HHS Public Access

Author manuscript

J Urol. Author manuscript; available in PMC 2020 March 01.

Published in final edited form as:

J Urol. 2019 March ; 201(3): 595–604. doi:10.1097/JU.0000000000000001.

Guidelines for reporting of statistics for clinical research in urology

Melissa Assel[1], **Daniel Sjoberg**[1], **Andrew Elders**[2], **Xuemei Wang**[3], **Dezheng Huo**[4], **Albert Botchway**[5], **Kristin Delfino**[5], **Yunhua Fan**[6], **Zhiguo Zhao**[7], **Tatsuki Koyama**[8], **Brent Hollenbeck**[9], **Rui Qin**[10], **Whitney Zahnd**[11], **Emily C. Zabor**[1], **Michael W. Kattan**[7], and **Andrew J. Vickers**[1]

[1]Memorial Sloan Kettering Cancer Center

[2]Glasgow Caledonian University

[3]The University of Texas, MD Anderson Cancer Center

[4]The University of Chicago

[5]Southern Illinois University School of Medicine

[6]University of Minnesota

[7]Cleveland Clinic

[8]Vanderbilt University Medical Center

[9]University of Michigan

[10]Janssen Research & Development

[11]University of South Carolina

Abstract

In an effort to improve the quality of statistics in the clinical urology literature, statisticians at *European Urology, The Journal of Urology, Urology* and BJUI came together to develop a set of guidelines to address common errors of statistical analysis, reporting and interpretation. Authors should “break any of the guidelines if it makes scientific sense to do so” but would need to provide a clear justification. Adoption of the guidelines will in our view not only increase the quality of published papers in our journals but improve statistical knowledge in our field in general.

It is widely acknowledged that the quality of statistics in the clinical research literature is poor. This is true for urology just as it is for other medical specialties. In 2005, Scales et al. published a systematic evaluation of the statistics in papers appearing in a single month in one of the four leading urology medical journals: *European Urology, The Journal of Urology®, Urology* and BJUI. They reported widespread errors, including 71% of papers with comparative statistics having at least one statistical flaw[1]. These findings mirror many

Correspondence to: Andrew J Vickers, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, 2nd Floor, New York, NY 10017, vickersa@mskcc.org.

Conflicts of interest: The authors have nothing to disclose.

others in the literature, see, for instance, the review given by Lang and Altman[2]. The quality of statistical reporting in urology journals has no doubt improved since 2005, but remains unsatisfactory.

The four urology journals in the Scales et al. review have come together to publish a shared set of statistical guidelines[3]. Statistical reviewers at the four journals will systematically assess submitted manuscripts using the guidelines to improve statistical analysis, reporting and interpretation. Adoption of the guidelines will, in our view, not only increase the quality of published papers in our journals but improve statistical knowledge in our field in general. Asking an author to follow a guideline about, say, the fallacy of accepting the null hypothesis, would no doubt result in a better paper, but we hope that it would also enhance the author's understanding of hypothesis tests.

The guidelines are didactic, based on the consensus of the statistical consultants to the journals. We avoided, where possible, making specific analytic recommendations and focused instead on analyses or methods of reporting statistics that should be avoided. We intend to update the guidelines over time and hence encourage readers who question the value or rationale of a guideline to write to the authors.

1. The golden rule: Break any of the guidelines if it makes scientific sense to do so.

Science varies too much to allow methodologic 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, ReMARK for marker studies, TRIPOD for prediction models, STROBE for observational studies, or AMSTAR for systematic reviews.

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 many other guidelines, covering a wide range of different types of study. Reporting guidelines can be downloaded from the Equator Web site (http://www.equator-network.org).

2.2. Describe cohort selection fully.

It is insufficient to state, for instance, "the study cohort consisted of 1144 patients treated for benign prostatic hyperplasia at our institution". The cohort needs to be defined in terms of dates (e.g. "presenting March 2013 to December 2017"), inclusion criteria (e.g. "IPSS > 12") and whether patients were selected to be included (e.g. for a research study) vs. being a consecutive series. Exclusions should be described one by one, with the number of patients omitted for each exclusion criterion to give the final cohort size (e.g. "patients with prior surgery (n=43), allergies to 5-ARIs (n=12) and missing data on baseline prostate volume (n=86) were excluded to give a final cohort for analysis of 1003 patients"). Note that inclusion criteria can be omitted if obvious from context (e.g. no need to state "undergoing

radical prostatectomy for histologically proven prostate cancer"); on the other hand, dates may need to be explained if their rationale could be questioned (e.g. "March 2013, when our specialist voiding clinic was established to December 2017").

