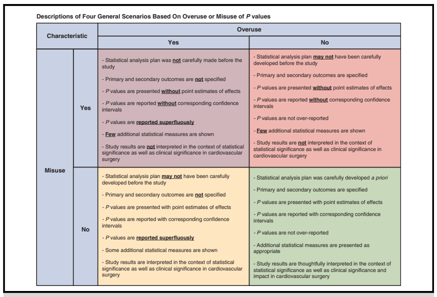
Guidelines for improving the use and presentation of *P* values

Steven J. Staffa, MS, and David Zurakowski, MS, PhD



Descriptions of Four General Scenarios Based On Overuse or Misuse of *P* values

Characteristic		Overuse	
		Yes	No
Misuse	Yes	- Statistical analysis plan was **not** carefully made before the study - Primary and secondary outcomes are **not** specified - *P* values are presented **without** point estimates of effects - *P* values are reported **without** corresponding confidence intervals - *P* values are **reported superfluously** - **Few** additional statistical measures are shown - Study results are **not** interpreted in the context of statistical significance as well as clinical significance in cardiovascular surgery	- Statistical analysis plan **may not** have been carefully developed before the study - Primary and secondary outcomes are specified - *P* values are presented **without** point estimates of effects - *P* values are reported **without** corresponding confidence intervals - *P* values are not over-reported - **Few** additional statistical measures are shown - Study results are **not** interpreted in the context of statistical significance as well as clinical significance in cardiovascular surgery
	No	- Statistical analysis plan **may not** have been carefully developed before the study - Primary and secondary outcomes are **not** specified - *P* values are presented with point estimates of effects - *P* values are reported with corresponding confidence intervals - *P* values are **reported superfluously** - Some additional statistical measures are shown - Study results are interpreted in the context of statistical significance as well as clinical significance in cardiovascular surgery	- Statistical analysis plan was carefully developed *a priori* - Primary and secondary outcomes are specified - *P* values are presented with point estimates of effects - *P* values are reported with corresponding confidence intervals - *P* values are not over-reported - Additional statistical measures are presented as appropriate - Study results are thoughtfully interpreted in the context of statistical significance as well as clinical significance and impact in cardiovascular surgery

Descriptions of 4 general scenarios based on overuse or misuse of *P* values. We have outlined 4 general scenarios of statistical design and reporting in cardiovascular surgery articles based on the overuse and misuse of *P* values. The *red box* describes the scenario where *P* values are overused and misused and considerable improvement is needed in the statistical reporting of *P* values. The *orange box* highlights the situation where *P* values have likely been misused. The *yellow box* contains reasons that are indicative of a scenario of *P* value overuse but not misuse. The *green box* represents the ideal scenario. These studies are the most impactful because of characteristics of robust and appropriate statistics without overuse or misuse of *P* values.

CENTRAL MESSAGE

Careful statistical analysis planning and focusing on study design and other statistical measures will help mitigate the misuse and overuse of *P* values.

This Invited Expert Opinion provides a perspective on the following paper: *N Engl J Med*. 2019;381(3):285-286. https://doi.org/10.1056/NEJMe1906559.

P values were originally developed for hypothesis testing; however, the use of the *P* value in cardiovascular and thoracic research has expanded as it has become the most commonly reported summary measure of statistical results. A *P* value measures whether or not the actual results of a study or trial are consistent with what would be expected by chance or if the results more likely indicate a real difference between the treatment groups or surgical approaches being studied. A small *P* value is often regarded as evidence that the surgeon can rule out the chance explanation for the observed differences between the study groups.

Recently, there has been discussion in the literature regarding *P* values in medical research as it pertains to reporting standards.[1,2] Although *P* values have value in providing a snapshot summary of the statistical evidence of group differences, treatment effects, or measures of association, there are times when *P* values have been misused or overused. The objective of this article is to provide a perspective regarding the use of *P* values in the *Journal* and to provide guidance in how to make the best use of *P* values particularly for common statistical study designs found in studies in the *Journal*. We provide a background regarding *P* values and guidance regarding the optimal use of *P* values for 3 specific common statistical study designs of manuscript submitted to the *Journal*.

