Journal of Pediatric Surgery 55 (2020) 1699–1705

Contents lists available at ScienceDirect

Journal of Pediatric Surgery

journal homepage: www.elsevier.com/locate/jpedsurg

Strategies in adjusting for multiple comparisons: A primer for pediatric surgeons☆

Steven J. Staffa *, David Zurakowski

Department of Surgery, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Article history:
Received 7 August 2019
Received in revised form 16 December 2019
Accepted 4 January 2020

Key words:
Multiple comparisons
Type I error
Multiplicity
P value
Study design
Bonferroni

ABSTRACT

Background/Purpose: In pediatric surgery research, the issue of multiple comparisons commonly arises when there are multiple patient or experimental groups being compared two at a time on an outcome of interest. Performing multiple statistical comparisons increases the likelihood of finding a false positive result when there truly are no statistically significant group differences (falsely rejecting the null hypothesis when it is true). In order to control for the risk of false positive results, there are several statistical approaches that surgeons should consider in collaboration with a biostatistician when performing a study that is prone to the issue of false discovery related to multiple comparisons. It is becoming increasingly more common for high impact journals to require authors to carefully consider multiplicity in their studies. Therefore, the objective of this primer is to provide surgeons with a useful guide and recommendations on how to go about taking multiple comparisons into account to keep false positive results at an acceptable level.

Methods: We provide background on the issue of multiple comparisons and risk of type I error and guidance on statistical approaches (i.e. multiple comparisons procedures) that can be implemented to control the type I false positive error rate based on the statistical analysis plan. These include, but are not limited to, the Bonferroni correction, the False Discovery Rate (FDR) approach, Tukey's procedure, Scheffé's procedure, Holm's procedure, and Dunnett's procedure.

Results: We present the results of the various approaches following one-way analysis of the variance (ANOVA) using a hypothetical surgical research example of the comparison between three experimental groups of rats on skin defect coverage for experimental spina bifida: the TRASCET group, sham control, and saline control. The ultimate decision in accounting for multiple comparisons is situation-dependent and surgeons should work with their statistical colleagues to ensure the best approach for controlling the type I error rate and interpreting the evidence when making multiple inferences and comparisons.

Conclusions: The risk of rejecting the null hypothesis increases when multiple hypotheses are tested using the same data. Surgeons should be aware of the available approaches and considerations to take into account multiplicity in the statistical plan or protocol of their clinical and basic science research studies. This strategy will improve their study design and ensure the most appropriate analysis of their data. Not adjusting for multiple comparisons can lead to misleading presentation of evidence to the surgical research community because of exaggerating treatment differences or effects.

Type of study: Review article.

Level of evidence: N/A

The issues of multiple comparisons and multiple outcomes are widespread in surgical research. For instance, an investigator may be interested in comparing multiple groups of animals on a specific outcome or two patients groups may be compared on multiple primary outcome variables. Examples in surgical research include the comparison of multiple surgical groups, such as laparoscopic, open, robotic and transanal mesorectal excision for rectal cancer [1], the comparison of multiple animal groups, such as treatment, active control and sham control, or the comparison of two study groups on multiple primary outcome variables such as mortality, readmission, and length of stay.

Traditionally, a two-tailed significance level of 0.05 (5% alpha level) is specified and used as the criterion for determining statistical significance for each hypothesis test or comparison. For a single statistical test, this assumption means that conditional on there being no effect (e.g. no treatment effect, no group difference, or no association), we are allowing a 5% risk of getting a false positive result. In other words,

☆ Declarations of interest: none.
Financial disclosures: none.

* Corresponding author at: Department of Surgery, Boston Children's Hospital, 300 Longwood Avenue, Boston, MA 02115. Tel.: +1 617 355 8254.
E-mail address: Steven.Staffa@childrens.harvard.edu (S.J. Staffa).

https://doi.org/10.1016/j.jpedsurg.2020.01.003
0022-3468/

there is a 5% chance that we will conclude there is a statistically significant effect when in reality there truly is not (i.e. that we will incorrectly reject the null hypothesis when in fact it is true). When more than one independent comparison is made, the risk of obtaining at least one false positive result is increased [2,3]. Under the condition that there truly is no statistically significant effect (i.e. assuming the null hypothesis is true), the chance that each comparison or test is not statistically significant is 95% (100% - alpha). Therefore, assuming that in reality there is no significant effect, the probability that 2 independent comparisons or hypothesis tests are both not statistically significant is 95% × 95%, or 90.25% [4]. The risk of at least one false positive result in these 2 independent comparisons is 100% − 90.25% = 9.75%. As the number of comparisons or tests increases, so does the risk of a false positive result which may lead to a false conclusion. The risk of obtaining at least one false positive result when C independent comparisons or tests are made follows the formula $1 - 0.95^C$. Fig. 1 shows the relationship between the number of hypothesis tests being performed in a study and the risk of yielding at least one false positive result.

