Checklist for Statistical Topics in *Anesthesia & Analgesia* Reviews

Franklin Dexter, MD, PhD

In my role as Statistics Editor for *Anesthesia & Analgesia*, I have reviewed >200 papers each year for the past 3 years. This editorial summarizes the recurring statistical issues identified in papers submitted to the journal. Table 1 is an annotated version of my checklist.[1–7] The checklist relies on some basic statistical definitions. Table 2 is an annotated glossary for words and phrases in Table 1.[8] Many of these definitions are taken verbatim from Wikipedia. They are shown in quotes.

Table 1 is not intended as a complete checklist. *Anesthesia & Analgesia* encourages authors to follow the CONSORT, STROBE, or PRISMA checklists, which cover basic topics. Table 1 includes topics not directly addressed in these checklists.

More than three-quarters of articles in general medical journals include relatively advanced topics such as those in Table 1.[9,10] *Anesthesia & Analgesia* has more mathematics and statistics than many clinical journals because of the nature of our research and practice. Performing the work in teams including analysts, statisticians, engineers, etc., helps assure that quantitative topics including those listed in Table 1 are addressed adequately.

From the Department of Anesthesia, University of Iowa, Iowa City, Iowa.

Accepted for publication April 18, 2011.

Supported by departmental funds.

Address correspondence and reprint requests to Franklin Dexter, MD, PhD, Department of Anesthesia, University of Iowa, 6JCP, 200 Hawkins Dr., Iowa City, IA 52242. Address e-mail to Franklin-Dexter@UIowa.edu or www.FranklinDexter.net.

DOI: 10.1213/ANE.0b013e3182204e95

DISCLOSURES

Name: Franklin Dexter, MD, PhD.
Role: This author helped write the manuscript.
Conflicts: Franklin Dexter reported no conflicts of interest.
Attestation: Franklin Dexter approved the final manuscript.

REFERENCES

1. Mascha EJ. Equivalence and noninferiority testing in anesthesiology research. Anesthesiology 2010;113:779–81
2. Austin PC, Grootendorst P, Norman SL, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. Stat Med 2007;26:754–68
3. Strum DP, Sampson AR, May JH, Vargas LG. Surgeon and type of anesthesia predict variability in surgical procedure times. Anesthesiology 2000;92:1454–67
4. Dexter F, Epstein RH, Lee JD, Ledolter J. Automatic updating of times remaining in surgical cases using Bayesian analysis of historical case duration data and instant messaging updates from anesthesia providers. Anesth Analg 2009;108:929–40
5. Smallman B, Dexter F. Optimizing the arrival, waiting, and NPO times of children on the day of pediatric endoscopy procedures. Anesth Analg 2010;110:879–87
6. Baker WL, White CM, Cappelleri JC, Kluger J, Coleman CI; Health Outcomes, Policy, and Economics (HOPE) Collaborative Group. Understanding heterogeneity in meta-analysis: the role of meta-regression. Int J Clin Pract 2009;63:1426–34
7. Cohen MM, O'Brien-Pallas LL, Copplestone C, Wall R, Porter J. Nursing workload associated with adverse events in the postanesthesia care unit. Anesthesiology 1999;91:1882–90
8. Silber JH, Rosenbaum PR, Zhang X, Even-Shoshan O. Influence of patient and hospital characteristics on anesthesia time in medicare patients undergoing general and orthopedic surgery. Anesthesiology 2007;106:356–64
9. Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H. The use of statistics in medical research: a comparison of The New England Journal of Medicine and Nature Medicine. Am Stat 2007;61:45–55
10. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA 2007;298:1010–22

Table 1. Recurring Statistical Issues in Manuscripts Submitted to *Anesthesia & Analgesia*, with Superscript Letters Indicating Words or Phrases Described in Table 2

The study has small *P* values (e.g., $P < 0.001$)[a] but does not include confidence intervals[b] and (when applicable) estimates of attributable risk.[c]

Explanation: A small *P* value does not identify an important finding. For example, if postoperative nausea and vomiting were studied in a double-blind, randomized, controlled trial with 1 million patients receiving 3 mg ondansetron and another 1 million receiving 4 mg ondansetron, likely the *P* value[a] would be tiny, far less than 0.01. However, *Anesthesia & Analgesia* would be unlikely to accept the clinically trivial paper. *P* values[a] should be accompanied by quantification of the practical or clinical importance of the estimated differences, generally confidence intervals[b] and/or attributable risk.[c]

The study suggests conclusion(s) of equality but lacks comparison of confidence interval[b] widths to an external standard.