BACKGROUND

The fundamental use of the *P* value is in the context of null hypothesis significance testing.[3] Under the null hypothesis (H_0) (in other words, assuming that the null hypothesis is true), the *P* value is an estimated probability of observing an effect as large or larger than the result calculated from the data. As an example, the null hypothesis may be that there is no difference in the rate of 30-day readmission between 2 surgical groups, A and B. According to the study data, surgical group A has a 30-day readmission rate that is 10% larger than that of surgical group B. The *P* value, in this example, is the probability that the difference in 30-day readmission rates between surgical groups A and B is 10% or more, assuming that the null hypothesis is true (that there is no difference between groups). If the *P* value is very small, then there is a very small probability of observing a result as extreme or more extreme than the one observed, and therefore the result is unlikely to be due to chance, and the result is declared statistically significant (usually the threshold of

From the Departments of [a]Surgery and [b]Anesthesiology, Critical Care and Pain Medicine, Boston Children's Hospital, Harvard Medical School, Boston, Mass.

Received for publication Nov 26, 2019; revisions received March 19, 2020; accepted for publication March 24, 2020; available ahead of print April 30, 2020.

Address for reprints: David Zurakowski, MS, PhD, Director of Biostatistics, Department of Surgery, Boston Children's Hospital, 300 Longwood Ave, Boston, MA 02115 (E-mail: David.Zurakowski@childrens.harvard.edu).

J Thorac Cardiovasc Surg 2021;161:1367-72
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.03.156

$P < .05$ is imposed). Very often, the goal in the mindset of a researcher is to be able to claim that a result is "statistically significant."[4] In some severe situations, investigators may be searching for statistically significant results in what is referred to as "*P*-Hacking."[5]

In modern-day medical research, investigators do not always formally write null and alternative hypotheses for all statistical comparisons being performed. Instead, researchers most often consider primary and secondary research questions. In the example of 30-day readmission rates discussed, the investigator may not consider a null hypothesis that the rates are equal in the 2 groups and an alternative hypothesis that the 30-day readmission rates are not equal in the 2 groups. Instead, the investigator would state the research question as, Is there a difference in the 30-day readmission rates between surgical groups A and B? The investigator may proceed to calculate an estimated treatment effect in terms of the difference in 30-day readmission rates and then may also compute a corresponding 95% confidence interval and *P* value. However, the investigator might not view that *P* value with the mindset that it was computed under the assumption that the "null hypothesis" is true. Instead, the investigator will interpret it to represent whether or not the observed difference between the 2 groups is statistically significant. Although this interpretation is not faulty, it is worth noting that often investigators lose sight of the fact that using the hypothesis testing framework involves the assumption that the null hypothesis is true (that there is no difference between the study groups).