Statistical strategies to account for multiplicity are intended to preserve the family-wise or study-wide type I error rate in a reasonable way. The objective of this review article is to provide surgeons a framework and some direction on how to take multiple comparisons into account when planning their clinical or basic science research studies. We present the key concepts and approaches for handling the problem of multiple comparisons and multiple outcomes or endpoints, while explaining the methods in the context of pediatric surgery research. To ensure appropriate interpretation of statistical analyses, a provision needs to be made for adjusting the significance level or *P* value criterion for multiple comparisons and multiple testing to minimize false positive results.

1. Multiple comparisons versus multiple outcome testing

Multiple comparisons and multiple outcome testing are similar issues that refer to the inflated risk of falsely rejecting the null hypothesis in the situation where no statistically significant differences truly exist. However, multiple comparisons refer to the circumstance where there are 3 or more separate groups being compared (two at a time) on a given outcome variable. Multiple comparisons methods following analysis of the variance (ANOVA) have been developed and are discussed further below. On the other hand, multiple testing refers to the situation where multiple related statistical tests are being performed in a research study, but they do not necessarily have to be between multiple groups on a particular outcome variable. Rather, the situation could arise where there are only two groups being compared in a surgical research study, for instance patients undergoing two different surgical approaches such as laparoscopic or open, and these two groups are being compared on multiple outcome variables, such as hospital length of stay, mortality, and readmission. Multiple testing is at play in this situation since 3 statistical tests will be performed to compare the two patient groups on each of these 3 outcome variables. Similar to the setting of multiple comparisons, considerations for an adjustment to control for the chance of false positive results should be made when performing a study involving the tests of multiple outcome variables.

1.1. Families of comparisons

Broadly, a family of comparisons is the set of related statistical comparisons being performed. This may be all or a subset of hypothesis tests performed in a given research study or experiment. When multiple statistical comparisons are being made using the same data, it is important to consider controlling the family-wise error rate (the error rate for a set of related hypothesis tests) or experiment wise or study-wide error rate (the error rate for an entire study or experiment). The 5% alpha level should not be used repeatedly for each separate hypothesis test, but rather the entire family of tests should preserve the 5% risk of false positive results (type I error). When an alpha level of 0.05 is used for each comparison, there will be greater than a 5% risk of observing at least one false positive result when multiple comparisons are being performed. The methods presented below are designed to protect against the chance of false positive results (when no effect truly exists) by preserving the experiment-wise error rate.

1.2. The Bonferroni adjustment

The Bonferroni correction is the most commonly used approach to control for the risk of false positive results. The Bonferroni method entails that a more conservative individual alpha level be used for each comparison in a family of tests. This is intended to preserve the study-wide error rate of 5% and in turn control the likelihood of false positive results when in reality there are no real group differences. Rather than using the traditional 5% alpha level, the Bonferroni adjustment requires the alpha level for each individual comparison be equal to 5% divided by the total number of those comparisons [5]. If C comparisons or tests are planned, then each comparison uses the alpha/C threshold to determine statistical significance. In a given study, the maximum number of possible group comparisons depends on the number of groups. If there are 3 groups, there are up to 3 possible pairwise comparisons. A "pairwise" comparison means the comparison of two independent study groups. If there are 4 groups, the number of possible comparisons rises to 6, and if there are 5 groups, the maximum number of comparisons

