Review Article

Statistics in clinical research: Important considerations

Howard Barkan

Affiliated Researcher and Consulting Statistician, School of Public Health, University of California Berkeley, Berkeley, CA 94704-7380, Saybrook University, Oakland, CA 94612, USA

ABSTRACT

Statistical analysis is one of the foundations of evidence-based clinical practice, a key in conducting new clinical research and in evaluating and applying prior research. In this paper, we review the choice of statistical procedures, analyses of the associations among variables and techniques used when the clinical processes being examined are still in process. We discuss methods for building predictive models in clinical situations, and ways to assess the stability of these models and other quantitative conclusions. Techniques for comparing independent events are distinguished from those used with events in a causal chain or otherwise linked. Attention then turns to study design, to the determination of the sample size needed to make a given comparison, and to statistically negative studies.

Received: 10-06-14
Accepted: 25-08-14

Key words: Clinical research, measurement types, predictive modelling, statistical analysis

INTRODUCTION

Clinicians examine and intervene with individual patients. The understandings of the clinical challenges they will need to address, of the likely past and future courses of the clinical conditions they are seeing, and evaluations of the effectiveness and risks of their clinical actions and strategies are all based on consideration of the characteristics and histories of clients similar to the one they're now seeing and with whom they may be about to intervene. Statistics is a key tool linking the multiplicity of potential observations of every client with the more abstract concepts of clinical entities, natural histories, clinical response and risks. These more abstract constructs are the foundation on which clinical decisions rest. On a more applied level, clinicians need to understand statistics well enough to follow and evaluate the empirical studies that provide an evidence base for clinical practices. Studies conducted decades ago found major lacunae in physicians' knowledge of statistics.[1-3] This is a problem more recent studies have found to be only somewhat reduced in magnitude.[4-6] It leads clinicians to mistrust, misunderstand and ignore the statistics in journal articles.[7]

There are several aspects of statistical concepts, methods and their application which are key to their understanding and interpretation. These have been presented for practitioners in major clinical journals by excellent clinicians and statisticians (for initial papers in such series, cf. e.g.[8-17]). We will present these concepts and methods with goals of strengthening clinicians' comprehension of statistical aspects of the clinical literature, their evaluation of the strengths and weaknesses of the analyses presented, and their active participation in research. The presentation in this paper is rooted in experience gained from studies conducted by the author[18-20] and the clinical literature. We hope to help make these inherently abstract statistical concepts and techniques more intelligible in the applied world of clinical practice. We will begin by discussing aspects of measurement, sampling, and analytic goal that guide the choice of statistical techniques. The discussion will then turn to aspects of analytic design and

Access this article online
Website: www.annals.in
DOI: 10.4103/0971-9784.148325
Quick Response Code:

Address for correspondence: Dr. Howard Barkan, 1411 Arch Street, Berkeley, CA 94708, USA. E-mail: howardbarkan@cs.com

conduct, which impact important finer details of the project's conduct and of the interpretation of its results. The reader is referred to the paper series referenced above [8–17] for more detailed discussions of particular statistical techniques.

Clinical processes are real world. Statistics is abstract. Two-way translation is important

Biological and clinical entities are complex and changing, multi-dimensional structures and processes, which evolve over time. All research works begin by selecting particular features of physical objects and segments of processes, which will be used in the research to represent those structures and processes.[11] These selected observations operationalize the abstract concept of a clinical entity or process into specified measurements. Statistics works with these operationalizations, modeling and analyzing properties and processes which are shared among groups of observations. Note that, the external validity of the results of statistical analyses, while key to the value of those results, is importantly a function of measurement and sampling - that is, what was measured how, in which subjects, at which times, and how well those selected measurements represent the clinical entities and processes which the research is investigating.[21] A perfectly chosen and executed analysis will be at best misleading if it is conducted of the wrong data or data collected using an inaccurate measurement technique, or at the wrong time, and so on. To quote the frequent aphorism from introductory statistics courses, "Garbage in, garbage out." We will discuss the analysis of appropriately selected and measured data. Evaluations of the validity of the measures collected as representations, of the modeling of causal processes, and of the generalizability of the results are all important to the value of statistical analyses but beyond the scope of this paper.

