Guidelines for Statistical Analysis Plans

David L. DeMets, PhD; Thomas D. Cook, PhD; Kevin A. Buhr, PhD

The emergence of the randomized clinical trial as the gold standard for the evaluation of new clinical interventions has been met by the emergence of a host of guidelines for the design, conduct, monitoring, analysis,[1-3] and reporting[4] of randomized clinical trials including guidance from regulatory authorities reviewing pivotal studies to support approval of products, such as drugs and devices.[5] Much of this guidance reflects the key scientific principle that comprehensive documentation should be prepared in advance, including a detailed protocol providing scientific rationale, clear statement of hypotheses, definitions of outcome measures, and description of trial design, as well as a statistical analysis plan (SAP) that provides technical detail about conventions and procedures for testing trial hypotheses using collected data.

Related article page 2337

These elements should be defined in advance to protect investigators and the scientific community from the temptation to "interrogate" the data intensively and extensively. The more outcomes, subgroups, and variant measures evaluated, the greater the chance of a false-positive finding, ie, nominally significant results that suggest treatment group differences when none exist. In particular, guidelines suggest when multiple outcomes are of interest, those outcomes should be specified in advance and ordered in importance, with clear designation of the "primary" outcome. Similarly, formal subgroup evaluations should be prespecified, limited in number, and ideally analyzed according to a formal testing procedure that considers their relative importance. Such an approach protects against data dredging to detect something, anything, resembling nominal statistical significance.

Moreover, any given outcome or hypothesis can be evaluated using several different statistical methods and sets of reasonable assumptions. To prevent "shopping around" for the best (ie, most statistically significant) result, the SAP must specify statistical analysis methods and associated assumptions in advance. Additional methods may be used but should be specifically identified as part of a secondary evaluation or sensitivity analysis, for example, approaches for dealing with multiple comparisons[2,5] and missing data.[6,7] On the other hand, while a comprehensive SAP supports transparency and reproducibility, trial integrity requires that results also be robust. That is, armed solely with the final data, the trial protocol, the case report form, in either paper or electronic format, used for data collection and suitable documentation of the data collection procedures, a competent analyst should be able to roughly reproduce the key primary and secondary analyses. However, the primary conclusions for studies for which results can change substantially with only minor and somewhat arbitrary changes in the analysis may be called into question.

The Special Communication by Gamble et al[8] in this issue of *JAMA* addresses the question of the content of the SAP. The authors carefully prepared a set of surveys to identify current guidance, assess current practice, and develop consensus on required content; collectively, these surveys were sent to colleagues in academia, industry, and regulatory agencies. Despite some limitations regarding response rates, the results provide useful insight: existing guidance is limited, current practice is highly variable with respect to content and level of detail, yet broad consensus is possible on a minimum set of required items. The authors offer 2 sets of general recommendations: what should be in the SAP and what should be provided elsewhere.

The cornerstone of the article is a list of 55 items recommended for inclusion in the SAP, the result of consensus achieved via a 2-stage Delphi survey. These items include administrative information (eg, trial registration number and SAP version history); a synopsis of the trial background, rationale, and objectives; a summary of the design with emphasis on statistical aspects (eg, randomization details, hypothesis testing framework); general statistical conventions; definition and characterization of the trial population; and outcome definitions and analysis methods. In particular, the items include designation of the primary outcome measure, a priori specification of subgroup analyses, and specific analysis methods including required statistical assumptions and tests of their validity. Detailed explanation and elaboration for all items are provided in the article's Supplement.[8]

The authors also provide a list of 17 items that, while deemed important in the Delphi process, were not determined to be required SAP components. These items are likely to be very useful for investigators interested in conducting secondary analyses to explore new hypotheses or independently verify previously published results, and they include file metadata, edit checking algorithms, and data archiving and sharing procedures.

The SAP also has implications for interim monitoring, including the operation and conduct of the data monitoring committee (DMC), where applicable. The protocol and SAP will largely determine the outcomes monitored for early evidence of harm, benefit, or futility, and the list of recommended items includes timing and nature of interim analyses and stopping guidelines. However, in the context of DMCs, experience suggests that the unexpected is almost to be expected,[9] and there is risk associated with excessive rigidity in interim monitoring. For example, recommended items and examples given suggest an inflexible approach with method and timing for sequential monitoring precisely specified in advance. In practice, if results at a prespecified interim analysis, for example, just miss or cross a group sequential boundary for benefit,[3] most experienced DMCs would want a subsequent, unscheduled analysis to reevalu-

ate the borderline result, and it would seem unethical to do otherwise. Fortunately, statistical methods, such as the alpha spending function, are available to allow for unscheduled interim analyses without prespecification of the time and number of interim evaluations.[2,3,10] For example, the CAST trial[11] is one example for which the DMC required flexibility in its approach to an unexpected interim result. The protocol was designed as a 1-sided test of treatment benefit for mortality reduction in a patient population with arrhythmias using a class of drugs that suppressed these arrhythmias. Unexpectedly, the mortality trend very early in CAST indicated an increase in mortality among those receiving active treatment but the protocol provided no prespecified monitoring mechanism for this outcome. Ultimately, the DMC applied a symmetric version of the prespecified 1-sided sequential boundary for benefit and terminated the trial for a harmful effect.

