available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Opinion

Guidelines for Reporting of Statistics in *European Urology*

Andrew J. Vickers [*], *Daniel D. Sjoberg*

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

The key conclusions of most medical studies are based on the results of statistical analysis, yet there is wide acknowledgment that standards of statistical analysis and reporting are far from ideal. It has been reported that a large proportion of studies do not include an author with formal quantitative training and that statistical errors are ubiquitous, with typically more than half of any given sample of papers including important statistical errors (some examples of this type of empirical research are included in the Appendix).

In this paper, we provide guidelines for statistical analysis and reporting. The guidelines were based on a review of close to 50 papers published in or submitted to *European Urology*. This review confirmed previous findings that published studies are prone to statistical errors: The first 14 published papers we reviewed had at least one statistical error.

These guidelines are very much directed toward the sort of papers typically submitted to *European Urology*. This means that there are several guidelines on prediction models but none on genomic discovery. Many of the guidelines specify analyses or methods of reporting statistics that should be avoided. We take the approach of focusing on the *don'ts* rather than the *do's* because although it is generally difficult to specify what constitutes good science, there is often widespread agreement on what would be bad science. As a simple example, there is considerable disagreement on the best methods for analysis of randomized trials with repeated measures of a continuous end point, such as when a generalized estimating equations approach should be used. What is not in doubt, however, is that regression approaches are preferable to unadjusted analyses, and that χ^2 has no role at all.

Another notable aspect of this paper is that it is without references. This is because many of the recommendations are seen as routine common sense by practicing statisticians, and citations, if available, would merely be to a second statistician's opinion. For instance, it is not clear to us how adding citations to a guideline to avoid inappropriate levels of precision would be of benefit. The paper is meant to be didactic: Readers who question the rationale of a guideline are welcome to write to the authors.

1. The golden rule

1.1. *Break any of the guidelines if it makes scientific sense to do so*

Science varies too much to allow methodological or reporting guidelines to apply universally.

2. Reporting of design and statistical analysis

2.1. *Follow existing reporting guidelines for the type of study you are reporting, such as CONSORT for randomized trials or ReMARK for marker studies*

Statisticians and methodologists have contributed extensively to a large number of reporting guidelines. The first is widely recognized to be the Consolidated Standards of Reporting Trials (CONSORT) statement on the reporting of randomized trials, but there are now a large number covering a wide variety of different types of study. Reporting guidelines can be downloaded from the Equator Web site (Appendix).

2.2. *Describe the practical steps of randomization in randomized trials*

Although this reporting guideline is part of the CONSORT statement, it is so critical and so widely misunderstood that

[*] Corresponding author. Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, 2nd Floor, New York, NY 10065, USA. Tel.: +1 646 735 8142; Fax: +1 646 735 0032.
E-mail address: vickersa@mskcc.org (A.J. Vickers).

http://dx.doi.org/10.1016/j.eururo.2014.06.024

it bears repeating. The purpose of randomization is to prevent patient selection. This can be achieved only if those consenting patients cannot guess a patient's allocation before registration in the trial or change it afterward. This safeguard is known as *allocation concealment*. Stating merely that "a randomization list was created by a statistician" or that "envelope randomization was used" does not ensure allocation concealment: A list could have been posted in the nurse's station for all to see; envelopes can be opened and resealed. Investigators need to specify the exact logistical steps taken to ensure allocation concealment. The best method is to use a password-protected computer database.

2.3. The statistical methods should describe the study questions and the statistical approaches used to address each question

Many statistical methods sections state something like, "Mann-Whitney was used for comparisons of continuous variables and Fisher's exact for comparisons of binary variables." This says little more than "the inference tests used were not grossly erroneous for the type of data." Instead, statistical methods sections should lay out each primary study question separately: Carefully detail the analysis associated with each, and describe the rationale for the analytic approach if this is not obvious or if there are reasonable alternatives.

2.4. The statistical methods should be described in sufficient detail to allow replication by an independent statistician given the same data set

Vague reference to "adjusting for confounders" or "nonlinear approaches" is insufficiently specific to allow replication, a cornerstone of science.