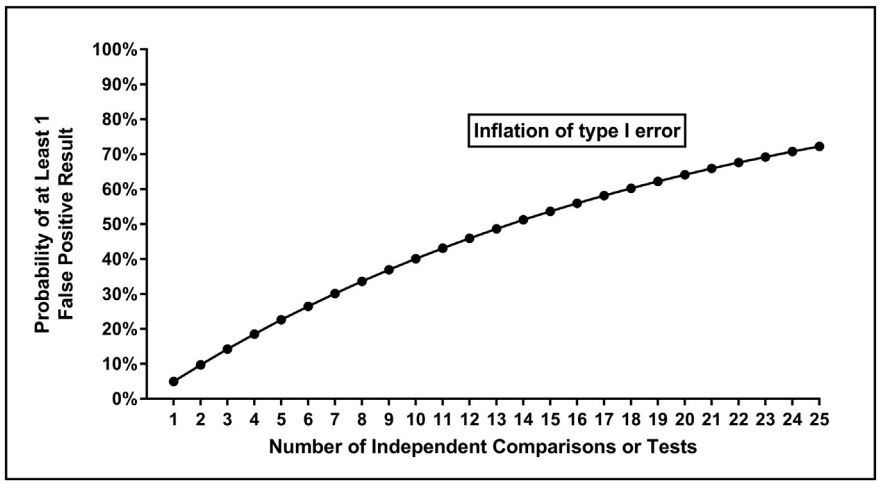

Fig. 1. Curve showing relationship between number of comparisons and risk of at least one false positive result. As the number of hypothesis tests performed increases, the chance of observing one or more false positive results increases, which is why multiple comparisons procedures need to be implemented to control for this risk.

jumps up to 10. Generally, if k is the number of groups, then the possible number of pairwise comparisons follows the formula $k(k-1)/2$. Fig. 2 illustrates the relationship between the number of groups and the maximum possible number of comparisons that can be performed.

As an example of a Bonferroni adjustment, consider a surgeon performing a comparison between 4 groups of mice being injected (experimental drug 1 treatment, experimental drug 2 treatment, saline-injected control, and injection-free control) on the outcome of overall survival. The surgeon is interested in comparing all combinations of groups, meaning that there would be 6 statistical tests performed. The Bonferroni approach would adjust the 0.05 alpha level or P value to $0.05/6 = 0.0083$. Then, in the comparison of mortality rates between each pair of treatment groups, the author would declare P values less than 0.0083 as statistically significant. In this scenario, if the Bonferroni correction was not applied, the risk of reporting at least one false positive result among the six group comparisons when there are truly no statistically significant associations would be $1 - 0.95^6 = 0.265$. Although they are not independent, it must be noted that this calculation is performed assuming independence among the six statistical tests.

However, if the surgeon has planned prior to the study to compare each of the other 3 groups to the injection-free control group, then only 3 comparisons would be made (drug 1 treatment vs injection-free control, drug 2 treatment vs injection-free control, and saline-control vs injection-free control). In this setting, the Bonferroni procedure would suggest that a P value or significance level of $0.05/3 = 0.017$ be considered "significant" for each of these three tests performed. The Bonferroni correction is flexible in that it can be accomplished regardless of the distribution of the outcome of interest and the analytic procedure employed, and for any number of specified group comparisons or tests. However, it will be overly conservative in the setting of a large number of hypothesis tests being performed and thus can decrease statistical power.

1.3. The false discovery rate (FDR)

The false discovery rate (FDR) approach sets a fraction of results that are expected to be false positives assuming there truly is no association. This is accomplished by setting a P value threshold, similarly to a significance level; however, the results with a P value under this threshold are called "discoveries" instead of "significant results". The FDR may be held as a constant threshold for all comparisons, or it may be adjusted using the method developed by Benjamini and Hochberg [6]. The FDR may be a suitable approach in studies with a very large number of comparisons or tests being performed, such as genomic microarray studies. While the FDR may be a useful approach to take into account multiple comparisons in certain situations when evaluating the notion of "discoveries", the remainder of our primer focuses on specific multiple comparisons procedures to control the risk of false positive results in the context of statistical significance.

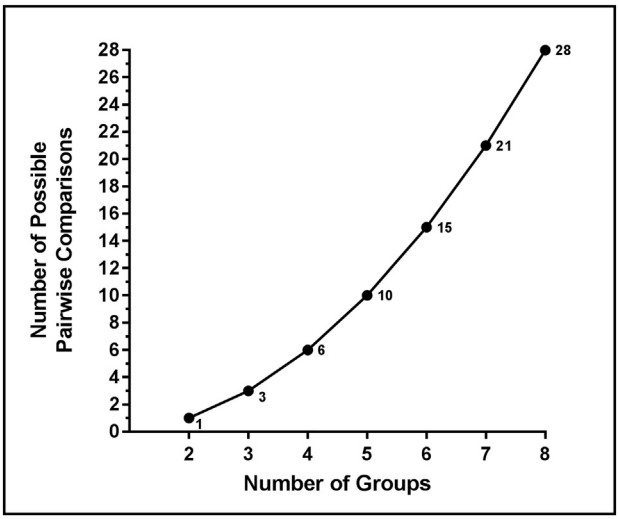

Fig. 2. Curve showing relationship between number of groups and possible number of group comparisons that can be made. For k unique groups, the maximum possible number of group comparisons follow the formula $k(k-1)/2$. Investigators should specify the number of planned comparisons in their study design.