Measurement scaling

Certain aspects of measurement and sampling are key to which statistical techniques are appropriate. The first attribute, which indicates the appropriateness of and hence guides the choice among statistical procedures, is the scaling of the measurements being treated as variables in the analysis. Statistics represents measurements as scales. In terms of the appropriateness of statistical techniques, the key differentiation among scaling techniques is mathematical: What each number represents and which mathematical analyses of those numbers are valid.[9,11,14,23,24]

Measurements can be classified as using nominal, ordinal, and interval scalings. Nominal scalings use distinct and mutually-exclusive numbers to name each category of observation. Nominal scalings only classify observations. The numbers assigned in a nominal scale carry no further information about magnitude. Set theory, which deals with which observations belong in which groups and with how groups overlay, is the only mathematics appropriate for nominal scales. Clinical examples of nominal scalings include any notation that a disease is (simply) present or absent, a the binary classification used in calculating incidence and prevalence rates and the sensitivity and specificity of diagnostic tests, demographic measures (such as gender and ethnic group), and disease classification systems such as the International Classification of Disease (ICD)-10 and the Diagnostic and Statistical Manual of Mental Disorders-5. Sensitivity and specificity, key indices of the strength of diagnostic findings as evidence of a disease, both begin by treating the finding and the disease as binomial nominal variables. Binomials are attributes that are either present or absent.[25-27]

Ordinal scalings are mathematically the next more complex. Ordinal scalings place observations in order-say from least to most-but are not able to specify or compare the differences between pairs of measurements. Many clinical measurements and indices and many psychological and attitude measurements are ordinally scaled: e.g. tumor grade, pain scales, and Likert attitude scales. Disease stage is an example of an ordinal scaling. Stage 4 cancers are "worse" than stage 3 cancers which are in turn worse than stage 2 cancers, but the ordinal scaling of staging does not indicate how much worse. It is impossible to say whether the difference between stage 4 and stage 3 is more or less than the difference between stage 3 and stage 2 based on the assigned stage alone. That is, the statement that one stage is "worse" than another derives from the association of stage differences with other factors such as duration of survival rather than on the measurement of stage itself. Ordinal scalings add the mathematics of inequalities to set theory as permissible mathematical operations.

Interval scalings are mathematically the most complex of the measurement scales used. Interval scales place observations in order and specify both the magnitude of individual measurements and the distance between pairs of measurements. Interval scalings permit all of

the basic arithmetic operations and the calculations based on those operations. Many widely used clinical observations are intervally scaled: e.g. anthropometric measurements of height and weight, blood pressure, and duration of time intervals.

Two other scaling options considerations are frequently mentioned for interval scales. The first is whether the source measurements are discreet (e.g. number of children in the household) or continuous (e.g. blood pressure). This distinction bears on the source measurement and may influence how collected data are displayed graphically, but has no influence on the choice or calculation of statistical analyses. The second difference among interval scales is whether or not the scale has a true "0" point. Those with a true "0" points are sometimes called ratio scales because the presence of a true "0" point makes division and hence the calculation of ratios possible. Consider, for example, temperature. The Kelvin scale has a true zero point at absolute zero and hence is a ratio scale. The Centigrade and Fahrenheit scales have a zero point that's mathematically arbitrary and hence are interval scales. This difference bears on which conclusions regarding these measurements are meaningful. For example, it is meaningful to say that the temperature of 30°K is half a temperature of 60°K while it is not valid to make the same statement regarding 30°F versus 60°F. This difference has no bearing on the choice of statistical procedures to analyze these data.

This mathematical type of scaling is one of the principal determinants of the appropriateness of a particular statistical analysis for a particular dataset.[28] In general, statistical analyses which can be conducted of mathematically simpler scales, say nominal scales, can also be conducted of more complex scales. For example, the mode, i.e. identification of the most frequent observation, which is the principal statistic describing central tendency for nominally scaled variables, can also be used to describe distributions of ordinally and intervally scaled variables. On the other hand, statistical analyses designed for more complex scales often cannot be applied to mathematically simpler scales. For example, calculation of an average depends on the ability to add and divide the observed measurements. These mathematical operations are not valid with ordinal and nominal scalings, making invalid the operations involved in calculating the average of such a scaling in a sample. Note that the greater power of the analyses available for interval scalings leads to a frequent temptation to treat measurements such as tumor stage which are appropriately scaled ordinally as though they were scaled intervally.

