Published in final edited form as:

Stat Biosci. 2011 December ; 3(2): 145–168. doi:10.1007/s12561-011-9042-5.

Innovative Clinical Trial Designs:

Toward a 21st-Century Health Care System

Tze L. Lai and

Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA

Philip W. Lavori

HRP Redwood Building, T152a, Stanford, CA 94305-5405, USA

Tze L. Lai: lait@stanford.edu; Philip W. Lavori: lavori@stanford.edu

Abstract

Whereas the 20th-century health care system sometimes seemed to be inhospitable to and unmoved by experimental research, its inefficiency and unaffordability have led to reforms that foreshadow a new health care system. We point out certain opportunities and transformational needs for innovations in study design offered by the 21st-century health care system, and describe some innovative clinical trial designs and novel design methods to address these needs and challenges.

Keywords

Cancer clinical trials; Comparative effectiveness research; Embedded experiments; Personalized therapies; Sequential design and randomization

1 Introduction

Clinical trials have played an important role in evidence-based medicine and in drug development. The statistical methodology underlying their design and analysis has witnessed important advances in the past 50 years and laid the foundation of today's standard designs of randomized clinical trials. This period also had many spectacular advances in the biomedical sciences, leading to treatments for many diseases and substantial increase in life expectancy. The next challenge is to come up with breakthroughs in developing treatments of complex diseases such as cancer and improving decision-making in the management of chronic diseases. In his 2010 budget request, the Director of the National Cancer Institute has earmarked "re-engineering" cancer clinical trials as a research initiative. We review commonly used designs in current cancer clinical trials in Sect. 2.2 and describe in Sect. 3 innovative design methodologies and emerging approaches to their re-engineering.

The second part of the title is inspired by Arrow et al. [3] whose eight-point plan for health care reform toward a 21st-century health care system includes the following:

Correspondence to: Tze L. Lai, lait@stanford.edu.

- Establish a securely funded, independent agency to sponsor and evaluate research on the comparative effectiveness of drugs, devices, and other medical interventions. (Point 2)
- Create a national health database with the participation of all payers, delivery systems, and others who own health care data. Agree on methods to make de-identified information from this database on clinical interventions, patient outcomes, and costs available to researchers. (Point 5)

The new health care system envisioned in [3] has a universal electronic health record, fully linked history of treatments and outcomes in the database, stored biomarkers together with genetic data and tissues, universal consent for observational research, and computer-assisted decision support for patient care that is adapted to the individual patient's evolving health status. Comparative effectiveness research (CER; see Point 2 above) in this new environment is a research activity to evaluate the effectiveness of approved initial treatments given to patients presenting with a new illness, which supports the coverage decisions of the new health care system. But it must also aim at developing the best ways to use adaptive (dynamic) treatment strategies for ongoing patient care across time as the patient's disease evolves. Most CER research does not involve randomized assignment, and relies instead on the presumed ability to adjust successfully for treatment selection effects (as in [3] above, and the recent [91]).

Despite considerable skepticism and warnings on the part of methodologists [73], most projections of CER in the future reflect even greater reliance on observational methods (statistical adjustment and instrumental variable methods). The low impact of recent large-scale effectiveness trials has prompted several calls for improvement of the trial designs. As will be shown in Sect. 2.1, traditional randomized clinical trial designs are widely viewed as too costly and inefficient for CER studies, especially by those who (perhaps optimistically) believe that statistical maneuvers can compensate for bias due to selection by indication. Section 4 describes some innovative design methods, with examples, that are promising to meet these new challenges. Some concluding remarks on new directions for 21st-century clinical trial designs are given in Sect. 5.

2 Limitations of Standard Clinical Trial Designs

2.1 Standard Randomized Clinical Trial Designs and an Illustrative Example

The typical late-phase randomized trial perfected in the 20th century chooses two or three treatments, randomizes the 1%–5% of eligible patients who consent, and spends the next 5 to 10 years in a constant struggle to increase lagging accrual and reduce protocol violations. At regular intervals there will be Data and Safety Monitoring Board meetings, often guided by some kind of group sequential rule for early termination. At the end of the trial, the analysis will follow the Intent-to-Treat (ITT) principle, counting all outcomes according to the randomization, regardless of intervening changes in adherence. There will follow months or years of arguments about generalizability (external validity), the effects of uncontrolled intervening treatments, the relevance of results after such a lapse of time since the study was proposed, and the validity and importance of subgroup analyses that appear to reveal different results.

The Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) exemplifies both an inspiring scientific and organizational triumph and a disappointing impact. ALLHAT was a randomized, double-blind, multi-center clinical trial sponsored by the National Heart Lung and Blood Institute in conjunction with the Department of Veteran's Affairs. Designed to recruit about 40,000 patients, its aim was to determine if the combined incidence of fatal coronary heart disease (CHD) and non-fatal myocardial infarction differs between diuretic (chlorthalidone) treatment and each of three alternative antihypertensive pharmacologic treatments: a calcium antagonist (amlodipine), an ACE inhibitor (lisinopril), and an alpha adrenergic blocker (doxazosin). A lipid-lowering subtrial (>10,000 patients) was designed to determine whether lowering cholesterol with an HMG Co-A reductase inhibitor (pravastatin), in comparison with usual care, reduced mortality in a moderately hy-percholesterolemic subset of participants. ALLHAT was the largest antihypertensive trial ever conducted, and the second largest lipid-lowering trial. It recruited many patients over age 65, women, African-Americans and patients with diabetes. The study was conducted between 1994 and 2002 largely in community practice settings. In ALLHAT, hypertensive patients were randomly assigned to receive one of four drugs in a double-blind design, and a limited choice of second-step agents were provided for patients not controlled on first-line medication. Patients were followed every three months for the first year and every four months thereafter for an average of six years of follow-up. This landmark study cost over $100 million, and the final results were presented in 2002 [1]. An accompanying editorial by Appel [2] led with "Quite simply, the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) is one of the most important trials of antihypertensive therapy."

Yet, Andrew Pollack of *The New York Times* on November 28, 2008, wrote under the headline "The Minimal Impact of A Big Hypertension Study":

> The findings, from one of the biggest clinical trials ever organized by the federal government, promised to save the nation billions of dollars in treating the tens of millions of Americans with hypertension. . . .

The article further quoted C. Furberg, chair of ALLHAT, as saying "The impact was disappointing." The reasons cited for this "blunted impact" include the difficulty of persuading doctors to change, scientific disagreement about the government's interpretation of the results, and heavy marketing by pharmaceutical companies of their own drugs, paying speakers to "publicly interpret the Allhat results in ways that made their products look better." Some quotes from leading experts included the following.

Dr. Sean Tunis: "There's a lot of magical thinking that it will all be science and won't be politics."

Dr. Robert Temple: "This is the largest and best attempt to compare outcomes we are ever going to see, and people are extremely doubtful about whether it has shown anything at all."

Dr. Carolyn Clancy: "While clinical trials are the gold standard, they are costly and time-consuming . . . . You might be answering a question that by the time you are done, no longer feels quite as relevant."