2. Multiple comparisons procedures following ANOVA

One-way ANOVA is used to test the global hypothesis that the means of a continuous outcome variable are equal across all of the groups (or levels) being analyzed. It is therefore an overall test and the results do not tell the investigator which specific groups differ from one another. If the overall F-test yields a statistically significant P value, the authors may be inclined to test groups in a pairwise fashion (two at a time). The comparisons of interest may be determined a priori or they may be determined following results of ANOVA as a global hypothesis test in a *post-hoc* fashion. A subset or all possible pairwise comparisons may be performed [7].

The following multiple comparisons procedures are designed to preserve the experiment-wise type I error rate at 5%. However, since these tests follow the parametric one-way ANOVA procedure, they rely on the assumption that the data follow a normal bell-shaped (Gaussian) distribution. These tests have been shown to be robust to deviation from normality, especially when sample sizes are equal across the groups [8]. In the presence of heavily skewed data, these tests may not be appropriate; however, when normality is verified, these tests are useful for taking into account multiple comparisons.

As we discuss the advantages of each of the following multiple comparisons procedures below, we will present the results of each approach in controlling the family-wise error rate using a pediatric surgical research example. In the example, a surgeon is evaluating the efficacy of a transamniotic stem cell therapy (TRASCET), in particular amniotic fluid-derived mesenchymal stem cells (afMSCs), in promoting or inducing skin defect coverage in a rat model of experimental spina bidifa [9]. Please note that the results presented are hypothetical and are intended to illustrate a multiple comparisons scenario. In our hypothetical results, defect coverage is a continuous variable ranging from 0% to 100% for each animal. There are three experimental groups each with 12 rats in each: the TRASCET group, sham control, and saline control. We will assume that among the rats in the TRASCET group, the mean defect coverage was 0.70 (standard deviation (SD) = 0.20), in the sham control group the mean defect coverage was 0.50 (SD = 0.22), and in the saline control group the mean defect coverage was 0.40 (SD = 0.17). Table 1 presents the results on defect coverage of the 3 pairwise group comparisons when applying each of the approaches presented below. Fig. 3 summarizes the guidelines and recommendations regarding advantages of which approach to use to adjust for multiple comparisons. A decision has to be made as to how conservative the surgeon wants to be at protecting against type I error since there is a tradeoff as this may lead to more false negative results with an overly conservative adjustment.

2.1. Tukey's procedure

Tukey's procedure, also known as Tukey's honestly significant difference (HSD), is a methodology to perform all pairwise comparisons of means between groups [10]. This method is developed for the situation of a balanced design with equal sample sizes per group, and its results will include adjusted 95% confidence intervals for the difference in means along with adjusted P values. In the situation where there are unequal sample sizes per group, it has been shown that the Tukey–Kramer procedure can be applied as a modification of Tukey's procedure [11,12]. In our surgical example, Tukey's procedure determines a statistically significant difference comparing TRASCET vs saline control (difference in means = 0.30; 95% CI: 0.10, 0.49; $P = 0.002$), and TRASCET vs

Table 1
Comparison of multiple comparison procedures following ANOVA in the surgical research example.

Procedure	Group Comparison	Difference in Mean Defect Coverage	95% CI	*P* value
Tukey	TRASCET vs Saline Control	0.30	(0.10, 0.49)	**0.002***
	TRASCET vs Sham Control	0.20	(0.01, 0.39)	**0.047***
	Sham Control vs Saline Control	0.10	(−0.09, 0.29)	0.438
Scheffé	TRASCET vs Saline Control	0.30	(0.09, 0.51)	**0.003***
	TRASCET vs Sham Control	0.20	(−0.01, 0.41)	0.059
	Sham Control vs Saline Control	0.10	(−0.11, 0.31)	0.471
Bonferroni	TRASCET vs Saline Control	0.30	(0.09, 0.50)	**0.002***
	TRASCET vs Sham Control	0.20	(−0.01, 0.40)	0.055
	Sham Control vs Saline Control	0.10	(−0.10, 0.30)	0.671
Holm	TRASCET vs Saline Control	0.30	.	<**0.001***
	TRASCET vs Sham Control	0.20	.	**0.018***
	Sham Control vs Saline Control	0.10	.	0.224
Dunnett[a]	TRASCET vs Saline Control	0.30	(0.11, 0.49)	**0.001***
	Sham Control vs Saline Control	0.10	(−0.09, 0.28)	0.367

ANOVA, analysis of the variance; CI, confidence interval.
* Statistically significant, after adjustment for multiple comparisons.
[a] Dunnett's procedure is shown for each group being compared to saline control. Another multiple comparisons procedure may be preferred because Dunnett's procedure does not allow for the comparison of TRASCET to sham control, which may certainly be of major interest.

sham control (difference in means = 0.20; 95% CI: 0.01, 0.39; $P = 0.047$), but not saline control vs sham control ($P = 0.438$). It should be noted that the Student–Newman–Keuls procedure and Duncan's procedure are methods related to Tukey's procedure [13,14].