2.3. 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 it bears repeating. The purpose of randomization is to prevent selection bias. This can be achieved only if those consenting patients cannot guess a patient's treatment 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 logistic steps taken to ensure allocation concealment. The best method is to use a password-protected computer database.

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

Many statistical methods sections state only 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, where this is not obvious or if there are reasonable alternatives. Special attention and description should be provided for rarely used statistical techniques.

2.5. 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 "non-linear approaches" is insufficiently specific to allow replication, a cornerstone of the scientific method. All statistical analyses should be specified in the Methods section, including details such as the covariates included in a multivariable model. All variables should be clearly defined where there is room for ambiguity. For instance, avoid saying that "Gleason grade was included in the model"; state instead "Gleason grade group was included in four categories 1, 2, 3 and 4 or 5".

3. Inference and p-values (see also "Use and interpretation of p-values" below)

3.1. Don't 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 not rejected. If the p-value is 0.05 or more, 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 statistically 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 the p-value isn’t moving). Alternative language might be: “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. P-values and 95% confidence intervals do not quantify the probability of a hypothesis.

A p-value of, say, 0.03 does not mean that there is 3% probability that the findings are due to chance. Additionally, a 95% confidence interval should not be interpreted as a 95% certainty the true parameter value is in the range of the 95% confidence interval. The correct interpretation of a p-value is the probability of finding the observed or more extreme results when the null hypothesis is true; the 95% confidence interval will contain the true parameter value 95% of the time were a study to be repeated many times using different samples.

3.4. Don’t use confidence intervals to test hypotheses.

Investigators often interpret confidence intervals in terms of hypotheses. For instance, investigators might claim that there is a statistically significant difference between groups because the 95% confidence interval for the odds ratio excludes 1. Such claims are problematic because confidence intervals are concerned with estimation, not inference. Moreover, the mathematical method to calculate confidence intervals may be different from those used to calculate p-values. It is perfectly possible to have a 95% confidence interval that includes no difference between groups even though the p-value is less than 0.05 or *vice versa*. For instance, in a study of 100 patients in two equal groups, with event rates of 70% and 50%, the p-value from Fisher’s exact test is 0.066 but the 95% C.I. for the odds ratio is 1.03 to 5.26. The 95% C.I. for the risk difference and risk ratio also exclude no difference between groups.

3.5. Take care interpreting results when reporting multiple p-values.

The more questions you ask, the more likely you are to get a spurious 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 endpoints, thereby testing 30 separate hypotheses, a p-value of 0.04 should not be interpreted in the same way as if study tested only a single hypothesis with a p-value of 0.04.

3.6. Do not report separate p-values for each of two different groups in order 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 (standard deviation 14) points in the drug group (p=0.03 by one-sample t-test) and 5 (standard deviation 15) points 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.

3.7. 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 another. One approach that is more appropriate 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 men, the model might be as follows:

$$\{Final\ Pain\ Score\} = \beta_0 + \beta_1\{Baseline\ Pain\ Score\} + \beta_2\{Drug\} + \beta_3\{Sex\} + \beta_4\{Drug\} \times \{Sex\}$$

It is sometimes appropriate to report estimates and confidence intervals within subgroups of interest, but p-values should be avoided.

3.8. Tests for change over time are generally uninteresting.

A common analysis is to conduct a paired t-test comparing, say, erectile function in older men at baseline with erectile function after 5 years of follow-up. The null hypothesis here 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% confidence interval.

3.9. Avoid using 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 Mann-Whitney, 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 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, only altering the analysis if the data unambiguously point to a better alternative.

3.10. When reporting p-values, be clear about the hypothesis tested and ensure that the hypothesis is a sensible one.

P-values test very specific hypotheses. When reporting a p-value in the results section, state the hypothesis being tested unless this is completely clear. Take, for instance, the statement "Pain scores were higher in group 1 and similar in groups 2 and 3 (p=0.02)". It is ambiguous whether the p-value of 0.02 is testing group 1 vs. groups 2 and 3 combined or the hypothesis that pain score is the same in all three groups. Clarity about the hypotheses being tested can help avoid the testing of inappropriate hypotheses. For instance, p-values for differences between groups at baseline in a randomized trial is testing a null hypothesis that is known to be true (informally, that any observed differences between groups are due to chance).

