CHEST

Statistical Analysis and Reporting Guidelines for *CHEST*

Michael W. Kattan, PhD; and Andrew J. Vickers, PhD

Considerable heterogeneity persists in the conduct and reporting of statistical analyses in the medical literature. Authors submitting manuscripts to *CHEST* are encouraged to adhere to the following guidelines where possible.

CHEST 2020; 158(1S):S3-S11

KEY WORDS: design and statistical analysis; *P* values; statistical reporting

Introduction

What follows are statistical reporting guidelines for the journal *CHEST*. Our aim with these guidelines is to improve the quality of manuscripts submitted as well as the review process by providing both clear recommendations and a relatively comprehensive list of the typical statistical errors that are seen in manuscripts. We recognize that there will be times when certain guidelines would not apply; science is too varied to be fit into rigid boxes. However, authors should be aware that papers submitted to *CHEST* will be reviewed using the statistical guidelines and that a compelling justification for not following the guidelines would be required in any response to peer review.

We previously published a similar version of these guidelines for urology researchers.[1] Those guidelines were found to be very helpful in standardizing and informing the review process. We have adapted them and changed the examples to suit the readers of *CHEST*. The guidelines are summarized in Table 1.

Reporting of Design and Statistical Analysis

Where Applicable, Follow Specific Reporting Guidelines for the Type of Study Being Described

These guidelines include Consolidated Standards of Reporting Trials (CONSORT) for randomized trials, Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) for marker studies, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) for prediction models, Strengthening the Reporting of Observational studies in Epidemiology (STROBE) for observational studies, and Assessing the Methodological Quality of Systematic Reviews (AMSTAR) for systematic reviews. These specific guidelines can be found at http://www.equator-network.org.

Describe Cohort Selection Completely

It is insufficient to state, for instance, "the study cohort consisted of 843 patients treated for lung cancer at our institution." The cohort

ABBREVIATIONS: LASSO = least absolute shrinkage and selection operator; ROC = receiver-operating characteristic

AFFILIATIONS: From the Department of Quantitative Health Sciences (Dr Kattan), Cleveland Clinic, Cleveland, OH; and the Department of Epidemiology & Biostatistics (Dr Vickers), Memorial Sloan Kettering Cancer Center, New York, NY.

FUNDING/SUPPORT: This work was supported in part by the National Institutes of Health/National Cancer Institute (NIH/NCI) with a Cancer Center Support Grant to Memorial Sloan Kettering Cancer Center [P30 CA008748].

CORRESPONDENCE TO: Michael W. Kattan, PhD, Department of Quantitative Health Sciences, Cleveland Clinic, 9500 Euclid Ave/JJN3-01, Cleveland, OH 44195; e-mail: kattanm@ccf.org

DOI: https://doi.org/10.1016/j.chest.2019.10.064

TABLE 1] Numbered Summary of the Statistical Reporting Guidelines for *CHEST*

2. Reporting of design and statistical analysis
2.1 Where applicable, follow specific reporting guidelines for the type of study being described
2.2 Describe cohort selection completely
2.3 Describe the study questions and the statistical approaches used to address each question in the statistical methods section
2.4 Describe the statistical methods with sufficient detail to allow replication by an independent statistician given the same dataset
2.5 Provide the sample size calculation for a clinical trial
3. Reporting of inference and *P* values
3.1 Do not accept the null hypothesis; it is either rejected or not rejected
3.2 Avoid stating that *P* values just above 5% are a trend or are moving
3.3 Do not quantify the probability of a hypothesis with *P* values or 95% CIs
3.4 Do not equate a statistically significant *P* value with clinical significance
3.5 Do not use CIs to test hypotheses.
3.6 Caution is warranted when reporting multiple *P* values
3.7 Do not report separate *P* values for each of two different groups to address the question of whether there is a difference between groups
3.8 Use interaction terms in place of subgroup analyses
3.9 Avoid using statistical tests to determine the type of analysis to be conducted
3.10 When reporting *P* values, be clear about the hypothesis tested and ensure that the hypothesis is sensible
4. Reporting of study estimates
4.1 Use appropriate levels of precision
4.2 Avoid redundant statistics in cohort descriptions
4.3 For descriptive statistics, median and quartiles are preferred over means and SDs
4.4 Report CIs for the main estimates of interest
4.5 Do not treat categorical variables as continuous
4.6 Avoid categorization of continuous variables unless there is a convincing rationale
4.7 Do not use statistical methods to obtain thresholds for clinical practice
4.8 Time-to-event analyses
4.8a Report the number of events but not the proportion
4.8b Report median follow-up separately for patients without the event or the number followed up without an event at a given follow-up time
4.8c Describe when the follow-up period started and when and how patients are censored
4.8d Avoid reporting mean follow-up, mean survival time, or estimates of survival in those who had the event of interest
4.8e Make sure that all predictors are known at baseline or consider alternative approaches such as a landmark analysis or time-dependent covariates
4.8f When presenting Kaplan-Meier figures, present the number at risk and truncate follow-up when low
5. Reporting of multivariable models and diagnostic tests
5.1 Do not assume that multivariable, propensity, and instrumental variable analyses will substitute for randomized trials
5.2 Avoid univariable screening and stepwise selection
5.3 When reporting the effects of continuous predictors, choose two clinically interesting predictor values or a clinically relevant range
5.4 Avoid reporting both univariable and multivariable analyses unless there is a good reason
5.5 Avoid ranking predictors in terms of strength
5.6 Be cautious when comparing models assessed on different datasets
5.7 Correct for overfit when conducting internal validation
5.8 Calibration for a prediction model should be presented graphically
5.9 Report the clinical consequences of using a test or a model