Descriptive statistics and measurement scaling: Single variables

Examinations of single variables use descriptive statistics to characterize the central tendency, the single best description of the sample of measurements, and variability. Descriptive statistics for single variables play important roles in research. Descriptive statistics summarize characteristics of the study and control groups in randomized trials.[18] To evaluate the baseline comparability of the an investigation's study and control groups, the proportions are examined when comparing nominally scaled variable such as gender. They are at the core of clinically relevant indices of prevalence and incidence, and of the evaluation of the sensitivity and specificity of diagnostic findings as evidence of particular conditions. The median is can also be examined when comparing the ordinally scaled urgency. While Averages are can be examined when comparing intervally scaled characteristics: e.g. groups members' age, serum albumin, and platelet count and other key hematologic indices.

In most general terms, a form of descriptive statistical analysis which is valid for simpler mathematical scalings can be used with mathematically more complex scalings. For example, the category containing the highest proportion of a nominal variable is termed the mode. The mode is a valid analysis of nominally scaled variables. We can count the number of patients assigned each ICD-9 coded diagnosis. We can then compare these counts to evaluate which diagnosis was most frequent that is, the mode. The mode can also be used to describe variables that are ordinally scaled - e.g. which stage of lung cancer is most frequent - and intervally scaled - e.g. what number of children/family is most frequent. In contrast, statistics designed specifically for more complex scalings may be invalid for measurements using mathematically simpler scalings. For example, it is valid to calculate the mean and standard deviation of the number of distant metastases/patient because our count of the number of distant metastases is intervally scaled: The difference between 0 and 1 distant metastases equals the difference between 3 and 4 distant metastases which equals one.[22] In contrast, we cannot calculate the average lung cancer stage because we cannot add or divide stage measurements: Is it at all meaningful to say that stage 2 lung cancer is twice stage 1 lung cancer? The situation is even more clouded with nominally scaled variables. The numbers used as codes in the ICD-9 carry

no direct implication of magnitude. It is not meaningful to say that the diagnosis of reticulosarcoma is twice the diagnosis of leptospirosis icterohemorrhagica because reticulosarcoma's ICD-9 code of 200.0 is twice, the leptospirosis ICD-9 code of 100.0.[23,29]

Descriptive statistics and measurement scaling: Multiple variables

Let us now turn our attention to the associations among variables, first paying attention to how we describe that association. The strength of the association between two variables is described by correlation coefficients.[30,31] Correlation coefficients describe the strength of association between two variables of the same mathematical type. Correlation coefficients typically range from "0" indicating no association to "−1" and "1", indicating perfect association. The square of the correlation coefficient can be interpreted as the proportion of the variance of one variable that is predicted by the other variable. The square of "1" equals the square of "−1" equals "1," indicating perfect association. The most frequently used correlation coefficients are phi and Cramer's V for nominal variables, Spearman's rho (or rank-order) correlation for ordinal variables, and Pearson's *r* (or product-moment) correlation for interval variables. Kappa is also often used for binomial nominal variables. Binomial variables are nominal variables with only two values: e.g. gender and the presence versus absence of a characteristic or disease. Kappa adjusts in its calculation for the agreement expected by chance alone.[32] This has made kappa a useful index in investigations of inter-observer agreement among radiologists and other clinicians (there has been some argument about this interpretation of kappa, cf.[33]). Note that agreement does not imply accuracy. Accuracy, assessed for binary classifications by sensitivity, specificity, and receiver operating characteristic curves, will not be discussed further in this paper.[25-27,34,35]

For all but nominal variables, the sign of the correlation coefficient indicates the direction of the association. Positive correlation coefficients describe situations in which increases in value of one of the variables are associated with increases in the other variable, while negative coefficients describe situations in which increases in one of the variables are associated with decreases in the other. Correlation-based analyses using techniques such as factor analysis can be used to examine the associations among multiple measures used to investigate single events or conditions.[36,37] This technique can identify groupings and key measures, potentially reducing the length and increasing the efficiency of diagnostic evaluations.[38,39] Patterns found in factor analysis can be helpful in exploring biological interactions and indicate particular groupings which may have clinical implication.[40-42]

