Top 10 mistakes made in research

Eric McGinnis,[1] Nancy M. Heddle,[2,3] and Andrew W. Shih [1,4]

Modern medical decision making is dependent on high-quality research to inform evidence-based practice. As the rate at which medical literature is produced increases, there is potential for the publication of studies with flawed research methods and inappropriate interpretations of data. In this article, we outline common mistakes made in research that, if avoided, can improve the quality of published literature.

MISTAKE 1. NOT DEVELOPING A GOOD RESEARCH QUESTION

The research question is the foundation on which the hypotheses and study protocols are designed and should be established early in the planning stages of a study. Selection of a good research question can be challenging and requires consideration of what is clinically or scientifically relevant and what can practically be investigated. Useful frameworks for developing and refining a research question include the PICOT format, commonly used in interventional studies, and the FINER criteria (Fig. 1).[1] Further discussion on research question development can be found in a previous Clinical Research Focus article.[2]

ABBREVIATIONS: ITT = intention to treat; PP = per protocol

From the [1]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada; [2]Department of Medicine, McMaster University, Hamilton, Ontario, Canada; [3]McMaster Centre for Transfusion Research, McMaster University, Hamilton, Ontario, Canada; and the [4]Vancouver Coastal Health Authority, Vancouver, British Columbia, Canada.

Address reprint requests to: Andrew Shih, MD, FRCPC, DRCPSC, MSc, Department of Pathology, Vancouver General Hospital, JPP1, 855 W 12th Ave, Room 1553, Vancouver, BC V5Z 1M9, Canada; e-mail: andrew.shih@medportal.ca.

Received for publication July 10, 2018; revision received August 30, 2018; and accepted September 2, 2018.

doi:10.1111/trf.14982

TRANSFUSION 2018;58;2478–2482

MISTAKE 2. NOT REVIEWING EXISTING LITERATURE BEFORE INITIATING A STUDY

A review of existing knowledge in the field is crucial in the planning stages of research. This should focus on high-quality methodological evidence, possibly including systematic reviews and meta-analyses (Fig. 2). Although systematic reviews and meta-analyses can be valuable tools, their quality and content depends on the expertise and biases of the authors, and the value of careful appraisal of the primary literature cannot be overemphasized. An in-depth review including gray literature (literature produced by entities other than dedicated publishers) and clinical trials registries can be informative as well.

Review of the literature provides important background for research question formulation and is helpful in considerations such as the following: whether similar studies have been done before, where knowledge gaps exist, appropriate sample sizes, and background information that might be important for readers. Failing to perform a search of existing literature and report relevant previous findings is poor research practice and is potentially unethical, given it may lead to needless duplication of clinical investigations. An example in transfusion medicine was inadequate citation of previous literature documenting efficacy of aprotinin for perioperative bleeding, which led to numerous similar randomized controlled trials that were unnecessary.[3]

MISTAKE 3. NOT CALCULATING (AND REPORTING) THE REQUIRED SAMPLE SIZE

When designing a study, it is important to consider the potential for the following:

- Type I error: inappropriately rejecting the null hypothesis (the hypothesis being tested, a statement generally accepted as true until proved otherwise), or rate of false positivity
  - Error rate of 5% is generally considered acceptable (as such, a p value of 0.05 or less is generally considered statistically significant)
- Type II error: inappropriately failing to reject the null hypothesis, or false negativity
  - Error rate of 10 to 20%, or power of 80 to 90% (where power = 1, type 2 error rate), is generally considered acceptable

PICOT format	FINER criteria
Population ❑ Patient characteristics, including (but not limited to) age, diagnosis, comorbidities **Intervention** ❑ Exposure to be investigated **Comparison / Control** ❑ Alternative to compare to the exposure **Outcome** ❑ Any relevant clinical outcomes, including intermediate outcomes **Timing** ❑ Duration of follow-up	**Feasible** ❑ Availability of subjects, funds, time, measurable outcomes **Interesting** ❑ To investigator and clinical / scientific community **Novel** ❑ Contributes in some way to current body of knowledge **Ethical** ❑ Can be investigated ethically **Relevant** ❑ Can provide information that is clinically or scientifically relevant

Fig. 1. The PICOT framework and FINER criteria[1] provide useful guidelines in developing a good research question.