(Continued)

TABLE 1] *(Continued)*

6. Discussion, interpretation, and conclusions
6.1 Draw a conclusion; do not just repeat the results
6.2 Avoid using words such as “may” or “might”
6.3 Avoid pseudo-limitations such as “small sample size” and “retrospective analysis,” and consider instead sources of potential bias and the mechanism for their effect on findings
6.4 Discuss the impacts of missing data and patient selection

needs to be defined in terms of dates (eg, “diagnosed March 2013 to December 2017”), inclusion criteria, and whether patients were selected to be included (eg, 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 (eg, “patients with prior surgery [n = 43], COPD [n = 12], or missing data on smoking [n = 86] were excluded to give a final cohort for analysis of 702 patients”).

Describe the Study Questions and the Statistical Approaches Used to Address Each Question in the Statistical Methods Section

Statistical methods sections should lay out each primary study question separately: carefully detail the analysis associated with each and describe the rationale for the analytical approach, where this is not obvious or if there are reasonable alternatives. Special attention and description should be provided for rarely used statistical techniques.

Describe the Statistical Methods With Sufficient Detail to Allow Replication by an Independent Statistician Given the Same Dataset

Vague references such as “adjusting for confounders” or “nonlinear approaches” are not sufficient 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 to avoid ambiguity. For instance, it is insufficient to say that cancer stage was used as a covariate in a study of localized lung cancer; the authors need to specify whether stage was entered as, say, I vs II or, alternatively, divided into IA vs IB vs IIA vs IIB.

Provide the Sample Size Calculation for a Clinical Trial

Presenting a sample size calculation is expected by *CHEST* to be included in a submitted manuscript for experimental studies. In particular, if a statistically significant difference is not found, the 95% CI for the effect should be examined to determine whether it includes a clinically meaningful difference.

Reporting of Inference and *P* Values

Do Not Accept the Null Hypothesis; It Is Either Rejected or Not Rejected

If the *P* value is ≥ .05, 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.”

Avoid Stating That P *Values Just Above 5% Are a Trend or Are Moving*

An example of accepted alternative language might be: “although we saw some evidence of improved response rates in patients receiving the new drug, differences between groups did not meet conventional levels of statistical significance.”

Do Not Quantify the Probability of a Hypothesis With P *Values or 95% CIs*

For example, a *P* value of .03 does not mean that there is 3% probability that the findings are due to chance. In addition, a 95% CI should not be interpreted as a 95% certainty the true parameter value is in the range of the 95% CI. 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% CI will contain the true parameter value 95% of the time were a study to be repeated many times using different samples.

Do Not Equate a Statistically Significant P Value With Clinical Significance

A small *P* value means only that the null hypothesis has been rejected. This finding may or may not have implications for clinical practice. For instance, the fact 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 one should be preferred over the other. Authors need to justify any clinical recommendations by carefully analyzing the clinical implications of their findings.

Do Not Use CIs to Test Hypotheses

Investigators often interpret CIs in terms of hypotheses. For instance, investigators might claim that there is a statistically significant difference between groups because the 95% CI for the OR excludes 1. Such claims are problematic because CIs are concerned with estimation, not inference. Moreover, the mathematical method to calculate CIs may be different from those used to calculate *P* values. It is possible to have a 95% CI that includes no difference between groups even though the *P* value is $< .05$ 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 the Fisher exact test is .066 but the 95% CI for the OR is 1.03 to 5.26. The 95% CI for the risk difference and risk ratio also excludes no difference between groups.