3. Inference and *p* values

3.1. Do not accept the null hypothesis

In a court case, defendants are declared guilty or not guilty; there is no verdict of "innocent." Similarly, in a statistical test, the null hypothesis is rejected or is not rejected. If the *p* value is 0.05 or higher, investigators should avoid conclusions such as "the drug was ineffective," "there was no difference between groups," or "response rates were unaffected." Instead, authors should use phrases such as "we did not see evidence of a drug effect," "we were unable to demonstrate a difference between groups," or simply "there was no significant difference in response rates."

3.2. P values just above 5% are not a trend, and they are not moving

Avoid saying that a *p* value such as 0.07 shows a "trend" (which is meaningless) or "approaches statistical significance" (because if you come back and look the next day, the *p* value will not be any closer to 0.05). Alternative language might state, "Although we saw some evidence of improved response rates in patients receiving the novel procedure, differences between groups did not meet conventional levels of statistical significance."

3.3. Take care when reporting multiple p values

The more questions you ask, the more likely you are to get a silly answer to at least one of them. For example, if you report *p* values for five independent true null hypotheses, the probability that you will falsely reject at least one is not 5% but >20%. Although formal adjustment of *p* values is appropriate in some specific cases, such as genomic studies, a more common approach is simply to interpret *p* values in the context of multiple testing. For instance, if an investigator examines the association of 10 variables with three different end points, thereby testing 30 separate hypotheses, a *p* value of 0.04 should not be interpreted in the same way as if a paper contained only a single *p* value of 0.04.

3.4. Do not report separate p values for each of two different groups to address the question of whether there is a difference between groups

One scientific question means one statistical hypothesis tested by one *p* value. To illustrate the error of using two *p* values to address one question, take the case of a randomized trial of drug versus placebo to reduce voiding symptoms, with 30 patients in each group. The authors might report that symptom scores improved by 6 points (standard deviation [SD]: 14) in the drug group ($p = 0.03$ by one-sample *t* test) and 5 points (SD: 15) in the placebo group ($p = 0.08$); however, the study hypothesis concerns the difference between drug and placebo. To test a single hypothesis, a single *p* value is needed. A two-sample *t* test for these data gives a *p* value for 0.8—unsurprising, given that the scores in each group were virtually the same—confirming that it would be unsound to conclude that the drug was effective based on the finding that change was significant in the drug group but not in placebo controls.

3.5. Use interaction terms in place of subgroup analyses

A similar error to the use of separate tests for a single hypothesis is when an intervention is shown to have a statistically significant effect in one group of patients but not in another. The correct approach is to use what is known as an *interaction term* in a statistical model. For instance, to determine whether a drug reduced pain scores more in women than in men, the model might be as follows: final pain score = b1 baseline pain score + b2 drug + b3 sex + b4 drug × sex.

3.6. Do not report p values or confidence intervals for differences in discrimination for nested prediction models

A common research question in urology is whether a novel predictor (eg, results of a genetic test) adds information to standard clinical predictors (eg, stage and tumor size). It is good practice to assess whether the new predictor is statistically significant in a multivariable model that includes the established predictors and to report a statistic such as the area under the curve (AUC) or the concordance index for both models. However, it is statistically unsound to use the standard method, known as *Delong, Delong and Clarke-Pearson*, to determine whether the increase in discrimination is statistically significant or to calculate a 95% confidence interval (CI) for the increment in discrimination.

3.7. *Tests for change over time are generally uninteresting*

A common analysis is to conduct a paired *t* test comparing, say, erectile function at baseline with erectile function after 5 yr of follow-up. The null hypothesis in this example is that "erectile function does not change over time," which is known to be false. Investigators are encouraged to focus on estimation rather than inference, reporting, for example, the mean change over time along with a 95% CI.

