ARTICLE IN PRESS

Journal of Cardiothoracic and Vascular Anesthesia ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

ScienceDirect

journal homepage: www.jcvaonline.com

Review Article
Biostatistics

Developing a Hypothesis and Statistical Planning

Sarah L. Nizamuddin, MD*, Junaid Nizamuddin, MD*,
Ariel Mueller, MPH†, Harish Ramakrishna, MD‡,
Sajid S. Shahul, MD, MPH*,[1]

*Department of Anesthesia and Critical Care, University of Chicago, Chicago, IL
†Department of Anesthesia, Critical Care, and Pain Medicine, Beth Israel Deaconess Medical Center, Boston, MA
‡Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Phoenix, AZ

Key Words: biostatistics; statistics; hypothesis development; clinical research; research question; null hypothesis; alpha and beta values; confidence intervals

WITH A SURPLUS OF questions and a desire to find meaningful answers, understanding the basics of performing clinical research is crucial. The first step is to develop a research question, which serves as the objective of a study. A concise question will give a clear aim to a study, narrow the vast amount of literature on the subject, and aid in developing a hypothesis.[1] When choosing a question, it is imperative to focus on a topic that is novel, interesting, and/or provides scientific value (see Table 1 and Fig 1).[2] Among topics already well researched, a good research question may add new cohorts, further expanding upon generalizability of the findings for the population with that condition.[3]

A thorough literature search is necessary to gain an understanding of what is already known, which areas need further investigation, and which areas have not been examined at all. This initial search can narrow down a research question that would add value to the scientific community. However, it is important to keep in mind that a literature search may not reveal all studies performed. Studies with negative results may not always be as well represented in the literature as positive studies (ie, they may be less likely to get published), and studies with positive results still may need further investigation. Furthermore, it is possible that a study with negative results may be due to issues related to lack of power or small number of subjects rather than the actual hypothesis being false.

After examining the current literature, clinicians may choose to expand or improve upon a question that already has been asked, or they may choose a question that is entirely novel. For example, interest in studying the use of dexmedetomidine in cardiac surgery would require a literature search that would lead to the study by Cheng et al, *The Effects of Dexmedetomidine on Outcomes of Cardiac Surgery in Elderly Patients*,[4] which found decreases in operative and in-hospital mortality as well as postoperative stroke and delirium after perioperative dexmedetomidine infusion. While reviewing this and other articles on this topic would elucidate the areas that already have been studied, they also may inspire further inquiry in an area that may need further investigation or that has not been well examined, such as the effects of intraoperative clonidine use in similar patient populations, or differences in anesthetic costs after dexmedetomidine use in cardiac surgery.

Although interest in a specific topic often leads to a search in the literature, at times, reviewing medical literature subsequently may lead to discovery of a topic of interest. Case reports may present novel treatments or management that can be invaluable in leading to further high-quality research questions in that area.[5] In the case report by Gutsche et al, *Treatment of Ventricular Assist-Device-Associated Gastrointestinal Bleeding with Hormonal Therapy*,[6] the authors described a case in which the use of ethinyl estradiol and norethindrone may have aided in the cessation of gastrointestinal bleeding in a patient with a ventricular assist device. The case report may lead to a research question: does hormone therapy with ethinyl estradiol and norethindrone assist in

The authors have no financial disclosures or conflicts of interest to report.

[1] *Address reprint requests to* Sajid S. Shahul, MD, MPH, 5841 S Maryland Ave MC 4028, Chicago, IL 60637.

E-mail address: sajid.shahul@gmail.com (S.S. Shahul).

http://dx.doi.org/10.1053/j.jvca.2017.04.020

Table 1
Characteristics of a High-Quality Research Question

Concise with clear aim
Novel or improving on prior research
Adds value to the current literature
Clinically relevant/meaningful
Feasible costs, patient variables, ethics

cessation of gastrointestinal bleeding in patients with ventricular assist devices?