2.2 Current Cancer Clinical Trial Designs and Their Limitations

The conventional "phase" designation arose in the context of the development of drug treatments for cancer, particularly cytotoxic chemotherapy. Traditionally, safety is evaluated in a Phase I dose-finding trial, efficacy is evaluated in terms of some early binary outcome in Phase II, often referred to as "response", and survival or disease-free survival time is evaluated in a randomized Phase III trial. This sequence has proven to be somewhat dysfunctional, and is breaking down with the development of new, often cytostatic rather than cytotoxic, targeted treatments. In typical Phase I studies in the development of relatively benign drugs, the drug is initiated at low doses and subsequently escalated to show safety at a level where some positive response occurs, and healthy volunteers are used as study subjects. This paradigm does not work for cytotoxic therapies for diseases like cancer, for which a non-negligible probability of severe toxic reaction has to be accepted to give the patient some chance of a favorable response to the treatment. Moreover, in many such situations, the benefits of a new therapy may not be known for a long time after enrollment but toxicities manifest themselves in a relatively short time period. Therefore patients (rather than healthy volunteers) are usually used as study subjects, and given the hoped-for (rather than observed) benefit for them, one aims at an acceptable level of toxic response in determining the dose. Current designs for most Phase I cancer trials are an ad hoc attempt to reconcile the objective of finding a *maximum tolerated dose* (MTD) with the ethical demands for protecting the study subjects from toxicities in excess of what they can tolerate. A commonly used design is to treat groups of three patients sequentially, starting with the smallest of an ordered set of doses. Escalation occurs if no toxicity is observed in all three patients; otherwise an additional three patients are treated at the same dose level. If only one of the six patients has toxicity, escalation again continues; otherwise the trial stops, with the current dose declared as the MTD if two of the six patients have toxicity, and with the lower dose declared as MTD if more than two of the six patients have toxicity. As pointed out by Storer [81], these designs, commonly referred to as 3-plus-3 designs, are difficult to analyze since even a strict quantitative definition of MTD is lacking, "although it should be taken to mean some percentile of a tolerance distribution with respect to some objective definition of clinical toxicity," and the "implicitly intended" percentile seems to be the 33rd percentile (related to 2/6). Numerous simulation studies have shown that they are inferior to the sequential designs described in Sect. 3.1 in terms of both safety and reliability in estimating the MTD.

Besides the ethical issue of safe treatment of patients currently in the trial, a traditional Phase I design also has the goal of determining the MTD for a future Phase II cancer trial, and needs an informative experimental design to meet this goal. Von Hoff and Turner [94] have documented that the overall response rates in Phase I trials are low and that substantial numbers of patients are treated at doses that are retrospectively found to be non-therapeutic. Eisenhauer et al. [29, p. 685] have pointed out that "with a plethora of molecularly defined antitumor targets and an increasingly clear description of tumor biology, there are now more antitumor candidate therapies requiring Phase I study than ever," and that "unless more efficient approaches are undertaken, Phase I trials may be a rate-limiting step in the process of evaluation of novel anticancer agents." To address this difficulty, they propose to develop and use (a) methods to determine more informative starting doses, (b) pharmacokinetics-

guided dose escalation methods, and (c) model-based methods for dose determination. There have been ongoing methodological developments along these lines and a comprehensive methodology is emerging, as will be described in Sect. 3.1.

Vickers et al. [93, p. 927] give the following description of a typical Phase II study of a novel cancer treatment:

> A cohort of patients is treated, and the outcomes are related to the prespecified target or bar. If the results meet or exceed the target, the treatment is declared worthy of further study; otherwise, further development is stopped. This has been referred to as the 'go/no-go' decision. Most often, the outcome specified is a measure of tumor response, e.g., complete or partial response using Response Evaluation Criteria in Solid Tumors, expressed as a proportion of the total number of patients. Response can also be defined in terms of the proportion who have not progressed or who are alive at a predetermined time (e.g., one year) after treatment is started.

The most widely used designs for these single-arm Phase II trials are Simon's [76] two-stage designs, which allow early stopping of the trial if the treatment has not shown beneficial effect that is measured by a Bernoulli proportion. These designs are optimal in the sense of minimizing the expected sample size under the null hypothesis of no viable treatment effect, subject to Type I and II error probability bounds. Given a maximum sample size M, Simon considered the design that stops for futility after $m < M$ patients if the number of patients exhibiting positive treatment effect is r_1 ($\leq m$) or fewer, and otherwise treats an additional $M - m$ patients and rejects the treatment if and only if the number of patients exhibiting positive treatment effect is r_2 ($\leq M$) or fewer. Simon's designs require that a null proportion p_0, representing some "uninteresting" level of positive treatment effect, and an alternative $p_1 > p_0$ be specified. The null hypothesis is $H_0 : p \leq p_0$, where p denotes the probability of positive treatment effect. The Type I and II error probabilities $P_{p_0}\{\text{Reject } H_0\}$, $P_{p_1}\{\text{Accept } H_0\}$ and the expected sample size $E_{p0} N$ can be computed for any design of this form, which can be represented by the parameter vector (m, M, r_1, r_2). Using computer search over these integer-valued parameters, Simon [76] tabulated the optimal designs for different values of (p_0, p_1). He also introduced minimax two-stage designs with the smallest maximum sample size subject to the error probability constraints. Note that these designs are group sequential designs with two groups and early stopping only for futility.

Whether the new treatment is declared promising in a Phase II trial depends strongly on the prescribed p_0 and p_1. In their systematic review of 134 papers reporting Phase II trials in *J. Clin. Oncology*, Vickers et al. [93] found 70 papers referring to historical data for their choice of the null or alternative response rate, and that nearly half (i.e., 32) of these papers did not cite the source of the historical data used, while only nine gave clearly a single historical estimate of their choice of p_0. Moreover, no study "incorporated any statistical method to account for the possibility of sampling error or for differences in case mix between the Phase II sample and the historical cohort." Trials that failed to cite prior data appropriately were significantly more likely to declare an agent to be active (83% versus 33%). They conclude that "more appropriate use of historical data in Phase II design will improve both the sensitivity and specificity of Phase II for eventual Phase III success,

avoiding both unnecessary definitive trials of ineffective agents and early termination of effective drugs for lack of apparent benefit." They also note that uncertainty in the choice of p_0 and p_1 can increase the likelihood that (a) a treatment with no viable positive treatment effect proceeds to Phase III, or (b) a treatment with positive treatment effect is prematurely abandoned at Phase II.

Besides the difficulties associated with the choice of p_0 in the typical Phase II trial, other important related problems are failure to account for known patient heterogeneity, artificially simplifying a complex outcome to a binary "response" that results in a loss of information, and using early outcomes that are convenient but not substantively associated with survival time. To circumvent the problem of choosing p_0, and some of the other difficulties described above, randomized Phase II designs have been advocated [72, 74]. In particular, it is argued that randomized Phase II trials are needed before proceeding to Phase III trials when (a) a known accurate historical control rate is not available, due to either incomparable controls, few control patients (so that the estimated control rate has large variation), or a different outcome than "antitumor activity"; or (b) the goal of Phase II is to select one from multiple candidate treatments or multiple doses for use in Phase III. Thall et al. [85] have extended Simon's two-stage design to randomized Phase II trials. However, because randomized designs typically require a much larger sample size than single-arm designs and there are multiple research protocols competing for patient recruitment, few Phase II cancer studies have been randomized with internal controls.

Whereas the endpoint of a Phase II cancer trial is response that can be measured within a relatively short period of time after treatment, the clinically definitive endpoint in Phase III cancer trials is usually time to event, such as time to death or time to progression, which is often of long latency. Interim reviews are routinely incorporated in the design and execution of long-term clinical trials with survival endpoints, at least for safety. The past three decades have witnessed major advances in the sequential design and analysis of clinical trials with failure-time endpoints and interim analyses. In particular, the technical difficulties in inference due to staggered patient entry and censoring at time of interim analysis have been resolved. On the other hand, it is well known that the success rate of Phase III cancer clinical trials is low [45]. Preliminary data at the end of the early-phase trials and the reported survival curves of related treatments in the literature are often inadequate to determine the sample size and duration of a Phase III trial, and whether it should even be launched at all.