The sample size needed to reject the null hypothesis depends on the following: the desired power of the study, the study hypothesis (i.e., superiority, equivalence, and noninferiority), prevalence of the outcome with current practice, expected magnitude of the effect of the intervention on the outcome, and degree of variability expected within groups. Pilot studies and literature reviews can be informative, and details of how the sample size was determined should be reported to aid in the reader's interpretation of results.

MISTAKE 4. NOT CONSIDERING THE IMPACTS OF INTENTION-TO-TREAT OR PER-PROTOCOL ANALYSES

Appropriate selection of the analytical approach, intention-to-treat (ITT) or per-protocol (PP) analysis, is an important element of study design (Table 1). In general, a conservative approach to analysis is favored for clinical studies to avoid falsely attributing benefit (or lack of harm) where it is not present. ITT is preferred for superiority studies because it provides the most conservative estimate of the treatment effect. Every subject randomized is included; hence, the analysis ignores anything that happens after randomization, such as protocol deviations, noncompliance, and withdrawals. PP analysis should also be reported for noninferiority studies because ITT analysis has greater potential to demonstrate false noninferiority if many of the study individuals deviate from their randomized protocol.

As an example of the importance of consideration of the analytic approach, consider a randomized controlled trial published in 2016, wherein PP analysis detected a significant reduction development of food allergies with early introduction of allergens that was not detected with ITT analysis.[4] With only 32% of participants in the intervention arm adhering to the protocol, the authors acknowledged that differential attrition may have been an important source of bias.

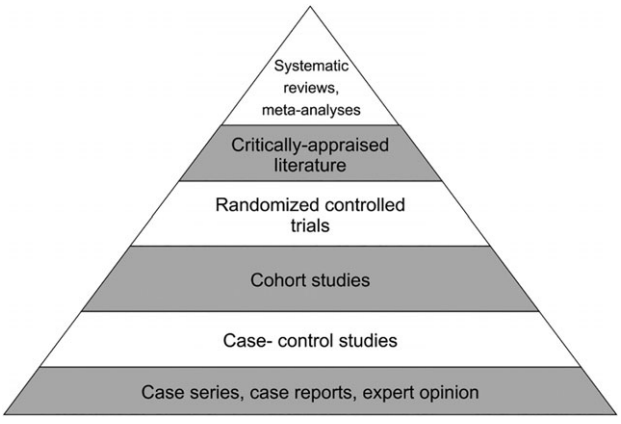

Fig. 2. Hierarchy of clinical evidence strength on the basis of study design. Weaker study designs form the base, with strength of evidence increasing toward the peak of the pyramid.

MISTAKE 5. NOT ADEQUATELY DEFINING VARIABLES AND DATA

Exposure and outcome variables can be categorized as different types of data, depending on the nature of the measured parameter and the method of measurement (Table 2). These are often referred to as independent and dependent variables, which is a different concept than independent and dependent data, explained later. Collecting continuous data tends to provide greater statistical power than categorical or interval data.