A literature search will aid in creating a research question that is desired by or would benefit the scientific community. However, a novel question may not lead necessarily to clinically meaningful results. Inquiring on the incidence of preoperative hiccups and postoperative outcome, although not studied before, may not yield a change in perioperative care by the patient's anesthesiologist or surgical team. What defines clinically meaningful remains subjective in nature, but must add value to the topic at hand, whether by increasing awareness on an important subject, by leading to a change in practice, or by inspiring further inquiry about a topic that may be fruitful after additional study.

Lastly, studies must be feasible to perform; not every relevant research question will lead to a study that is feasible. For instance, a study examining vapocoolants and lidocaine infiltration for pain control in radial artery cannulation performed by Rusch et al[7] involved careful consideration of patient population when designing the study. This prospective trial was performed on patients undergoing *elective* cardiac surgery or carotid endarterectomy, which allowed for proper patient enrollment and blinding. A similar study designed to examine this question in patients receiving emergent ruptured aneurysm repair would be difficult to perform. Feasibility of a study carries an element of subjectivity, but requires that a study question will be able to be examined properly in the population desired, considering costs, patient variables, and ethical factors.

Developing a Hypothesis

Transforming a research question into a hypothesis is the next step in designing a research study. In a study by Haanschoten et al titled *Use of Postoperative Peak Arterial Lactate Level to Predict Outcome After Cardiac Surgery,*[8] a research question may postulate, "do peak postoperative lactate levels aid in predicting mortality after cardiac surgery?" This question then can be transformed into the research hypothesis: the peak postoperative lactate level is a predictive factor of mortality for patients undergoing cardiac surgery.

A research hypothesis compares what is observed in the data to what would be expected if the null hypothesis (H_0) is true. The goal of hypothesis testing is to measure the consistency of the observed data with the null hypothesis.[9,10] The null hypothesis (Ho) assumes that there is no association between exposure (or treatment) and the outcome.[11] In the current example, the null hypothesis may be: there is no association between the postoperative lactate level and mortality in

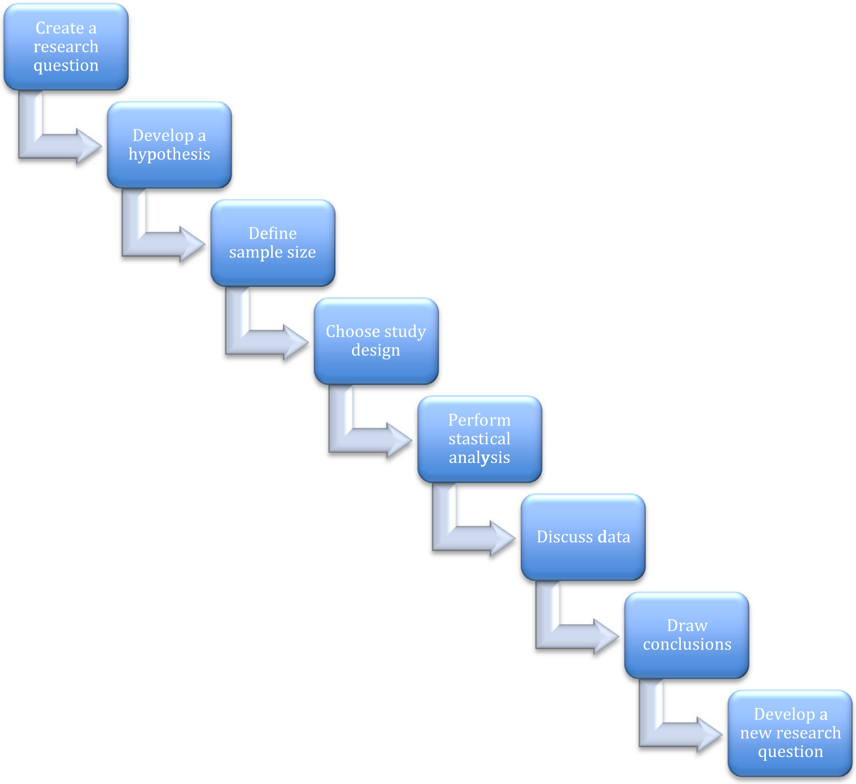

Fig 1. Steps for creating and designing a study.