3.8. *Do not apply statistical tests to determine the type of analysis to be conducted*

Numerous statistical tests are available that can be used to determine how a hypothesis test should be conducted. For instance, investigators might conduct a Shapiro-Wilk test for normality to determine whether to use a *t* test or a Mann-Whitney test, calculate Cochran's Q to decide whether to use a fixed- or random-effects approach in a meta-analysis, or use a *t* test for between-group differences in a covariate to determine whether that covariate should be included a multivariable model. The problem with these sorts of approaches is that they are often testing a null hypothesis that is known to be false; for instance, no data set perfectly follows a normal distribution. Moreover, it is often questionable that changing the statistical approach in the light of the test is actually of benefit: Statisticians are far from unanimous as to whether Mann-Whitney is always superior to a *t* test when data are non-normal, or that fixed effects are invalid under study heterogeneity, or that the criterion of adjusting for a variable should be whether it is significantly different between groups. Investigators should generally follow a prespecified analytic plan, altering the analysis only if the data unambiguously point to a better alternative.

4. Reporting of study estimates

4.1. *Use appropriate levels of precision*

Reporting a *p* value of 0.7345 suggests that there is an appreciable difference between *p* values of 0.7344 and 0.7346. Reporting that 16.9% of 83 patients responded entails a precision (to the nearest 0.1%) that is nearly 200 times greater than the width of the confidence interval (10%–27%). Reporting in a clinical study that the mean calorie consumption was 2069.9 suggests that calorie consumption can be measured that precisely by a food questionnaire. Some might argue that being overly precise is irrelevant because the extra numbers can always be ignored. The counterargument is that investigators should think very hard about every number they report rather than just carelessly cutting and pasting numbers from the statistical software printout. As a general rule:

- Report *p* values to a single significant figure unless the *p* value is close to 0.05, in which case, it is reasonable to report two significant figures. Do not report "not significant" for *p* values of 0.05 or higher. Very low *p* values can be reported as $p < 0.001$ or similar. A *p* value can indeed be 1, although some investigators prefer to report this as >0.9. For instance, the following *p* values are reported to appropriate precision: 0.3, 0.004, <0.001, 0.045, 0.13, 1, 0.5.
- Report percentages, rates, and probabilities to two significant figures, for example, 75%, 3.4%, 0.13%.
- There is usually no need to report estimates to more than three significant figures.
- Hazard and odds ratios are normally reported to two decimal places, although this can be avoided for high odds ratios (eg, 18.2 rather than 18.17).

4.2. *Do not treat categorical variables as continuous*

A variable such as Gleason score can be scored as 5, 6, 7, 8, 9, and 10, but it is not true that the difference between 8 and 7 is half as great as the difference between 7 and 5. Variables such as Gleason score should be reported as categories (eg, 40% Gleason ≤6, 40% Gleason 7, 20% Gleason ≥8) rather than as a continuous variable (eg, mean Gleason score of 7.2). Similarly, categorical variables such as Gleason score should be entered into regression models not as a single variable (eg, a hazard ratio of 1.5 per 1-point increase in Gleason score) but as multiple categories (eg, hazard ratio of 1.6 comparing Gleason 7 with Gleason 6 and hazard ratio of 3.9 comparing Gleason 8 with Gleason 6).

4.3. *For time-to-event variables, report the number of events but not the rate*

Take the case of a study reporting that "of 60 patients accrued, 10 (17%) died." Although it is important to report the number of events, patients entered the study at different times and were followed for different periods; therefore, the reported 10% rate is meaningless. The standard statistical approach to time-to-event variables is to calculate probabilities, such as the risk of death being 60% by 5 yr, or the median survival—the time at which the probability of survival first drops below 50%—being 52 mo.

4.4. *In time-to-event analyses, report median follow-up for patients without the event or the number followed without an event at a given follow-up time*

It is often useful to describe how long a cohort has been followed. To illustrate the appropriate methods of doing so, take the case of a cohort of 1000 pediatric cancer patients treated in 1970 and followed to 2010. If the cure rate was only 40%, the median follow-up for all patients might be only a few years, whereas the median follow-up for patients who survived was 40 yr. This latter statistic gives a much better impression of how long the cohort had been followed. Now assume that in 2009, a second cohort of 2000 patients was added to the study. The median follow-up for survivors will now be around a year, which, again, is misleading. An alternative would be to report a statistic such as "300 patients have been followed for more than 35 years."

4.5. *For time-to-event analyses, avoid reporting mean follow-up or survival time, or estimates of survival in those who had the event*

All three estimates are problematic in the context of censored data.

