures and tables

Tables and figures are effective data presentation that should be properly managed.[102–105] Figures should be selected based on the type of variable(s) and appropriately scaled. The error bar graph as an illustration can be used for displaying the mean and CI. It is inappropriate to give a bar chart with an SE bar superimposed instead (the so-called 'dynamite plunger plot'[105]). Tables should be able to stand on their own and include sufficient details such as labels, units and values.

ITEMS 24–30: INTERPRETATION

Item 24: interpreting the results based on association measures and 95% CIs along with p values and correctly interpreting large p values as indecisive results, not evidence of absence of an effect

The study results should be interpreted in light of the point estimate of appropriate association measures such as mean difference and 95% CI as well as precise p values. When testing a null hypothesis of no treatment effect, the p value is the probability the statistical association would be as extreme as observed or more extreme, assuming that null hypothesis and all assumptions used for the test are correct. P values for non-null effect sizes can also be computed. The point estimate is the effect size most compatible with the data in that it has p value=1.00, while the 95% CI shows the range of effect sizes reasonably compatible with the data in the sense of having p value >0.05.[97] We should judge the clinical importance and statistical reliability of the results by examining both of the 95% CIs as well as looking at precise p values, not just whether a p value crosses a threshold or not.[28 106] It is incorrect to interpret p value >0.05 as showing no treatment effect; instead, it represents an ambiguous outcome.[107 108] It is not evidence that the effect is unimportant ('absence of evidence is not evidence of absence'); inferring unimportance requires that every effect size inside the CI be considered unimportant.[97]

Item 25: using CIs rather than post hoc power analysis for interpreting the results of studies

Conceptually, it is not valid to interpret power as if it pertains to the observed study results.[109–111] Rather, power should be treated as part of the study rationale and design before actual conduct begins, for example, as in sample size calculations. Power does not correctly account for the observations that follow; for example, a study could have high power and observe a high p value, yet still favour the alternative hypothesis over the null hypothesis.[111] The precision of results should be gauged using CIs.

Item 26: correctly interpreting occurrence or association measures

It will be crucial to interpret occurrence and association measures correctly. ORs commonly provide examples of misinterpretation: if the event is rare, they can approximate risk ratios, but they are not conceptually the same and will differ considerably if the event is common.[112 113] In a study with a risk of 60% in an exposed group and 40% in an unexposed group, the error in interpreting the OR (2.25) as a risk ratio (1.5) is considerable. Prevalence in cross-sectional studies is another example, which sometimes has been incorrectly called 'risk'.

Item 27: distinguishing causation from association and correlation

We should be cautious about the correct use of technical terms such as effect, association and correlation. Association, meaning no independence, does not imply causation (and effect). Causal effect estimation requires measurement of exposure before outcome (temporality) as well as confounding adjustment. The correlation refers to a monotonic association between two variables. Therefore, no correlation does not imply no association.

Item 28: results of prespecified analyses are distinguished from the results of exploratory analyses in the interpretation

The results obtained from the prespecified (a priori) analyses that have been already designed and mentioned in a protocol are much more reliable than the results obtained after data dredging (data-derived or post hoc analysis).

Item 29: appropriate discussion of the study methodological limitations

The methodological limitations of the study design and analysis should be discussed. Ideally, the probabilistic bias analysis, in which a probability distribution is assumed for the bias parameters and bias is probabilistically accounted for using Monte Carlo sensitivity or Bayesian analysis, should be performed for adjustment of uncontrolled confounding (eg, due to an unmeasured confounder), selection bias (eg, through missing outcome data) and measurement bias (eg, subsequent to measurement error in the exposure).[114–116] The authors should at least qualitatively discuss the main sources of bias and their impact on the study results.[117 118]

Item 30: drawing only conclusions supported by the statistical analysis and no generalisation of the results to subjects outside the target population

The study interpretation must be based not only on the results but also in the light of the study population as well as any limitation in the design and analysis.[82] For example, if the study has been done in women, it cannot necessarily be generalised to a population of men and women.

CONCLUSION

The important role of sound statistics and methodology in medical research cannot be overstated. We strongly encourage authors to adhere to CHAMP for carrying out and reporting medical research, and to editors and reviewers for assisting the evaluation of manuscripts for potential publication. We have only covered some basic items, and each type of study or statistical model (eg, randomised trial, prediction model) has its own issues that ideally require statistical expertise. We appreciate that for some items in the checklist there is no unequivocal answer, and thus assessing the statistics of a paper may involve some subjectivity. Moreover, the questions in the checklist are not equally important for example, papers with serious errors in design are statistically unacceptable regardless of how the data were analysed and aspects of presentations are clearly less important than other elements of the checklist. It is important to note that statistical review, carried out by experienced statisticians, is the preferred way of reviewing statistics in research papers, more so than what any checklist can achieve. We hope CHAMP improves authors' practice in their use of statistics in medical research and serves as a useful handy reference for editors and referees during the statistical assessment of medical papers.