Caution Is Warranted When Reporting Multiple P Values

If there is interest in whether any gene is associated with the development of lung cancer, and many genes are tested, it is likely to find at least one that is declared associated just by chance. The more hypotheses tested (eg, is this marker predictive?), the more likely it is that a spurious answer to at least one of them is obtained (eg, marker 23 is associated with mortality). 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 .04 should not be interpreted in the same way as if the study tested only a single hypothesis with a *P* value of .04.

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 allows for one statistical hypothesis to be 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 a drug vs placebo to reduce COPD symptoms, with 30 patients in each group. The authors might report that symptom scores improved by 6 (SD, 14) points in the drug group ($P = .03$ by one-sample Student *t* test) and 5 (SD, 15) points in the placebo group ($P = .08$). 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 Student *t* test for these data gives a *P* value of .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 more effective than placebo based on the finding that change was significant in the drug group but not in the placebo controls.

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 has a statistically significant effect in one group of patients but not another. A more appropriate approach would be to use an “interaction term” in a statistical model. Although reporting estimates and CIs within subgroups of interest are appropriate in some cases (eg, a prespecified subgroup based on a compelling rationale) reporting *P* values should be avoided.

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 Student *t* test or Mann-Whitney, Cochran’s Q to decide whether to use a fixed effects or random effects approach in a meta-analysis, or use a Student *t* test for between-group differences in a covariate to determine whether that covariate should be included in a multivariable model. The problem with these approaches is that the null hypothesis tested is known to be false. For instance, no dataset 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 disagree as to whether the Mann-Whitney test is always superior to the Student *t* test when data are nonnormal, 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 analytical plan, only altering the analysis if the data unambiguously point to a more appropriate alternative.

When Reporting P Values, Be Clear About the Hypothesis Tested and Ensure That the Hypothesis Is Sensible

P values test very specific hypotheses. When reporting a *P* value in the Results section, state the hypothesis being tested. Take, for instance, the statement "Pain scores were higher in group 1 and similar in groups 2 and 3 ($P = .02$)." It is ambiguous whether the *P* value of .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 are testing a null hypothesis that is known to be true, and this should not be reported.

Reporting of Study Estimates

Use Appropriate Levels of Precision

Reporting a *P* value of .7345 suggests that there is an appreciable difference between *P* values of .7344 and .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 CI (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 using a food questionnaire. The specific guidelines for precision are as follows:

1. Report *P* values to a single significant figure unless the *P* is close to .05 (eg, .01 to .2), in which case, report two significant figures. Do not report "not significant" ("NS") for *P* values of $\geq .05$. Very low *P* values can be reported as $P < .001$, while very high *P* values can be reported as $> .9$. For instance, the following *P* values are reported to appropriate precision: $< .001$, .004, .045, .13, and .3.
2. Report percentages, rates and probabilities to two significant figures; for example, 75%, 3.4%, and 0.13%.
3. There is generally no need to report estimates to more than three significant figures.
4. Hazard ratios and ORs are normally reported to two decimal places, although this can be avoided for high ORs (eg, 18.2 rather than 18.17).

Avoid Redundant Statistics in Cohort Descriptions

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 that 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, for example, the median age is 60 years in group 1 and 62 years in group 2, we do not need to be told that the median age in the cohort as a whole is close to 61 years.

For Descriptive Statistics, Median and Quartiles Are Preferred Over Means and SDs

The median and quartiles provide all sorts of useful information; for example, that 50% of patients had values above the median or between the quartiles. Also, it is generally better to report the SD rather than the SE. The SD quantifies the variability of observations from the sample mean, which is usually of interest. The SE quantifies the uncertainty regarding the true value of the population mean, which is usually not of primary interest.

Report CIs for the Main Estimates of Interest

Authors should generally report a 95% CI 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% CI (eg, 1-9). CIs should not be reported for the estimates within each group (eg, adverse event rate in group A of 10%; 95% CI, 7-13). Similarly, CIs should not be given for descriptive statistics such as mean age or sex ratio as these are unrelated to the research questions addressed in the study.

Do Not Treat Categorical Variables as Continuous

Similarly, categorical variables should be entered into regression models not as a single variable but as multiple categories so that the estimated effect is not held constant across categories.

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 (aged ≥ 60 years) or young (aged < 60 years) and then enter age into analyses as a categorical variable, reporting, for example, that "patients aged ≥ 60 years had twice the risk of an operative complication than patients aged < 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 have equal weight. For instance, a patient aged 61 years is unlikely to have the same risk as a patient aged 90 years, or have a risk very different from a patient aged 59 years. It is preferable to leave variables in a continuous form, reporting, for instance, how risk changes with a 10-year increase in age. Nonlinear terms can also be used, to avoid the assumption that the association between age and risk follows a straight line.