Measurement timing

The discussion so far has carried the implicit assumption that we are able to measure the entire course of the events we are studying. That may be true for many of the acute clinical events and processes in which cardiac anesthesiology plays a major role. However, this is clearly true neither for all long term processes in cardiac anesthesia nor for those iatrogenic effects whose appearance is delayed, nor for cardiology, nor for clinical processes generally. Clinical and research data are often gathered within a limited time frame while the processes to which clinical attention is being given, and those which are being studied continue beyond that time frame's boundaries. The techniques of survival analysis and life-table statistics have been developed to address these challenges presented by what is termed "right censoring."[43-46] Right censoring exists when a study is investigating a process that has reached a conclusion in some, but not all of the subjects when the study ends hence censoring information about that outcome. In situations such as this, the sample size of those at risk for a study's terminal event varies over the course of the study because that size is reduced by "1" every time one of the study's terminal those events (say tumor recurrence or mortality) occurs, removing the person experiencing the event from the group at risk for it. Life-table analyses typically examine median time to the target event to avoid being biased by the long times to event of those in the sample who have not experienced the event by the time the study concludes and whose experience is right-censored. Life-table experience is typically depicted using Kaplan-Meier survival curves, where "survival time" is taken to signify time to the process designated's final effect (e.g. re-infection, tumor recurrence or mortality). Appropriate evaluation of statistical significance also uses techniques discussed below which take this right-censorship into account. It is important that studies whose samples are right-censored use such life-table based techniques. Studies in that situation that calculate survival time by averaging time to the terminal events which have occurred will produce biased estimates unless all of those terminal events have occurred because right-censorship will be excluding those with the potentially longest survival times.

Modeling associations and prediction

Correlations measure the strength and, for all types except nominal variables, the direction of associations

between variables. Regression modeling provides the tools for making those predictions from one or more independent variables to the dependent variable.[30,31,47-50] The measurement and the completeness of the measurement of the dependent variable indicate which form of regression modeling is appropriate. If the dependent variable is a binomial, that is, a nominal variable with only two values, and it is known whether or not each member of the sample experienced that outcome, multiple logistic regression is used to model the effects of the independent variables on the odds ratio of experiencing that outcome.[51,52] When the outcome condition is relatively rare and with some other constraints, these odds ratios can be treated as estimates of the relative risk each independent variable carries for the outcome. This model is appropriate for outcomes in, say, a study of surgical intervention in which the outcome of interest is short term and can be predicted to have occurred before discharge from hospital. In contrast, the Cox proportional hazards model and regression is are used when the outcome data are right censored., that is, when the outcome status of all subjects is not known (often because insufficient time has passed for the outcome to have occurred in all subjects in whom it may eventually occur). This is likely to be the case, for example, if the study is investigating delayed effects after therapeutic interventions such as postsurgical survival in cancer patients. Cox regression models the risk of the target outcome as a hazard function which is a function of time and of the independent variables included in the model. The final principal form of regression modeling is (multiple) linear regression, which predicts a dependent variable measured on an interval scale based on the values of one or more predictors. For example, linear regression can be used to model the association of the natural log of urea with age[30] (taking the natural log of urea made the relationship of urea with age a straight line). Linear regressions predict straight lines (or planes or their multi-dimensional analogs). There are constraints on the type of distribution and on the associations among variables suitable for linear regression analysis.[30] Many clinical variables have exponential or other nonlinear associations. Discussion of the regression modeling of these processes and of their associations is beyond the scope of this paper.[53]

Result likelihood and stability

Clinical decisions and research need to move beyond the initial sample of measurements (of say the initial patient or group of patients) to reach more generalized conclusions. Say a change is noted in laboratory measurement following an operative procedure. How likely is it that other patients undergoing that procedure will experience the same change? Is that change other than the difference that would be seen in patients with the same clinical condition who are measured twice, but who do not undergo that procedure? What is the range of change in that laboratory measurement, which can be expected in future patients who do and who do not undergo that procedure?

These questions explore the extent to which we can generalize from our particular clinical observations and the trustworthiness of those generalizations. These questions are in the arena of statistical inference. There have been many presentations of the general logic underlying statistical inference (cf. e.g.[54-57]). The reader is referred to those sources, and to any classical statistics or biostatistics text for the logic underlying classical tests of statistical significance. We will now first discuss alternatives to the point comparison represented by classical significance testing. Then, given the widespread use of classical significance testing, we will discuss several modifications necessary for its appropriate use in clinical studies.

Classical tests of significance assess the likelihood of the study's actual results given a set of assumptions about the sources of the measures being compared. The tests are designed to support a point judgment about the likelihood of those source groups being identical. The statistical significance test result evaluates the likelihood of the results obtained were the data drawn from identical groups, saying nothing about the magnitude or stability of any differences that were actually found. Further, these tests refer to an arbitrary cut-point (usually $P < 0.05$) to support conclusions about similarity versus difference. There is a long-standing argument that analyses should estimate the range of inter-group differences consistent with the collected data rather than ending with a single statement regarding statistical significance.[58-63] Given that significance tests provide a point statement, while confidence intervals express a range of estimation, some advocate reporting both (e.g.[60,61]). Confidence intervals can also be calculated using what are termed "Bayesian" techniques. These techniques initially presented by Thomas Bayes (1702–1761) treat probability as a statement of degree of belief in a statement rather than as an estimate of the frequency. In clinical practice, Bayesian techniques are used to calculate the predictive value (positive)