Consideration should be given to whether the data are independent or dependent—i.e., does the measured parameter depend on other measurements in the same set of data, as can be seen between two measurements on the same individual separated by time (e.g., different hospital admissions), or between multiple measured parameters from the same individual (e.g., repeated height and weight measurements in given subjects). For example, in a study measuring the frequency of transfusion reactions in which some

TABLE 1. Comparison of intention-to-treat and per-protocol analyses in clinical studies

Variable	Intention-to-treat analysis	Per-protocol analysis
Principle	Participants analyzed in arm randomized to whether protocol was completed or not	Participants analyzed in arm randomized to only if protocol is strictly followed
Advantages and disadvantages	• More representative of "real-life" situations where compliance is not ensured • Maintains group comparability initially achieved by randomization • Maintains study power • Less likely to falsely identify effect if bias is introduced by poorly designed study	• Representative of the "true effect" of the intervention by excluding deviations from protocol • Less likely to demonstrate false noninferiority, with significant protocol deviations • Group comparability and study power may be reduced by dropouts and deviations

individuals were multiply transfused, these subjects would contribute multiple outcomes to the data, each of which are dependent on that subject's other contributions. Whether data are dependent or independent will affect the choice of appropriate methods for statistical analysis. For example, comparing the frequency of unnecessary transfusions between two unrelated groups of physicians (independent data) may be suited to an unpaired t test, whereas comparing the frequency of unnecessary transfusions in one group of physicians before and after an educational program (dependent data) may be more appropriately performed with a paired t test.

Appropriate measured variables and surrogate markers for outcomes of interest should be carefully considered. If possible, collaboration between content experts and biostatisticians can help navigate these complexities.

MISTAKE 6. NOT DIFFERENTIATING BETWEEN ALLOCATION CONCEALMENT AND BLINDING

- Allocation concealment aims to prevent selection bias by concealing the randomization scheme until the time that a patient is assigned to a treatment. Concealment should be possible 100% of the time: the person assessing eligibility, recruiting, and enrolling patients MUST NOT be aware of the allocation sequence.
- Blinding refers to ensuring study participants, care providers, data collectors, and often data analysts remain unaware of which group participants are randomized to. Only participants are blinded in single-blinded studies, whereas study personnel and participants are blinded in double-blinded studies.

The goal of both approaches is to minimize bias, and the highest level of blinding feasible (i.e., participants, providers, and data collectors) is desirable. In the absence of blinding, allocation concealment, at a minimum, should be performed in all cases.[5]

MISTAKE 7. NOT USING APPROPRIATE DESCRIPTIVE STATISTICS

Descriptive statistics (e.g., mean and median) summarize and describe the distribution of a set of data. These form the basis for the quantitative assessment of data, such as the statistical tests used to detect differences and draw conclusions in clinical research.

The mean and SD are commonly used to represent characteristics of a set of data but may not be an accurate representation if the data are not normally distributed. For skewed data, where extreme outliers can easily shift the mean and SD, the median and interquartile range may provide a more accurate description. Visual depiction of data (e.g., in a histogram or quantile-quantile plot) can help to determine if data are normally distributed or skewed (Fig. 3) and guide the selection of the appropriate statistical tests. Various tests to supplement visual assessment of normality are available, such as the Shapiro-Wilk, Anderson-Darling, and Kolmogorov-Smirnov tests.

TABLE 2. Types of data commonly collected and examples of their use

Type of data	Description	Example of use
Continuous	Numerical, can take any value within a range	Hemoglobin (68-127g/L), partial thromboplastin time (34-59 seconds)
Discrete	Numerical, can take limited number of values	Number of transfusion events in a patient (one, two, three)
Nominal	Categorical, value only assigns name	Sex, blood group (A+, O-)
Ordinal	Categorical, value assigns order but interval between cannot be interpreted	Self-reported likelihood to transfuse (not likely, very likely)
Interval	Categorical, provides information on order with interpretable intervals	Age of blood product (0-7, 8-14, or 15-21 days)

MISTAKE 8. NOT ACCOUNTING FOR ERROR INTRODUCED BY PERFORMING MULTIPLE TESTS OF SIGNIFICANCE

When multiple tests of significance are performed on the same data, the overall probability of detecting statistical significance by random chance alone increases. For example, take an investigation in which the threshold of statistical significance is considered $p < 0.05$ and, thus, the probability of no false-positive result for each test performed is 95%; the total probability of no false positives being detected can be calculated as 0.95 raised to the power of the number of statistical tests. If three independent tests of significance are done, the probability of false positives by chance for the study increases to 14% (probability of false positive = $1 - 0.95^3 = 0.14$).