The lack of information also makes it difficult to decide which survival endpoint, overall survival or progression-free survival (PFS), should be used. A recent example from a joint ECOG/Genentech study illustrates the difficulty of this issue. In 2008 the FDA granted approval for the use of the drug bevacizumab (Avastin) in combination with paclitaxel for advanced breast cancer, under the accelerated approval mechanism, which requires ongoing study. In July 2010 an FDA panel (the Oncology Drugs Advisory Committee, ODAC) recommended 12 to 1 to remove the breast cancer indication. The initial FDA panel review in December 2007 resulted in a split decision narrowly recommending against approval, but the FDA reversed the panel in February of 2008. The 2007 panel reviewed results from a clinical trial involving 722 women with recurrent/metastatic breast cancer. Adding bevacizumab to paclitaxel improved median PFS to 11.3 months, versus 5.8 in the paclitaxel

alone arm. Survival, however, was not significantly improved: median 26.5 months with bevacizumab and paclitaxel, compared with 24.8 months for paclitaxel alone ($P = 0.14$). In the subsequent review, based on two new studies, the PFS advantage narrowed, and no survival benefit was seen. On December 16, 2010 the FDA announced that it is "recommending removing the breast cancer indication from the label for bevacizumab (Avastin) because the drug has not been shown to be safe and effective for that use." The generic problem exposed by the debate on the proper endpoint is that the effects of salvage treatment at disease progression are confounded with front-line treatment effects (as assessed by the intention-to-treat method). We note that this motivates consideration of multi-stage treatment trials, discussed below.

3 Innovative Designs Toward Re-engineering Cancer Clinical Trials

3.1 Single-Arm Dose-Finding Studies

While investigators writing Phase I cancer trial protocols find the traditional 3-plus-3 design mentioned in Sect. 2.2 and various step-up/down variants in the literature within their comfort zone to gain IRB approval to try the new treatment on human subjects and thereby obtain some publishable data and experience, there is the ethical dilemma that patients in the trial are treated at sub-therapeutic albeit safe doses. As pointed out by Bartroff and Lai [9, 10], there are two conflicting goals in a Phase I cancer trial: (a) determination of the MTD for a future Phase II trial, which they call "collective ethics," and (b) safe treatment of current patients in the trial, preferably at doses near the unknown MTD (to improve the chances of benefit), which they call "individual ethics." An efficient Phase I design should strike a good balance between these two conflicting goals. Because of funding and time constraints, Phase I trials typically involve a relatively small number (between 20 and 30) of patients. In view of this sample size, efficiency seems to be achievable only by a model-based approach. In model-based methods, a patient's dose-limiting toxicity (DLT) y for treatment at dose level x is usually modeled by a binary random variable taking values 0 or 1, such that $y = 1$ indicates a DLT g and whose distribution depends on x and an unknown vector θ of parameters through the function $F_\theta(x) = P(y = 1|\text{dose} = x)$. The MTD is then the pth quantile of F_θ, i.e., $\text{MTD}=F_\theta^{-1}(p)$. A widely used working model for F_θ is the two-parameter logistic regression model

$$F_\theta(x)=1/(1+e^{-(\alpha+\beta x)}),$$

where $\beta > 0$ and $\theta = (\alpha, \beta)$, for which the MTD is equal to $[\log(p/(1 - p)) - \alpha]/\beta$.

Several Bayesian model-based methods have been proposed in the literature, assuming a prior distribution Π_0 of θ and using the posterior distribution Π_k of θ based on $(x_1, y_1), \ldots, (x_k, y_k)$ to estimate the MTD and determine the dose x_{k+1} for the next patient. Denoting the MTD by η, the Bayes estimate of η with respect to squared error loss is the posterior mean $E_{\Pi_k}(\eta)$, which O'Quigley et al. [70] proposed to use in their "continual reassessment method" (CRM). Babb et al. [5] pointed out that the symmetric nature of the squared error loss or its close relative, the absolute error loss, may not be appropriate for modeling the

toxic response to a cancer treatment, and proposed the "escalation with overdose control" (EWOC) method, which is a Bayesian design with respect to the asymmetric loss function

$$h(x)=\begin{cases} \omega(\mathrm{MTD}-x) & \text{if } x \le \mathrm{MTD}, \\ (1-\omega)(x-\mathrm{MTD}) & \text{if } x \ge \mathrm{MTD}, \end{cases}$$

where the chosen constant $0 < \omega < 1/2$ is called the "feasibility bound." Note that this loss function penalizes an overdose $x = \mathrm{MTD} + \delta$ more than an underdose $x = \mathrm{MTD} - \delta$ of the same magnitude $\delta > 0$. EWOC can be shown to be equivalent to estimating the MTD at each stage by the ωth quantile of the posterior distribution of the MTD. A sequence of doses x_n is called *Bayesian feasible* at level $1 - \omega$ if $P_{\Pi_{k-1}}(\eta \ge x) \ge 1 - \omega$ for all $n \ge 1$, and the EWOC doses have been shown to be optimal among the Bayesian-feasible ones.

CRM and EWOC can be regarded as focusing on individual ethics as they treat the next patient at the dose x that minimizes $E_{\Pi_k}[l(\eta, x)]$, where l is a loss function, which is the squared error loss for CRM and the piecewise linear loss for EWOC. Designs that focus on collective ethics instead have also been proposed. Noting that the explicitly stated objective of a Phase I cancer trial is to estimate the MTD, Whitehead and Brunier [97] considered Bayesian sequential designs that are optimal, in some sense, for this estimation problem. Haines et al. [37] made use of the theory of optimal design of experiments [4, 24, 31] to construct Bayesian *c*- and *D*-optimal designs, and further imposed a relaxed Bayesian feasibility constraint on the design to avoid highly toxic doses. Optimal design theory involves a design measure ξ on the dose space X, and a sequential design updates the empirical design measure ξ_{n-1} at stage n by changing it to ξ_n with the addition of the dose x_n. The empirical measure ξ_n of the doses $x_1, \ldots, x_n$ up to stage n can be represented by $\xi_n = n^{-1}\sum_{i=1}^{n}\delta_{x_i}$, where δ_x is the probability measure degenerate at x.

Combining individual and collective ethics, Bartroff and Lai [9] consider the overall Bayes risk $E_{\Pi_0}[\sum_{i=1}^{n} l(x_i, \eta) + g(\hat{\eta}, \eta)]$ of a Phase I trial involving n patients. This measures the effect of the dose x_k on the kth patient through $l(x_k, \eta)$, its effect on future patients in the trial through $\sum_{i=k+1}^{n} l(x_i, \eta)$, and its effect on the post-trial estimate $\hat{\eta}$ through $g(\eta, \hat{\eta})$, thereby incorporating the dilemma between safe treatment of current patients in the study and efficient experimentation to gather information about η for future patients. By making use of recent advances in approximate dynamic programming to tackle the problem, they find an approximate solution of the Bayesian optimal design. The resulting design is a convex combination of a "treatment" design, such as EWOC, and a "learning" design, such as a Bayesian *c*-optimal design, thus directly addressing the treatment versus experimentation dilemma inherent in Phase I trials and providing a simple and intuitive design for clinical use. Instead of using dynamic programming to minimize the overall Bayes risk, Bartroff and Lai [10] introduced an alternative approach that involves a simpler loss function composed of two terms, with one representing the individual risk of the current dose and the other representing its impact on the collective risk. Note that these ideas apply even if the goal of the study is not the MTD but is some other target such as the dose yielding a prescribed

probability of response. There will still be a trade-off between the dose that is apparently best for the current patient and the dose chosen for an efficient design to estimate the target dose.