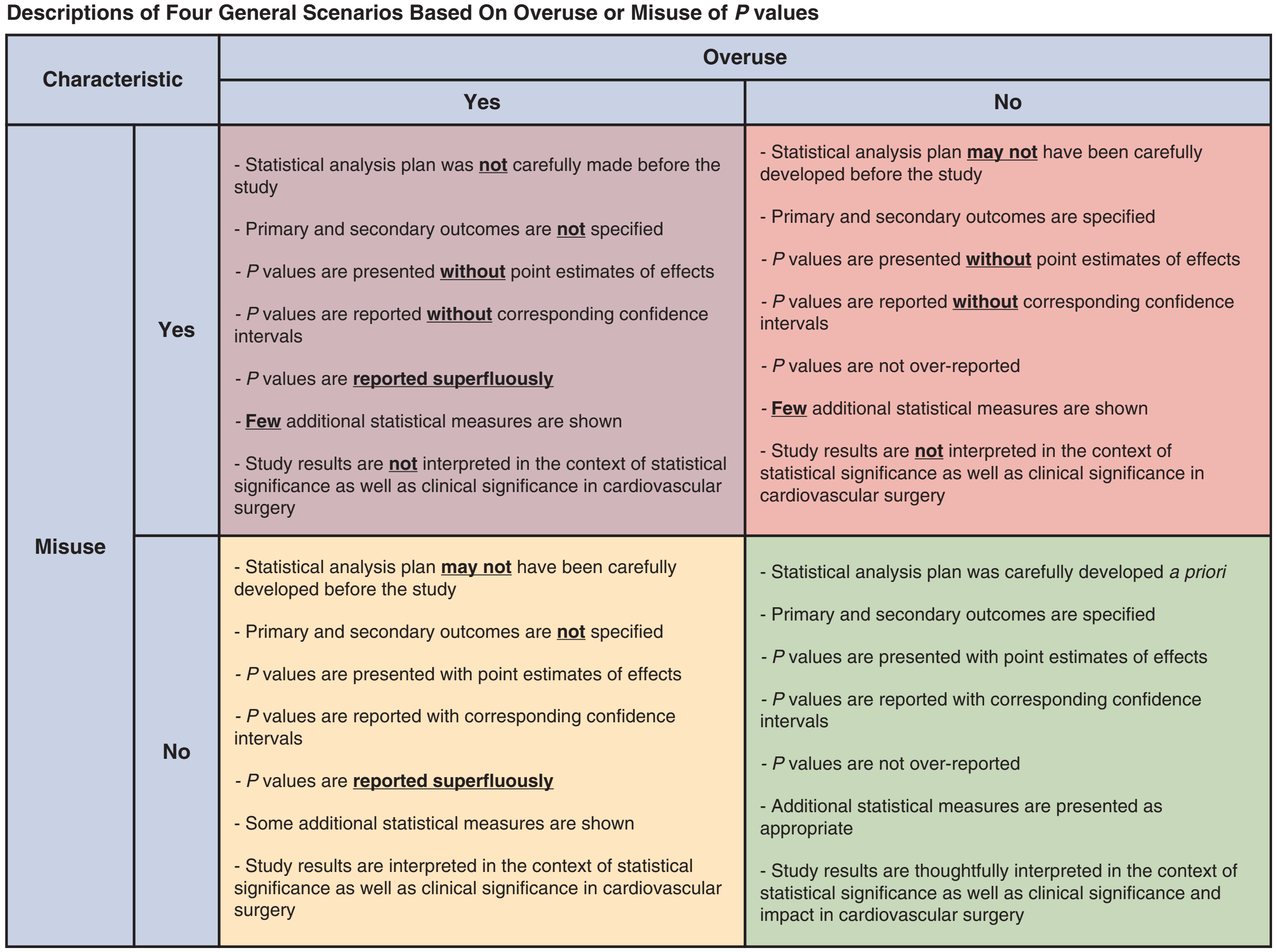
GUIDANCE REGARDING STUDY DESIGN

The issue of overuse of *P* values potentially can be addressed with careful statistical and design planning before performing the study. Thorough and thoughtful statistical study design in specifying primary and secondary study objectives along with determining the statistical analyses that will be performed for planned comparisons in advance of a study will help authors to judiciously report *P* values. In statistical analysis planning with preliminary discussions with a statistician, the specification of primary outcomes and planned comparison will give the authors the knowledge of where *P* values will be reported in their final analyses. This will enable investigators, in collaboration with a statistician, to astutely preserve the study wide alpha level (ie, false-positive error rate or type I error rate) while also purposefully reporting *P* values and taking multiplicity into account (ie, multiple testing). It should be noted that there are several approaches to account and adjust for multiplicity,[6,7] with the most common method being the Bonferroni correction[8]; however, further detail on multiple testing is beyond the scope of this article. In this framework, *P* values may only be reported for the a priori planned comparisons and analyses of primary study outcomes. This will simultaneously make *P* values more meaningful with fewer reported, because each will be used for summarizing the significance of the key statistical results of a study, likely pertaining to the study's Figure 1.