Subgroup analyses and post hoc analyses are not exempt from the increased likelihood of false identification of statistical significance and should be considered independent statistical tests requiring adjustment of the critical value for significance. Primary and preplanned secondary outcomes should be clearly stated, and it must be considered whether the study was designed to answer the question being asked in any post hoc analyses. It has been shown that discrepancies between initially established and published primary outcomes in randomized controlled trials are common, frequently favoring publication of statistically significant results.[6]

For example, consider the Pragmatic, Randomized Optimal Platelet and Plasma Ratios (PROPPR) trial comparing component transfusion ratios in trauma resuscitation: no significant differences were detected in primary mortality outcomes, but post hoc analysis identified a significant reduction in the number of deaths attributable to exsanguination in 24 hours, without having performed adjustment for multiple tests of significance.[7] When interpreting these results, it is important to consider the possibility of statistical significance identified by random chance.

Methods are available to account for multiple tests of significance on the same data, such as the Bonferroni method, where the critical p value is divided by the number of tests of significance performed to modify the critical value required to reach statistical significance.

MISTAKE 9. REPORTING RESULTS ONLY AS POSITIVE/NEGATIVE (P VALUES) AND OVERINTERPRETING STATISTICAL SIGNIFICANCE

Reporting of 95% confidence intervals, wherein there is a 95% probability that the reported interval includes the true value, can be useful in demonstrating the spread of data observed and the inherent uncertainty of point estimates. In contrast to the binary indication of significance or nonsignificance defined by a p value, confidence intervals provide indications of significance as well as direction, range, and magnitude of effect, which are often more useful in clinical decision making.

It is important to remember that statistical significance can be achieved without clinical significance, and the two should not be equated in all situations. Clinical significance usually depends on the magnitude of difference between exposures, whereas statistical significance can be achieved with relatively small absolute differences when sample sizes are large. A study with very large sample sizes might detect statistical significance between outcomes in two groups that is clinically meaningless. A p value of 0.05 is selected as the threshold for statistical significance solely as a matter of convention, reflecting the probability of a falsely positive test, and borderline negative tests of significance should not necessarily be discounted in the presence of other evidence in support of a particular finding.

Interpretation of significant findings also requires caution, because the presence of a relationship between two phenomena does not necessarily imply causation. Various frameworks exist for determining whether causation is likely when interpreting observational data, such as the commonly used Bradford Hill criteria.[8]

MISTAKE 10. DRAWING INAPPROPRIATE CONCLUSIONS FROM NONINFERIORITY AND SUPERIORITY STUDIES

Most clinical studies aim to determine superiority or noninferiority of an intervention relative to standard practice