As noted in Sect. 2.2, the MTD determined at the end of a Phase I study is used in a single-arm Phase II trial to test the response rate of the new treatment at the MTD. Although the patients in the Phase II trial also have toxicity data, these data are not used to improve the estimate of MTD for the subsequent Phase III trial. Similarly, the patients in the Phase I study also have response data, but these data are not used, even though the sample size for the Phase II trial is typically below 60, because their associated doses are not set at the MTD estimate. A more efficient method to determine the dose for the subsequent Phase III trial is to use both response and toxicity data from all patients studied, as in the Phase I–II designs proposed by Thall and Russel [84], O'Quigley et al. [71], Braun [16], Ivanova [41], Thall and Cook [83], Bekele and Shen [11], and Ivanova and Wang [42]. These designs use joint modeling of (x_i , y_i , z_i), in which y_i is the same as in Sect. 3.1 and $z_i = 1$ or 0 according to whether the ith subject responds to the treatment. For cytotoxic cancer treatments, both the dose-toxicity curve $P(y_i = 1|x_i = x)$ and the dose-response curve $P(z_i = 1|x_i = x)$ increase with the dose x, and therefore the MTD is the most efficacious dose subject to a prespecified probability of DLT.

The advances in dose-finding studies above are Phase I–II in the sense that they integrate the toxicity data of traditional Phase I studies and the response data of Phase II trials. They do not consider testing the efficacy hypothesis that is the purpose of typical Phase II cancer studies, for which the standard practice is to use Simon's two-stage test. Bryant and Day [18] have extended Simon's design to incorporate toxicity outcomes in the Phase II trial by stopping the trial after the first stage if either the observed response rate is inadequate or the number of observed toxicities is excessive, and by recommending the treatment at the end of the Phase II trial only if there are both a sufficient number of responses and an acceptably small number of toxicities. We are working with our medical colleagues to develop new Phase I–II designs for cytotoxic chemotherapies that start with a Phase I design and then seamlessly merge it into a Phase II trial to test the null hypothesis that the probability $p(\eta)$ of response at the actual MTD η is $\leq p_0$. Note that traditional Phase II trials are designed to test null hypothesis $p(\hat{\eta}) \leq p_0$, where $\hat{\eta}$ is the estimate of η at the end of the Phase I trial. The preceding reformulation of the null hypothesis enables us to update the estimates of throughout the course of the Phase I–II trial.

Another direction of our ongoing research is related to cytostatic cancer treatments. As noted by Ma et al. [62], targeted agents that are "predominantly antiproliferative, rather than cytotoxic" and that "produce less acute toxicity than cytotoxic agents, resulting in a broad therapeutic margin" pose new challenges in Phase I and Phase II trial designs. Several designs that use both toxicity and efficacy outcomes for these targeted agents have been proposed; see [25, 26, 39, 46, 88, 100]. In particular, the last two references consider patient-specific dose-finding designs and dose-finding for drug combinations, respectively. In their recent survey article, LoRusso et al. [59, p. 1714] have noted that the selection of agents for combination studies "remains very difficult, especially as preclinical studies are not well-validated" and recommend that "seamless phase I–II designs, especially for studies

combining therapeutics, should be considered." They also point out that "toxicity remains a relevant endpoint, as does defining an MTD, recognizing that the RD (recommended dose) or BAD (biologically active dose) may be different from the MTD." Because targeted agents may show a plateau on the dose-efficacy curve, pushing a treatment to the MTD "may be to the detriment of the patient and the development of the drug, and may make evaluation of treatment combinations difficult." Instead of dose escalation to determine the MTD, they recommend using plasma drug levels for toxicity evaluation via pharmacokinetics and the inclusion of "relevant blood, tissue, imaging, and physiological biomarkers" as surrogate endpoints for toxicity and for measuring inhibition of a target.

3.2 Biomarker-Based Personalized Therapies: Development and Testing

The development of imatinib (Gleevec), the first drug to target the genetic defects of a particular cancer while leaving healthy cells unharmed, has revolutionized the treatment of cancer. A Phase I clinical trial treating CML (chronic myeloid leukemia) patients with the drug began in June 1998, and within six months remissions had occurred in all patients as determined by their white blood cell counts returning to normal. Note in this connection that Phase I trials may already have highly informative efficacy data, which is one of the motivations of the Phase I–II trials in the preceding section. In a subsequent five-year study on survival, which followed 553 CML patients who had received imatinib as their primary therapy, only 5% of the patients died from CML and 11% died from all causes during the five-year period. Moreover, there were few significant side effects; see Druker et al. [28]. Such remarkable success of targeted therapies has led to hundreds of kinase inhibitors and other targeted drugs that are in various stages of development in the present anti-cancer drug pipeline. However, most new targeted treatments have resulted in only modest clinical benefit, with less than 50% remission rates and less than one year of progression-free survival, unlike a few cases such as trastuzumab in HER2-positive breast cancer, imatinib in CML and GIST, and gefitinib and erlotinib in non-small cell lung cancer. While the targeted treatments are devised to attack specific targets, the "one size fits all" treatment regimens commonly used may have diminished their effectiveness, and genomic-guided and risk-adapted personalized therapies that are tailored for individual patients are expected to substantially improve the effectiveness of these treatments. Personalized medicine provides a way to figure out what is driving the growth of cancer in an individual patient and to ultimately match the patient with the right targeted therapy; see [27].

To achieve this potential for personalized therapies, the first step is to identify and measure the relevant biomarkers. The markers can be individual genes or proteins or gene expression signatures. How to measure them conveniently from patients is also an important consideration. One can use tissue samples from the tumor—fixed versus fresh, or circulating tumor cells, or from serum. One has to determine which biomarker measurement technology to use and its reliability: quantitative rt-PCR, immunohistochemical, phospho-flow, etc. The next step is to select drugs (standard cytotoxins, monoclonal antibodies, kinase inhibitors and other targeted drugs) based on the genetics of the disease in individual patients and biomarkers of drug sensitivity and resistance. The third step is to design clinical trials to provide data for the development and verification of personalized therapies.

Simon [77, 79] and others note the distinctions between biomarker discovery, identification, validation, and subsequent use of a validated biomarker for individualized treatment selection. Different designs are appropriate for each of these steps, and in particular the identification/validation steps may be iterative, as described below. The "targeted designs" described by Maitournam and Simon [63] led to a class of designs aimed at testing the effectiveness of biomarker-guided personalized therapies. In these clinical trials eligibility is restricted to patients who are predicted to respond to the therapy being tested by using genomic technologies. In [77] they are compared with traditional randomized designs having broader eligibility criteria, with regard to the number of patients required for randomization and for screening. Simon [79] described the design of validation studies for comparing a biomarker-based treatment strategy to "standard of care" (a somewhat indefinite strategy making no use of the biomarkers to select treatment), noting the inefficiency of the straightforward randomized comparison of patients receiving treatments based on their biomarkers to those whose treatment selection does not depend on the biomarkers, due to the overlap of treatments actually received by the two groups (Fig. 1 of [79]). Following Freidlin et al. [34], such designs are now commonly called "biomarker-strategy designs." (An example of a biomarker-strategy design trial, comparing a biomarker-directed chemotherapy versus physicians' choice in patients with recurrent platinum-resistant ovarian cancer, is given in [21].)

Simon proposed a design (Fig. 2 of [79]) which excludes patients for whom the biomarker-guided and standard of care treatment choices agree and then randomizes the patients for whom the two strategies make different recommendations to the respective treatments. Such designs are now called "enrichment designs" [34, 64, 80], and presumably offer greater efficiency than biomarker-strategy designs, as long as one can identify in advance of randomization the treatment that will be offered by the standard of care. Another design, called "biomarker-stratified design" in [34] and "all-comers design" in [64], randomizes patients to the treatments themselves, treating the biomarker status as a baseline stratification factor rather than a determinant of treatment, but the analysis plan focuses on the dependence of the treatment effects on the biomarker status.

