CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

1

Clinical Trial Basics

We will present here a brief review of some of the key aspects of therapeutic clinical trials. In oncology, clinical trials are often categorized into phase I, phase II, and phase III trials. Phase I trials are conducted to determine the maximum dose at which a new drug can be delivered in a defined schedule of administration before dose-limiting toxicity occurs. Phase I trials may also evaluate the pharmacokinetics of the drug administration schedule and the pharmacodynamics of whether the drug inhibits its molecular target. Phase II trials are conducted to identify whether a new drug has antitumor activity when administered as a single agent or whether it contributes to the antitumor activity of other drugs. Phase II trials are traditionally conducted in patients with a particular histologic diagnosis and stage of disease. Traditionally, phase II trials of chemotherapeutic drugs are conducted in a wide range of types of cancer to screen for activity sufficiently great to warrant a phase III trial. With the advent of molecularly targeted drugs, an increasingly important objective of phase II trials is to develop a pretreatment biological measurement, that is, a *biomarker*, that can be used to identify the patients whose tumors are the most likely to benefit from the drug. Phase II trials do not generally establish the medical utility of a new drug or new regimen; that is the role of phase III trials. Phase II trials generally use an intermediate end-point that reflects antitumor activity but has not been established as a valid measure of patient benefit. The phase II trials determine whether the new drug is sufficiently promising to evaluate in a larger phase III trial and, if so, what the target population should be and how the new drug should be administered.

Phase III clinical trials are generally large studies in which patients who satisfy predefined eligibility criteria are randomized to receive either

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

the new regimen or a control regimen, usually representing a standard of care. There is generally a primary end-point, or measure of effectiveness, that represents a direct measure of patient benefit such as survival or survival without evidence of disease. The end-point is the same for both the new treatment being evaluated and the control, and great care is taken to plan follow-up surveillance of the patients so that the end-point can be evaluated equivalently for the two treatment arms. The plans for accruing, evaluating, and treating patients, collecting data, and performing data analyses are rigorously specified in a written protocol so that a medically meaningful and statistically reliable assessment of a well-defined population of patients, carefully staged and homogeneously treated, can result (Simon, 2011, 828; Piantadosi, 2005; Crowley and Hoering, 2012).

Statistically, most phase III clinical trials have been structured to test a single null hypothesis that the distribution of outcomes for the new treatment is equivalent to that of the control with regard to the primary end-point overall for all randomized patients. The *intention to treat principle* is observed in the analysis; that is, all eligible randomized patients are included in the primary null hypothesis test regardless of whether or not they received their assigned treatment as defined in the protocol. The intention to treat principle is counterintuitive to many physicians. Why retain a patient in the treatment arm to which he or she was randomized if he or she did not receive that treatment or if the treatment wasn't administered in the way it was intended (Piantadosi, 2005, 829; Green, Benedetti, & Crowley, 2003, 322)? The purpose of the intention to treat principle is to ensure that the prognostic comparability created by randomization is not destroyed by the exclusion of prognostically unfavorable patients from one arm more than the other. By comparing the outcomes of the patients as randomized, one avoids false positive findings resulting from biased exclusions. The intention to treat principle may increase the false negative rate (i.e., decrease the statistical power) of the trial. In oncology clinical trials, it is common to exclude patients who were not actually eligible for the clinical trial but not to exclude any eligible randomized patients. If there are numerous major treatment violations for eligible randomized patients, the credibility of the trial may be compromised to an extent that cannot be rescued by statistical analysis.

CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

At the time of the final analysis, a single null hypothesis of no average treatment effect is tested by computing a test statistic that summarizes a difference in average outcomes between the group randomized to the new treatment and the control group. With survival or disease-free survival data, a log-rank statistic is used. The probability of obtaining a value of the test statistic as great as that computed from the data is calculated under the assumption that the null hypothesis is true. That probability is called the *p value* or the statistical significance level. A one-sided p value is the probability, under the null hypothesis, of obtaining a value of the test statistic as great as that computed from the data and in the direction favoring the new treatment. A two-sided p value is the probability under the null hypothesis of obtaining a value of the test statistic as great in absolute value as that computed from the data in either direction. Although most phase III clinical trials compare a new treatment to a control, results are usually only reported as statistically significant if the two-sided p value is less than 0.05. This assures that only 2.5% of phase III clinical trials will be reported as finding new treatments statistically significantly better than control groups.