Author affiliations
[1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
[2]Sports Medicine Research Center, Neuroscience Institute, Tehran University of Medical Sciences, Tehran, Iran
[3]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[4]National Institute for Health Research Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK
[5]Department of Public Health, Section for Sports Science, Aarhus University, Aarhus, Denmark
[6]Research Unit for General Practice, Aarhus, Denmark
[7]Psychosocial Health Research Institute, Iran University of Medical Sciences, Tehran, Iran
[8]Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, UK
[9]Division of Epidemiology & Biostatistics, School of Public Health, University of California, Berkeley, California, USA
[10]ScHARR, University of Sheffield, Sheffield, UK

Twitter Rasmus Oestergaard Nielsen @RUNSAFE_Rasmus

Acknowledgements We thank Sander Greenland, Stephen Senn and Richard Riley for their valuable comments on an earlier draft of this paper.

Contributors MAM, MN and DGA produced the first draft. GSC, RON, NPJ and MJC suggested revisions. All authors approved the final version.

Funding GSC was supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme grant: C49297/A27294).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Mohammad Ali Mansournia http://orcid.org/0000-0003-3343-2718
Rasmus Oestergaard Nielsen http://orcid.org/0000-0001-5757-1806

REFERENCES

1 Altman DG. *Practical statistics for medical research*. CRC press, 1990.
2 Thiese MS, Arnold ZC, Walker SD. The misuse and abuse of statistics in biomedical research. *Biochem Med* 2015;25:5–11.
3 Thiese MS, Walker S, Lindsey J. Truths, lies, and statistics. *J Thorac Dis* 2017;9:4117–23.
4 Altman DG. The scandal of poor medical research. *BMJ* 1994;308:283–4.
5 Nielsen RO, Shrier I, Casals M, *et al*. Statement on methods in sport injury research from the 1st methods matter meeting, Copenhagen, 2019. *Br J Sports Med* 2020;54:941.
6 Moher D, Hopewell S, Schulz KF, *et al*. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Int J Surg* 2012;10:28–55.
7 Vandenbroucke JP, Von Elm E, Altman DG. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Ann Intern Med* 2007;147:W.
8 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med* 2003;138:W1–12.
9 Altman DG, McShane LM, Sauerbrei W, *et al*. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Med* 2012;10:51.
10 Moons KGM, Altman DG, Reitsma JB, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;162:W1–73.
11 Altman DG. Statistical reviewing for medical journals. *Stat Med* 1998;17:2661–74.
12 Goodman SN, Altman DG, George SL. Statistical reviewing policies of medical journals. *J Gen Intern Med* 1998;13:753–6.
13 Nielsen Rasmus Østergaard, Shrier I, Casals M, *et al*. Statement on methods in sport injury research from the first methods matter meeting, Copenhagen, 2019. *J Orthop Sports Phys Ther* 2020;50:226–33.
14 Verhagen E, Stovitz SD, Mansournia MA, *et al*. *BJSM* educational editorials: methods matter. *Br J Sports Med* 2018;52:1159–60.
15 Lang T, Altman D. Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A, eds. *Handbook, European association of science editors*, 2013.
16 Assel M, Sjoberg D, Elders A, *et al*. Guidelines for reporting of statistics for clinical research in urology. *Eur Urol* 2019;75:358–67.
17 Gardner MJ, Machin D, Campbell MJ. Use of check Lists in assessing the statistical content of medical studies. *BMJ* 1986;292:810–2.
18 Mansournia MA, Collins GS, Nielsen RO, *et al*. CHecklist for statistical Assessment of Medical Papers: the CHAMP statement. *Br J Sports Med* 2021;55:1007–8.
19 Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *CHANCE* 2019;32:42–9.
20 Guimaraes GV, de Barros Cruz LG, Fernandes-Silva MM, *et al*. Heated water-based exercise training reduces 24-hour ambulatory blood pressure levels in resistant hypertensive patients: a randomized controlled trial (HEx trial). *Int J Cardiol* 2014;172:434–41.