2.2. Scheffé's procedure

Scheffé's procedure is flexible because it allows for the analysis of all linear contrasts [15,16]. Scheffé's procedure may be a reasonable alternative to Tukey's procedure when there is an imbalanced study design (i.e. there are unequal sample sizes in the groups), and the Tukey–Kramer procedure is also reasonable in the scenario of imbalanced study design. However, owing to its flexibility, Scheffé's procedure may have less power than Dunnett's procedure (see below) or Tukey's procedure in the settings where those tests are most appropriate. Using the multiple comparison procedure of Scheffé's procedure, we observe a statistically significant difference between TRASCET and saline control (difference in means = 0.30; 95% CI: 0.09, 0.51; $P = 0.003$). The other pairwise comparisons are not statistically significant in the hypothetical data.

2.3. Bonferroni's procedure

Among the approaches to performing multiple comparisons following ANOVA, Bonferroni's procedure (also called the Bonferroni (Dunn) procedure) is the most conservative, especially if the goal is perform all pairwise group comparisons [17]. The results from Bonferroni's procedure following ANOVA will include adjusted 95% confidence intervals

Multiple Comparisons Procedure	Advantages, Considerations, and Related Procedures
Bonferroni Adjustment	- Simple and easy to implement regardless of the nature of the data - Sets a Bonferroni-adjusted alpha level of 0.05 divided by the number of comparisons or tests - May be an overly conservative adjustment in the presence of many comparisons or tests
False Discovery Rate (FDR)	- Sets a threshold for deeming a result as a "discovery" - Often used when a large set of comparisons are performed - Used in genomic studies and statistical genetics
Procedures following ANOVA	
Tukey's procedure	- Makes all pairwise group comparisons - Originally used for analysis featuring balanced design with equal sample sizes per group - The Tukey-Kramer procedure is a modification of the test for imbalanced designs (unequal sample sizes per group)
Scheffe's procedure	- Allows for analysis of any set of specific contrasts (specific pairwise comparisons) - Can be implemented for imbalanced designs - May have less power than Tukey' procedure or Dunnett's procedure when those tests are more suitable
Bonferroni's procedure	- Generally the most conservative multiple comparison procedure - Intuitive to understand the correction - May be less powerful than the other procedure following ANOVA
Holm's procedure	- Modified version of Bonferroni's procedure with more power to detect differences - Less conservative than Bonferroni's procedure - Does not provide confidence intervals for the average group difference
Dunnett's procedure	- Compares each group to a single reference group (does not make all possible comparisons) - Decision to make all comparisons to a single reference group should be made *a priori* - Provides more power for detecting group differences in this situation
This list of multiple comparisons procedures is not exhaustive. The determination of how to proceed with adjusting for multiple comparisons is study- and situation-dependent. The best strategy is to adhere to the statistical plan which should include a pre-specified number of group comparisons and a suitable procedure to protect against type I errors.	

Fig. 3. Recommendations and guidelines for procedures to use to adjust for multiple comparisons or multiple testing. Each approach may have certain advantages relative to the other approaches. The ultimate decision in accounting for multiple comparisons is situation-dependent and can present challenges. However, surgeons should work with their statistical colleagues to ensure the best approach for controlling the type I error rate and interpreting the evidence. An overly conservative adjustment to protect against type I error (false positive results) may lead to an unwanted increase in the type II error rate (false negative results).

and adjusted *P* values. Please see the Bonferroni Adjustment section above for further information on the Bonferroni approach. When applying Bonferroni's procedure to our surgical research example, we find a statistically significant difference in defect coverage only between TRASCET and saline control (difference in means = 0.30; 95% CI: 0.09, 0.50; $P = 0.002$).

2.4. Holm's procedure

Holm's procedure (also known as the Bonferroni–Holm correction) is a modified version of the Bonferroni correction which keeps the experiment-wise error rate below alpha (the type I error rate) [18]. However, this method does not provide confidence intervals in its output. In our TRASCET example, this results in statistically significant differences between the TRASCET group and the saline control group ($P < 0.001$) as well as between the TRASCET group and the sham control group ($P = 0.018$). It should be also noted that the Holm–Sidak procedure has been developed as a more powerful modification of Holm's procedure.