The size and duration of phase III clinical trials are usually planned so that if the true degree of benefit of the new treatment versus control is of a prespecified magnitude Δ, then the probability of obtaining a statistically significant result will be large, usually 80% or 90%. The Δ value is called the *treatment effect* to be detected and the 80% or 90% is called the *statistical power*. The power is a function of the Δ value, the number of patients, and the follow-up time. For survival or disease-free survival data, a commonly used formula is

$$E_{\text{tot}} = 4\frac{\left(k_{1-\alpha} + k_{1-\beta}\right)^2}{\Delta^2}. \tag{1.1}$$

In formula (1.1), E_{tot} denotes the total number of events to be detected at the time of the final analysis. If the primary end-point is survival, then an event is death. If the primary end-point is disease-free survival, then an event is the earlier of disease recurrence or death. The constants $k_{1-\alpha}$ and $k_{1-\beta}$ in the numerator of (1.1) are percentiles of the standard normal distribution, α is the desired one-sided significance level of the test (usually 0.025), and $1-\beta$ is the desired statistical power. For

CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

$\alpha = 0.025$ and $1 - \beta = 0.80, k_{1-0.025} = 1.96$ and $k_{0.80} = 0.84$. For 90% power, the latter constant becomes 1.28. This approach to sample size planning is based on the assumption of proportional hazards for survival or disease-free survival data. Variable Δ represents the natural logarithm of the ratio of the hazard of death for a patient on the new treatment relative to the hazard of death for a patient on the control. With a proportional hazards model, the ratio of hazard of death at time t for a patient on the new treatment relative to a patient on control is the same for all times t. Phase III clinical trials are often planned to detect reductions in the hazard by 25–40%. A 33% reduction in hazard corresponds to $\Delta = \log(0.67) = -0.40$.

For proportional hazards models, the statistical power is determined by the number of events at the time of final analysis rather than by the number of patients. The number of total events at any time is, however, a function of both the number of patients and the follow-up time relative to the survival distributions. For such survival studies, the timing of the final analysis is best indicated in terms of number of events, not absolute calendar time.

A very important feature of (1.1) is that the required number of events is proportional to the reciprocal of the square of the size of the treatment effect to be detected, Δ. This is generally true, not just for proportional hazard models, nor just for survival data. A small increase in the size of the treatment effect to be detected results in a large decrease in the required size of the study. This provides an important motivation for the search for predictive biomarkers that will enable the eligibility criteria to be restricted to patients for whom the treatment effect is likely to be large. Use of even an imperfect predictive biomarker can result in a large reduction in the required size of the study. The size of the treatment effect can be increased either by excluding patients unlikely to do well on the new treatment or by excluding patients who do very well on the control.

The process of planning the size of a clinical trial sometimes ignores the interim analyses that will be conducted. The statistical features of the interim analyses must, however, be detailed in the protocol. The type I error of the clinical trial is the probability that the null hypothesis is falsely rejected at any analysis, interim or final. Consequently, the threshold

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

significance levels for declaring statistical significance at individual interim and final levels must be reduced below 0.05 if the total type I error is to be limited to 0.05 (Pocock, 1982, 425; Fleming, 1989, 303; Jennison & Turnbull, 1999, 344). One conservative approach would be to have each threshold be $0.05/(I + 1)$, where I is the number of interim analyses. This approach is rarely used in oncology trials. A more common approach is to use a threshold of 0.045 for the final analysis and to use very extreme thresholds like 0.001 for the interim analyses (Haybittle, 1971, 332; O'Brien & Fleming, 1979, 406). This has little effect on the statistical power computed ignoring the interim analyses but provides relatively little likelihood of stopping early. For large multicenter clinical trials, the latter philosophy is often desired because the interim analyses are conducted with incompletely quality-controlled data.

For most phase III clinical trials, the results of the interim analyses are kept blinded to investigators entering patients in the trial and are reviewed by a data safety monitoring committee consisting of individuals with no conflict of interest and who are not entering patients in the trial (Smith et al., 1997, 504; Ellenberg, Fleming, & DeMets, 2002, 292). The purpose of the data monitoring committee is to maintain the equipoise of physicians who enter patients in the clinical trial while ensuring that the patients are protected.