4.6. *For descriptive statistics, median with quartiles is preferred; range should be avoided*

The median and quartiles provide all sorts of useful information, for instance, that 50% of patients had values above the median or between the quartiles. The range gives the values of just two patients and so is generally uninformative about the data distribution.

4.7. *Avoid categorization of continuous variables*

A common approach to a variable such as age is to define patients as either *old* (aged $\geq$60 yr) or *young* (aged $<$ 60 yr) and then enter age into analyses as a categorical variable, reporting, for example, that "patients aged 60 years and older had twice the risk of an operative complication than patients younger than 60 years." In epidemiologic and marker studies, a common approach is to divide a variable into quartiles and report a statistic such as a hazard ratio for each quartile compared with the lowest (reference) quartile. This is problematic because it assumes that all values of a variable within a category are the same. For instance, it is likely not the case that a patient aged 65 yr has the same risk as a patient aged 90 yr but a different risk than a patient aged 64 yr. It is generally preferable to leave variables in continuous form, reporting, for instance, how risk changes with a 10-yr increase in age.

4.8. *The association between a continuous predictor and outcome can be demonstrated graphically, particularly by using nonlinear modeling*

Much high school math is based around the relationship between *y* and *x*, plotted graphically as a line, with a scatterplot added in some cases. This also holds true for many scientific studies. In the case of a study of age and complication rates, for instance, an investigator could plot age on the *x*-axis against risk of a complication on the *y*-axis and show a regression line, perhaps with a 95% CI. Nonlinear modeling is often useful because it avoids assuming a linear relationship and allows the investigator to determine questions such as whether risk starts to increase disproportionately beyond a given age.

4.9. *Report confidence intervals for the main estimates of interest*

A clinical study typically focuses on a limited number of scientific questions, each associated with an estimate, such as the AUC of a statistical model to predict biopsy outcome, the difference in rates of an adverse event comparing two different surgical techniques, or the hazard ratio of a risk factor for disease recurrence. Authors should generally report a 95% CI around these key estimates but not other estimates given in a paper. For instance, in the study comparing the two surgical techniques, the authors might report adverse event rates of 10% and 15%; however, the key estimate in this case is the difference between groups, so this estimate, 5%, should be reported along with a 95% CI (eg, 1–9%). Confidence intervals should not be reported for the estimates within each group (eg, rate in group A of 10% with 95% CI of 7%–13%). Similarly, confidence intervals should not be given for statistics such as average age or gender ratio.

4.10. *Consider the impact of missing data and patient selection*

It is rare that data are obtained from all patients in a study. A typical paper might report, for instance, that of 200 patients, 8 had data missing on important baseline variables and 34 did not complete the end-of-study questionnaire, leading to a final data set of 158. Similarly, many studies include a relatively narrow subset of patients, such as 50 patients referred for imaging before surgery of the 500 treated surgically during that time frame. In both cases, it is worth considering analyses to investigate whether patients with missing data or who were not selected for treatment were different in some way from those who were included in the analyses. Although statistical adjustment for missing data is complex and is warranted only in a limited set of circumstances, basic analyses to understand the characteristics of patients with missing data are relatively straightforward and are often helpful.

5. Multivariable models and diagnostic tests

5.1. *Multivariable analysis is not a magic wand*

Some investigators assume that multivariable adjustment "removes confounding" or "makes groups similar." There are two problems. First, the value of a variable recorded in a data set is often approximate and so may mask differences between groups. For instance, clinical stage might be used as a covariate in a study comparing treatments for localized prostate cancer, but stage T2c might constitute a small nodule on each prostate lobe or, alternatively, most of the prostate consisting of a large, hard mass. The key point is that if one group has more T2c disease than the other, it is also likely that those with T2c disease in that group will fall toward the more aggressive end of the T2c spectrum. Multivariable adjustment has the effect of making the rates of T2c in each group the same but does not ensure that the type of T2c is identical. Second, a model adjusts for only a small number of measured covariates. That does not exclude the possibility of important differences in unmeasured (or even unmeasurable) covariates.