2.5. Dunnett's procedure

Dunnett's procedure following ANOVA is used to compare each group with a single reference group, and provides the most power in this setting [19]. However, this decision should be made prior to the study as to whether all pairwise group comparisons will be performed or if the only comparisons of interest are between each group to the control group. In our example, this would mean the study was designed to compare defect coverage between TRASCET and saline control, and between saline and sham control. The results of Dunnett's procedure include adjusted confidence intervals and adjusted *P* values. Dunnett's procedure finds a statistically significant difference in average skin defect coverage for spina bifida between the TRASCET and saline control groups (difference in means = 0.30; 95% CI: 0.11, 0.48; $P = 0.001$) and no statistically significant difference between sham and saline control ($P = 0.367$). In our hypothetical example, Dunnett's procedure may not be desirable because it does not allow for the statistical comparison of TRASCET versus sham control. The comparison of TRASCET to sham control may certainly be of major interest since the sham control group may differ from the saline control group. Surgical investigators who are designing a similar study with two control groups may wish to choose a multiple comparisons procedure other than Dunnett's procedure because it requires the specification of one reference group to which all other groups are compared. This illustrates the importance of careful and thoughtful selection of the particular multiple comparison procedure.

In the analyses performed by the methods above in the hypothetical data, all methods determined a statistically significant difference in mean skin defect coverage for experimental spina bifida comparing TRASCET and saline control; however, only Tukey's procedure and Holm's procedure determined a statistically significant difference between TRASCET and sham control (Table 1).

3. Discussion

We have presented several useful statistical procedures for handling multiple comparisons in surgical research studies. These methods control the likelihood of declaring false positive results which may lead to misleading interpretations and inferences. It is debatable what the correct approach for handling multiple comparisons may be in certain situations [20,21]. Many of the approaches assume that the data conform to a normal (Gaussian) distribution. These approaches consider that parametric testing, usually ANOVA and Student t-tests, is being performed to compare multiple groups on an outcome variable. However, if the data are not amenable to analysis by ANOVA because the data are not normally distributed or if binary or count data are being considered, or if a Bonferroni correction is not reasonable because it would be overly conservative, then it may be difficult to determine how to proceed. It is important to understand that multiple comparisons or testing may be an issue, and therefore it is important for the reader to understand how it was considered. One may argue that no statistical adjustment for multiple comparisons is needed if the reader understands that no correction was done and interprets the results presented with that understanding [5,22]. In the situation where the usual multiple comparison procedures following ANOVA do not apply, and the Bonferroni correction or FDR approach may be considered overly conservative [23], authors can consider setting a generally more conservative alpha level of 0.01 as an alternative to the traditional 0.05 level, in order to control the type I error across the set of comparisons or hypothesis tests. It is worth noting that while *P* values are useful in surgical research, they should not stand alone but should be accompanied by treatment effect estimates and corresponding confidence interval. While it is beyond the scope of this article, there is ongoing discussion regarding the usefulness of *P* values in the context of deriving a dichotomous test results [2,24] and in the context of multiplicity adjustments [21–23].

The distinction needs to be drawn between statistical significance and clinical significance when considering multiple comparisons. In a highly powered study with large sample sizes, a statistical test may obtain a very small (i.e. highly statistically significant) *P* value for a relatively small difference between the groups. In this situation, the surgeon needs to make a decision regarding clinical significance of the difference when interpreting the evidence. A meaningful effect size based on clinical or scientific ground should be decided prior to the analysis so that interpretations of group differences are within the relevant clinical setting.

Multiple comparison procedures provide more certainty that a statistically significant result observed is not a false positive result. However, adjusting the alpha level in order to be more conservative to account for multiple comparisons will in turn have implications on the power and sample size considerations for a study. For instance, when a smaller (Bonferroni-adjusted) alpha level is specified, in order to preserve 80% power for detecting a fixed clinically meaningful effect size, the sample size requirements per group will be increased. Surgical researchers should take the adjusted type I error rate (alpha level) into consideration in the design phase of their study when performing power analysis and sample size calculations. Practical consideration and thought should be put towards specifying which comparisons are of interest a priori and power calculations should be modified accordingly. Furthermore, investigators should be cautious when applying conservative corrections for multiple comparisons because this will provide a less powerful analysis and increase the risk of false-negative results, or Type II errors [22].