of a diagnostic finding given prior beliefs about the finding's sensitivity and specificity and about the prevalence of the diseases being considered.[27] In the context of statistical inference, Bayesian techniques take into account prior beliefs about the statistics being compared by the test of significance. This is in contrast to classical tests of statistical significance and calculations of confidence intervals that are based only on the sets of actual measurements and assumptions about the underlying population distributions.[63-65] The continual reassessment method, first proposed by O'Quigley in 1990, applies Bayesean techniques to toxicity data from dose-finding trials.[66] Bayesian techniques are used to reapply new trial data cyclically to prior toxicity estimates (from the trial or initially from elsewhere) to re-estimate dose-toxicity curves and estimate the optimal dose in Phase 1 clinical trials.[67-71]

The above paragraph noted that confidence intervals can be used to evaluate a likelihood of inter-group differences. These confidence intervals estimating the magnitude of the inter-group difference go beyond traditional point computations of statistical significance which only refer to the likelihood of the particular difference tested to estimate the magnitude of the inter-group difference. They also estimate the expected stability of associations between variables. Please note that confidence intervals can also be calculated around other statistics, ranging from the proportions and means calculated as descriptive statistics through correlation coefficients to regression coefficients. In each case, the confidence interval predicts the stability of the point statistic calculated using a defined sample. The confidence interval estimates the boundaries likely to include (desired target) proportions (often 95%) of future similar measurements made from that statistical population.

Independent versus paired measurement

While there is serious discussion about alternatives to classical tests of statistical significance as evaluations of the generalizability of findings as noted above, these classical tests continue to be widely used.[72-74] Several issues regarding the conduct of these tests and the interpretation of their findings recur repeatedly. The first issue is whether or not the measures being compared are independent.[54,75,76] Tests of the statistical significance of differences in paired measurements differ from tests of independent measurements because in paired observations the first set of measurements is a precise prediction against which the second measurement is compared. Any difference or any difference in a specified direction is potentially of interest when comparing independent samples. The sets of measurements in repeat measurement of the same subjects are obviously related, with the second measurements being departures from first measurements that are already in the sample study. Paired analyses are also needed when the selection of samples is matched. Matching is often used in epidemiological studies to maximize comparability of the samples on all factors other than the factor whose influence is being compared (i.e. a risk factor in a cohort study or clinical outcome in a study using a case–control design).

Adjustment for multiple outcomes

Classical tests of statistical significance assume there has been only a single examination of the relationship being investigated. This assumption is often violated. It is violated when there are a series of separate examinations of the association of a single dependent variable with multiple potential independent variables, or of a single independent variable with multiple potential effects.[77,78] This can also happen by design in randomized controlled trials, when the Data Safety Monitoring Committee by protocol reviews the data at prespecified intervals. Associations can also be examined during the study's initial design phase then reexamined in the full study. This is problematic because each analysis in multiple comparison which uses a $P < 0.05$ threshold has a 1 in 20 chance of producing a false positive result. In essence, this means that if 20 tests are performed there's a virtual certainty that at least one will yield a false positive result.[79] This risk of a false positive can be mitigated in the design by adjusting the threshold for declaring statistical significance. The simplest but most conservative approach, the Bonferroni adjustment, divides the target P value by the number of comparisons made. Equally rigorous but less stringent techniques such as the false detection rate are now in use.[80] All these techniques adjust individual comparison thresholds so the final statistical significance for all comparisons combined is $P < 0.05$.

Statistical power and negative studies

Clinical studies can only be effective if the sample size is large enough to give the study a reasonable chance of finding the association as hypothesized by the study's designers which it is investigating. The chance of a study yielding a statistically significant result if its

hypothesis is supported is termed its statistical power. There are established methods for calculating statistical power for studies given the planned analysis, sample size, and assumptions about the population from which the sample will be drawn.[81,82]

If studies achieve statistically significant results, the question of statistical power is moot. The power was de facto adequate. The real challenge is when study results fail to reach statistical significance. Over a period of decades, examinations of studies with statistically nonsignificant results have found the studies to have been underpowered.[83-85] Paralleling Freiman's *et al*.[85] earlier study, Moher *et al*.[84] reviewed 383 randomized trials published in three major journals, finding 102 which had failed to reach statistical significance. Of the 70 of these negative trials which examined binary or intervally scaled primary outcomes, only 16 (22.9%) had 80% power to detect a 25% difference in outcome rates, and only 36 (51.4%) had 80% power to detect the easier to find 50% difference in outcome rates. This problem continues. In a recent study examining papers published in British orthopedic journals, Sexton *et al*.[83] found 49 papers reporting findings that failed to reach statistical significance. Only three (6.1%) of those papers reported a statistical power analysis and had a sample size large enough to give the study adequate statistical power.

