Key Statistical Concepts in Cancer Research

Qian Shi, PhD, and Daniel J. Sargent, PhD

The authors are affiliated with the Department of Health Science Research at the Mayo Clinic in Rochester, Minnesota. Dr Shi is an associate professor of biostatistics and Dr Sargent is a professor of biostatistics.

Corresponding author:
Daniel J. Sargent, PhD
Professor of Biostatistics
Department of Health Science Research
Mayo Clinic
200 First Street SW
Rochester, MN 55905
Tel: 507-284-5380
E-mail: sargent.daniel@mayo.edu

Keywords
Clinical trials, endpoints, statistical analysis, study design

Abstract: In this article, we provide a high-level overview of statistical concepts related to study design and data analysis in oncology research. These concepts are discussed for 2 main types of clinical research: (1) observational studies, which focus on biomarker discovery in order to predict disease risk and prognosis, and (2) prospectively designed, well-controlled clinical trials, which are critical for the development of new cancer treatments. Throughout the article, we emphasize the importance of appropriate design and prospectively determined analysis plans. We also hope to promote effective collaboration between oncology investigators and statisticians who center their research on the development of cancer treatments.

Introduction

Oncology research is a highly active field of discovery that has substantial challenges and, most importantly, huge unmet needs in patient care. In this setting, it is critical to perform scientifically sound studies that are reproducible and generalizable, have a high degree of credibility, and can be applied efficiently to real-world practice. This need has led to the integration of statistical expertise into multiple aspects of oncology research.

When research—or even a single experiment—involves human beings, substantial complexities exist owing to multidimensional variations in genetic, behavioral, environmental, and sociological factors. Many of these factors are uncontrollable or even unobservable, and can have unpredictable interactions with each other, creating even more complexity. In statistics, the term *error* is used to describe this variation in patient outcomes due to unknown or uncontrollable factors. Generally speaking, there are 2 types of error: random error and bias.

Random error, which is purely due to chance, commonly is caused by sampling variability, measurement error, and other sources of "noise." Random error can be quantified by determining the variability in patient outcomes that exists among similar patients. By applying the knowledge of probability and statistical theory, the magnitude and likelihood of the errors resulting from chance can be estimated. In general, the impact of random errors on a study can be reduced by using more participants.

Unlike random errors, which have no preferred direction, a bias represents a distortion of a true state and is not a consequence of chance alone. For example, differences in baseline disease characteristics between 2 treatment groups can cause a bias. Unlike random error, averaging after repetition or recording additional observations cannot reduce this bias. There are many statistical methods that can be applied to correct for bias and thus allow accurate inferences to be drawn from the results; for example, using multivariable modeling to adjust for confounding factors in retrospective studies. However, the ability to reduce bias using statistical methods is limited to factors that are known and measured. If bias from unknown sources is greater than bias from known sources, these statistical methods will not be useful. Randomization is one of the fundamental principles of prospective clinical trials because it balances known and unknown factors between comparison groups, thereby reducing bias.

In any medical research, reducing random error and controlling systematic bias are essential to providing valid and generalizable results. These goals cannot be achieved without collaboration between clinical and statistical experts. In this article, we provide an overview of statistical methods relevant to 2 major types of oncology research: observational studies and clinical trials. In observational studies, the primary factor of interest (ie, the explanatory or independent variable) cannot be manipulated; instead these studies are used to assess potential associations between risk or prognostic factors and disease outcome. By contrast, in clinical trials the treatments under evaluation are fully defined by the study and can be manipulated. Our intention is to provide the reader with insight regarding the use of statistics (and collaborations with statisticians) in the design and analysis of medical studies, rather than to provide all the details required for an investigator to perform his or her own analyses. Several excellent statistical textbooks provide the technical details, and we cite several such references.

Observational Studies

Overview

Oncology research is entering a new era that is characterized by increased understanding of biological mechanisms. Therefore, the identification of disease-related biomarkers and their underlying mechanism of action is a critical step in cancer treatment development, especially in regard to personalized treatment. Examples include gene mapping to better understand the risk of developing various cancers, development of risk classification tools to better predict patient outcomes, and molecular discovery for targeted therapies. These studies can be described as association studies, which evaluate the relationships between explanatory factors and disease outcomes. The distinguishing feature of observational studies is that these explanatory factors have their own natural course; ie, the investigator can observe but cannot intentionally alter the factors' status. These studies commonly are based on hospital or institutional cohorts, translational studies derived from completed clinical trials, or meta-analyses that combine data from multiple studies.