Explanation: A *P* value[a] that is not statistically significant (e.g., $P > 0.10$) shows failure to achieve a significant difference, which could be attributable to many reasons. A common example is a sample size that was too small. A *P* value[a] cannot be used to conclude that the study shows that groups are the same. Include calculation of confidence interval(s)[b] to determine the limits of the magnitude of the effect that can be excluded by the study findings.

The study has a primary or secondary end point that is either cost or time but does not analyze it using methods suitable to estimate the mean of skewed[d] data.

Explanation: Time is a common surrogate for cost in anesthesia (e.g., time to start a case, time to extubation, and hospital length of stay). Total cost is proportional to total time and thus should usually be analyzed by means and reported as confidence intervals of absolute or relative differences in the mean. Variables involving time are typically skewed.[d] There are often also many patients with a value of zero (e.g., when studying patient time in the postanesthesia care unit, analyses often need to model the proportions of patients who bypass the postanesthesia care unit and go directly to the home-going unit or who instead go to the intensive care unit). Rank-based[e] methods are rarely appropriate because the sustained cost (e.g., of time in the operating room) is not proportional to the rank[e] time. Transformation[f] is rarely sufficient because the public does not pay transformed[f] dollars.

The study has an intervention variable that was not randomized, such as choice of drug dose, but does not include propensity scores[g] analysis or analogous technique (e.g., logistic regression[h]).

Explanation: Usually independent variables should include a model for the different practice styles of providers (e.g., physician or hospital treated as a fixed effect[i] or as a random effect[j]). For binary end points, propensity score[g] analysis and logistic regression[h] are not interchangeable.[2] When propensity score[g] analysis is used to create matched subjects in different treatment groups, include an assessment of the achieved similarity between groups. Consider use of different methods for matching subjects as a sensitivity analysis[k] (e.g., matching pairs in a 1:1 manner versus stratified sampling[l] on quantiles[m] of the propensity score[g]).

The study has a dependent variable that may differ among providers, such as physicians, but that dependent variable is not analyzed by stratification[l] or mixed models.[n]

Explanation: In double-blind[o] randomized clinical trials, the influence of providers is generally fully controlled. In retrospective studies, usually this is not so. When differences among providers can have substantive influences on dependent variables, the influence of provider needs to be included in the statistical model.[p] This is especially common for studies in economics and education, because differences among providers are relatively large.[3–5]

The study has multiple observations measured over time, but the statistical model is not fully specified and/or justified.

Explanation: Mixed effect[n] models are frequently used to assess repeated observations of the same items over time. Examples are mixed models[n] for repeated measures[q] for blood pressure measured every 15 minutes and nonlinear[r] mixed effects models for pharmacokinetic investigations. Such models should be described with sufficient detail for reviewers and readers to understand the model and its rationale completely, including assumptions for the variance-covariance[s] structure(s).This is often easily done by including the statistical model,[p] expressed using equations or the statistical package's model statement.

The study uses meta-regression[t] without complete study data and with an independent variable(s) that is not the basis for randomization.[6]

Explanation: Meta-regression[t] is a tool to understand the relationship between study design characteristics and end points. Suppose that a randomized trial compares the mortality rate between patients receiving 0 mg versus 10 mg of a study drug. Another trial compares mortality between patients receiving 0 mg versus 20 mg of study drug. Then, analysis is straightforward because complete data are known for all patients (i.e., 0 mg, 10 mg, and 20 mg are known without error). When only means and standard deviations are known for the independent variables (e.g., patient age), secondary data and analyses (e.g., in Appendix) are needed to validate the statistical methodology.[6]

Table 2. Definitions of Some Words and Phrases Used in Table 1, with Locations Specified by Superscript Letters

(a) *P* value. The *P* value is the probability of obtaining the statistical test's statistic at least as large as the one observed, assuming that the null hypothesis is true. For example, consider a null hypothesis that the probability is 0.50 of a coin flip to land heads, with the alternative being tails. After 10 flips, there have been 10 heads. The 2-sided *P* value for obtaining 10 flips of all heads or all tails equals $0.50^{10} + 0.50^{10} = 0.002$. See http://en.wikipedia.org/wiki/P-value.

(b) Confidence interval. "A confidence interval is an interval estimate of a population parameter" (e.g., mean). If the data analysis were repeated many times, the "confidence level" (e.g., 95%) would specify the "proportion of such intervals that contain the true value of the parameter." See http://en.wikipedia.org/wiki/Confidence_interval.