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% to 27%). Reporting in a clinical study that the mean calorie consumption was 2069.9 suggests that calorie consumption can be measured extremely precisely by a food questionnaire. Some might argue that being overly precise is irrelevant, because the extra numbers can always be ignored. The counter-argument 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. The specific guidelines for precision are as follows:

- Report p-values to a single significant figure unless the p is close to 0.05, in which case, report two significant figures. Do not report "NS" for p-values of 0.05 or above. 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: <0.001, 0.004, 0.045, 0.13, 0.3, 1.
- Report percentages, rates and probabilities to 2 significant figures, e.g. 75%, 3.4%, 0.13%.
- Do not report p-values of zero, as any experimental result has a non-zero probability.
- Do not give decimal places if a probability or proportion is 1 (e.g. a p-value of 1.00 or a percentage of 100.00%). The decimal places suggest it is possible to have, say, a p-value 1.05. There is a similar consideration for data that can only

take integer values. It makes sense to state that, for instance, the mean number of pregnancies was 2.4, but not that 29% of women reported 1.0 pregnancies.

- There is generally 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 (e.g. 18.2 rather than 18.17).

4.2. Avoid redundant statistics in cohort descriptions.

Authors should be selective about the descriptive statistics reported and ensure that each and every number provides unique information. Authors should avoid reporting descriptive statistics that can be readily derived from data that have already been provided. For instance, there is no need to state 40% of a cohort were men and 60% were women, choose one or the other. Another common error is to include a column of descriptive statistics for two groups separately and then the whole cohort combined. If, say, the median age is 60 in group 1 and 62 in group 2, we do not need to be told that the median age in the cohort as a whole is close to 61.

4.3. For descriptive statistics, median and quartiles are preferred over means and standard deviations (or standard errors); 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 of the data distribution.

4.4. Report estimates for the main study questions.

A clinical study typically focuses on a limited number of scientific questions. Authors should generally provide an estimate for each of these questions. In a study comparing two groups, for instance, authors should give an estimate of the difference between groups, and avoid giving only data on each group separately or, simply saying that the difference was or was not significant. In a study of a prognostic factor, authors should give an estimate of the strength of the prognostic factor, such as an odds ratio or hazard ratio, as well as reporting a p-value testing the null hypothesis of no association between the prognostic factor and outcome.

4.5. Report confidence intervals for the main estimates of interest.

Authors should generally report a 95% confidence interval around the estimates relating to the key research questions, but not other estimates given in a paper. For instance, in a study comparing 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% confidence interval (e.g. 1% to 9%). Confidence intervals should not be reported for the estimates within each group (e.g. adverse event rate in group A of 10%, 95% CI 7% to 13%). Similarly, confidence intervals should not be given for statistics such as mean age or gender ratio.

4.6. Do not treat categorical variables as continuous.

A variable such as Gleason grade groups are scored 1 – 5, but it is not true that the difference between group 3 and 4 is half as great as the difference between group 2 and 4. Variables such as Gleason grade group should be reported as categories (e.g. 40% grade group 1, 20% group 2, 20% group 3, 20% group 4 and 5) rather than as a continuous variable (e.g. mean Gleason score of 2.4). Similarly, categorical variables such as Gleason should be entered into regression models not as a single variable (e.g. a hazard ratio of 1.5 per 1-point increase in Gleason grade group) but as multiple categories (e.g. hazard ratio of 1.6 comparing Gleason grade group 2 to group 1 and hazard ratio of 3.9 comparing group 3 to group 1).

4.7. Avoid categorization of continuous variables unless there is a convincing rationale.

A common approach to a variable such as age is to define patients as either old (≥ 60) or young (<60) and then enter age into analyses as a categorical variable, reporting, for example, that “patients aged 60 and over had twice the risk of an operative complication than patients aged less than 60”. 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 to 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 has the same risk as a patient aged 90, but a very different risk to a patient aged 64. It is generally preferable to leave variables in a continuous form, reporting, for instance, how risk changes with a 10-year increase in age. Non-linear terms can also be used, to avoid the assumption that the association between age and risk follows a straight line.