If the null hypothesis is rejected at an interim or final analysis, then generally, the new treatment is recommended for all future patients who satisfy the eligibility criteria for the study. If the null hypothesis is not rejected, then the new treatment is not recommended for regulatory approval or future use outside clinical trials. Although subset analyses are often performed, those results are generally viewed skeptically by statisticians (Fleming, 1995, 300; Pocock et al., 2002, 426). Some statisticians use the rule of thumb of not believing the subset analysis unless the primary overall null hypothesis is rejected. This reflects the importance to statisticians of protecting the overall type I error for the study. It also reflects an implicit prior belief that new treatments usually do not work, so it is better to believe that some subsets do not benefit when the overall null hypothesis is rejected than it is to believe that some subsets do benefit when the overall null hypothesis is not rejected. This focus on the overall analysis was reflective of an era of blockbuster drugs for

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

homogeneous diseases. This approach has protected physicians from false positive results based on post-hoc subset analyses with ineffective drugs. The approach has led to overtreatment of many patients based on statistically significant but small average treatment effects in large clinical trials with broad eligibility criteria. For study of molecularly heterogeneous diseases in which treatment effects are expected to vary among patients, the methods described in this monograph for development of companion diagnostics for refining therapeutic decision making take on increased importance.

© in this web service Cambridge University Press www.cambridge.org

2

Actionable Prognostic Biomarkers

Biological measurements used to inform treatment selection are sometimes called biomarkers, but the term invites misinterpretation. Many people think of biomarkers as measures of disease activity, increasing as the disease progresses and decreasing as the disease responds. Such disease biomarkers would have considerable utility as surrogate end-points for clinical trials. Regulatory agencies are, of course, very concerned about accepting a surrogate end-point as a basis for drug approval. Although biomarkers are commonly used as end-points in phase I and phase II clinical trials to establish that a drug inhibits its target or has antidisease activity and for selecting among doses, very stringent criteria have been established for validating surrogate end-points for use in phase III clinical trials. It is generally very difficult to establish that a biological measurement is a valid "disease biomarker." Our focus here is on prognostic and predictive baseline biomarkers, not on surrogate end-points.

Prognostic markers are pretreatment measurements that provide information about long-term outcome for patients who are either untreated or receive standard treatment. Prognostic markers often reflect a combination of intrinsic disease factors and sensitivity to standard therapy. Predictive biomarkers identify patients who are likely or unlikely to benefit from a specific treatment. For example, HER2 amplification is a predictive biomarker for benefit from trastuzumab. A predictive biomarker may be used to identify patients who are poor candidates for a particular drug; for example, colorectal cancer patients whose tumors have KRAS mutations are poor candidates for treatment with anti-EGFR monoclonal antibodies. Most of the following chapters address the development and validation of predictive biomarkers for guiding the use of

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

a new treatment. In this chapter, however, we discuss the development and validation of prognostic biomarkers that have medical utility for informing treatment decisions.

Although many diseases feature a large literature of prognostic factor studies, relatively few such tests are recommended by professional societies, reimbursed by payers, or widely ordered by practicing physicians (Pusztai, 2004, 641). Published studies are rarely planned with an intended use in mind. The cases included represent a convenience sample of patients for whom preserved tissue is available. These patients are often heterogeneous with regard to treatment, staging procedures used, and extent of disease. The prognostic factors identified are often not helpful in making treatment decisions. The studies are often motivated by the hope of better understanding the pathogenesis of the disease rather than by plans for developing a test with medical utility.

In discussing medical tests, three types of validity can be distinguished: *analytical validity*, *clinical validity*, and *medical utility* (Simon, Paik, & Hayes, 2009, 684). Analytical validity originally meant that the assay provides an accurate measurement of the quantity that it claims to measure. In some cases, however, there is no gold-standard measurement on which to base the comparison. In such cases, analytical validity is taken to mean that the measurement is reproducible and robust over time, within and between laboratories (Cronin et al., 2007, 806).

Clinical validity means that the test result correlates well with some clinical end-point. Most prognostic factor publications demonstrate some form of clinical validity but not analytical validity or medical utility. Medical utility means that the test result is actionable in a way that results in patient benefit. The action that a test result can inform is often treatment selection. If the test tells something about the patient's disease but there is nothing one can do with that knowledge, then there is no medical utility. In some cases, the patient might want to have that increased knowledge of his or her prognosis, but few payers reimburse for prognostic measurements that are not medically indicated. If the knowledge is actionable but could have been obtained from standard prognostic factors, then there is no incremental medical utility.