ACCOMPANYING THE *P* VALUE

Along with purposeful and diligent study designing to help mitigate overuse, the *P* value should be presented in a comprehensive fashion to improve proper use of the *P* value. Sometimes, investigators may rely heavily on *P* values because they are not aware or comfortable in interpreting alternative and additional statistics that may also be suitable. In isolation, the *P* value does not provide a comprehensive summary of the results because it does not show the magnitude or directionality of treatment effects or measures of association. The importance of reporting magnitude and directionality of treatment effects must be emphasized because this allows interpretations of the results to be made in terms of both statistical and clinical significance. Failing to include this can be dangerous and misleading to the reader. Larger sample sizes do not protect the investigator when reporting *P* values. Rather, in situations with large sample sizes, particularly in the situation of a large number of events for binary outcomes, small effects may be declared as statistically significant even if the effect itself it not clinically important. Therefore, it is advisable to accompany the *P* value with the effect estimate (ie, odds ratio, risk ratio, difference in means or proportions) with a corresponding confidence interval.[9] The point estimate and confidence interval are more informative and interpretable than the *P* value in isolation. Going beyond the *P* value and using other statistical metrics are important for determining if results are clinically meaningful. Readers may interpret a *P* value less than .05 for determining statistical significance; however, because *P* values are heavily dependent on sample size, a statistically significant result with a *P* value less than .05 may not correspond to a clinically meaningful difference or effect. Furthermore, the threshold of a *P* value less than .05 has been seemingly arbitrarily agreed upon to denote statistical significance, and investigators will provide starkly different interpretations of their analysis results for the scenario with a *P* value equal to .04 as opposed to a *P* value equal to .06. What is considered statistically significant in terms of a *P* value versus what is regarded as clinically important to the surgeon are separate questions and must be evaluated separately when interpreting the results of any research study. Complementing all *P* values with estimated treatment effects and measurements of variability like confidence intervals should be a mandatory requirement for all submitted manuscripts because this will provide improved transparency and more comprehensive presentation of the statistical results.

Guidance regarding reporting of *P* values is presented in Figure 2, and the characteristics regarding misuse and



Descriptions of Four General Scenarios Based On Overuse or Misuse of *P* values

Characteristic		Overuse	
		Yes	No
Misuse	Yes	- Statistical analysis plan was **not** carefully made before the study - Primary and secondary outcomes are **not** specified - *P* values are presented **without** point estimates of effects - *P* values are reported **without** corresponding confidence intervals - *P* values are **reported superfluously** - **Few** additional statistical measures are shown - Study results are **not** interpreted in the context of statistical significance as well as clinical significance in cardiovascular surgery	- Statistical analysis plan **may not** have been carefully developed before the study - Primary and secondary outcomes are specified - *P* values are presented **without** point estimates of effects - *P* values are reported **without** corresponding confidence intervals - *P* values are not over-reported - **Few** additional statistical measures are shown - Study results are **not** interpreted in the context of statistical significance as well as clinical significance in cardiovascular surgery
	No	- Statistical analysis plan **may not** have been carefully developed before the study - Primary and secondary outcomes are **not** specified - *P* values are presented with point estimates of effects - *P* values are reported with corresponding confidence intervals - *P* values are **reported superfluously** - Some additional statistical measures are shown - Study results are interpreted in the context of statistical significance as well as clinical significance in cardiovascular surgery	- Statistical analysis plan was carefully developed *a priori* - Primary and secondary outcomes are specified - *P* values are presented with point estimates of effects - *P* values are reported with corresponding confidence intervals - *P* values are not over-reported - Additional statistical measures are presented as appropriate - Study results are thoughtfully interpreted in the context of statistical significance as well as clinical significance and impact in cardiovascular surgery

FIGURE 1. Descriptions of 4 general scenarios based on overuse or misuse of *P* values. We have outlined 4 general scenarios of statistical design and reporting in cardiovascular surgery articles based on the overuse and misuse of *P* values. The *red box* describes the scenario where *P* values are overused and misused and considerable improvement is needed in the statistical reporting of *P* values. The *orange box* highlights the situation where *P* values have likely been misused. The *yellow box* contains reasons that are indicative of a scenario of *P* value overuse but not misuse. The *green box* represents the ideal scenario. These studies are the most impactful because of characteristics of robust and appropriate statistics without overuse or misuse of *P* values.

overuse of *P* values are outlined in Figure 1. The strategies in Figure 2 are aimed to achieve the optimal presentation of *P* values and a more impactful manuscript, as represented by the best case scenario in the green box in Figure 1.

COMMON CARDIOVASCULAR SURGERY EXAMPLES

Next, we will discuss *P* values in the context of 3 common statistical techniques that cardiovascular surgeons are acquainted with: time-to-event survival analysis, logistic regression, and propensity score matching.