5.2. *Propensity and instrumental variable approaches are not a magic wand*

A common assumption is that propensity methods somehow provide better adjustment for confounding than traditional multivariable methods. Except in certain rare circumstances, such as when the number of covariates is large relative to the number of events, propensity methods give extremely similar results to multivariable regression. Similarly, instrumental variables analyses depend on the availability of a good instrument, which is less common than is often assumed. In many cases, the instrument is not strongly associated with the intervention, leading to underestimate of treatment effects, or the analysis leads to a very large increase in the 95% CI, leading to loss of precision.

5.3. *Discrimination is a property not of a multivariable model but rather of the predictors and the data set*

Model building is generally seen as a process of fitting coefficients; however, discrimination is largely a property of what predictors are available. For instance, we have excellent models for prostate cancer outcome primarily because Gleason score is very strongly associated with malignant potential. In addition, discrimination is highly dependent on how much a predictor varies in the data set. As an example, a model to predict erectile dysfunction that includes age will have much higher discrimination for a population sample of adult men than for a group of older men presenting at a urology clinic, where the variation in lower than the population as a whole. Authors need to consider these points when drawing conclusions about the discrimination of models.

5.4. *Correction for overfit is strongly recommended for internal validation*

In the same way that it is easy to predict last week's weather, a prediction model generally has very good properties when evaluated on the same data set used to create the model. This problem is generally described as *overfit*. Various methods are available to correct for overfit, including cross-validation and bootstrap resampling. Note that such methods should include all steps of model building. For instance, if an investigator uses stepwise methods to choose which predictors should go into the model and then fits the coefficients, a typical cross-validation approach would be (1) to split the data into 10 groups, (2) to use stepwise methods to select predictors using the first 9 groups, (3) to fit coefficients using the first 9 groups, (4) to apply the model to the 10th group to obtain predicted probabilities, and (5) to repeat steps 2–4 until all patients in the data set have predicted probability derived from a model fitted to a data set that did not include that patient's data. Statistics such as the AUC are then calculated using the predicted probabilities directly.

5.5. *Calibration is nearly always excellent on internal validation*

It is rarely worth reporting calibration for a model created and tested on the same data set, even if techniques such as cross-validation are used.

5.6. *Calibration is a property not of a multivariable model but rather of the relationship between the model and a data set*

A model cannot be inherently "well calibrated." All that can be said is that predicted and observed risk are close in a given data set, representative of a given type of population.

5.7. *The optimal cut-point for a test or model cannot be derived from the receiver operating characteristic curve*

It is sometimes claimed that the optimal cut-point for a test is the one closest to the top left-hand corner of the receiver operating characteristic (ROC) curve. The problem with this approach is that the ROC curve assumes that sensitivity and specificity are equally important, but this is rarely, if ever, the case.

5.8. *Avoid reporting sensitivity and specificity for continuous predictors or a model*

Investigators often report sensitivity and specificity at a given cut-point for a continuous predictor (such as prostate-specific antigen [PSA] of 10 ng/ml) or report specificity at a given sensitivity (eg, 90%). Reporting sensitivity and specificity is not of value because it is unclear how high sensitivity or specificity would have to be in order to be "high enough" to justify clinical use. Similarly, it is very difficult to determine which of two tests, one with a higher sensitivity and the other with a higher specificity, is preferable because clinical value depends on the prevalence of disease and the relative harms of a false-positive result compared with a false-negative result. In the case of reporting specificities at fixed sensitivity, or vice versa, it is all but impossible to choose the specific sensitivity rationally. For instance, a team of investigators may state that they want to know specificity at 80% sensitivity because they want to ensure they catch 80% of cases. But 80% might be too low if prevalence is high or too high if prevalence is low.

5.9. *Report the clinical consequences of using a test or a model*

In place of statistical abstractions such as sensitivity and specificity, or a ROC curve, authors are encouraged to choose illustrative cut-points and then report results in terms of clinical consequences. As an example, consider a study in which a marker is measured in a group of patients undergoing biopsy. Authors could report that if a given level of the marker had been used to determine biopsy, then a certain number of biopsies would have been conducted and a certain number of cancers found and missed.