Comments

Clinicians practice with individual patients, while conclusions about care practices almost always involve considerations of aspects of the clinical courses followed by many. Statistics is one of the important tools to help bridge this gap. This paper has reviewed certain selected key aspects of the statistical approach to clinical events and care. Please note that many of the studies used as examples are clinically illuminating and methodologically sound. However, there are also aspects of the design and execution which were the subject to recurring methodological weaknesses. These include statistical power analysis and sample size planning and the selection and conduct of appropriate analyses in light of the sampling and measurements used. Routine conduct of pilot studies before full studies are initiated could help strengthen study designs and lessen the threat of such methodological weaknesses.

Hopefully the clinical reader will use these tools to understand the strengths and weaknesses of past work. One central goal in conducting methodologically robust studies is to build a sound evidence base for clinical care. These quantitative tools can contribute to building such a solid foundation.

ACKNOWLEDGMENTS

The authors acknowledge the sincere efforts of Dr. Dave Nicholas in reviewing and developing the manuscript.

REFERENCES

1. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? Stat Med 1987;6:3-10.
2. Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. Am J Med 1981;71:991-8.
3. Weiss ST, Samet JM. An assessment of physician knowledge of epidemiology and biostatistics. J Med Educ 1980;55:692-7.
4. Best AM, Laskin DM. Oral and maxillofacial surgery residents have poor understanding of biostatistics. J Oral Maxillofac Surg 2013;71:227-34.
5. Bookstaver PB, Miller AD, Felder TM, Tice DL, Norris LB, Sutton SS. Assessing pharmacy residents' knowledge of biostatistics and research study design. Ann Pharmacother 2012;46:991-9.
6. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. JAMA 2007;298:1010-22.
7. Hack JB, Bakhtiari P, O'Brien K. Emergency medicine residents and statistics: What is the confidence? J Emerg Med 2009;37:313-8.
8. du Prel J, Rohrig B, Blettner M. Statistical methods in medical research. Dtsch Arztebl Int 2009;106:99.
9. Curran-Everett D. Explorations in statistics: Standard deviations and standard errors. Adv Physiol Educ 2008;32:203-8.
10. Overholser BR, Sowinski KM. Biostatistics primer: Part I. Nutr Clin Pract 2007;22:629-35.
11. Applegate KE, Crewson PE. An introduction to biostatistics. Radiology 2002;225:318-22.
12. Proto AV. Radiology 2002 – Statistical concepts series. Radiology 2002;225:317.
13. Whitley E, Ball J. Introducing the Critical Care Forum's ongoing review of medical statistics. Crit Care Forum 2002;6:3-4.
14. Driscoll P, Lecky F, Crosby M. An introduction to everyday statistics – 1. J Accid Emerg Med 2000;17:205-11.
15. Healy MJ. Populations and samples. Arch Dis Child 1991;66:1355-6.
16. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995;152:27-32.
17. Altman DG. Statistics and ethics in medical research: V – Analysing data. Br Med J 1980;281:1473-5.
18. Hecht-Dolnik M, Barkan H, Taharka A, Loftus J. Hetastarch increases the risk of bleeding complications in patients after off-pump coronary bypass surgery: A randomized clinical trial. J Thorac Cardiovasc Surg 2009;138:703-11.
19. Spring DB, Barkan HE, Pruyn SC. Potential therapeutic effects of contrast materials in hysterosalpingography: A prospective randomized clinical trial. Kaiser Permanente Infertility Work Group. Radiology 2000;214:53-7.
20. Farley M, Barkan H. Somatization, dissociation, and tension-reducing behaviors in psychiatric outpatients. Psychother Psychosom 1997;66:133-40.
21. Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin Company; 2001.
22. Rasmussen F, Lindequist S, Larsen C, Justesen P. Therapeutic effect of hysterosalpingography: Oil- versus water-soluble contrast media – A randomized prospective study. Radiology 1991;179:75-8.