Example 1: Time-to-Event Analysis

Survival analysis using Kaplan–Meier curves and Cox proportional hazards regression modeling is a cornerstone of cardiovascular and thoracic surgery clinical research studies.[10,11] Many outcome variables in our specialty are time-to-event end points, with examples including time to mortality, reoperation, readmission after surgery, or freedom from heart transplant. Traditionally, Kaplan–Meier curves are used to present the comparison of outcomes between groups, with the statistical comparison accomplished using the log-rank test.[12] The log-rank test *P* value alone does not adequately describe the study findings in terms of which of the comparison groups demonstrated better survival or freedom from the event of interest, nor does it describe the variability of those estimates. The log-rank test *P* value provides a summary of the statistical significance in comparing Kaplan–Meier curves across the follow-up time course, and although relevant it should not be presented in isolation. It should accompany the figure with the Kaplan–Meier curves with confidence bands and numbers at risk.[13] The investigator should also consider presenting in the text or figure legend

Guidance on Reporting *P* values Appropriately and Integrating Other Statistical Measures

1. Specify the primary outcomes and carefully plan the statistical analyses before conducting the study (*a priori*).

2. Discuss with a statistician before the study to determine the analysis plan including *P* values and complementary statistics that are suitable.

3. Reserve *P* values for reporting the results of primary analyses or primary comparisons of interest.

4. Communicate statistical significance, while paying close attention to clinical significance, and make appropriate *P* value adjustments to avoid Type I error (false-positive findings) due to multiple comparisons.

5. Report *P* values along with effect estimates (e.g. odds ratios, hazard ratios, differences in proportions) and measurements of variability such as confidence intervals.

6. Integrate alternative statistics where appropriate, such as the number at risk for Kaplan-Meier curves and the standardized mean difference in propensity score matching studies.

7. Interpret the statistical significance as well as the clinical significance of the results.

FIGURE 2. Guidance on reporting *P* values while integrating other statistical measures. There are several general recommendations and strategies that surgeons should integrate to help improve the quality of their manuscripts in terms of reporting *P* values. These guidelines will help with mitigating the misuse and overuse of *P* values in submissions to the *Journal*. Good balance of presentation of *P* values and other statistical measures is crucial.

the Kaplan–Meier estimates at specific time points (eg, 1 year, 3 years, and 5 years postsurgery) with corresponding confidence intervals, particularly if confidence bands are visually cluttering. The log-rank test does not compare estimates at specific individual time point between groups, and therefore presenting numeric estimates with confidence intervals at time points of interest is useful. Furthermore, the log-rank test *P* value ideally should be presented only for the analysis of the primary outcome as specified in the study design and statistical analysis plan. These strategies can help to reduce the overuse of *P* values in time-to-event analyses.

Cox multivariable regression analysis can be used to determine independent risk factors associated with the risk of the event of interest in a time-to-event analysis. Hazard ratios are obtained as point estimates for the measure of increased or decreased risk of the outcome associated with each predictor variable. Each estimated hazard ratio will have a corresponding confidence interval and *P* value. Here, the authors should save the use of *P* values in a manuscript or report for the adjusted hazard ratio for the primary exposure of interest (eg, the hazard ratio associated with surgical group). Careful study design and statistical analysis planning regarding the primary exposure (predictor variable) and primary end point will help the researcher decide where to present *P* values in a Cox regression analysis. Furthermore, reporting *P* values from Cox multivariable regression analysis without the corresponding hazard ratio and confidence interval should be avoided. *P* values should be reported with other statistics related to treatment evidence of risk factors.

Specific Example of Misuse of *P* Values in Time-to-Event Analysis

To illustrate a time-to-event analysis scenario with poor statistical reporting and improvement needed regarding misuse of *P* values, consider the Kaplan–Meier curves shown in Figure 3. In this hypothetical analysis, the investigator is trying to determine survival over the course of 1 year after the Fontan operation, while comparing preterm versus full-term patients. In Figure 3, a *P* value is reported to compare survival between the 2 groups. Although *P* values are not overused in this scenario, this is an example of misuse of *P* values because the *P* value is not presented with the necessary accompanying statistics. There are no confidence bands around the Kaplan–Meier curves, which are important for showing the level of precision of the Kaplan–Meier survival estimates. There are no numbers at risk shown in Figure 3 to depict the sample sizes of the number of patients alive and followed over the course of follow-up. In this analysis, the *P* value is presumed as being calculated using the log-rank test; however, the method for calculating the *P* values is not stated in Figure 3. This specific example of *P* value misuse can be improved by providing additional statistics and including estimates of variability using confidence bands.