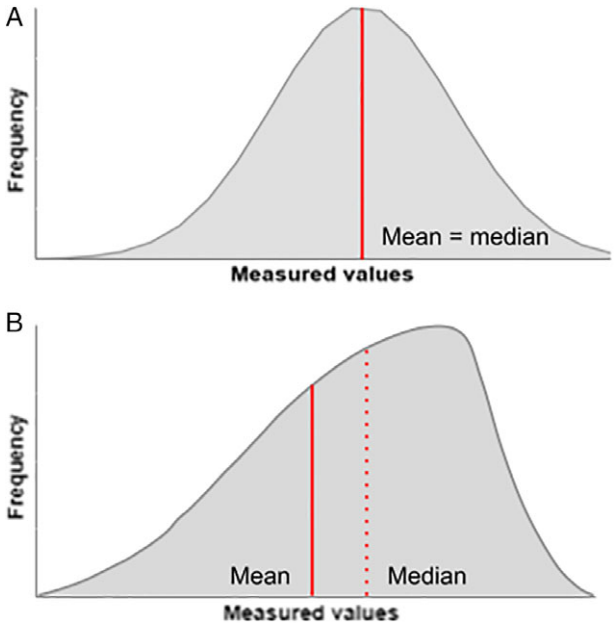

Fig. 3. Comparison of mean and median as measures of central tendency. In normally distributed data (A), the mean accurately represents the central tendency, whereas in skewed data (B), the mean can be altered significantly by large outliers, and the median is likely to be more representative. [Color figure can be viewed at wileyonlinelibrary.com]

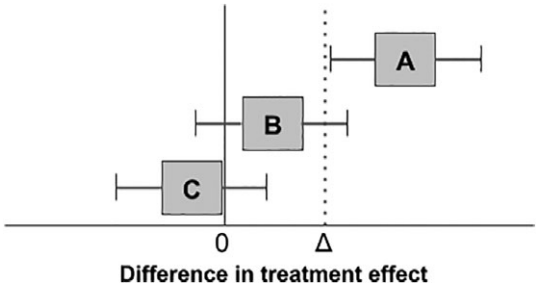

Fig. 4. Representations of accepting or rejecting the null hypothesis in noninferiority studies. Noninferiority studies require preselection of a margin of difference between exposures that can be considered noninferior, which will depend on the context of the measured parameter (clinically and statistically). Zero represents no difference, and Δ represents the margin of noninferiority. A falls beyond the limit of noninferiority and is inferior to the comparison exposure. B is inconclusive, with confidence intervals spanning 0 and Δ. C is noninferior to the comparison exposure.

(control). Noninferiority studies typically require a larger sample size than superiority studies; hence, studies designed to answer a superiority question are unlikely to have adequate power to make any inferences about noninferiority. When interpreting the results of superiority studies, it is important to remember that failure to prove superiority cannot be interpreted as noninferiority or equivalence. Similarly, demonstration of noninferiority should not be interpreted as equivalence (Fig. 4). On occasion, when noninferiority is not shown, authors may do a test of superiority and claim that one exposure is superior to another. In this situation, the analysis must be considered a post hoc analysis unless it was specified in the original study protocol.

CONCLUSIONS

Clinical decision making is dependent on high-quality medical literature. It is critical to maintain a high caliber of research, both to contribute to clinical and scientific knowledge and to support decisions that are in the best interests of patients. Awareness and avoidance of the common mistakes outlined above can provide a solid foundation for the performance of higher-quality research.

CONFLICT OF INTEREST

The authors have disclosed no conflicts of interest.

REFERENCES

1. Cummings S, Browner WS, Hulley SB. Conceiving the research question. In: Stephen B. Hulley, Steven R. Cummings, Warren S. Browner, Deborah Grady, Thomas B. Newman (Eds.) *Designing clinical research*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2007. p. 17-26.
2. Heddle NM. The research question. Transfusion 2007;47:15-7.
3. Fergusson D, Glass KC, Hutton B, et al. Randomized controlled trials of aprotinin in cardiac surgery: could clinical equipoise have stopped the bleeding? Clin Trials 2005;2:218–29; discussion 29–32.
4. Perkin MR, Logan K, Tseng A, et al. Randomized trial of introduction of allergenic foods in breast-fed infants. N Engl J Med 2016;374:1733–43.
5. Schulz KF, Altman DG, Moher D, et al. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332.
6. Mathieu S, Boutron I, Moher D, et al. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009;302:977–84.
7. Holcomb JB, Tilley BC, Baraniuk S, et al. Transfusion of plasma, platelets, and red blood cells in a 1:1:1 vs a 1:1:2 ratio and mortality in patients with severe trauma: the PROPPR randomized clinical trial. JAMA 2015;313:471–82.
8. Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Bradford Hill's "guidelines for causation" contribute? J R Soc Med 2009;102:186-94.