Owing to the large number of potential hypotheses and the discovery-based nature of clinical studies, investigators may be overwhelmed by the large number of potential analyses possible for a data set, or they can be distracted by signals that may not be true (ie, false-positives). This is why clinically and statistically sound study design and prospectively defined analysis plans are essential. The study design includes many aspects, such as: (1) relevant, focused, and precise objectives; (2) effective and feasible sampling schema (eg, correctly targeted populations, sufficient sample size, and technique-appropriate selection procedures); (3) clinically and statistically relevant outcomes; (4) careful control of confounding and bias; and (5) rigorous data collection and quality control. Prospectively defining a statistical analysis plan is a critical step in creating an executable study, because it allows researchers to discover and address the potential flaws and inefficiencies in the study design. This plan describes the patient selection procedures, data to be collected, proper statistical methods, analytical steps, data presentation methods, and data interpretation principles. It is vital that thorough planning is done before beginning a study, because, as noted by Steyerberg,[1] "[a] sophisticated analysis cannot salvage a poorly designed study, or poor data collection procedures."

Basic Statistical Methods of Testing Associations

There are 3 main types of observational studies: cross-sectional, cohort, and case-control. In cross-sectional studies, the collected data (eg, environmental exposure and disease status) are assessed at a single preselected time, and their association is commonly measured by a correlation coefficient. Cohort studies, in which subjects are selected based on risk factors (eg, environmental exposure) and are followed for disease status until a future point, generally provide more comprehensive data than cross-sectional studies. An important association measurement in cohort studies is relative risk (RR), which is a ratio of the disease incidence rate in exposed vs unexposed subjects. When disease incidence is rare, long follow-up times or very large sample sizes are needed to conduct a cohort study. In this situation, a case-control study can be considered. Subjects in a case-control study are selected based on disease status, with or without matching known risk factors. The potential risk factors under consideration

are ascertained by looking back in time. Because subjects are selected based on their disease outcomes, the exposure-specific disease rates and RR cannot be estimated. Instead, an odds ratio (OR)—the ratio of the odds of exposure among disease cases vs nondisease cases—is commonly used to quantify the association between the risk factor and the disease rate. The OR can be estimated regardless of the study design. If the disease is rare, the OR can be used as an approximation of the RR.

Confidence intervals for these measurements can be calculated and used to determine whether the observed association is statistically significant (ie, different from the null hypothesis, which is that no association exists). Different statistical methods are appropriate depending on the nature of the risk factors (eg, continuous, count, nominal [without order between categories], or ordinal [with order between categories] variables). Hypothesis testing methods, such as the Chi-square test, commonly are used when the risk factor and outcome are both categorical. The Chi-square test conveys only the existence or nonexistence of an association between an exposure and an outcome, and not its nature or strength. With an ordinal risk factor (eg, the categories of nonsmoker vs former smoker vs current smoker), the Cochran-Armitage trend test provides better power to detect a linear trend in disease rate according to the levels of the risk factor. When there is more than 1 risk factor of interest, regression analysis (eg, logistic regression) is appropriate, because it adjusts for confounders and identifies effect modifiers. We refer readers to the categorical data analysis textbook of Agresti[2] for more details.

Prognostic studies in cancer research commonly are used to identify disease-related markers that can predict a patient's prognosis. Time-to-event analysis (or survival analysis) is a statistical method used to assess this marker-prognosis association. Time-to-event outcomes are continuous variables defined as the time from the beginning of observation (eg, diagnosis date or surgery date) to the occurrence of the relevant event(s) (eg, disease recurrence or death). Analyses of these studies differ from other types of statistical analyses because the event may not be observable for all subjects owing to loss of follow-up, competing risks, or termination of follow-up because of financial, logistical, or study duration considerations. Although the event is not observed in some subjects, partial information that the subject was event-free until the last known date still can be valuable. These data are called censored data. A common way to summarize censored survival data is to estimate the Kaplan-Meier curve,[3] which shows the proportion of subjects who are event-free at each point that an event is observed. The Kaplan-Meier curve can provide an estimate of the event-free rate at any point during the follow-up period, and allows for censoring and varying lengths of follow-up. Common descriptive statistics associated with Kaplan-Meier curves are median survival time and survival rates for a specific time. Confidence intervals also are used to interpret these results. Some types of censoring may require advanced analytic methods; for example, the competing risk model is used when the event of interest is not observed owing to the occurrence of another competing event.