Traditional designs require large sample sizes for these clinical trials; moreover, they cannot adapt to evolving knowledge about biomarkers. Innovative clinical trial design, "which allows researchers to avoid being locked into a single, static protocol of the trial," can "yield breakthroughs, but must be handled with care" to ensure that "the risk of reaching a false positive conclusion" is not inflated, as pointed out in the recent editorial in *Nature* (April 2010, vol. 464), in which two clinical trials of biomarker-guided personalized therapies, BATTLE and I-SPY2, are reported [56].

The clinical trials BATTLE [57, 102] and I-SPY2 [6] use Bayesian adaptive randomization designs and perform Bayesian inference from the posterior distributions, although the frequentist operating characteristics of the Bayes tests are also assessed by simulation studies under certain assumed models. "The approach has been controversial, but is catching on with both researchers and regulators as companies struggle to combat the nearly 50% failure rate of (cancer) drugs in large, late-stage trials," says Ledford [56]. The Bayesian designs are indeed more efficient than conventional randomized clinical trials under the

model assumptions and accepting the Bayesian inferences, and it is hoped "to drive down the cost of clinical trials 50-fold" for the development of personalized medicine, otherwise "drug companies (won't) be interested in taking the risk of developing a drug for these small numbers of patients," as mentioned in Ledford's article. Since the posterior distributions can be defined irrespective of whether the observations are generated adaptively or by independent sampling from some population, Bayesian inference can be carried out in the same way for sequential/adaptive samples as in the conventional fixed samples. The controversy lies in whether the frequentist type I error rate is inflated and in the assumptions of prior distributions and parametric models in the Bayesian approach. Moreover, the complexity of the design requires computationally intensive Markov chain Monte Carlo methods to implement the Bayesian approach.

We are currently working with our medical colleagues on clinical trial designs for the development and validation of biomarker-based personalized therapies. One such trial being planned involves personalized therapies for treating patients with recurrent, platinum-resistant ovarian cancers. Standard chemotherapy drugs (liposomal doxorubicin, abbreviated LD; topotecan, or Top; docetaxel, or Dxl) have modest activity (15–20% rates of remission within one year) in this clinical setting. These drugs have diverse targets (TOPO2A for LD, TOPO1 for Top, and TUBB3 for Dxl), and several potential genomic biomarkers related to drug mechanisms of action and resistance (e.g., ABCB1 for resistance to Dxl) have been identified but not clinically validated for these agents. Tumor tissue is available from the initial staging and de-bulking surgery of the patients, and confirmatory tumor specimens are also available in a large fraction of patients at relapse. Published data on expression of these genes in ovarian cancers, extensive laboratory data from our collaborators, and a recent study by Tothill et al. [90] have identified four genes, ABCB1, TOPO2A, TOPO1, and TUBB3, and thresholds for their expression levels to predict resistance or sensitivity to the three drugs. This suggests a personalized therapy that classifies patients into six predictive subgroups based on their biomarker data (sensitive to one drug only, or to two of the three drugs) besides a subgroup for which the biomarkers cannot make a recommendation for or against any of the three treatments. The sensitivity or resistance is based on the targets of the drugs and the expression levels of these genomic biomarkers. Our approach is to use generalized likelihood ratio tests for concurrent testing of a "strategy null hypothesis" which is associated with the biomarker-based personalized therapy and its validation, and an "intersection null hypothesis" whose rejection demonstrates the efficacy of some biomarker-based therapy, not necessarily the one being considered but still using the same biomarker-classified patient subgroups.

There is a growing literature on the issues of randomized clinical trial designs to develop and test personalized therapies; see [21, 34, 58, 79, 80], and [64]. There is also tension between the adaptive Bayesian approach, as in BATTLE and I-SPY2, and more traditional frequentist approaches to development and validation in this literature. We are working toward an innovative design that can be viewed as a hybrid of both approaches, capturing their individual strengths and resolving the controversies between them.

3.3 Seamless Phase II–III Randomized Clinical Trials

Ellenberg and Eisenberger [30] pointed out the dilemma that although most clinical investigators are aware of the "unreliability of data" from small single-arm Phase II cancer trials, they cannot commit the resources needed for comparative controlled trials that are thought to require much larger sample sizes, until the new treatment has some promising results. Being able to include the Phase II study as an internal pilot of the confirmatory Phase III trial may be the only way that a randomized Phase II cancer trial with substantially larger sample size than currently used in single-arm Phase II trials can be conducted. In standard clinical trial designs, the sample size is determined by the power at a given alternative, and an obvious method to determine a realistic alternative on which sample size calculation can be based is to carry out a preliminary pilot study. Noting that the results from a small pilot study are often difficult to interpret and apply, Wittes and Brittain [98] proposed to use an adaptive design, whose first stage serves as an internal pilot from which the overall sample size of the study can be estimated. Bartroff and Lai [7] focus on tumor response as the primary endpoint so that Phase II–III designs for this endpoint can be embedded into group sequential designs, with the first group representing the Phase II component. Instead of a conventional group sequential design, they use an adaptive design which allows stopping early for efficacy, in addition to futility, in Phase II as an internal pilot, and which also adaptively chooses the next group size based on the observed data. Despite the data-dependent sample size and the inherent complexity of the adaptive design, the usual generalized likelihood ratio (GLR) statistics can still be used to test for differences in the response rates of the two treatments, as the Markov property can be used to compute error probabilities in group sequential or adaptive designs; see [48] and [7, 8].

As noted in Sect. 2.2, although tumor response is an unequivocally important treatment outcome and most targeted therapies are designed to generate the response, time to event is usually the clinically definitive endpoint in Phase III cancer trials. Because of the long latency of the clinical failure-time endpoints, the patients treated in a randomized Phase II trial carry the most mature definitive outcomes if they are also followed in the Phase III trial. Seamless Phase II–III trials with bivariate endpoints consisting of tumor response and time to event are therefore an attractive idea. Inoue et al. [40] and Huang et al. [38] have introduced a Bayesian approach to the design of Phase II–III trials. The approach is based on a parametric mixture model that relates survival to response. Let J_i denote the treatment indicator (0 = control, 1 = experimental), T_i denote survival time, and Z_i denote the binary response for patient i. Assume that the responses Z_i are independent Bernoulli variables and that the survival time T_i given Z_i follows an exponential distribution, denoted Exp(λ), in which $1/\lambda$ is the mean:

$$Z_i|J_i=j \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_j),$$
$$T_i|\{Z_i=z,\ J_i=j\} \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda_{j,z}).$$

Then the conditional distribution of T_i given J_i is a mixture of exponentials:

$$T_i|J_i=j \overset{\text{i.i.d.}}{\sim} \pi_j\text{Exp}(\lambda_{j,1})+(1-\pi_j)\text{Exp}(\lambda_{j,0}).$$

Let $\mu_j = E(T_i \mid J_i = j)$ denote the mean survival time in treatment group j . Inoue's et al. Phase II–III design is based on the parametric model above and assumes independent prior gamma distributions for $\lambda_{j,0}$ and $\lambda_{j,1}$ ($j = 0, 1$) and beta prior distributions for π_0 and π_1. Each interim analysis involves updating the posterior probability $p \hat{=} P(\mu_1 > \mu_0 \mid \text{data})$ and checking whether p êxceeds a prescribed upper bound p_U or falls below a prescribed lower bound p_L, which is less than p_U . If $p \hat{>} p_U$ (or $p \hat{<} p_L$), then the trial is terminated, rejecting (accepting) the null hypothesis that the experimental treatment is not better than the standard treatment; otherwise the study continues until the next interim analysis or until the scheduled end of the study. The posterior probabilities are computed by Markov chain Monte Carlo, and simulation studies of the frequentist operating characteristics under different scenarios are used to determine the maximum sample size, study duration and the thresholds p_L and p_U . Huang et al. [38] recently introduced a more elaborate design that uses the posterior probability p âfter an interim analysis for outcome-adaptive random allocation of patients to treatment arms until the next interim analysis.