4.8. Do not use statistical methods to obtain cut-points for clinical practice.

There are various statistical methods available to dichotomize a continuous variable. For instance, outcomes can be compared either side of several different cut-points, and the optimal cut-point chosen as the one associated with the smallest p-value. Alternatively, investigators might choose a cut-point that leads to the highest value of sensitivity + specificity, that is, the point closest to the top left-hand corner of a Receiver Operating Curve (ROC). Such methods are inappropriate for determining clinical cut-points because they do not consider clinical consequences. The ROC curve approach, for instance, assumes that sensitivity and specificity are of equal value, whereas it is generally worse to miss disease than to treat unnecessarily. The smallest p-value approach tests strength of evidence against the null hypothesis, which has little to do with the relative benefits and harms of a treatment or further diagnostic work up.

4.9. The association between a continuous predictor and outcome can be demonstrated graphically, particularly by using non-linear modeling.

In high-school math we often thought about the relationship between y and x by plotting a line on a graph, 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% confidence interval. Non-linear 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.

4.10. Do not ignore significant heterogeneity in meta-analyses.

Informally speaking, heterogeneity statistics test whether variations between the results of different studies in a meta-analysis are consistent with chance, or whether such variation reflects, at least in part, true differences between studies. If heterogeneity is present, authors need to do more than merely report the p-value and focus on the random-effects estimate. Authors should investigate the sources of heterogeneity and try to determine the factors that lead to differences in study results, for example, by identifying common features of studies with similar findings or idiosyncratic aspects of studies with outlying results.

4.11. For time-to-event variables, report the number of events but not the proportion.

Take the case of a study that reported: “of 60 patients accrued, 10 (17%) died”. While it is important to report the number of events, patients entered the study at different times and were followed for different periods, so the reported proportion of 17% is meaningless. The standard statistical approach to time-to-event variables is to calculate probabilities, such as the risk of death being 60% by five years or the median survival – the time at which the probability of survival first drops below 50% - being 52 months.

4.12. For 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 1,000 pediatric cancer patients treated in 1970 and followed to 2010. If the cure rate was only 40%, median follow-up for all patients might only be a few years, whilst the median follow-up for patients who survived was 40 years. 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 is again misleading. An alternative would be to report a statistic such as “312 patients have been followed without an event for at least 35 years”.

4.13. For time-to-event analyses, describe when follow-up starts and when and how patients are censored.

A common error is that investigators use a censoring date which leads to an overestimate of survival. For example, when assessing the metastasis-free survival a patient without a record of metastasis should be censored on the date of the last time the patient was known to be free of metastasis (e.g. negative bone scan, undetectable PSA), not at the date of last patient contact (which may not have involved assessment of metastasis). For overall survival, date of last patient contact would be an acceptable censoring date because the patient was indeed known to be event-free at that time. When assessing cause-specific endpoints, special consideration should be given the cause of death. The endpoints “disease-specific survival” and “disease-free survival” have specific definitions and require careful attention to methods.

With disease-specific survival, authors need to consider carefully how to handle death due to other causes. One approach is to censor patients at the time of death, but this can lead to bias in certain circumstances, such as when the predictor of interest is associated with other cause death and the probability of other cause death is moderate or high. Competing risk analysis is appropriate in these situations. With disease-free survival, both evidence of disease (e.g. disease recurrence) and death from any cause are counted as events, and so censoring at the time of other cause death is inappropriate. If investigators are specifically interested only in the former, and wish to censor deaths from other causes, they should define their endpoint as "freedom from progression".

4.14. 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.

4.15. For time-to-event analyses, make sure that all predictors are known at time zero or consider alternative approaches such as a landmark analysis or time-dependent covariates.

In many cases, variables of interest vary over time. As a simple example, imagine we were interested in whether PSA velocity predicted time to progression in prostate cancer patients on active surveillance. The problem is that PSA is measured at various times after diagnosis. Unless they were being careful, investigators might use time from diagnosis in a Kaplan-Meier or Cox regression but use PSA velocity calculated on PSAs measured at one and two-year follow-up. As another example, investigators might determine whether response to chemotherapy predicts cancer survival, but measure survival from the time of the first dose, before response is known. It is obviously invalid to use information only known "after the clock starts". There are two main approaches to this problem. A "landmark analysis" is often used when the variable of interest is generally known within a short and well-defined period of time, such as adjuvant therapy or chemotherapy response. In brief, the investigators start the clock at a fixed "landmark" (e.g. 6 months after surgery). Patients are only eligible if they are still at risk at the landmark (e.g. patients who recur before six months are excluded) and the status of the variable is fixed at that time (e.g. a patient who gets chemotherapy at 7 months is defined as being in the no adjuvant group). Alternatively, investigators can use a time-dependent variable approach. In brief, this "resets the clock" each time new information is available about a variable. This would be the approach most typically used for the PSA velocity and progression example.