(c) Attributable risk. The population attributable risk refers to "the reduction in incidence that would be observed if the population was entirely unexposed compared with its current (actual) exposure pattern." For example, suppose that all adverse events in the postanesthesia care unit were eliminated.[7] Then, the percentage reduction in the number of unexpected intensive care unit admissions would be an attributable risk. Frequently, we are also interested in the attributable excess cost or time. For example, if all adverse events in the postanesthesia care unit were eliminated, the percentage reduction in total nursing labor would be reduced.[7] See http://en.wikipedia.org/wiki/Attributable_risk.

(d) Skewed. A probability distribution has zero skewness when the "values are relatively evenly distributed on both sides of the mean," and a positive skewness when "the tail on the right side is longer than the left side and the bulk of the values lie to the left of the mean." For example, operating room times for a procedure are right skewed.[3,4] Typical case durations might be 20th percentile 1.5 hours, 50th percentile 2.0 hours, and 80th percentile 4.0 hours. The difference between the 80th percentile and the 50th percentile is far larger than the difference between the 50th percentile and the 20th percentile. See http://en.wikipedia.org/wiki/Skewness.

(e) Rank-based. The "nonparametric" statistical tests generally analyze the ranks of values instead of the values themselves. The American Society of Anesthesiologists' physical status is appropriately analyzed as a ranked value. For example, patients with a score of 4 are sicker than those with a score of 2, but not twice as sick. The rank-based tests typically make few assumptions about the probability distributions of the variables being assessed. For example, some tests assume only that the probability distribution is not skewed.[e] See http://en.wikipedia.org/wiki/Non-parametric_statistics.

(f) Transformed. Data transformation is often applied for the data "to more closely meet the assumptions of" the statistical test "or to improve the interpretability or appearance of graphs." One of the most common data transformations is to take the logarithm of each value. Logarithmic transformation reduces positive skewness[d] in the data. See http://en.wikipedia.org/wiki/Data_transformation_(statistics).

(g) Propensity scores. "A propensity score is the probability of a" patient "being assigned to a particular" treatment "given a set of known covariates. Propensity score" modeling is "used to reduce selection bias" from the nonrandomized assignment of patients to treatments. In observational studies using propensity scores, the score is used to match patients to one another. See http://en.wikipedia.org/wiki/Propensity_score.

(h) Logistic regression. "Logistic regression is used to predict the probability of occurrence of an event by fitting data to" the "logit function," a transformation[f] of the probability. The logit function is predicted using regression because "it can take as an input any value from negative infinity to positive infinity, whereas" its output representing the probability of occurrence of an event "is confined to values between 0 and 1." Logistic regression can include different types of explanatory variables, both "numerical or categorical. For example, the probability that a person has a heart attack within a specified time period might be predicted from knowledge of the person's age, sex and body mass index." Importantly, logistic regression is nonlinear regression and many of the simple properties of linear regression do not apply.[2] Logistic regression is often the method used to model the propensity score.[g] See http://en.wikipedia.org/wiki/Logistic_regression.

(i) Fixed effect. "A fixed effects model is a statistical model[p] that represents … observed quantities" (e.g., mortality) "in terms of explanatory variables that are" themselves treated as nonrandom. For example, comparison of mortality at the 3 hospitals in 1 city would treat the hospitals as a fixed effect. The fixed effects model differs from "random effects models[j] and mixed models[n] in which either all or some of the explanatory variables are treated as if they arise from" random causes. See http://en.wikipedia.org/wiki/Fixed_effects_model.

(j) Random effect. Suppose that 100 multiple specialty hospitals are chosen randomly from among the thousands in a country. Suppose also that the data from 50 Medicare patients who underwent laparoscopic cholecystectomy are chosen randomly from the data for each hospital. The billed anesthesia minutes are obtained.[8] The statistical model[p] estimates the anesthesia minutes for each patient as equaling an overall constant plus a patient effect plus a hospital effect. The hospital effect is modeled as a random effect. The hospital effect is estimated for each hospital by taking the difference between the mean anesthesia minutes for all sampled patients at the hospital and the mean anesthesia minutes nationwide. The hospital effect is said to be "random" because the hospital was selected at random from among all of the multiple specialty hospitals in the country. See http://en.wikipedia.org/wiki/Random_effects_model.

(k) Sensitivity analysis. "Sensitivity analysis is the study of how the variation … in the output of a mathematical model can be apportioned … to different sources of variation in the input of the model." Sensitivity analysis "is a technique for systematically changing parameters in a model to determine the effects of such changes." "Sensitivity analysis investigates the robustness of a study" including the mathematical model. For example, the total cost of an anesthetic is highly sensitive to the compensation of the anesthesiologist. See http://en.wikipedia.org/wiki/Sensitivity_analysis.