Although the mixture model nicely captures the response-survival relationship, the Bayesian framework in the preceding section assumes restrictive parametric survival models and precludes one from using standard methods in cancer biostatistics to analyze survival data semiparametrically and GLR tests for sample proportions. Moreover, the posterior probability p that is central to the Bayesian stopping and terminal decision rules is targeted toward comparing mean survival times instead of the commonly used hazard ratios. Furthermore, semiparametric methods such as Cox regression are usually preferred to parametric exponential models for reproducibility considerations and because of the relatively large sample sizes in Phase III studies. We have recently developed an alternative seamless Phase II–III design that uses a semiparametric model to relate survival to response. It is targeted toward frequentist testing with generalized likelihood ratio (GLR) or partial likelihood ratio statistics, using the ideas of Gu and Lai [35] and Lai and Shih [48] for sequential logrank and GLR tests with modified Haybittle–Peto boundaries. The details and a software package using R for the design and analysis are available at the website http://med.stanford.edu/biostatistics/ClinicalTrialMethodology.html. We are exploring with our medical colleagues how these seamless Phase II–III trials can address some long-standing difficulties in designing confirmatory clinical trials.

4 New Clinical Trial Design Methods for CER Studies

4.1 Equipoise-Stratified Randomization

We describe this innovative randomization method in the context of the NIMH Sequenced Alternatives to Relieve Depression (STAR*D), which was a multisite, prospective, randomized, multistep clinical trial of outpatients with nonpsychotic major depressive disorder [75]. The study compared seven treatment options in patients who did not attain a satisfactory response with citalopram, a selective serotonin re-uptake inhibitor antidepressant. After receiving citalopram participants without sufficient symptomatic benefit were eligible for randomization among four switch options (sertraline, bupropion, venlafaxine, cognitive therapy) and three citalopram augment options (bupropion, buspirone, cognitive therapy). It was clear to the study designers that few patients would be

willing to be randomized among all seven options, so other design options were considered. One possibility was to randomize patients between two overall strategies: “switch” or “augment”, allowing physician choice to determine which specific option would be implemented. A prototype for this kind of design was the NHLBI AFFIRM study which contrasted broad options of “rhythm control” versus “rate control” in patients with atrial fibrillation, with physician’s choice of specific method (e.g., sotalol or amiodarone for rhythm control). But for various reasons this design was discarded (a lucky choice, given the outcomes). Instead, the designers chose to ascertain before randomization the set of options that the patient-clinician dyad considered to be equally reasonable, given the patient’s preferences, and his or her state after a trial of citalopram. This set of options characterizes the patient’s Equipoise Stratum (ES). A total of 1429 patients were randomized under this scheme. The largest ES were the “Medication Switch Only” group, allowing randomization among the three medications (40%) and the “Medication Augmentation Only”, allowing randomization between two options (29%). The “Any Augmentation” (10%) and “Any Switch” (7%) were the next largest, and only 5% of patients were randomized among options that contrasted a switch and augment condition. In retrospect, it became clear that patients (and their clinicians) were sorted into two groups, roughly, those who obtained partial benefit from citalopram and therefore were interested in augmentation, and those who obtained no benefit and were interested only in switching. Thus, the ES design allowed the study to self-design in assigning patients to the parts of the experiment that were relevant to current practice and to patient preferences.

4.2 Sequential Multiple-Assignment Randomization (SMAR)

The STAR*D project had subsequent randomization levels, to attempt to improve outcomes in patients who were not brought to full remission by initial levels of treatment. Thus it can be regarded as an approximation to a Sequential Multiple Assignment Randomized Trial (SMART). Such trials have been proposed and conducted [52, 61, 66, 68, 86, 87, 99] to study *adaptive treatment strategies*, which are rules for adapting the current treatment to the patient’s history of past treatments and responses to those treatments. This clinical usage of the word “adaptive” should not be confused with adaptation of assignment or other methods in trials; here it is the treatment that is adapting.

Implicit or informal adaptive treatment strategies arise naturally in the management of patients whose disease has a chronic dimension, neither acutely fatal nor completely curable. They are usually cobbled together from disparate sources of information. For example, the choice of drugs for the initial treatment of newly diagnosed hypertension or bipolar manic disorder may be well-informed by carefully controlled trials conducted as part of the registration process or comparative effectiveness studies. However, the choice of a second-line drug if the first drug fails to bring the hypertension or mania under control may not have such a strong evidence base, and as the patient experiences multiple failures of disease control, the basis for making such decisions may thin out completely. Furthermore, the best way to start treating the disease may depend on the downstream options. If a highly effective but risky treatment can be deployed successfully as a salvage treatment after failure of a less effective but non-toxic treatment, then it may produce an overall better long-term benefit to use it in that way. But if the progression of disease makes the riskier treatment less effective

in the salvage role, the conclusion may be reversed. In other words, the “myopic” outcomes over the short term after a treatment is used may not be a good guide to the role it would play in an adaptive treatment strategy.

A simple two-stage example is the Habermann et al. [36] study of induction and maintenance rituximab on survival in diffuse large B-cell lymphoma. In this study, patients were randomized to induction by standard chemotherapy (CHOP) or chemotherapy with the addition of rituximab (R-CHOP). Those who developed a remission were again randomized to standard maintenance treatment or rituximab maintenance. The results in all patients were analyzed to contrast the survival rates on each of the four strategies defined by induction and maintenance choices. Wahed and Tsiatis [96] had devised optimal estimators for survival outcomes in two-stage trials of this type.

The SMAR design offered the ability to distinguish between short-term (“myopic”) outcome differences (on rates of remission, by initial induction option) and long-term survival differences. The use of SMAR designs is crucial when the long-term, clinically important outcome is not simply determined by the myopic results. For example, it was considered possible that even if R-CHOP was more successful as an induction, rituximab maintenance in remitters on R-CHOP might do less well than rituximab maintenance in remitters on CHOP, enough so that the overall strategy of waiting to use rituximab in maintenance would prove superior, despite its initial myopic deficit. The idea goes back to Bellman’s principle of dynamic programming, which has been mentioned in Sect. 3.1. In the event, it turned out that the initial treatment with the best “myopic” outcome (R-CHOP) was also the right way to start the optimal strategy (which then added rituximab maintenance).

Thall et al. [87] describe a two-stage randomized trial of a total of 12 different strategies of first-line and second-line treatments for androgen-independent prostate cancer. After randomization to one of four treatments, patients who responded were continued on the initial treatment, and those who did not were randomized among the other three treatments not assigned initially. The intent of this Phase II SMART design was to select a candidate treatment for evaluation in a Phase III trial. The treatment with the best initial success rate was the combination TEC (weekly paclitaxel, estramjustine, and carboplatin), and the authors proposed that it be taken forward to Phase III testing. In a companion piece, Bembom and van der Laan [12] reviewed methods for analyzing such SMAR methods, and found that the best overall strategy seemed to be CVD (cyclophosphamide, vincristine, and dexamethasone) followed by TEC if CVD did not achieve success. The small study sample size and some problems of dropout prevent a definitive conclusion about the strategy comparisons, but the results illustrate an apparent instance of the phenomenon described above: the myopically best initial treatment does not necessarily initiate an optimal strategy.

Another example concerns the use of compliance enhancements. In clinical practice, patients often fail to comply adequately with treatment recommendations. The clinician may consider switching treatments, or instead adding a “treatment compliance” component, which might mean another treatment to reduce a side effect or a behavioral intervention. When combined with the choice of initial treatment, the result is a multi-stage adaptive treatment, with at least two options at each stage.