Example 2: Logistic Regression

Logistic regression analysis may be used for determining the association between a set of predictor variables and a dichotomous outcome variable. In cardiovascular surgery research, binary outcome variables are common.[14,15] Examples include 30-day mortality and postoperative complications (without incorporating the time to event if perhaps a time window is defined or time is not relevant). Similar to reporting *P* values in the context of Cox regression modeling, *P* values in the logistic regression framework should be reported with accompanying confidence intervals for the point estimates. Univariate logistic regression may be used as a screening for potentially important predictor variables, and that decision can be made using *P* values. *P* values may also be used in a stepwise logistic regression

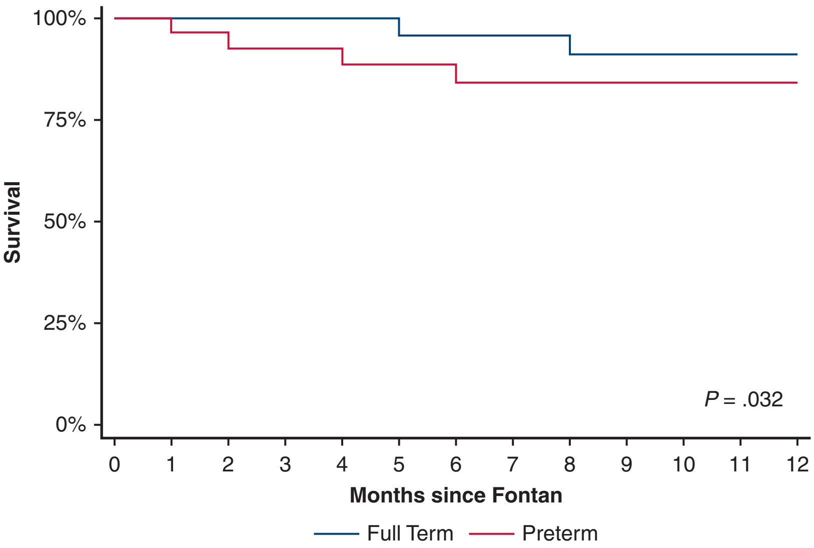


FIGURE 3. Misuse of *P* values in Kaplan–Meier curve analysis. The *P* value is not presented with the necessary accompanying statistics. There are no confidence bands around the Kaplan–Meier curves and no numbers at risk shown in the figure to depict the sample sizes of the number of patients alive and followed over the course of follow-up period. The method for calculating the *P* values is not stated on the figure itself. This specific example of *P* value misuse can be improved by providing additional statistics and including estimates of variability using confidence bands.

model building, such as backwards elimination for covariate removal. However, the emphasis of the presentation of study results should be placed on the multivariable logistic regression model and the *P* values pertaining to the primary exposure variable in this model. In addition to presenting *P* values, the c-index (area under the curve) is a useful metric for evaluating the performance of a multivariable logistic regression model.

Specific Example of Misuse and Overuse of *P* Values in Logistic Regression Analysis

The following text is a hypothetical example misuse and overuse of *P* values in the reporting of results of multivariable logistic regression:

> *"Multivariable logistic regression identified the following significant risk factors for reoperation after arterial switch operation: gestational age (*P *= .002), sex (*P *= .04), and crossclamp time (*P *<.001)."*

In this hypothetical scenario, *P* values are overused and misused for several reasons. The *P* values are reported in isolation, without effect estimator or confidence intervals, which limits the ability of the reader to understand the clinical meaningfulness of the results. It is unclear if gestational age and crossclamp duration are treated as continuous predictor variables or if a dichotomization was performed. Because there are no effect estimates reported, it is unclear whether larger or smaller values for gestational age and crossclamp time are associated with increased or decreased odds of reoperation. Also, it cannot be determined from what is reported whether men or women are at higher risk of reoperation. If the adjusted odds ratio for male sex (vs female) was 1.04 with a 95% confidence interval 1.02 to 1.07, then it could be argued that this statistically significant result with *P* less than .05 is not clinically significant. In this example, *P* values are reported superfluously, but without more information such as effect estimates with corresponding confidence intervals, the results cannot be used to generate any meaningful clinical insight.