21 Centre for Evidence-Based Medicine. Study designs, 2020. Available: https://www.cebm.net/2014/04/study-designs/
22 Machin D, Campbell MJ. *The design of studies for medical research*. John Wiley & Sons, 2005.
23 Etminan M, Collins GS, Mansournia MA. Using causal diagrams to improve the design and interpretation of medical research. *Chest* 2020;158:S21–8.
24 Stovitz SD, Verhagen E, Shrier I. Distinguishing between causal and non-causal associations: implications for sports medicine clinicians. *Br J Sports Med* 2019;53:398–9.
25 Etminan M, Nazemipour M, Sodhi M, *et al*. Potential biases in studies of acid suppressing drugs and COVID-19 infection. *Gastroenterology* 2020. doi:10.1053/j.gastro.2020.11.053. [Epub ahead of print: 16 Dec 2020].
26 Mansournia MA, Danaei G, Forouzanfar MH, *et al*. Effect of physical activity on functional performance and knee pain in patients with osteoarthritis : analysis with marginal structural models. *Epidemiology* 2012;23:631–40.
27 Machin D, Campbell MJ, Tan SB. *Sample sizes for clinical laboratory and epidemiology studies*. John Wiley & Sons, 2018.
28 Mansournia MA, Altman DG. Invited commentary: methodological issues in the design and analysis of randomised trials. *Br J Sports Med* 2018;52:553–5.
29 Cook JA, Julious SA, Sones W, *et al*. DELTA^2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ* 2018;363:k3750.
30 Bland JM. The tyranny of power: is there a better way to calculate sample size? *BMJ* 2009;339:b3985.
31 Rothman KJ, Greenland S. Planning study size based on precision rather than power. *Epidemiology* 2018;29:599–603.
32 Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
33 Riley RD, Snell KIE, Ensor J, *et al*. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019;38:1276–96.
34 Riley RD, Snell KIE, Ensor J, *et al*. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Stat Med* 2019;38:1262–75.
35 Jungmalm J, Bertelsen ML, Nielsen RO. What proportion of athletes sustained an injury during a prospective study? Censored observations matter. *Br J Sports Med* 2020;54:70–1.
36 Nielsen RO, Bertelsen ML, Ramskov D, *et al*. Randomised controlled trials (RCTs) in sports injury research: authors-please report the compliance with the intervention. *Br J Sports Med* 2020;54:51–7.
37 Edouard P, Steffen K, Navarro L, *et al*. Methods matter: dealing with low compliance in sports injury trials analyses using instrumental variable analysis. *Br J Sports Med* 2021;55:1002–4.
38 Mansournia MA, Nazemipour M, Naimi AI, *et al*. Reflections on modern methods: demystifying robust standard errors for epidemiologists. *Int J Epidemiol* 2020;318.
39 Altman DG, Bland JM. Statistics notes: the normal distribution. *BMJ* 1995;310:298.
40 Senn S. The t-test tool. *Significance* 2008;5:40–1.
41 Heeren T, D'Agostino R. Robustness of the two independent samplest-test when applied to ordinal scaled data. *Stat Med* 1987;6:79–90.
42 Greenland S, Mansournia MA, Altman DG. Sparse data bias: a problem hiding in plain sight. *BMJ* 2016;352:i1981.
43 Mansournia MA, Geroldinger A, Greenland S, *et al*. Separation in logistic regression: causes, consequences, and control. *Am J Epidemiol* 2018;187:864–70.
44 Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med* 2015;34:3133–43.
45 Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. John Wiley & Sons, 2012.
46 Mansournia MA, Altman DG. Some methodological issues in the design and analysis of cluster randomised trials. *Br J Sports Med* 2019;53:573–5.
47 Korn EL, Graubard BI. *Analysis of health surveys*. John Wiley & Sons, 2011.
48 Mansournia MA, Hernán MA, Greenland S. Matched designs and causal diagrams. *Int J Epidemiol* 2013;42:860–9.
49 Mansournia MA, Jewell NP, Greenland S. Case–control matching: effects, misconceptions, and recommendations. *Eur J Epidemiol* 2018;33:5–14.
50 Greenland S, Jewell NP, Mansournia MA. Theory and methodology: essential tools that can become dangerous belief systems. *Eur J Epidemiol* 2018;33:503–6.
51 Mansournia MA, Altman DG. Population attributable fraction. *BMJ* 2018;360:k757.
52 Khosravi A, Nielsen RO, Mansournia MA. Methods matter: population attributable fraction (PAF) in sport and exercise medicine. *Br J Sports Med* 2020;54:1049–54.
53 Riley RD, van der Windt D, Croft P. *Prognosis research in healthcare: concepts, methods, and impact*. Oxford University Press, 2019.
54 Steyerberg EW. *Clinical prediction models*. Springer, 2019.
55 Doosti-Irani A, Nazemipour M, Mansournia MA. What are network meta-analyses (NMAs)? A primer with four tips for clinicians who read NMAs and who perform them (methods matter series). *Br J Sports Med* 2021;55:520–1.
56 Bland JM, Peacock JL. Interpreting statistics with confidence. *The Obstetrician Gynaecologist* 2002;4:176–80.