Please cite this article as: Nizamuddin SL, et al. (2017), http://dx.doi.org/10.1053/j.jvca.2017.04.020

patients undergoing cardiac surgery. The alternate hypothesis assumes that there is an association between the exposure and the outcome.[12] There are 4 principle types of alternative hypotheses: point, 1-tailed and 2-tailed directional, and nondirectional. Point alternative hypotheses involve a population that is fully defined with no unknown parameters. One-tailed versus 2-tailed alternative hypotheses depend on whether there is concern with rejection of 1 or both tails of the sampling distribution.[2] For example, stating a drug leads to increased incidence of hyperglycemia compared with the standard drug would be a 1-tailed hypothesis, whereas stating the drug is associated with a change in blood glucose would be a 2-tailed hypothesis. Lastly, a nondirectional hypothesis simply is concerned that the null hypothesis is not true. A 1-tailed alternate hypothesis in the current example may be: the peak postoperative lactate level is associated positively with increased mortality after cardiac surgery. Acceptance of an alternate hypothesis implies that the observed findings are not due to chance, bias, or confounding variables.

Another important factor to consider when developing a hypothesis is the type of trial that will be conducted in order to test the hypothesis. Three trial types include noninferiority, equivalence, and superiority (Fig 2). A superiority study aims to demonstrate that the new therapy is more efficacious, whereas a noninferiority test aims to say that the new therapy is not inferior to the conventional therapy, without proving efficacy.[13] An equivalence study, on the other hand, is seeking to prove that the new therapy has the *same* results as the other therapy, within a specified clinically acceptable margin. It is important that all results be interpreted based off of the trial that was conducted, as failing to determine superiority does not imply equivalence.

Differentiating among these study types is easier when considering the study of a new drug. Perhaps this new drug is significantly less expensive than the current drug that is used. In this case, investigators may be interested in simply proving that the new drug is not inferior to the current drug. In this scenario, a noninferiority study would be appropriate. On the other hand, if it were desired to prove that a new drug is better (ie, leads to fewer adverse reactions) than the currently used drug, then a superiority study would be more appropriate. Determining which trial type and, if applicable, which clinically acceptable margin, can dramatically affect the sample size required in order to test the hypothesis. For example, when testing for equivalence, the clinically acceptable threshold is often small, as the researcher is attempting to demonstrate that the new therapy is the same as the current therapy, therefore necessitating a large sample size.

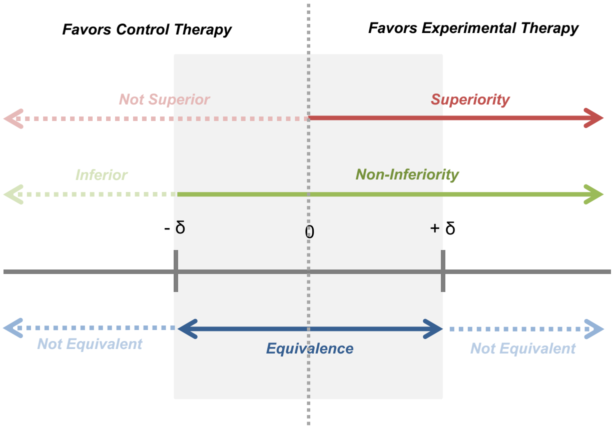

Fig 2. Types of trials. Depicted above are 3 trial types that may be used to test a hypothesis. A superiority study would test the null hypothesis that there is no difference between therapies against the alternative hypothesis that there is a difference between the 2 therapies. This can be further specified as either a 1-sided (depicted above) or 2-sided superiority trial. A noninferiority study, however, tests whether or not the experimental therapy is inferior to the control therapy, allowing for a clinically acceptable margin (denoted δ). An equivalence study seeks to demonstrate that the experimental therapy is the same, or within a clinically acceptable margin (-δ to δ), of the control therapy.