23. Spriestersbach A, Röhrig B, du Prel JB, Gerhold-Ay A, Blettner M. Descriptive statistics: The specification of statistical measures and their presentation in tables and graphs. Part 7 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009;106:578-83.
24. Whitley E, Ball J. Statistics review 1: Presenting and summarising data. Crit Care 2002;6:66-71.
25. Langlotz CP. Fundamental measures of diagnostic examination performance: Usefulness for clinical decision making and research. Radiology 2003;228:3-9.
26. Brismar J, Jacobsson B. Definition of terms used to judge the efficacy of diagnostic tests: A graphic approach. AJR Am J Roentgenol 1990;155:621-3.
27. McNeil BJ, Keller E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975;293:211-5.
28. du Prel JB, Röhrig B, Hommel G, Blettner M. Choosing statistical tests: Part 12 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010;107:343-8.
29. Greenhalgh T. How to read a paper. Statistics for the non-statistician. I: Different types of data need different statistical tests. BMJ 1997;315:364-6.
30. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. Crit Care 2003;7:451-9.
31. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology 2003;227:617-22.
32. Kundel HL, Polansky M. Measurement of observer agreement. Radiology 2003;228:303-8.
33. Uebersax JS. Modeling approaches for the analysis of observer agreement. Invest Radiol 1992;27:738-43.
34. Bewick V, Cheek L, Ball J. Statistics review 13: Receiver operating characteristic curves. Crit Care 2004;8:508-12.
35. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229:3-8.
36. Ismail K. Unravelling factor analysis. Evid Based Ment Health 2008;11:99-102.
37. Genser B, Cooper PJ, Yazdanbakhsh M, Barreto ML, Rodrigues LC. A guide to modern statistical analysis of immunological data. BMC Immunol 2007;8:27.
38. Akeroyd MA, Guy FH, Harrison DL, Suller SL. A factor analysis of the SSQ (Speech, Spatial, and Qualities of Hearing Scale). Int J Audiol 2014;53:101-14.
39. Stein DJ, Rothbaum BO, Baldwin DS, Szumski A, Pedersen R, Davidson JR. A factor analysis of posttraumatic stress disorder symptoms using data pooled from two venlafaxine extended-release clinical trials. Brain Behav 2013;3:738-46.
40. Manhenke C, Ørn S, von Haehling S, Wollert KC, Ueland T, Aukrust P, *et al.* Clustering of 37 circulating biomarkers by exploratory factor analysis in patients following complicated acute myocardial infarction. Int J Cardiol 2013;166:729-35.
41. Dossus L, Lukanova A, Rinaldi S, Allen N, Cust AE, Becker S, *et al.* Hormonal, metabolic, and inflammatory profiles and endometrial cancer risk within the EPIC cohort – A factor analysis. Am J Epidemiol 2013;177:787-99.
42. Zandieh A, Kahaki ZZ, Sadeghian H, Pourashraf M, Parviz S, Ghaffarpour M, *et al.* The underlying factor structure of National Institutes of Health Stroke scale: An exploratory factor analysis. Int J Neurosci 2012;122:140-4.
43. Abd ElHafeez S, Torino C, D'Arrigo G, Bolignano D, Provenzano F, Mattace-Raso F, *et al.* An overview on standard statistical methods for assessing exposure-outcome link in survival analysis (Part II): The Kaplan-Meier analysis and the Cox regression method. Aging Clin Exp Res 2012;24:203-6.
44. Tripepi G, Torino C, D'Arrigo G, Bolignano D, Provenzano F, Zoccali C. An overview of standard statistical methods for assessing exposure-outcome link in survival analysis (Part I): Basic concepts. Aging Clin Exp Res 2012;24:109-12.
45. Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan-Meier curves. Otolaryngol Head Neck Surg 2010;143:331-6.
46. Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. Crit Care 2004;8:389-94.
47. Curran-Everett D. Explorations in statistics: Regression. Adv Physiol Educ 2011;35:347-52.
48. Slinker BK, Glantz SA. Multiple linear regression: Accounting for multiple simultaneous determinants of a continuous dependent variable. Circulation 2008;117:1732-7.
49. Gareen IF, Gatsonis C. Primer on multiple regression models for diagnostic imaging research. Radiology 2003;229:305-10.
50. Guyatt G, Walter S, Shannon H, Cook D, Jaeschke R, Heddle N. Basic statistics for clinicians: 4. Correlation and regression. CMAJ 1995;152:497-504.
51. Cummings P. The relative merits of risk ratios and odds ratios. Arch Pediatr Adolesc Med 2009;163:438-45.
52. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Crit Care 2005;9:112-8.
53. Chen HC, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC Med Res Methodol 2012;12:102.
54. Zou KH, Fielding JR, Silverman SG, Tempany CM. Hypothesis testing I: Proportions. Radiology 2003;226:609-13.
55. Tello R, Crewson PE. Hypothesis testing II: Means. Radiology 2003;227:1-4.
56. Whitley E, Ball J. Statistics review 3: Hypothesis testing and *P* values. Crit Care 2002;6:222-5.
57. Driscoll P, Lecky F, Crosby M. An introduction to statistical inference – 3. J Accid Emerg Med 2000;17:357-63.
58. Curran-Everett D. Explorations in statistics: Confidence intervals. Adv Physiol Educ 2009;33:87-90.
59. du Prel JB, Hommel G, Röhrig B, Blettner M. Confidence interval or *P* value?: Part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009;106:335-9.
60. Greenhalgh T. How to read a paper. Statistics for the non-statistician. II: "Significant" relations and their pitfalls. BMJ 1997;315:422-5.
61. Guyatt G, Jaeschke R, Heddle N, Cook D, Shannon H, Walter S. Basic statistics for clinicians: 2. Interpreting study results: Confidence intervals. CMAJ 1995;152:169-73.
62. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: Estimation rather than hypothesis testing. Br Med J (Clin Res Ed) 1986;292:746-50.
63. Rothman KJ. A show of confidence. N Engl J Med 1978;299:1362-3.
64. Goodman SN. Toward evidence-based medical statistics 1: The *P* value fallacy. Ann Intern Med 1999;130:995-1004.
65. Goodman SN. Toward evidence-based medical statistics 2: The Bayes factor. Ann Intern Med 1999;130:1005-13.
66. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for phase 1 clinical trials in cancer. Biometrics 1990;46:33-48.
67. Iasonos A, O'Quigley J. Adaptive dose-finding studies: A review of model-guided phase I clinical trials. J Clin Oncol 2014;32:2505-511.
68. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual reassessment method for phase I studies. Stat Med 1995;14:1149-61.
69. Møller S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. Stat Med 1995-30;14:911-22.
70. Piantadosi S, Fisher JD, Grossman S. Practical implementation of a modified continual reassessment method for dose-finding trials. Cancer Chemother Pharmacol 1998;41:429-36.
71. Storer BE. An evaluation of phase I clinical trial designs in the continuous dose-response setting. Stat Med 2001;20:2399-408.
72. Arnold LD, Braganza M, Salih R, Colditz GA. Statistical trends in the Journal of the American Medical Association and implications for training across the continuum of medical education. PLoS One 2013;8:e77301.