© in this web service Cambridge University Press www.cambridge.org

9 Actionable Prognostic Biomarkers

How can a new prognostic marker have medical utility? When the standard of care (SOC) is intensive therapy, this can be accomplished by identifying patients who have such good prognosis with conservative treatment that they may choose to forgo more intensive therapy. For example, the Oncotype DX recurrence score was developed for women with newly diagnosed breast cancer that expressed hormonal receptors and had not apparently disseminated to the lymph nodes or beyond (Paik et al., 2004, 411). At the time of development, the SOC for such women was treatment with hormonal therapy and cytotoxic chemotherapy. The test was developed to identify which breast cancer patients had such good prognoses on hormonal therapy alone that they could opt to forgo chemotherapy. The action involved was withholding chemotherapy, and the benefit to the patient was having an excellent outcome while avoiding the toxicity of chemotherapy.

Developing Oncotype DX as a prognostic marker that would inform treatment decisions required studying patients whose tumors expressed hormonal receptors and had not disseminated to lymph nodes or beyond and who received hormonal therapy as their only systemic treatment. The intended use of the test determined both the selection of patients and the analysis of the study performed for developing the test. Such focused development is unusual and is the primary reason why prognostic markers with medical utility are so unusual.

The analysis and interpretation of the Oncotype DX prognostic study was also unusual. Most prognostic factor studies are analyzed as statistical exercises in significance testing and reporting hazard ratios. Such methods are mostly irrelevant for identifying medical utility, where the key issue is whether the marker identifies a subset of patients who have such excellent prognoses on conservative treatment that they are unlikely to benefit to a meaningful degree from higher-intensity regimens. A more appropriate analysis would simply involve computing the relationship of biomarker value to prognosis with confidence limits. For example, suppose that the marker identifies a subset of patients with an apparent cure rate of more than 95%. Because their outcome on low-intensity treatment is so good, they cannot benefit much, in absolute terms, from higher-intensity treatment. A reduction in hazard of failure of 30%, which is substantial for cancer

CAMBRIDGE

Cambridge University Press
978-1-107-00880-9 - Genomic Clinical Trials and Predictive Medicine
Richard M. Simon
Excerpt
More information

treatments, would provide an increase in the cure rate of only 5% × 30%, or 1.5% (from 95% to 96.5%), in absolute terms, and this may not be worth the side effects of chemotherapy for many patients. If, however, the cure rate for the good-prognosis group were only 85%, then the potential benefit of chemotherapy might be 15% × 30%, or 4.5 % (from 85% to 89.5%), in absolute terms, and the prognostic marker would have less certain utility for informing a decision to withhold more intensive treatment, particularly when uncertainty in the estimate of the cure rate is taken into account.

In discussing prognostic markers, treatment is often ignored. However, for developing a prognostic marker with medical utility, treatment is of fundamental importance. Utility for withholding intensive therapy can only be established if two conditions prevail: (1) the SOC for the target population is intensive treatment and (2) the prognostic analysis is based on patients who did not receive the intensive treatment. These two conditions prevailed in the development of Oncotype DX. At the time of development, the SOC for most stage I breast cancer patients with hormone-receptor-positive tumors was hormonal treatment plus chemotherapy, but the development was based on data from a clinical trial performed years earlier, when the SOC was hormonal therapy alone. For examples of much less effective or actionable prognostic classifiers, see the review by Subramanian and Simon (2010, 719–720) of prognostic gene expression signatures for early lung cancer.

The Oncotype DX recurrence score is a weighted average of the expression levels of 21 genes. The MammaPrint score is a classifier based on the expression levels of 70 genes (van-de-Vijver et al., 2002, 103; van't-Veer et al., 2002, 90). We include such composite tests as prognostic markers because the manner in which a marker is used in a validation clinical trial is independent of whether it is based on a single measurement or a summary of multiple measurements. For composite markers like Oncotype DX and MammaPrint, however, it is essential that the way the multiple measurements are combined is completely specified, that is, "locked down." It is not enough just to specify the genes involved; the weighting factors used in combining the genes and the cut points, if any, used for treatment decisions must also be specified. Oncotype DX is based on a weighted average of 21 genes, and so the way those expression

© in this web service Cambridge University Press www.cambridge.org