Statistical Considerations for Hypothesis Testing

Integral to study design are selecting an appropriate alpha (a) and beta (β); this requires an understanding of their meanings as they are often misunderstood.[14–16] A commonly used a value of 0.05 means that the researcher has set a maximum of a 5-percent chance that the null hypothesis will be rejected when it is in fact true, or a type-I error.[17,18] More simply stated, it is the probability of stating an association exists between the variable and the outcome when that correlation was actually due to chance alone.

Beta is the probability of failing to reject the null hypothesis when it is false, also known as a type-II error.[19,20] A commonly set value for beta is 0.2. The power of a study is the probability of correctly rejecting the null hypothesis when an association does exist between the tested variable and outcome, or 1-B.[21] The power of a study is related to many factors, including sample size.[11]

A test statistic is calculated from the study data and is used to test the study hypothesis. Test statistics are chosen based on the outcome variable and help determine if the test statistic's value is typical when H_0 is true. The statistical analysis plan is chosen prior to carrying out the study based on the study design and primary outcome in order to decrease bias and confounders. A p value is the probability that the observed findings of the test statistic could be equal to or more extreme than that which is actually observed if the H_0 is true.[22,23] For example, a p value of 0.05 means that the probability of rejecting the null hypothesis when it is in fact true (false positive) is 5 percent. If the p value is less than alpha, the H_0 will be rejected. Therefore, a smaller p value denotes that data like that which were observed do not occur often when the null hypothesis is true.[20,24,25] For example, a very small p value of 0.0001 means that the probability of rejecting the null hypothesis when it is in fact true is 1 in 10,000. A larger p value signifies that the data observed have a higher chance of occurring even when the null hypothesis is true, therefore supporting the null hypothesis. While a p value that is less than 0.05 long has been considered statistically significant, many members of the scientific and statistical community argue that a much smaller p value is necessary to consider study results statistically significant.[26] It is important to note that

Please cite this article as: Nizamuddin SL, et al. (2017), http://dx.doi.org/10.1053/j.jvca.2017.04.020

determining results to be *clinically* significant or meaningful is distinct from statistical significance and takes into consideration other factors as well.

The study by Greenberg et al, *Rainy Days for the Society of Pediatric Anesthesia,* found a statistically significant (p value = 0.006) correlation with the incidence of rain on the opening day of Society of Pediatric Anesthesia meetings.[27] While this finding may be considered statistically significant if evaluating the p value alone, what are the chances that this correlation in results would be reproduced after another ten years? In other words, what are the chances that Society of Pediatric Anesthesia meetings truly have a causative effect on rainfall in the cities in which they are held? Reproducibility is a factor to consider when analyzing and interpreting presented data. While a $p < 0.05$ may appear to prove strong evidence in favor of a specific result in a study, the ability to reproduce those results with a $p < 0.05$ on a subsequent study may be surprisingly rare. The study by Greenberg et al nicely demonstrated the ability of a small p value (0.006) to be deceiving in the interpretation of results. Reproducibility is also study specific and should be considered when testing or interpreting results of a study with multiple hypotheses, as the probability of finding a significant result can increase, especially when using a threshold such as 0.05 to determine significance. To increase the chance of reproducible results in a study, the threshhold needs a much smaller p value, as low as 0.0001.[28,29]