4.16. When presenting Kaplan-Meier figures, present the number at risk and truncate follow-up when numbers are low.

Giving the number of risk is useful for helping to understand when patients were censored. When presenting Kaplan-Meier figures a good rule of thumb is to truncate follow-up when the number at risk in any group falls below 5 (or even 10) as the tail of a Kaplan-Meier distribution is very unstable.

5. Multivariable models and diagnostic tests

5.1. Multivariable, propensity and instrumental variable analyses are not a magic wand.

Some investigators assume that multivariable adjustment “removes confounding”, “makes groups similar” or “mimics a randomized trial”. There are two problems with such claims. 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 the T2c’s in that group will fall towards the more aggressive end of the 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 only adjusts for a small number of measured covariates. That does not exclude the possibility of important differences in unmeasured (or even unmeasurable) covariates. 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 a large increase in the 95% confidence interval or, in some cases, an underestimate of treatment effects.

5.2. Avoid stepwise selection.

Investigators commonly choose which variables to include in a multivariable model by first determining which variables are statistically significant on univariable analysis; alternatively, they may include all variables in a single model but then remove any that are not significant. This type of data-dependent variable selection in regression models has several undesirable properties, increasing the risk of overfit and making many statistics, such as the 95% confidence interval, 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.

5.3. 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 and a distraction from the key finding, the hazard ratio between the two types of surgery.

5.4. Rescale predictors to obtain interpretable estimates.

Predictors sometimes have a moderate association with outcome and can take a large range of values. This can lead to uninterpretable estimates. For instance, the odds ratio for cancer 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 ten years of age. This is simply achieved by creating a new variable equal to age divided by ten to obtain an odds ratio of 1.16 (95% CI 1.10, 1.22; $p<0.0001$) per 10-year difference in age.

5.5. 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 African-Americans is explained by socioeconomic factors. However, the routine reporting of estimates from both univariate and multivariable analysis is discouraged.

5.6. Avoid ranking predictors in terms of strength.

It is tempting for authors to rank predictors in a model, claiming, for instance, “the novel marker was the strongest predictor of recurrence”. Most commonly, this type of claim is based on comparisons of odds or hazard ratios. Such rankings are not meaningful since, among other reasons, it depends on how variables are coded. For instance, the odds ratio for hK2, and hence whether or not it is an apparently “stronger” predictor than PSA, will depend on whether it is entered in nanograms or picograms per ml. Further, it is unclear how one should compare model coefficients when both categorical and continuous variables are included. Finally, the prevalence of a categorical predictor also matters: a predictor with an odds ratio is 3.5 but a prevalence if 0.1% is less important that one with a 50% prevalence and an odds ratio of 2.0.

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

Although model building is generally seen as a process of fitting coefficients, 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, because there is a greater variation in age in the population sample. Authors need to consider these points when drawing conclusions about the discrimination of models. This is also why authors should be cautious about comparing the discrimination of different multivariable models where these were assessed in different datasets.

5.8. 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 crossvalidation 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 crossvalidation approach would be to: (1) split the data into ten groups, (2) use stepwise methods to select predictors using the first nine groups, (3) fit coefficients using the first nine groups, (4) apply the model to the 10^{th} group to obtain predicted probabilities, and (5) repeat steps 2–4 until all patients in the data set have a 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.

5.9. Calibration should be reported and interpreted correctly.

Calibration is a critical component of a statistical model: the main concern for any patient is whether the risk given by a model is close to his or her true risk. It is rarely worth reporting calibration for a model created and tested on the same data set, even if techniques such as crossvalidation are used. This is because calibration is nearly always excellent on internal validation. Where a pre-specified model is tested on an independent data set, calibration should be displayed graphically in a calibration plot. The Hosmer-Lemeshow test addresses an inappropriate null hypothesis and should be avoided. Note also that calibration depends upon both the model coefficients and the dataset being examined. A model cannot be inherently "well calibrated." All that can be said is that predicted and observed risk are close in a specific data set, representative of a given population.