Thus, when there are multiple stages of treatment (e.g., in chronic disease management), the right way to start may depend on the way that subsequent options deal with the state of the patients after the initial treatments. In addition to the methods for comparing strategies described above, Zhao et al. [101] used Q-learning methods to discover optimal treatments from data arising from a SMART design. As medical care for cancer, AIDS, and other life-threatening diseases evolves, they become chronic diseases that require adaptive treatment strategies. Therefore, we can expect to encounter questions about the value of an entire adaptive treatment strategy more often in the future, leading to the greater use of SMAR designs. There has been considerable recent development in the statistical underpinnings (design and analysis) of SMAR trials, and growing use by clinical investigators [17, 22, 23, 44, 50, 51, 53–55, 60 , 65–67, 69, 82, 92, 95].

4.3 Embedded Experiments to Close the Knowledge-Action Gap

Eight decades ago, Thompson [89] proposed to randomize patients to a treatment with probability equal to the current posterior probability that the treatment was superior to the alternative. His proposal was motivated by ethical considerations, and sought to minimize the number of patients exposed to the inferior treatment during the trial. Since then, the Bayesian approach to outcome-adaptive randomization has been further developed, especially by Berry [13], Berry and Eick [15], and Cheng and Berry [19]. But there is another advantage (noted by Berry) to point-of-care randomization: it tends to automatically close the gap between knowledge and implementation, at least at the institutions that participate in a trial. That is, if one treatment is better, the randomization is biased more and more toward that treatment, as the posterior probability tends to 1.

The “implementation gap,” as it is now called in discussions of CER and evidence-based medicine, refers to the undeniable fact that there is a systematic failure to obtain full value from comparative trials because of the low rate of uptake in clinical practice. The pilot’s checklist entered aviation in 1935, after the fatal crash of the B-17 prototype, and is still in wide use today. Similar checklists for surgery, infection control, and critical care have been tested over the past 40 years, with nearly universal positive results. Yet their implementation into routine practice has been painfully slow. There is now a branch of CER called “implementation science,” which studies the effectiveness of methods for disseminating and putting into practice new or old knowledge that improves care. An embedded clinical trial with outcome-adaptive randomization can bring the benefits of automatic, statistically valid learning to the participating health care systems, and improve care without having to mount a separate implementation strategy.

Such automatic learning may be crucial to taking full advantage of comparative effectiveness research. The US Department of Veteran’s Affairs is capitalizing on its superior informatics to launch an effort directed at fostering embedded clinical trials, in which the option to randomize is offered as part of the automated system of guidelines and computer-aided ordering of treatments; see [32]. The first trial in this program began enrollment in 2010 and aims at defining the optimal insulin dosing strategy for hospitalized diabetic patients. By combining automatic assessment of endpoints from the electronic health record, the randomization can adapt to the outcomes, as Thompson proposed in 1933.

Statistical issues raised by this approach include the desire to make valid frequentist inference while using Bayesian or other methodology to adapt the randomization, and our current research aims at developing methods to address these issues.

Challenges to the validity of the adaptive randomization scheme have been raised; see, e.g., [43]. In the presence of substantial drift in the characteristics of the patient sample over time, it may be subject to bias. It is also less efficient than fixed equal randomization, and has a chance of converging on the wrong treatment. Two of these issues (drift bias and inefficiency) can be addressed by a combination of statistical analysis and limitation on the extremes to which the randomization can be tilted, and the third by adjusting the usual error rates for adaptive randomization, using methods similar to those of Lai and Shih [48], Bartroff and Lai [7, 8] and Lai et al. [49]. The jury is still out on the value of adaptive randomization, and much more practical experience is needed in particular circumstances. All perturbations away from the fixed-sample, equal-randomization, double-blind, placebo-controlled trial come with some defect. But this argument can "make the best the enemy of the good." Much of the proposed research agenda for CER begins with the premise that the "pure" RCT relevant to an effectiveness question will be either too expensive, insufficiently generalizable, or too time-consuming to be of much use, and thus sweeps away any serious use of experiment in the process. The alternative then becomes observational data analysis, which is vulnerable to the well-known biases of selection. Furthermore, the use of adaptive randomization in the embedded clinical trial may provide crucial help in closing the "implementation gap," as discussed below. This benefit may justify increased resources devoted to preventing the problems described above.

The clearest motivation for embedded clinical trials comes from the somewhat disappointing history of standard CER, in which closing the implementation gap is not considered part of the research itself. There are many reasons that closing the gap is hard (sometimes harder than doing the research). They include resistance by clinicians who do not see the research as applying to their patients, and efforts by interested parties in deflecting implementation. The latter is particularly troublesome in the US, and we have learned to our dismay that CER often leads to a result whose implementation would threaten the livelihood of some organized group (a company, or a group of clinicians), and this often evokes considerable (and successful) opposition. That opposition is not confined to scientific discourse, but plays out in lobbying government, attacks in the media, and even worse. Embedded clinical trials offer a way to keep some of the implementation process under the control of clinicians interested in improving care and their home institutions. Eventually, it is hoped that health care systems that engage in such statistically valid learning will have a competitive advantage in an outcome-based reimbursement system, and patients will come to understand that advantage as well. The VA is uniquely positioned (in the US) to lead in this area, but other systems around the world with similar or better infrastructure may also take note.

5 Discussion

It is widely believed by CER practitioners that standard randomized clinical trial designs cannot meet the challenges of new developments and emerging trends in clinical medicine and health care in the 21st century. While the premise may be true if the emphasis is on the

word “standard,” it has led many to reject randomization in favor of observational methods. We believe that such a limited view of randomization will often lead to inappropriate reliance on statistical adjustment for selection effects instead of robust experimental methods. We have outlined above some innovative designs and novel design methods to address the challenges. These designs can resolve the dilemma between individual ethics (for each patient being treated in the trial) and collective ethics (for future patients), adapt to accruing information on treatment effects during the course of the trial, find the trial that fits the patient by using equipoise-stratified randomization, extend to adaptive treatment strategies via sequential multiple-assignment randomization, and close the knowledge-action gap by embedded experiments.

Luce et al. [60] have argued for transformational change in randomized clinical trials as they are “the most rigorous method of generating comparative effectiveness evidence and will necessarily occupy a central role in an expanded national CER agenda,” but “as currently designed and conducted, are ill-suited to meet the evidentiary needs implicit in the IOM definition of CER.” They also point out the usefulness of adaptive approaches, but that “with traditional trials and analytical methods, it is difficult to make optimal use of relevant existing, ancillary or new evidence as it arises during a trial.” Although they, and Berry [14] earlier, argue that Bayesian designs and analyses should be used in conjunction with adaptation, we want to point out that frequentist methods are also available to implement such adaptation. In our opinion, many modern developments in statistical methodology, such as resampling, likelihood theory and martingales, generalized linear mixed models, censored survival data and semiparametric inference, and boundary-crossing probabilities for Markov chains can all be incorporated in the adaptive and sequential approaches, as shown recently by Bartroff and Lai [7, 8], Lai et al. [49], Lai and Li [47], and Lai and Shih [48]. What we recommend is to make use of the advances in statistical modeling in the past two decades, and to use the frequentist or Bayesian approach depending on which is more appropriate for the situation.