5.10. Avoid stepwise selection

Data-dependent variable selection in regression models has a number of undesirable properties, increasing the risk of overfit and making many statistics, such as the 95% CI, highly questionable. Use of stepwise selection should be restricted to a limited number of circumstances, such as during the initial stages of developing a model, if there is poor knowledge of what variables might be predictive.

5.11. Avoid reporting estimates such as odds or hazard ratios for covariates when examining the effects of interventions

In a typical observational study, an investigator might explore the effects of two different approaches to radical prostatectomy on recurrence while adjusting for covariates such as stage, grade, and PSA. It is rarely worth reporting estimates such as odds or hazard ratios for the covariates. For instance, it is well known that a high Gleason score is strongly associated with recurrence; reporting a hazard ratio of say, 4.23, is not helpful.

5.12. Rescale predictors to obtain interpretable estimates

Several predictors have moderate association with outcome and can take a large range of values. This can lead to uninterpretable estimates. For instance, the odds ratio per year of age might be given as 1.02 (95% CI, 1.01–1.02; $p < 0.0001$). It is not helpful to have the upper bound of a confidence interval be equivalent to the central estimate; a better alternative would be to report an odds ratio per 10 yr of age. This is simply achieved by creating a new variable equal to age divided by 10 to obtain an odds ratio of 1.16 (95% CI, 1.10–1.22; $p < 0.0001$) per 10-yr difference in age.

5.13. Avoid reporting both univariate and multivariable analyses unless there is a good reason

Comparison of univariate and multivariable models can be of interest when trying to understand mechanisms. For instance, if race is a predictor of outcome on univariate analysis but not after adjustment for income and access to care, one might conclude that poor outcome in black patients is explained by socioeconomic factors. However, the routine reporting of estimates from both univariate and multivariable analysis is discouraged.

5.14. The net reclassification improvement is an invalid statistic for the evaluation of markers and models

It has been amply demonstrated in the methodological literature than the net reclassification improvement provides faulty inference and an estimate that is difficult to interpret.

5.15. Interpret decision curves with careful reference to threshold probabilities

It is insufficient merely to report that, for instance, "the marker model had highest net benefit for threshold probabilities of 35–65%". Authors need to consider whether those threshold probabilities are rational. If the study reporting benefit between 35–65% concerned detection of high-grade prostate cancer, few if any urologists would demand that a patient have at least a one-in-three chance of high-grade disease before recommending biopsy. The authors would need to conclude that the model was not of benefit.

6. Conclusions and interpretation

6.1. Draw a conclusion

Conclusion sections are often simply a restatement of the results. For instance, "A statistically significant relationship was found between body mass index (BMI) and disease outcome" is not a conclusion. Authors instead need to state the implications for research and/or clinical practice. For instance, a conclusion section might call for research to determine whether the association between BMI is causal or make a recommendation for more aggressive treatment of patients with higher BMI.

6.2. Avoid words such as "may" or "might" in conclusions

A conclusion that a novel treatment "may" be of benefit would be untrue only if had been proven that the treatment was ineffective. Indeed, that the treatment *may* help would have been the rationale for the study in the first place. Using words such as *may* in the conclusion is equivalent to stating, "We know no more at the end of this study than we knew at the beginning"—reason enough to reject a paper for publication.

6.3. Do not confuse statistical and clinical significance

A small p value means only that the null hypothesis has been rejected. This may or may not have implications for clinical practice. That a marker is a statistically significant predictor of outcome does not imply that treatment decisions should be made on the basis of that marker. Similarly, a statistically significant difference between two treatments does not necessarily mean that the former should be preferred to the latter. Authors need to justify any clinical recommendations by carefully analyzing the clinical implications of their findings.

6.4. Avoid pseudo-limitations

Authors commonly describe study limitations in a rather superficial way, such as, "Small sample size and retrospective analysis are limitations." But a small sample size may be immaterial if the results of the study are clear. For instance, if a treatment or predictor is associated with a very large odds ratio, a large sample size might be unnecessary.

Similarly, a retrospective design might be entirely appropriate, as in the case of a marker study with very long-term follow-up, and have no discernible disadvantages compared with a prospective study. Discussion of limitations should include both the likelihood and the effect size of possible bias.