5.10. 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 a PSA of 10 ng /mL), or report specificity at a given sensitivity (such as 90%). Reporting sensitivity and specificity is not of value because it is unclear how high sensitivity or specificity would have to be so as 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 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.

5.11. Report the clinical consequences of using a test or a model.

In place of statistical abstractions such as sensitivity and specificity, or an 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.

5.12. 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 therefore need to conclude that the model was not of benefit.

6. Conclusions and interpretation

6.1. Draw a conclusion, don't just repeat the results.

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 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 using words such as "may" or "might".

A conclusion such as that a novel treatment "may" be of benefit would only be untrue if it 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. A statistically significant p-value does not imply 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. For instance, 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 such as "small sample size" and "retrospective analysis", consider instead sources of potential bias and the mechanism for their effect on findings.

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 to a prospective study. Discussion of limitations should include both the likelihood and effect size of possible bias.

6.5. Consider the impact of missing data and patient selection.

It is rare that complete data is 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, out of the 500 treated surgically during that timeframe. 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.

6.6. Consider the possibility and impact of ascertainment bias

Ascertainment bias occurs when an outcome depends on a test, and the propensity for a patient to be tested is associated with the predictor. PSA screening provides a classic example: prostate cancer is found by biopsy, but the main reason why men are biopsied is because of an elevated PSA. A study in a population subject to PSA screening will therefore overestimate the association between PSA and prostate cancer. Ascertainment bias can also be caused by the timing of assessments. For instance, frequency of biopsy in prostate cancer active surveillance will depend on prior biopsy results and PSA level, and this induces an association between those predictors and time to progression.

6.7. Do not confuse outcome with response among subgroups of patients undergoing the same treatment: patients with poorer outcomes may still be good candidates for that treatment.

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 is to confuse differences between patients for 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.

6.8. Be cautious about causal attribution: correlation does not imply causation.

It is well-known that “correlation does not imply causation” but authors often slip into this error in making conclusions. The introduction and methods section 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, for instance, more frequent treatment would improve response rates.

Use and interpretation of p-values

That p-values are widely misused and misunderstood is apparent from even the most cursory reading of the medical literature. One of the most common errors is accepting the null hypothesis, for instance, concluding from a p-value of 0.07 that a drug is ineffective or that two surgical techniques are equivalent. This particular error is described in detail in guideline 3.1.

The more general problem, which we address here, is that p-values are often given excessive weight in the interpretation of a study. Indeed, studies are often classed by investigators into “positive” or “negative” based on statistical significance. Gross misuse of p-values has led some to advocate banning the use of p-values completely[4].

We follow the American Statistical Association statement on p-values and encourage all researchers to read either the full statement[5] or the summary[6]. In particular, we emphasize that the p-value is just one statistic that helps interpret a study, it does not determine our interpretations. Drawing conclusions for research or clinical practice from a clinical research study requires evaluation of the strengths and weakness of study methodology, the results of other pertinent data published in the literature, biological plausibility and effect size. Sound and nuanced scientific judgment cannot be replaced by just checking whether one of the many statistics in a paper is or is not less than 0.05.

Concluding remarks

These guidelines are not intended to cover all medical statistics but rather the statistical approaches most commonly used in clinical research papers in 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 analysis, reporting and interpretation of clinical urologic research will be improved by adherence to these guidelines.

Funding support:

Supported in part by 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.

References

[1]. Scales CD Jr., Norris RD, Peterson BL, Preminger GM, Dahm P. Clinical research and statistical methods in the urology literature. The Journal of urology. 2005;174:1374–9. [PubMed: 16145441]

[2]. Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the “Statistical Analyses and Methods in the Published Literature” or the SAMPL Guidelines. International journal of nursing studies. 2015;52:5–9. [PubMed: 25441757]

[3]. Vickers AJ, Sjoberg DD, European U. Guidelines for reporting of statistics in European Urology. European urology. 2015;67:181–7. [PubMed: 25037638]

[4]. Woolston C Psychology journal bans P-values. Nature. Nature: Nature; 2015.

[5]. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016;70:129–33.

[6]. [10/22/2018] https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf Accessed