Do Not Use Statistical Methods to Obtain Thresholds 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 thresholds, and the optimal threshold chosen as the one associated with the smallest *P* value. Alternatively, investigators might choose a threshold that leads to the highest value of sensitivity and specificity; that is, the point closest to the top left-hand corner of a receiver-operating characteristic (ROC) curve. Such methods are inappropriate for determining clinical thresholds 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 evaluation. We recommend decision analytic, rather than purely statistical, methods to determine thresholds for clinical practice.

Time-to-Event Analyses

Report the Number of Events but not the Proportion: As an example, consider a study that reported: "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 up for different periods, so the reported proportion of 17% is meaningless. The standard statistical approach to time-to-event variables is to calculate probabilities at certain time points, such as the risk of death being 60% by 5 years. Also, one might report the median survival, the time at which the probability of survival first drops below 50%.

Report Median Follow-up Separately for Patients Without the Event or the Number Followed Up Without an Event at a Given Follow-up Time: For example, consider the case of a cohort of 1,000 pediatric patients with cancer treated in 1970 and followed up until 2010. If the cure rate was only 40%, the median follow-up for all patients might only be a few years, while 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 up. Reporting the median follow-up for the group as a whole penalizes cohorts with many events that happen early in follow-up, despite the fact that those are the most informative data. It is the degree of follow-up on those observations that are censored that is important to recognize.

Describe When the Follow-up Period Started and When and How Patients Are Censored: A common error is that investigators use a censoring date, which leads to an overestimate of survival, because not all patients were assessed for the outcome of interest on that date. For example, when assessing 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, 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 end points, special consideration should be given to the cause of death. The end points "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 method can lead to bias in certain circumstances, such as when the predictor of interest is associated with other cause of death and the probability of other cause of death is moderate or high. Competing risk analysis is appropriate in these situations. With disease-free survival, both evidence of disease (eg, disease recurrence) and death from any cause are counted as events, and thus censoring at the time of other cause of death is inappropriate. If investigators are specifically interested only in the former, and wish to censor deaths from other causes, they should define their end point as "freedom from progression."

Avoid Reporting Mean Follow-up, Mean Survival Time, or Estimates of Survival in Those who had the Event of Interest: All three estimates are problematic in the context of censored data.

Make Sure That All Predictors Are Known at Baseline or Consider Alternative Approaches Such as a Landmark Analysis or Time-Dependent Covariates: In many cases, variables of interest vary over time. As an 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" (eg, 6 months following surgery). Patients are only eligible if they are still at risk at the landmark (eg, patients who recur prior to 6 months are excluded). Then, the status of the variable is fixed at that time (eg, a patient who undergoes chemotherapy at 7 months is defined as being in the no adjuvant group, their 6-month landmark status). Alternatively, investigators can use a time-dependent variable approach. In brief, this "resets the clock" each time new information is available about a variable.

When Presenting Kaplan-Meier Figures, Present the Number at Risk and Truncate Follow-up When Low: Providing the number at risk over time is useful for helping to understand when patients were censored. When the number at risk in any group falls below five (or even 10), the tail of a Kaplan-Meier distribution is very unstable.

Reporting of Multivariable Models and Diagnostic Tests

Do Not Assume That Multivariable, Propensity, and Instrumental Variable Analyses Will Substitute for Randomized Trials

Some investigators assume that multivariable adjustment "removes confounding" or "makes groups similar" and therefore leads to a study that "mimics a randomized trial." However, we can never have all the variables necessary and measured sufficiently accurately to remove all confounding. A common assumption is that propensity methods 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 nearly identical results to multivariable regression. Moreover, instrumental variable analysis depends 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% CI or, in some cases, an underestimate of treatment effects. For these reasons, authors need to be cautious regarding causal inference from observational studies even when the methods listed here are used.

Avoid Univariable Screening and Stepwise Selection

Investigators commonly choose which variables to include in a multivariable model by first determining which variables are statistically significant based on a univariable analysis; another approach 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 (ie, modeling the data too closely, such that future generalizability is reduced) 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. Ideally, a biologically based theoretical model best guides variable selection.

When Reporting the Effects of Continuous Predictors, Choose Two Clinically Interesting Predictor Values or a Clinically Relevant Range

For instance, the OR for cancer per year of age might be given as 1.02 (95% CI, 1.01-1.02; $P < .001$). It is not helpful to have the upper bound of a CI be equivalent to the central estimate; a better alternative would be to report an OR per 10 years of age. This is simply achieved by creating a new variable equal to age divided by 10 to obtain an OR of 1.16 (95% CI, 1.10-1.22; $P < .001$) per 10-year difference in age. Alternatively, one could calculate the risks for two values of clinical relevance (eg, age 65 years vs 40 years) for a patient who had average values of other predictors.