When comparing time-to-event outcomes between patient groups, the most common method used is the log-rank test. The log-rank test determines whether the hazard rates are different between 2 or more groups. Here, the hazard rate refers to the rate of change in the cumulative probability of an event happening at a given point relative to the corresponding event-free probability. To quantify the association between a prognostic factor and the survival outcome, the hazard ratio (ie, the ratio of the hazard rates of 2 populations) can be estimated by a Cox proportional hazards model.[4] Cox regression is a powerful method that can assess the impact of multiple factors on survival outcomes simultaneously. This model also can control for confounders, identify interaction effects, and provide risk predictions. We refer readers to a survival data analysis textbook[5] for additional details.

It is important to point out that no matter which study design is used to assess potential disease risk or prognostic factors, a significant association in an observational study is not sufficient to prove a causal relationship. For example, an association between gene expression and disease risk may actually be noncausal owing to underlying factors, such as linkage disequilibrium between the studied gene and the truly causal gene, or an unobserved intermediate association between a gene and disease. However, by appropriately controlling for confounders and reducing biases, observational studies can provide sufficient evidence to support further biological studies on a disease risk or prognostic factor.

Individualized Risk Classifications and Outcome Predictions

Association studies provide population-level results, and do not necessarily show the absolute risk prediction of an individual patient. In clinical practice, there are many predictors of patient survival that are based on a combination of variables, such as patient characteristics (eg, race, sex, age, and genetics), disease characteristics (eg, anatomic involvement, disease severity, and tumor gene mutation), and other factors (eg, family history, environmental exposure, and behavioral factors). These factors have an impact on individual patient care, such as the choice of treatment. Integration of all these factors can be achieved through clinical prediction models, which are designed to estimate a diagnostic or prognostic outcome

for an individual patient. In this section, we discuss general concepts regarding the development of clinical prediction models. Suggested further reading on this topic includes textbooks by Steyrberg[1] and Harrell.[6]

The development of a clinical prediction model involves 3 components: building, validation, and presentation. Both building and validation are guided by model prediction performance evaluations. Building a prediction model can be viewed as a process of conducting association analyses on many factors simultaneously within a systematic framework. This process starts with the selection of candidate predictors based on clinical relevance, statistical strength, and practical usefulness. This selection should consider: (1) newly discovered factors or markers with strong preliminary data suggesting an impact on disease prognosis and (2) previously established risk factors or prognostic markers that could be confounders or effect modifiers. When data on many factors are needed, missing or insufficient data likely will exist. Careful data inspection and coding of the potential predictors is important. To increase the model's predictive power, an optimal functional form (eg, linear, quadratic) for each given predictor is determined while separately assessing the association between the individual factor and outcome.

The next step, model specification and estimation, is the most critical. In this step, the question of which variables are of greatest importance to predict the outcome is examined through multivariable regression modeling procedures, which are often automatic algorithms such as stepwise selection. Another aspect of model specification is testing for interactions; ie, that the effect of one predictor depends on the value of another predictor.

During the model-build process, exploratory searching associated with data-driven procedures is always involved. This increases the chance of false discoveries. Therefore, validation of the findings is a critical step for achieving clinical and practical utility. Internal validity refers to reproducibility of the model; this is often studied by assessing the validity of a data set that came from the same source as the development data. Cross-validation[7] is one of the common methods. More critical is external validity, which should be assessed on a different, independent data set from a plausibly related population. Nomograms or web calculators are commonly used presentations for a prediction model.

Depending on the ultimate usage of the prediction model, different aspects associated with specific performance measures can be assessed. For example, the coefficient of determination, R^2, is a useful overall performance measure. It can be interpreted as the amount of variability seen in an outcome that can be explained by the predictors included in the model. For example, in oncology an important application of a prediction model is to classify patients into high-risk vs low-risk groups. In this case, the concordance statistic, which is a measure of the predictive accuracy of the model, is the primary requirement for model performance assessment.

Clinical Trials

Design of Clinical Trials

Observational studies are useful for advancing our knowledge of disease risk and prognosis, but treatment development requires rigorously designed experiments to test the safety and efficacy of a proposed regimen. This is achieved using clinical trials. Unlike observational studies, clinical trials contain a variable of interest (typically a treatment) that is manipulated by the investigators using a prespecified trial design. To ensure consistent trial conduct across sites and patients, the study protocol is rigorously defined for application of treatment, ascertainment of outcomes, safety monitoring, decision rules, analysis plans, and specimen collection.[8] Analysis of clinical trials is often simplified because these have a focused primary hypothesis, optimal sample size and power determinations, a properly defined data collection process, and the ability to control for confounders and reduce bias and variability. By using simple statistical analysis methods with fewer assumptions, clinical trials can deliver stronger and more convincing evidence than observational studies.[8]