6.5. *Do not confuse outcome with response*

Investigators often compare outcomes in different subgroups of patients all receiving the same treatment. A common error is to conclude that patients with poor outcome are not good candidates for that treatment and should receive an alternative approach. This recommendation confuses differences between patients and differences between treatments. As a simple example, patients with large tumors are more likely to recur after surgery than patients with small tumors, but that cannot be taken to suggest that resection is not indicated for patients with tumors greater than a certain size. Indeed, surgery is generally more strongly indicated for patients with aggressive (but localized) disease, and such patients are unlikely to do well on surveillance.

6.6. *Do not draw conclusions based on investigator-specified cut-points*

Investigators often compound the problem of categorization of continuous variables by drawing conclusions based on their categorization. For instance, an investigator might examine surgical outcomes by PSA using a cut-off of 10 ng/ml. If differences were shown between high- and low-PSA subgroups, it would be unsound to then conclude that patients with PSA $\geq$10 ng/ml should be subject to different treatment than those with PSA $<$ 10 ng/ml, because it was the investigator who chose the cut-point. Similar results would likely have been found for a cut-point of 9 ng/ml, 11 ng/ml, or even 13.67 ng/ml, but that does not make those justifiable clinical decision rules. A similar error has been made for studies of the learning curve, dividing patients into two groups depending on whether the surgeon had $>$100 prior cases, finding differences between groups, and then concluding that "the learning curve is 100." Volume–outcome studies have also been beset by the problem of investigator-determined cut-points. For instance, annual surgical volume might be divided into quartiles, with the upper quartile being about 10 cases. It would be unjustified, however, to state that urologists should conduct at least 10 cases per year.

6.7. *Be cautious about causal attribution*

It is well known that "correlation does not imply causation," but authors often slip into this error in making conclusions. The introduction and methods sections might insist that the purpose of the study is merely to determine whether there is an association between, say, treatment frequency and treatment response, but the conclusions may imply that more frequent treatment would improve response rates.

7. Concluding remarks

These guidelines are not intended to cover all medical statistics but rather the statistical approaches most commonly used in papers submitted to *European Urology*. It is quite possible for a paper to follow all of the guidelines yet be statistically flawed or to break numerous guidelines and still be statistically sound. On balance, however, the reporting, analysis, and interpretation of clinical urologic research will be improved by adherence to these guidelines.

Conflicts of interest: The authors have nothing to disclose.

Funding support: Supported in part by funds from David H. Koch provided through the Prostate Cancer Foundation, the Sidney Kimmel Center for Prostate and Urologic Cancers, P50-CA92629 SPORE grant from the National Cancer Institute to Dr. H. Scher, and the P30-CA008748 NIH/NCI Cancer Center Support Grant to Memorial Sloan-Kettering Cancer Center.

Appendix A. Further reading

A classic study documenting statistical errors in published papers
Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg 1985;64:607–11.

A similar study in the urology literature
Scales CD Jr, Norris RD, Peterson BL, Preminger GM, Dahm P. Clinical research and statistical methods in the urology literature. J Urol 2005;174:1374–9.

EQUATOR network for reporting guidelines
Enhancing the QUAlity and Transparency Of health Research Web site. http://www.equator-network.org.

Statisticians often not included in study teams
Altman DG, Goodman SN, Schroter S. How statistical expertise is used in medical research. JAMA 2002;287:2817–20. http://jama.jamanetwork.com/article.aspx?articleid= 194983.

Evaluating prediction models
Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). Urology 2010;76:1298–30. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2997853/.

General advice on writing clinical research
Vickers AJ. Writing up clinical research: a statistician's view. J Gen Intern Med 2013;28:1127–9. http://www.ncbi. nlm.nih.gov/pmc/articles/PMC3744309 [available at PubMed Central September 1, 2014].

Randomization
Vickers AJ. How to randomize. J Soc Integr Oncol 2006;4:194–8. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2596474.

Please cite this article in press as: Vickers AJ, Sjoberg DD. Guidelines for Reporting of Statistics in *European Urology*. Eur Urol (2014), http://dx.doi.org/10.1016/j.eururo.2014.06.024