Avoid Reporting Both Univariable and Multivariable Analyses Unless There Is a Good Reason

Comparison of univariable and multivariable models can be of interest when trying to understand mechanisms. For instance, if race is a predictor of outcome on univariable analysis, but not following adjustment for income and access to care, one might conclude that poor outcome in African-American subjects is explained by socioeconomic factors. However, univariable results should not routinely be reported when a multivariable analysis is presented.

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 ORs or hazard ratios. Such rankings are not meaningful for several reasons, including dependence on how variables are coded and the units of measurement. Furthermore, it is unclear how one should compare model coefficients when both categorical and continuous variables are included. Finally, the prevalence of a categorical predictor affects its importance: a predictor with an OR of 3.5 but a prevalence of 0.1% is less important than one with a 50% prevalence and an OR of 2.0. Finally, the prevalence of a categorical predictor affects its importance: a predictor with an OR of 3.5 but a prevalence of 0.1% is less important than one with a 50% prevalence and an OR of 2.0.

Be Cautious When Comparing Models Assessed on Different Datasets

A model applied to a dataset will usually show better discrimination in the dataset with higher variability. The reason for this finding is that it is more difficult to discriminate among patients who are more similar, so discrimination will not be as high. Thus, it is generally not helpful to compare models that were assessed in different datasets.

Correct for Overfit When Conducting 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 dataset 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 least absolute shrinkage and selection operator (LASSO) methods to choose which predictors should go into the model and then fits the coefficients, a typical cross-validation approach would be to: (1) split the data into 10 groups; (2) use LASSO to select predictors using the first nine groups; (3) fit coefficients using the first nine groups; (4) apply the model to the 10th group to obtain predicted probabilities; and (5) repeat steps 2 through 4 until all patients in the dataset have a predicted probability derived from a model fitted to a dataset that did not include that patient's data. It is a mistake to perform the LASSO method prior to the data split because LASSO is seeing all the data. If a variable selection procedure such as LASSO is being performed, it should be done within each cross-validation iteration.

Calibration for a Prediction Model Should Be Presented Graphically

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. Where a prespecified model is tested on an independent dataset, 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 on 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 dataset, representative of a given population.

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 thresholds and then report results in terms of clinical consequences. As an example, consider the use of an integrated proteomic classifier in identifying benign nodules in patients with a high prior probability of having lung cancer. The authors might conclude, for instance, that if the classifier were adopted, 40% fewer procedures would be performed on benign nodules, and 3% of malignant nodules would be misclassified.

Discussion, Interpretation, and Conclusions

Draw a Conclusion; Do Not Just Repeat the Results

Conclusion sections are often simply a restatement of the results. For instance, "a statistically significant relationship was found between BMI and disease outcome" is not a conclusion. Authors instead need to state implications for research and/or clinical practice, being careful to ensure these implications are supported by study results. 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.

Avoid Using Words Such as "May" or "Might"

That a study hypothesis "may" be true is the rationale for doing a study in the first place. Moreover, this will always be true except if a hypothesis is proven to be false, and it is difficult to prove a negative in science. In place of statements such as "novel drug X may reduce the incidence of hospitalizations for asthma," we recommend language such as "we have provided evidence that drug X has a clinically relevant effect on hospitalization. A randomized trial comparing X vs the current standard of care drug Y is indicated."

Avoid Pseudo-limitations Such as "Small Sample Size" and "Retrospective Analysis," and Consider Instead Sources of Potential Bias and the Mechanism for Their Effect on Findings

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 OR, 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 effect size of possible bias.

Discuss the Impacts of Missing Data and Patient Selection

Complete data are rarely obtained for all patients in a study. A typical paper might report, for instance, that of 200 patients, eight had data missing on important baseline variables and 34 did not complete the end-of-study questionnaire, leading to a final dataset 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 often helpful.

Conclusions

These guidelines are not intended to cover all medical statistics but rather the statistical approaches most commonly used in clinical research papers in thoracic medicine.

Acknowledgments

Financial/nonfinancial disclosures: None declared.

Role of sponsors: The sponsor had no role in the design of the study, the collection and analysis of the data, or the preparation of the manuscript.

Reference

1. Assel M, Sjoberg D, Elders A, et al. Guidelines for reporting of statistics for clinical research in urology. *BJU Int*. 2019;123(3):401-410.