With increasing skepticism of the current emphasis on p values, confidence intervals (CI) are a vitally important tool. A CI is the range of values with a specified probability that the parameter lies within it. For example, a 95% CI for central venous pressure after using a novel drug in a coronary artery bypass graft procedure may be reported as 8 to10. This would mean that if this study were performed 100 times, the actual mean central venous pressure would be contained in the CI 95 times out of 100. CIs can be a useful tool in interpreting study results, as they are better correlated with increased chances of replication.[30] According to Cumming, there is an 83% chance of replicating a mean that falls within the 95% CI of the initial experiment.[30] There has been a strong push by many researchers and statisticians in the scientific community to move toward a heavier reliance on CI over p values as a tool to evaluate the significance of study findings.[31]

The study by Komatsu et al, *Etomidate and the Risk of Complications After Cardiac Surgery: A Retrospective Cohort Analysis*, nicely demonstrated the value of CIs. They sought to evaluate the incidence of atrial arrhythmias in patients undergoing coronary artery bypass grafting after induction with etomidate versus standard induction agents.[32] While the incidence of atrial arrhythmias was slightly higher in the etomidate group versus the standard group (33.4% *v* 31.5%), the 98.3% CI of the odds ratio was reported at 0.92-to-1.23 (p = 0.29). The fact that 1 was included in this confidence interval showed that there is poor support of a claim that etomidate would increase or decrease risk of atrial arrhythmias. On the other hand, when evaluating secondary outcomes, the authors found an association with the use of packed red cells and the use of etomidate for induction (p = 0.002), with an odds ratio of 1.32. The 99.6% CI for the odds ratio was 1.02-to-1.70, with no overlap of 1 in the interval.

The fragility index of a study is another useful tool to evaluate statistical data, particularly as interest in p values alone has waned. The fragility index is the number of patients who would have to change outcomes to turn a statistically significant result to a nonsignificant result.[33] It is another method of measuring statistical significance, going a step beyond the p values and CIs reported with the results. Two recent studies have looked at the fragility index of previous trials, particularly landmark trials that changed practice. Walsh et al examined 399 randomized controlled trials; 25% of trials had a fragility index of 3 or less, meaning that moving 3 patients from one outcome to another outcome would render a statistically significant outcome nonsignificant, and more than half of the trials had fragility indices that were less than the number of patients lost to follow up.[34] Ridgeon et al examined 56 trials in critical care medicine; the median fragility index was 2 and more than 40% of trials had a fragility index of 1.[35] While the fragility index has not been adopted widely, it is another statistical parameter that can guide interpretation of results.

Conclusion

A well-designed study requires thoughtful preparation and careful attention to detail so that the results can be valuable to the research community. The development of a meaningful research question, a clear hypothesis, and reasonable alpha and beta values are the initial steps necessary to begin a study. While p values are used widely to evaluate the significance of a study's results, they increasingly have been misused and misinterpreted. CIs may be a better tool in assessing the results of a study and chances of reproducibility.

References

1 Tully MP. Research: Articulating questions, generating hypotheses, and choosing study designs. Can J Hosp Pharm 2014;67:31–4.

2 Hulley S, Cummings S, Browner W, et al. Designing clinical research, ed 3. Lippincot Wiliams and Wilkins; 2007.

3 Lipowski EE. Developing great research questions. Am J Health Syst Pharm 2008;65:1667–70.

4 Cheng H, Li Z, Young N, et al. The effect of dexmedetomidine on outcomes of cardiac surgery in elderly patients. J Cardiothorac Vasc Anesth 2016;30:1502–8.

5 Hessel EA 2nd. Why we should continue to publish case reports. J Cardiothorac Vasc Anesth 2013;27:825–7.

6 Gutsche JT, Atluri P, Augoustides JG. Treatment of ventricular assist-device-associated gastrointestinal bleeding with hormonal therapy. J Cardiothorac Vasc Anesth 2013;27:939–43.

7 Rusch D, Koch T, Seel F, et al. Vapocoolant spray versus lidocaine infiltration for radial artery cannulation: A prospective, randomized, controlled clinical trial. J Cardiothorac Vasc Anesth 2017;31:77–83.