It should be noted that there are specialized procedures and adjustments designed specifically for randomized controlled trials (RCTs) which feature sequential designs and interim analyses with stopping rules [25]. Examples of such methods for controlling the type I error probability (also known as alpha spending functions) include the O'Brien and Fleming method [26] and the Pocock method [27]. In this article, we focus on the common issue of multiple group comparisons and multiple testing in cross-sectional observational and these specialized methods pertaining to sequential trials are beyond the scope of this review [28].

Investigators may be inclined to perform all comparisons or hypothesis tests and subsequently decide which to include in the family of tests of multiplicity adjustment. The integrity of all statistical tests being performed and presented is not upheld if this is done. It must be emphasized that surgeons need to uphold honesty and transparency when performing an analysis that involves the adjustment for multiplicity or multiple comparisons. All statistical results should be reported transparently, allowing the reader to interpret and judge the findings.

4. Conclusions

Procedures for reducing the risk of false positive results and false discovery owing to multiplicity are valuable. These procedures protect the

study-wide or family-wise error rate when performing several group comparisons between groups or analyzing multiple endpoints. We provide guidance on strategies for how to handle multiplicity and multiple significance testing in surgical research studies. Surgeons need to be aware that multiple comparisons and multiple outcomes increase the likelihood of false positive errors even in well-designed studies and experiments. Appropriate adjustments are needed to control type I errors and to ensure the most valid analyses, interpretations and reporting of the results.

Appendix A. Glossary of statistical terms

Alpha level: Also known at the type I error rate or α, is the chance of a false positive result in a family of comparisons that the investigator is willing to accept. In a somewhat different sense, an alpha level is more often the *P* value threshold below which statistical significance for a single test is determined, and it is traditionally set at 5%.

Alternative hypothesis: The alternative hypothesis states that there is a difference between the groups or a significant association.

Beta level: Also known as the type II error rate, this is the chance of false negative results. Beta is traditionally set to be 20%. Power is equal to 1 − beta. Thus, when beta = 20%, power is 80%.

Bonferroni's adjustment: With a Bonferroni adjustment, if C comparisons or test are planned, then each comparisons uses alpha/C as the threshold for determining statistical significance. This may be an overly conservative adjustment, but it is commonly used and widely applied.

Bonferroni's procedure (following ANOVA): A multiple comparisons procedure which is the most conservative adjustment for reducing the risk of type I error. Results of Bonferroni's procedure include confidence intervals.

Conservative adjustment: An adjustment like the Bonferroni correction is considered to be conservative when it very strongly minimizes the alpha level. Any results determined as statistically significant following a conservative adjustment have a very low likelihood of being false positive results. Performing a conservative adjustment makes it more difficult to observe significant results, and therefore it may not be preferred in certain situations.

Dunnett's procedure: A multiple comparisons procedure following ANOVA used to compare each group with a single reference group. Dunnett's procedure provides the most power to detect group difference under this analysis plan. Results of Dunnett's procedure include confidence intervals.

Experiment-wise error rate: The experiment-wise or study-wide error rate is the type I error rate usually assumed to be 5% which is applied to a set of comparisons or tests in a given study. When multiple comparisons or multiple testing occur, the experiment-wise error rate needs to be adjusted in order to control for the risk of false positive results.

Family of comparisons: A family of comparisons is the general conceptual framework of the set of related statistical comparisons or statistical tests in an experiment or a clinical study. The family comparisons share a family-wise error rate.

False positive result: A false positive result occurs when the investigator finds a statistically significant result when there truly are no significant associations or group differences. In statistical terminology, this is when an investigator rejects the null hypothesis when in fact it is true (type I error).

False negative result: A false negative result occurs when the investigator does not find a statistically significant result when there truly exist significant associations or group differences. Failing to reject the null hypothesis when the alternative hypothesis is true constitutes a false negative result (type II error).

Holm's procedure: A modified version of the Bonferroni correction which keeps the experiment-wise error rate below alpha. The Holm–Sidak procedure has been developed as a more powerful modification of Holm's procedure. The results do not include confidence intervals.

Multiplicity: Multiplicity is the issue of an elevated risk of false positive results (type I error) owing to multiple group comparisons or multiple testing. An adjustment of the experiment-wise error rate (alpha) needs to be made in order to correct for the issue of multiplicity.

Multiple comparisons: Multiple comparisons refer to the situation where multiple groups are being compared two at a time, which lead to an increased chance of false positive results when there truly are no significant associations or group differences. An adjustment of the experiment-wise error rate (alpha) needs to be made in order to handle the issue of multiple comparisons.