57 Austin PC, Hux JE. A brief note on overlapping confidence intervals. *J Vasc Surg* 2002;36:194–5.
58 Mittal N, Bhandari M, Kumbhare D. A tale of confusion from overlapping confidence intervals. *Am J Phys Med Rehabil* 2019;98:81–3.
59 Matthews JNS, Altman DG. Statistics notes: interaction 2: compare effect sizes not P values. *BMJ* 1996;313:808.
60 Knol MJ, Pestman WR, Grobbee DE. The (mis)use of overlap of confidence intervals to assess effect modification. *Eur J Epidemiol* 2011;26:253–4.
61 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
62 Andersen PK, Skovgaard LT. *Regression with linear predictors*. Springer, 2010.
63 Royston P, Sauerbrei W. *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*. John Wiley & Sons, 2008.
64 Harrell Jr FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
65 Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. *Stat Med* 2013;32:3788–803.
66 Mardani M, Rahnavardi M, Rajaeinejad M, *et al*. Crimean-Congo hemorrhagic fever among health care workers in Iran: a seroprevalence study in two endemic regions. *Am J Trop Med Hyg* 2007;76:443–5.
67 Senn S. Testing for baseline balance in clinical trials. *Stat Med* 1994;13:1715–26.
68 Suzuki E, Tsuda T, Mitsuhashi T, *et al*. Errors in causal inference: an organizational schema for systematic error and random error. *Ann Epidemiol* 2016;26:788–93.
69 Greenland S, Mansournia MA. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *Eur J Epidemiol* 2015;30:1101–10.
70 Mansournia MA, Higgins JPT, Sterne JAC, *et al*. Biases in randomized trials: a conversation between Trialists and epidemiologists. *Epidemiology* 2017;28:54.
71 Mansournia MA, Greenland S. The relation of collapsibility and confounding to faithfulness and stability. *Epidemiology* 2015;26:466–72.
72 Almasi-Hashiani A, Nedjat S, Mansournia MA. Causal methods for observational research: a primer. *Arch Iran Med* 2018;21:164-169.
73 Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. *Eur Heart J* 2011;32:1704–8.
74 Gharibzadeh S, Mohammad K, Rahimiforoushani A, *et al*. Standardization as a tool for causal inference in medical research. *Arch Iran Med* 2016;19:666–70.
75 Nielsen RO, Bertelsen ML, Møller M, *et al*. Training load and structure-specific load: applications for sport injury causality and data analyses. *Br J Sports Med* 2018;52:1016–7.
76 Nielsen RO, Bertelsen ML, Møller M, *et al*. Methods matter: exploring the 'too much, too soon' theory, part 1: causal questions in sports injury research. *Br J Sports Med* 2020;54:1119–22.
77 Nielsen RO, Simonsen NS, Casals M, *et al*. Methods matter and the 'too much, too soon' theory (part 2): what is the goal of your sports injury research? Are you describing, predicting or drawing a causal inference? *Br J Sports Med* 2020;54:1307–9.
78 Nielsen RO, Bertelsen ML, Ramskov D, *et al*. Time-to-event analysis for sports injury research Part 1: time-varying exposures. *Br J Sports Med* 2019;53:61–8.
79 Nielsen RO, Bertelsen ML, Ramskov D, *et al*. Time-to-event analysis for sports injury research Part 2: time-varying outcomes. *Br J Sports Med* 2019;53:70–8.
80 Mansournia MA, Etminan M, Danaei G, *et al*. Handling time varying confounding in observational research. *BMJ* 2017;359:j4587.
81 Mansournia MA, Naimi AI, Greenland S. The implications of using Lagged and baseline exposure terms in longitudinal causal and regression models. *Am J Epidemiol* 2019;188:753–9.
82 Altman DG, Bland JM. Generalisation and extrapolation. *BMJ* 1998;317:409–10.
83 Altman DG, Bland JM. Missing data. *BMJ* 2007;334:424–24.
84 Vickers AJ, Altman DG. Statistics notes: missing outcomes in randomised trials. *BMJ* 2013;346:f3438.
85 Mansournia MA, Altman DG. Inverse probability weighting. *BMJ* 2016;352:i189.
86 Sterne JAC, White IR, Carlin JB, *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
87 Altman DG, Bland JM. Standard deviations and standard errors. *BMJ* 2005;331:903.
88 Campbell MJ, Swinscow TDV. *Statistics at square one*. John Wiley & Sons, 2011.
89 Altman DG, Bland JM. Detecting skewness from summary information. *BMJ* 1996;313:1200–1.
90 Nielsen RO, Debes-Kristensen K, Hulme A, *et al*. Are prevalence measures better than incidence measures in sports injury research? *Br J Sports Med* 2019;53:396–7.
91 Nielsen RO, Bertelsen ML, Verhagen E, *et al*. When is a study result important for athletes, clinicians and team coaches/staff? *Br J Sports Med* 2017;51:1454–5.
92 Pourahmadi M, Koes BW, Nazemipour M, *et al*. It is time to change our Mindset and perform more high-quality research in low back pain. *Spine* 2021;46:69–71.
93 Stovitz SD, Verhagen E, Shrier I. Misinterpretations of the 'p value': a brief primer for academic sports medicine. *Br J Sports Med* 2017;51:1176–7.