Clinical trials generally are classified as phase 1, 2, or 3 according to their primary aims and stage in the timeline of drug development. Phase 1 trials are used to determine an optimal dose level and/or treatment schedules, which is required before a regimen can be tested for efficacy. In order to determine dosage, researchers often identify a maximum tolerated dose (MTD) that produces the maximal treatment benefit while protecting patients from severe toxicities. These severe toxicities, called dose-limiting toxicities (DLTs), are study-specific and prespecified. In general, DLTs are defined as serious or fatal side effects of the regimen being tested, and commonly are measured during the first cycle of the treatment. A traditional and commonly used phase 1 design for determining the appropriate dosage in oncology is the "3+3" design. Three patients per cohort are treated per dose level, and the dose decreases or increases in subsequent cohorts based on the number of DLTs. This design is more appropriate for cytotoxic agents than for biologic agents. The fundamental assumption in this design is that both treatment benefit and toxicity monotonically increase as the dose increases; however, many targeted compounds and novel therapies with cytostatic mechanisms of action do not share the same dose-response assumption of cytotoxic agents. In the modern

era of therapeutic development in cancer treatment, dose-finding strategies of more than 1 dimension are needed for testing combinations of compounds. In addition, innovative designs that also account for efficacy data in dose finding may be more appropriate in many cases than the traditional 3+3 design. Braun's recent review on phase 1 clinical trials provides an excellent overview of the methodological advances in this area.[9]

Before moving into large confirmatory studies, new treatments or regimens are quickly screened based on early evidence of efficacy in phase 2 trials. These studies usually have small sample sizes and target large treatment effects. In the past, single-arm designs were commonly used, in which the effect of the new regimen is compared with historical data. As oncology research has advanced, the randomized phase 2 design that includes a concurrently randomized control patient group has become a more frequent choice. Starting in phase 2 trials, controlling type I and type II error rates becomes critical. If investigators incorrectly conclude that there is a treatment effect when none exists, the result is a "false-positive," or type I error. If treatment effects exist, but investigators fail to detect them, this result is a "false-negative," or type II error. When designing a trial, a balance between acceptable levels of type I and type II error must be considered. This involves an assessment of the risk levels that investigators are willing to accept for false-positive and false-negative findings. Typically, in phase 2 trials the type I and type II error rates are prespecified to be within a range of 10% to 20%.

For ethical reasons, it is desirable to minimize the number of patients treated with ineffective or inferior treatments; therefore many trials use interim analyses to determine if they can be stopped before their scheduled completion. From a statistical perspective, there are 2 reasons to stop a trial early: superiority (ie, efficacy) and inferiority (ie, futility). The first situation applies when there is overwhelming evidence that the experimental treatment is superior to control. In the second situation, trials are stopped when it is highly unlikely that a trial will achieve the target treatment effect, even if all the patients are enrolled. For single-arm phase 2 studies, the Fleming's 2-stage design[10] provides the ability to stop the trial early (after approximately half of the patients are accrued) if the results are either overwhelmingly positive or negative. Owing to the small sample size of phase 2 trials, a substantially large treatment effect—usually beyond what is realistic—generally will be required to meet the early stopping criteria for efficacy. However, this design raised concerns about unreliable estimates of the treatment effect. Simon's 2-stage design (optimal or minimax)[11] restricts the interim analysis to only evaluate futility, and has become the standard single-arm phase 2 design.

To provide definitive evidence to move a new regimen or modality into patient care, large confirmatory studies are needed. These phase 3 randomized controlled trials (RCTs) are designed to be comparative between concurrent arms. Some fundamental principles of RCTs are randomization, stratification, and blinding. Randomization of patients into treatment groups is a critical tool to prevent biases that could occur if treatment selection were based on patients' prognostic factors or other confounding factors.[12,13] Stratification is an effective and practical procedure to ensure the success of the randomization by assigning patients with certain characteristics equally to each treatment group (eg, assigning equal numbers of males and females to each group).[14] Blinding is a procedure that withholds information from specific groups of individuals (patients and/or healthcare providers) to reduce the response bias associated with the psychological impact of being treated with an intervention perceived as superior to a control treatment.[15] Blinding is particularly important when the endpoint measurement is subjective or semisubjective.