Multiple comparisons procedures: Multiple comparisons procedures are statistical approaches designed to adjust the experiment-wise error rate in order to control for the risk of false positive results when no differences truly exist. These include but are not limited to the Bonferroni correction, the False Discovery Rate approach, Tukey's test, Scheffé's test, Holm's test, and Dunnett's test.

Multiple testing: Multiple testing refers to the general issue of multiple related statistical tests being performed (for example two groups being compared on multiple outcome variables) which leads to an increased risk of false positive results. An adjustment of the experiment-wise error rate (alpha) needs to be made in order to take multiple testing into account in study design and analysis.

Null hypothesis: The null hypothesis assumes no group differences on the outcome or no significant associations.

One-way ANOVA: Also called one-factor analysis of the variance, this statistical test is used as a global or omnibus test to determine any differences in means of a continuous variable across 3 or more groups.

Power: Statistical power is the probability of detecting significant group differences when differences truly exist. In other words, power is the probability of rejecting the null hypothesis when it is false. Power is equal to 1 − beta, and usually desired to be 80% or 90%.

***P* value:** the probability that the observed effect is owing to chance given that the null hypothesis is true. A small *P* value less than the alpha level is determined as being statistically significant.

Scheffé's procedure: A multiple comparisons procedure following ANOVA which allows of the analysis of any set of linear contrasts. Scheffé's procedure may be reasonable in the situation of an imbalanced study design with unequal sample sizes per group. Results of Scheffé's procedure include confidence intervals.

Tukey's procedure: A multiple comparisons procedure following ANOVA developed for the situation of balanced design with equal sample sizes per group with results that include confidence intervals.

Type I error: Rejecting the null hypothesis when it is true (false positive result).

Type II error: Failing to reject the null hypothesis when it is false (false negative result).

References

[1] Simillis C, Lal N, Thoukididou SN, et al. Open versus laparoscopic versus robotic versus transanal mesorectal excision for rectal cancer. Ann Surg 2019;270(1):59–68.
[2] Harrington D, D'Agostino RB, Gatsonis C, et al. New guidelines for statistical reporting in the journal. N Engl J Med 2019;381(3):285–6.
[3] Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54(4):343–9.
[4] Motulsky H. Multiple comparisons concepts. Intuitive biostatistics: a nonmathematical guide to statistical thinking. 2nd ed. New York: Oxford University Press; 2010; 159–67.
[5] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995; 310(6973):170.
[6] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Series B 1995;57(1):289–300.
[7] Cabral HJ. Multiple comparisons procedures. Circulation 2008;117(5):698–701.
[8] Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med 1990;9(7):811–8.
[9] Fauza DO. Transamniotic stem cell therapy: a novel strategy for the prenatal management of congenital anomalies. Pediatr Res 2018;83(1–2):241–8.
[10] Tukey J. Comparing individual means in the analysis of variance. Biometrics 1949;5 (2):99–114.
[11] Brillinger DR, John W. Tukey: his life and professional contributions. Ann Statist 2002;30(6):1535–75.
[12] Hayter AJ. A proof of the conjecture that the Tukey–Kramer multiple comparisons procedure is conservative. Ann Statist 1984;12(1):61–75.
[13] McHugh ML. Multiple comparison analysis testing in ANOVA. Biochem Med 2011;21 (3):203–9.

[14] Duncan DB. Multiple range and multiple F tests. Biometrics 1955;11(1):1–42.
[15] Scheffe H. The analysis of variance. New York: John Wiley and Sons; 1959.
[16] Rosner B. Fundamentals of biostatistics. 6th ed. Thomson Brooks/Cole: Belmont, California; 2006.
[17] Sedgwick P. Multiple significance tests: the Bonferroni correction. BMJ 2012;344:e509.
[18] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6(2):65–70.
[19] Dunnett C. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc 1955;50:1096–121.
[20] Althouse AD. Adjust for multiple comparisons? It's not that simple. Ann Thorac Surg 2016;101(5):1644–5.
[21] Feise RJ. Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol 2002;2:8.
[22] Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1(1):43–6.
[23] Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316(7139):1236–8.
[24] Amrhein V, Gleenland S, McShane B. Scientists rise up against statistical significance. Nature 2019;567(7748):305–7.
[25] Dmitrenko A, D'Agostino RB. Multiplicity considerations in clinical trials. N Engl J Med 2018;378(22):2115–22.
[26] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35(3):549–56.
[27] Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64(2):191–9.
[28] Albers C. The problem with unadjusted multiple and sequential statistical testing. Nat Commun 2019;10(1):1921.