Example 3: Propensity Score Matching

Propensity score matching is used in observational (nonrandomized) studies to balance treatment groups on a set of variables to obtain a more valid comparison.[16,17] The comparison of the 2 groups may involve any type of outcome data, such as time-to-event outcomes, continuous outcomes, or binary outcomes. Groups should be compared on the matching variables before and after propensity score matching is performed, and investigators often use *P* values as a measure of determining whether significance imbalances occur pre- and postmatching. However, the matched sample will necessarily have a smaller sample size than the unmatched sample. Because *P* values are heavily dependent on sample size, in propensity score matching studies it is advisable to use standardized mean differences to determine balance between the 2 groups for each matching covariate. Standardized mean differences with an absolute value less than 0.10 are often taken as indicating good balance between the 2 groups achieved by propensity score matching.[18] The standardized mean difference can be calculated for continuous or binary variables.[18,19]

P values should be reserved for the statistical analysis based on postmatching data. As an example, once balance has been demonstrated on the confounders and matching variable between 2 surgical groups implying that groups are comparable, the log-rank test *P* value can be presented along with the Kaplan–Meier curves comparing survival between 2 surgical groups using the matched dataset.

CONCLUSIONS

P values are misused and overused in surgical research studies as well as all other areas of research. Used appropriately, *P* values can provide information regarding the statistical significance of a given treatment effect. However, they are often misused and even used in isolation, which may convey significant results that may not be clinically significant as well. Careful and thoughtful study design and statistical analysis planning can help investigators to reserve *P* values for the most important analysis results. Although the appropriate reporting and use of *P* values depend on the specific study design being considered, it is important to include effect estimates with measures of variability (eg, confidence intervals). Presenting the magnitude and directionality of effect estimates is crucial for interpreting the results in terms of statistical significance and clinical significance. Improvements regarding use of *P* values in manuscripts submitted to the *Journal* will lead to more robust statistical reporting and more impactful published studies in the *Journal*.

Conflict of Interest Statement

The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

References

1. Harrington D, D'Agostino RB, Gatsonis C, Hogan JW, Hunter DJ, Normand ST, et al. New guidelines for statistical reporting in the Journal. *N Engl J Med*. 2019;381:285-6.
2. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567:305-7.
3. Lehmann EL, Romano JP. *Testing Statistical Hypotheses*. 3rd ed. New York: Springer; 2005.
4. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *J R Stat Soc*. 1988;151:419-63.
5. Head ML, Holman L, Lanfear R, Rahn AT, Jennions MD. The extent and consequences of P-hacking in science. *PLoS Biol*. 2015;13:e1002106.
6. Althouse AD. Adjust for multiple comparisons? It's not that simple. *Ann Thorac Surg*. 2016;101:1644-5.
7. Bender R, Lange S. Adjusting for multiple testing—when and how? *J Clin Epidemiol*. 2001;54:343-9.
8. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995;310:170.
9. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting P values in the biomedical literature, 1990-2015. *JAMA*. 2016;315:1141-8.
10. Muralidaran A, Detterbeck FC, Boffa DJ, Wang Z, Kim AW. Long-term survival after lung resection for non-small cell lung cancer with circulatory bypass: a systematic review. *J Thorac Cardiovasc Surg*. 2011;142:1137-42.
11. Akins CW, Miller DC, Turina MI, Kouchoukos NT, Blackstone EH, Grunkemeier GL, et al. Guidelines for reporting mortality and morbidity after cardiac valve interventions. *J Thorac Cardiovasc Surg*. 2008;135:732-8.
12. Wormuth DW. Actuarial and Kaplan-Meier survival analysis: there is a difference. *J Thorac Cardiovasc Surg*. 1999;118:973.
13. Hosmer DW, Lemeshow S. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: John Wiley & Sons; 1999.
14. Blackstone EH. Sufficient data. *J Thorac Cardiovasc Surg*. 2016;152:1235-6.
15. Gandhi R, Almond C, Singh TP, Gauvreau K, Piercey G, Thiagarajan RR. Factors associated with in-hospital mortality in infants undergoing heart transplantation in the United States. *J Thorac Cardiovasc Surg*. 2011;141:531-6.
16. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399-424.
17. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
18. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput*. 2009;38:1228-34.
19. Flury BK, Riedwyl H. Standard distance in univariate and multivariate analysis. *Am Stat*. 1986;40:249-51.