94 Windt J, Nielsen RO, Zumbo BD. Picking the right tools for the job: opening up the statistical toolkit to build a compelling case in sport and exercise medicine research. *Br J Sports Med* 2019;53:987–8.
95 Nielsen RO, Chapman CM, Louis WR, *et al*. Seven SINS when interpreting statistics in sports injury science. *Br J Sports Med* 2018;52:1410–2.
96 McShane BB, Gal D, Gelman A, *et al*. Abandon statistical significance. *The American Statistician*2019;73:235–45.
97 Greenland S, Senn SJ, Rothman KJ, *et al*. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31:337–50.
98 Rothman KJ, Greenland S, Lash TL. Precision and Statistics in Epidemiologic Studies. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 148–67.
99 Altman DG, Bland JM. Statistics notes: presentation of numerical data. *BMJ* 1996;312:572.
100 Kordi R, Mansournia MA, Rostami M, *et al*. Troublesome decimals; a hidden problem in the sports medicine literature. *Scand J Med Sci Sports* 2011;21:335–6.
101 Higgins J, Wells G. *Cochrane Handbook for systematic reviews of interventions*, 2011.
102 Schriger DL, Sinha R, Schroter S, *et al*. From submission to publication: a retrospective review of the tables and figures in a cohort of randomized controlled trials submitted to the British medical Journal. *Ann Emerg Med* 2006;48:750–6.
103 Morris TP, Jarvis CI, Cragg W, *et al*. Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views: KMunicate. *BMJ Open*2019;9:e030215.
104 Vickers AJ, Assel MJ, Sjoberg DD, *et al*. Guidelines for reporting of figures and tables for clinical research in urology. *Eur Urol* 2020;78:97–109.
105 Freeman JV, Walters SJ, Campbell MJ. *How to display data*. Wiley, 2009.
106 Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research*. John Wiley & Sons, 2008.
107 Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305–7.
108 Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol* 2020;20:244.
109 Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 2001;55:19–24.
110 Bacchetti P. Peer review of statistics in medical research: the other problem. *BMJ* 2002;324:1271–3.
111 Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. *Ann Epidemiol* 2012;22:364–8.
112 Janani L, Mansournia MA, Nourijeylani K, *et al*. Statistical issues in estimation of adjusted risk ratio in prospective studies. *Arch Iran Med* 2015;18:713–9.
113 Talebi SS, Mohammad K, Rasekhi A, *et al*. Risk ratio estimation in longitudinal studies. *Arch Iran Med* 2019;22:46–9.
114 Lash TL, Fox MP, Fink AK. *Applying quantitative bias analysis to epidemiologic data*. Springer Science & Business Media, 2011.
115 Greenland S, Lash TL. Bias analysis. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins, 2008: 345–80.
116 Lash TL, Fox MP, MacLehose RF, *et al*. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43:1969–85.
117 Altman DG, Bland JM. Uncertainty beyond sampling error. *BMJ* 2014;349:g7065.
118 Altman DG, Bland JM. Uncertainty and sampling error. *BMJ* 2014;349:g7064.