In the AHEFT trial,[12] the primary outcome was a composite event outcome including a mortality component. The trial was designed on the basis of the composite outcome with interim analyses prespecified accordingly. Yet, mortality alone unexpectedly became highly statistically significant during the trial, something not anticipated by the prespecified analyses. Ultimately, the DMC felt an ethical imperative to react to this result and terminated the trial early for overwhelming mortality benefit.

Even in the context of the final analysis, there remain substantial challenges in SAP development and execution. When all outcomes are consistent with respect to the intervention effect, there is seldom difficulty in judging trial results to be solid and credible, but if there are differences, interpretation becomes more challenging. Although a prespecified framework for outcome evaluation set forth in the SAP is helpful in assessing the credibility of results, on occasion, the prespecified result may not be consistent formally with the balance of the trial data. For example, in the APEX trial, the primary analysis was conducted in an enrichment subgroup within a larger trial.[13] At trial completion, the result in this subgroup was not statistically significant (relative risk, 0.81; P = .054) as per the SAP, technically prohibiting testing of the entire cohort, which included the enrichment cohort but showed a significant benefit (relative risk, 0.75; P = .006). In the APPROVe trial[14] of a drug, rofecoxib, for prevention of colon cancer, the trial was terminated early for evidence of increased cardiovascular risk (ie, mortality, nonfatal myocardial infarction, nonfatal stroke). Data were analyzed according to a prespecified SAP. Based on the resulting SAP analysis, the Kaplan-Meier curves suggested that the increased risk did not emerge until after 18 months of treatment. However, the SAP clearly indicated that patients would not be followed up after being off treatment for 14 days. Because of controversy, an additional 1 year of follow-up was done for all patients, with the results clearly indicating that cardiovascular risk increased from the beginning of randomization, an excellent example of informative censoring.

Moreover, data do not always meet the assumptions of prescribed analytical methods, and it is impossible to anticipate every eventuality. In the MERIT-HF Trial (terminated early due to a very strong mortality benefit), all primary and secondary outcomes were positive, and all predefined subgroups were consistent in direction of effect.[15] Yet, one geographic subgroup not prespecified, namely the United States, did not show benefit. No final SAP anticipated that possibility of regional differences, yet this result became a major issue in regulatory review.[15]

Ultimately, a prespecified SAP is necessary to ensure interpretability and integrity of final results. However, there must be room to address unanticipated but important results. Such results should be reported in the article in the context of the SAP and accurately described in the context of the analytic approach (eg, post hoc) and should not be dismissed as irrelevant or meaningless without serious scientific discussion by investigators, sponsors, and regulators, taking into account the totality of evidence in the trial.

ARTICLE INFORMATION

Author Affiliations: Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin–Madison.

Corresponding Author: David L. DeMets, PhD, Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin–Madison, Room 201, WARF Bldg, 610 Walnut St, Madison, WI 53726 (demets@biostat.wisc.edu).

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr DeMets reported receiving personal fees from Frontier Science, Actelion Pharmaceuticals, Mesoblast, Population Health Research Institute, Amgen, DalCor, AstraZeneca, Portola, Duke/Patient-Centered Outcome Research Institute, GlaxoSmithKline, Duke Clinical Research Institute/Luipoid, and Duke Clinical Research Institute/National Heart, Lung, and Blood Institute. Dr Cook reported receiving compensation for serving on industry and National Institutes of Health data monitoring committees (DMCs), and his institution receives funds from industry through contracts for DMC statistical support. Dr Buhr reported serving on DMCs and acting as the independent statistician supporting DMCs for multiple industry-sponsored clinical trial programs.

REFERENCES

1. Friedman L, Furberg C, DeMets D, Grainger C, Reboussin D. *Fundamentals of Clinical Trials*. 5th ed. New York, NY: Springer Science+Business Media LLC;2015.

2. Cook T, DeMets DL. *Introduction to Statistical Methods for Clinical Trials, Chapman & Hall/CRC*. Boca Raton, FL: Taylor & Francis Group, LLC; 2008.

3. Ellenberg S, Fleming T, DeMets D. *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. West Sussex, England: John Wiley & Sons Ltd; 2002.

4. Schultz K, Altman D, Moher D; CONSORT group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. *BMC Med*. 2010;8:18.

5. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; US Department of Health and Human Services; Food and Drug Administration; Center for Drug Evaluation and Research (CDER); Center for Biologics Evaluation and Research (CBER). Guidance for industry: E9 statistical principles for clinical trials. https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf. Published September 1998. Accessed November 4, 2017.

6. European Medicines Agency. EMA/CPMP/EWP/1776/99 Rev 1: Committee for Medicinal Products for Human Use (CHMP): Guideline on missing data in confirmatory clinical trials.

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf. Published July 1, 2010. Accessed November 5, 2017.

7. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: The National Academies Press; 2010.

8. Gamble C, Krishan A, Stocken D, et al. Guidelines for the content of statistical analysis plans in clinical trials. *JAMA*. doi:10.1001/jama.2017.18556

9. DeMets DL, Ellenberg SS. Data monitoring committees: expect the unexpected. *N Engl J Med*. 2016;375(14):1365-1371.

10. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659-663.

11. Echt DS, Liebson PR, Mitchell LB, et al; CAST Investigators. Mortality and morbidity in patients receiving encainide, flecainide, or placebo: the Cardiac Arrhythmia Suppression Trial. *N Engl J Med*. 1991;324(12):781-788.

12. Taylor AL, Ziesche S, Yancy C, et al; African-American Heart Failure Trial Investigators. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med*. 2004;351(20):2049-2057.

13. Cohen AT, Harrington RA, Goldhaber SZ, et al; APEX Investigators. Extended thromboprophylaxis with betrixaban in acutely ill medical patients. *N Engl J Med*. 2016;375(6):534-544.

14. Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. *Lancet*. 2008;372(9651):1756-1764.

15. Wedel H, Demets D, Deedwania P, et al; MERIT-HF Study Group. Challenges of subgroup analyses in multinational clinical trials: experiences from the Merit-HF Trial. *Am Heart J*. 2001;142(3):502-511.

Is Medical Education a Public or a Private Good? Insights From the Numbers

Catherine R. Lucey, MD

The US educational system has 2, at times competing, goals. Education is commonly viewed as a public good, designed to prepare the workforce that the country needs and to educate citizens who contribute to the health of the US democracy. However, education is also seen as a private good, geared toward helping the individual maximize social mobility and personal success.[1] From Virchow in the 19th century to Frenk in the 21st century, thought leaders have embraced the view of medical education as a predominantly public good rather than a private one, maintaining that the purpose of medical education is to improve the health of communities and to decrease the burden of illness and disease.[2,3] The annual *JAMA* publication of data describing the demographic composition, geographic distribution, and specialty focus of learners and programs in US undergraduate medical education[4] and graduate medical education[5] provides an opportunity for the medical profession to once again consider whether the medical education community is designed to strike the appropriate balance between providing a public good and a private good.

Related articles pages 2360 and 2368

A medical education pipeline focused on public good would be designed to prepare the physician workforce that can effectively meet the needs of 21st-century patients and communities for high-quality, evidence-based, and patient-centered care. That ideal workforce would comprise sufficient numbers of primary care and generalist physicians, distributed throughout communities in the United States, so that everyone had ready access to the types of preventive, diagnostic, and therapeutic care they need.[6] It would be composed of physicians whose gender, race, ethnicity, religion, sexual orientation, disabilities, and other aspects of diversity mirrored that of the populations for whom they care. That workforce would have a sufficient number and diversity (both discipline and demographics) of physician scientists working to advance understanding of the most challenging health care and biomedical problems. It would be constituted to reflect an awareness of current and anticipated unmet health care needs driven by changes in populations (such as advancing age of the society), illnesses (such as mental illness and opioid addiction), scientific advances (such as genomics and molecular diagnoses), or changes in the way physicians work (such as the ability of technology and big data to change care delivery, research, and education). The data published in this issue of *JAMA* provide some insights into and many questions about the extent to which the systems of education are prepared to meet the expectations of a public good.

It is generally accepted that the United States will face a shortfall of physicians but also is experiencing geographic maldistribution of physicians.[7] The data presented in this issue demonstrate strides in addressing the projected shortfall of physicians. Between 2006-2007 and 2016-2017, the number of medical schools increased by 20, from 125 to 145, and the number of graduating medical students entering residency training by 3524 (15%), from 15 007 to 18 261.[4] Between 2011-2012 and 2016-2017, the number of residency programs increased by 1232 (from 9111 to 10 343) and the number of program year 1 (PGY1) residents by 2400 (from 25 538 to 27 938).[5] The data are less clear on whether there has been success in resolving the geographic maldistribution of physicians. Many new medical schools are designed and located to address the needs of geographic regions that are medically underserved.[8] For example, Central Michigan University College of Medicine and Geisinger Commonwealth Medical College are located in Mount Pleasant, Michigan, and Scranton, Pennsylvania, respectively. Given that residents are more likely to practice in a region close to the one in which they complete their