73. Hellems MA, Gurka MJ, Hayden GF. Statistical literacy for readers of Pediatrics: A moving target. Pediatrics 2007;119:1083-8.
74. Horton NJ, Switzer SS. Statistical methods in the journal. N Engl J Med 2005;353:1977-9.
75. Whitley E, Ball J. Statistics review 5: Comparison of means. Crit Care 2002;6:424-8.
76. Whitley E, Ball J. Statistics review 6: Nonparametric methods. Crit Care 2002;6:509-13.
77. Tyler KM, Normand SL, Horton NJ. The use and abuse of multiple outcomes in randomized controlled depression trials. Contemp Clin Trials 2011;32:299-304.
78. Curran-Everett D. Multiple comparisons: Philosophies and illustrations. Am J Physiol Regul Integr Comp Physiol 2000;279:R1-8.
79. Bland JM, Altman DG. Multiple significance tests: The Bonferroni method. BMJ 1995;310:170.
80. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B 1995;57:289-300.
81. Whitley E, Ball J. Statistics review 4: Sample size calculations. Crit Care 2002;6:335-41.
82. Curran-Everett D. Explorations in statistics: Power. Adv Physiol Educ 2010;34:41-3.
83. Sexton SA, Ferguson N, Pearce C, Ricketts DM. The misuse of 'no significant difference' in British orthopaedic literature. Ann R Coll Surg Engl 2008;90:58-61.
84. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994;272:122-4.
85. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med 1978;299:690-4.Vastrit patquam culintem sper licit igit ve, vervivit,

Cite this article as: Barkan H. Statistics in clinical research: Important considerations. Ann Card Anaesth 2015;18:74-82.

Source of Support: Nil, **Conflict of Interest:** None declared.