(l) Stratified sampling. Surgical cases are randomized to receive an extra circulating nurse. Two procedures are studied, adenoidectomy and lung resection. Adding one circulating nurse to each case is expected to reduce operating room time by 10 minutes. If there were 2 groups (presence and absence of the extra nurse), the sample size per group would need to be in the hundreds because of the large standard deviations of operating room times within groups, because of the mixture of 2 procedures. In contrast, if the time difference were taken for each procedure and then the mean differences pooled, the appropriate sample size would be smaller. Such "stratification is the process of dividing members of a population into homogeneous subgroups before sampling. The strata [are] mutually exclusive. Every" patient "must be assigned to only one stratum." No patient "can be excluded." "Random or systematic sampling is applied within each stratum. This … can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population." See http://en.wikipedia.org/wiki/Stratified_sampling.

(*Continued*)

Table 2. (Continued)

(m) Quantiles. The quantiles divide "ordered data into q essentially equal-sized data subsets … The quantiles are the data values marking the boundaries between consecutive subsets." For example, the 2nd "quantile is the median" and the 5 quantiles "are called quintiles." "The 2nd of the 5 quantiles for a random variable is the value x such that the probability that the random variable will be less than x is at most 2/5 and the probability that the random variable will be more than x is at most 3/5." See http://en.wikipedia.org/wiki/Quantile.

(n) Mixed model. "A mixed model is a statistical model containing both [one or more] fixed effects[i] and random effects."[j] "They are particularly useful in settings where repeated measurements[q] are made on the same" patients "or where measurements are made on clusters of related statistical units" (e.g., hospitals). See http://en.wikipedia.org/wiki/Mixed_model.

(o) Double-blind. A "blinded experiment is a scientific experiment where some of the persons involved are prevented from knowing certain information that might lead to conscious or unconscious bias on their part, invalidating the results. For example, when asking consumers to compare the tastes of different brands of a product, the identities of the [brands] should be concealed, otherwise consumers will generally tend to prefer the brand they are familiar." "In a double-blind experiment, neither the individuals nor the researchers know who belongs to the control group and the experimental group." See http://en.wikipedia.org/wiki/Blind_experiment.

(p) Statistical model. "A deterministic system is a system in which no randomness is involved in the development of future states of the system. A deterministic model will thus always produce the same output from a given starting condition or initial state." "A statistical model" is the opposite, with "one or more random variables … related to one or more random variables. The model is 'statistical' as the variables are not deterministically but stochastically related." "Most statistical tests" are "described in the form of a statistical model." See http://en.wikipedia.org/wiki/Statistical_model.

(q) Repeated measures. "A repeated measures design refers to studies in which the same measures are collected multiple times for each subject but under different conditions." For example, "repeated measures are collected … in which change over time is assessed. Other studies compare the same measure under two or more different conditions." For example, "to test the effects of caffeine on cognitive function, a subject's math ability might be tested once after they consume caffeine and another time when they consume a placebo." For such a study, patients are randomized to different sequences. See http://en.wikipedia.org/wiki/Repeated_measures_design. "Repeated measures analysis of variance" is more specific than "repeated measures," because there are analysis of variance assumptions plus the assumption of sphericity. See http://en.wikipedia.org/wiki/Mauchly's_sphericity_test. The sphericity assumption specifies that the variance[s] of the difference of scores between any 2 times or conditions is the same as the variance of the population difference scores for any other 2 times or conditions. This assumption makes sense when the repeated measure is over different conditions, but less so when repeatedly measured over sequential time.

(s) Variance-covariance. The correlation coefficient between 2 variables equals 1 when they are identical and 0 when independent. The covariance measures "how much two variables change together," but without the normalization of the correlation coefficient. Whereas the correlation coefficient is unitless, the covariance has the original units squared. See http://en.wikipedia.org/wiki/Covariance. "Variance is a special case of the covariance when the two variables are identical." When a repeated-measures[q] analysis is performed, the statistical model[p] includes the covariance of measurements at all times. See http://en.wikipedia.org/wiki/Covariance_matrix.

(t) Meta-regression. "A meta-analysis combines the results of several studies that address a set of related research hypotheses." For example, "a weighted" mean may be taken of each study's mean value, with "the weighting … related to sample sizes within the individual studies." Meta-regression "uses different characteristics of the study" to explore heterogeneity in the mean value among studies. See http://en.wikipedia.org/wiki/Meta-analysis.