8 Haanschoten MC, Kreeftenberg HG, Arthur Bouwman R, et al. Use of postoperative peak arterial lactate level to predict outcome after cardiac surgery. J Cardiothorac Vasc Anesth 2017;31:45–53.

9 Nelson AA Jr. Developing research hypotheses. Am J Hosp Pharm 1980;37:264–5.

Please cite this article as: Nizamuddin SL, et al. (2017), http://dx.doi.org/10.1053/j.jvca.2017.04.020

10 Cohen HW. P values: Use and misuse in medical literature. Am J Hypertens 2011;24:18–23.

11 Dawson B, Trapp RG. Basic & clinical biostatistics. New York, NY: Lange Medical Books/McGraw-Hill, Medical Pub Division, 2004, p 438.

12 Davis RB, Mukamal KJ. Hypothesis testing: means. Circulation 2006;114: 1078–82.

13 Christensen E. Methodology of superiority vs. equivalence trials and non-inferiority trials. J Hepatol 2007;46:947–54.

14 Mark DB, Lee KL, Harrell FE Jr. Understanding the role of p values and hypothesis tests in clinical research. JAMA Cardiol 2016;1:1048–54.

15 Goodman S. A dirty dozen: Twelve p-value misconceptions. Semin Hematol 2008;45:135–40.

16 Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999;130:995–1004.

17 Lopes RD, Harrington RA. Understanding clinical research. New York, NY: McGraw-Hill. 2313, p 245.

18 Kelen GD, Brown CG, Ashton J. Statistical reasoning in clinical trials: Hypothesis testing. Am J Emerg Med 1988;6:52–61.

19 Sedgwick P. Pitfalls of statistical hypothesis testing: Type I and type II errors. BMJ 2014;349:g4287.

20 Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ 2001;322:226–31.

21 Banerjee A, Chitnis UB, Jadhav SL, et al. Hypothesis testing, type I and type II errors. Ind Psychiatry J 2009;18:127–31.

22 Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995;152:27–32.

23 Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing: An explanation for new researchers. Clin Orthop Relat Res 2010;468:885–92.

24 Pocock SJ, McMurray JJ, Collier TJ. Making sense of statistics in clinical trial reports: Part 1 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol 2015;66:2536–49.

25 Pocock SJ, Stone GW. The primary outcome fails - what next? N Engl J Med 2016;375:861–70.

26 Pocock SJ, Stone GW. The primary outcome is positive - is that good enough? N Engl J Med 2016;375:971–9.

27 Greenberg RS, Bembea M, Heitmiller E. Rainy days for the Society for Pediatric Anesthesia. Anesth Analg 2012;114:1102–3.

28 Shafer SL, Dexter F. Publication bias, retrospective bias, and reproducibility of significant results in observational studies. Anesth Analg 2012;114: 931–2.

29 Goodman SN. A comment on replication, p-values and evidence. Stat Med 1992;11:875–9.

30 Cumming G. Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 2008;3:286–300.

31 Belia S, Fidler F, Williams J, et al. Researchers misunderstand confidence intervals and standard error bars. Psychol Methods 2005;10:389–96.

32 Komatsu R, Makarova N, You J, et al. Etomidate and the risk of complications after cardiac surgery: A retrospective cohort analysis. J Cardiothorac Vasc Anesth 2016;30:1516–22.

33 Feinstein AR. The unit fragility index: An additional appraisal of "statistical significance" for a contrast of two proportions. J Clin Epidemiol 1990;43:201–9.

34 Walsh M, Srinathan SK, McAuley DF, et al. The statistical significance of randomized controlled trial results is frequently fragile: A case for a fragility index. J Clin Epidemiol 2014;67:622–8.

35 Ridgeon EE, Young PJ, Bellomo R, et al. The fragility index in multicenter randomized controlled critical care trials. Crit Care Med 2016;44:1278–84.

Please cite this article as: Nizamuddin SL, et al. (2017), http://dx.doi.org/10.1053/j.jvca.2017.04.020