More advanced designs, such as outcome-adaptive designs, are increasingly used in modern oncology trials. Trials with adaptive designs formally and statistically incorporate ongoing modification during the course of the trial based on accumulating outcome data.[16] For example, a design can gradually assign newly enrolled patients to treatment arms that demonstrate better treatment effects based on these continuous evaluations. If there are multiple subpopulations defined by biomarkers that can potentially react differently to the treatment, adaptive design can be used to isolate the responder population.[17] The I-SPY2 (Investigation of Serial Studies to Predict Your Therapeutic Response With Imaging and Molecular Analysis 2) trial is an example of a study with such an adaptive design.[18] This ongoing phase 2 confirmatory trial screens pairs of compounds and biomarkers based on the predictive probability of each regimen being successful. For example, 1 analysis showed that veliparib and carboplatin along with standard chemotherapy improves outcome in women with triple-negative breast cancer.[19]

Analyses of Clinical Trials

Many references provide technical details on the analysis of clinical trials.[20-22] One common problem in clinical trials is analyzing data from patients who do not adhere to protocol treatment or who have nonprotocol treatment crossovers. Incorporating these imperfections in trial conduct into the resulting data potentially can impact the study findings and interpretation of results. One intuitive approach is to analyze patients according to the treatment they actually received, regardless of the original treatment group. This approach has been shown to be potentially

misleading (especially when the intention is to show difference between treatments), because there might be confounders associated with patient adherence to the treatment, such as other treatments, disease status, or lifestyle.[23] To minimize this bias, intention-to-treat analysis is used. This method reduces confounders by analyzing patients according to their original randomization assignment, regardless the treatment they actually received.

Conclusions

This article provided a high-level review of statistical design and analysis for both observational studies and clinical trials. For observational studies, factors to consider when selecting a statistical method include the questions to be addressed, the study design, and the types of variables to be analyzed. For clinical trials, the different phases of studies and the principles of RCTs were discussed. As is true of any research, fighting cancer is a process of continually updating knowledge about the disease using both biological and statistical points of view to allow evidence-based decisions for better outcomes. The sound integration of biology and statistics provides fertile ground for continual innovation in oncology research.

Disclosures

The authors have disclosed no financial conflicts of interest.

References

1. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.
2. Agresti A. *Categorical Data Analysis.* 2nd ed. Hoboken, NJ: Wiley; 2002.
3. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
4. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc, B*. 1972;34:187-220.
5. Kleinbaum DG, Klein, M. *Survival Analysis: A Self-Learning Text.* New York, NY: Springer; 1996.
6. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 1st ed. New York, NY: Springer; 2001.
7. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. 1st ed. Boca Raton, FL: Chapman & Hall/CRC; 1993.
8. Piantadosi S. *Clinical Trials: A Methodologic Perspective.* Hoboken, NJ: Wiley; 1997.
9. Braun TM. The current design of oncology phase I clinical trials: progressing from algorithms to statistical models. *Chin Clin Oncol*. 2014;3(1):2-13.
10. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics*. 1982;38(1):143-151.
11. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10(1):1-10.
12. Zelen M. The randomization and stratification of patients to clinical trials. *J Chronic Dis*. 1974;27(7-8):365-375.
13. Buyse M. Centralized treatment allocation in comparative clinical trials. *Applied Clin Trials.* 2000;9(6):32-37.
14. Lachin JM, Bautista OM. Stratified-adjusted versus unstratified assessment of sample size and power for analyses of proportions. *Cancer Treat Res*. 1995;75:203-223.
15. Psaty BM, Prentice RL. Minimizing bias in randomized trials: the importance of blinding. *JAMA*. 2010;304(7):793-794.
16. Berry DA. Adaptive clinical trials in oncology. *Nat Rev Clin Oncol*. 2012;9(4):199-207.
17. Chang M. *Adaptive Design Theory and Implementation Using SAS and R.* Boca Raton, FL: Chapman & Hall/CRC; 2008.
18. Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin Pharmacol Ther*. 2009;86(1):97-100.
19. Printz C. I-SPY2 trial yields first results on combination therapy for triple-negative breast cancer. *Cancer*. 2014;120(6):773-773.
20. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research.* 3rd ed. Malden, MA: Blackwell Science; 1994.
21. Everitt BS. *Statistical Methods for Medical Investigations.* New York, NY: Oxford University Press; 1989.
22. Campbell MJ, Machin D. *Medical Statistics: A Commonsense Approach.* Chichester, United Kingdom: Wiley; 1990.
23. Fisher LD, Dixon DO, Herson J, et al. Intention-to-treat in clinical trials. In: Peace KE, ed. *Statistical Issues in Drug Research and Development.* New York, NY: Marcel Dekker; 1990:331-350.