A case in point is the pivotal Phase II–III design of a clinical trial to gain the FDA’s approval of a new cancer drug in Sect. 3.3, or the design of a validation trial for a biomarker-based personalized targeted therapy in Sect. 3.2. Whereas the Bayesian approach requires precise assumptions on the data-generating mechanism besides prior distributions on the unknown parameters, the validation trial should be free of the assumptions used to develop the personalized therapy and a frequentist approach to confirmatory testing is the widely accepted mode of inference, especially with regulatory agencies. It is important to ensure that the flexible and adaptive features offered by the innovative clinical trial designs do not inflate the probability of a false positive conclusion, i.e., the type I error of the confirmatory test; see for example the European Medicines Agency [20] and the [33] guidelines on adaptive trial designs. On the other hand, Bayesian modeling and decisions can be very effective in the development of personalized therapies and in clinical development plans that integrate different phases of drug development.

As mentioned in the last paragraph of Sect. 1, the new health care system for the 21st century advocated by Arrow et al. [3] has the infrastructure to support the innovative study designs mentioned above and other innovations in clinical trial designs. We envision point-

of-care decision support for outcome-adaptive randomization, web-based “experimental embedding” services, a portfolio of CER “open questions” and initial observational studies, recruitment of “first adopter” provider systems, involvement of community and patient groups in planning and oversight, consciousness-raising on the value of experiments, and integration of experimental findings into point-of-care decision support. These innovations in the design and conduct of experiments will make use of major advances in statistical methodologies for their analysis, and will in turn lead to new frontiers for statistics in the biosciences.

Acknowledgments

T.L. Lai supported in part by NIH grant 1 P30 CA124435-01 and NSF grant DMS 0805879.

P.W. Lavori supported in part by NIH grant R01 MH051481 and clinical and translational science award 1 UL1 RR025744.

References

1. ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). J Am Med Assoc. 2002; 288(23):2981–2997.
2. Appel L. The verdict from ALLHAT—Thiazide diuretics are the preferred initial therapy for hypertension. J Am Med Assoc. 2002; 288(23):3039–3042.
3. Arrow K, Auerbach A, Bertko J, Brownlee S, Casalino LP, Cooper J, Crosson FJ, Enthoven A, Falcone E, Feldman RC, Fuchs VR, Garber AM, Gold MR, Goldman D, Hadfield GK, Hall MA, Horwitz RI, Hooven M, Jacobson PD, Jost TS, Kotlikoff LJ, Levin J, Levine S, Levy R, Linscott K, Luft HS, Mashal R, McFadden D, Mechanic D, Meltzer D, Newhouse JP, Noll RG, Pietzsch JB, Pizzo P, Reischauer RD, Rosenbaum S, Sage W, Schaeffer LD, Sheen E, Silber M, Skinner J, Shortell SM, Thier SO, Tunis S, Wulsin L Jr, Yock P, Bin Nun G, Bryan S, Luxenburg O, van de Ven WPMM. Toward a 21st-century health care system: recommendations for health care reform. Ann Intern Med. 2009; 150(7):493–495. [PubMed: 19258550]
4. Atkinson A, Donev A. Experimental designs optimally balanced for trend. Technometrics. 1996; 38(4):333–341.
5. Babb J, Rogatko A, Zacks S. Cancer Phase I clinical trials: Efficient dose escalation with overdose control. Stat Med. 1998; 17(10):1103–1120. [PubMed: 9618772]
6. 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):97–100.10.1038/clpt.2009.68 [PubMed: 19440188]
7. Bartroff J, Lai TL. Efficient adaptive designs with mid-course sample size adjustment in clinical trials. Stat Med. 2008; 27(10):1593–1611.10.1002/sim.3201 [PubMed: 18275090]
8. Bartroff J, Lai TL. Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. Seq Anal. 2008; 27(3):254–276.
9. Bartroff J, Lai TL. Approximate dynamic programming and its applications to the design of Phase I cancer trials. Stat Sci. 2010; 25:245–257.
10. Bartroff J, Lai TL. Incorporating individual and collective ethics into Phase I cancer trial designs. Biometrics. 2010; 67:596–603. [PubMed: 20731643]
11. Bekele BN, Shen Y. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. Biometrics. 2005; 61(2):344–354.10.1111/j.1541-0420.2005.00314.x
12. Bembom O, van der Laan MJ. Statistical methods for analyzing sequentially randomized trials. J Natl Cancer Inst. 2007; 99(21):1577–1582.10.1093/jnci/djm185 [PubMed: 17971533]

13. Berry, D. Statistical innovations in cancer research. In: Holland, J.; Frei, T., et al., editors. Cancer medicine. 6. BC Decker; London: 2003. p. 465-478.
14. Berry D. Bayesian statistics and the efficiency and ethics of clinical trials. Stat Sci. 2004; 19(1): 175–187.10.1214/088342304000000044
15. Berry D, Eick S. Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. Stat Med. 1995; 14(3):231–246. [PubMed: 7724909]
16. Braun T. The bivariate continual reassessment method: Extending the CRM to phase I trials of two competing outcomes. Control Clin Trials. 2002; 23(3):240–256. [PubMed: 12057877]
17. Brooner R, Kidorf M, King V, Stoller K, Peirce J, Bigelow G, Kolodner K. Behavioral contingencies improve counseling attendance in an adaptive treatment model. J Subst Abuse Treat. 2004; 27(3):223–232.10.1016/j.jsat.2004.07.005 [PubMed: 15501375]
18. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. 1995; 51(4):1372–1383.10.2307/2533268 [PubMed: 8589229]
19. Cheng Y, Berry DA. Optimal adaptive randomized designs for clinical trials. Biometrika. 2007; 94(3):673–689.10.1093/biomet/asm049
20. Committee for Medicinal Products for Human Use (CHMP). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency; 2007. http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf
21. Cree IA, Kurbacher CM, Lamont A, Hindley AC, Love S. TCA Ovarian Cancer Trial Group. A prospective randomized controlled trial of tumour chemosensitivity assay directed chemotherapy versus physician's choice in patients with recurrent platinum-resistant ovarian cancer. Anti-Cancer Drugs. 2007; 18(9):1093–1101. [PubMed: 17704660]
22. Dawson R, Lavori PW. Sequential causal inference: Application to randomized trials of adaptive treatment strategies. Stat Med. 2008; 27(10):1626–1645.10.1002/sim.3039 [PubMed: 17914714]
23. Dawson R, Green AI, Drake RE, McGlashan TH, Schanzer B, Lavori PW. Developing and testing adaptive treatment strategies using substance-induced psychosis as an example. Psychopharmacol Bull. 2008; 41(3):51–67. [PubMed: 18779776]
24. Dette H, Melas VB, Pepelyshev A. Optimal designs for a class of nonlinear regression models. Ann Stat. 2004; 32(5):2142–2167.10.1214/009053604000000382
25. Dragalin V, Fedorov V. Adaptive designs for dose-finding based on efficacy-toxicity response. J Stat Plan Inference. 2006; 136(6):1800–1823.10.1016/j.jspi.2005.08.005
26. Dragalin V, Fedorov V, Wu Y. Adaptive designs for selecting drug combinations based on efficacy-toxicity response. J Stat Plan Inference. 2008; 138(2):352–373.10.1016/j.jspi.2007.06.017
27. Druker B. Circumventing resistance to kinase-inhibitor therapy. N Engl J Med. 2006; 354(24): 2594–2596. [PubMed: 16775240]
28. Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, Deininger MWN, Silver RT, Goldman JM, Stone RM, Cervantes F, Hochhaus A, Powell BL, Gabrilove JL, Rousselot P, Reiffers J, Cornelissen JJ, Hughes T, Agis H, Fischer T, Verhoef G, Shepherd J, Saglio G, Gratwohl A, Nielsen JL, Radich JP, Simonsson B, Taylor K, Baccarani M, So C, Letvak L, Larson RA. IRIS Investigators. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. N Engl J Med. 2006; 355(23):2408–2417. [PubMed: 17151364]
29. Eisenhauer EA, O'Dwyer PJ, Christian M, Humphrey JS. Phase I clinical trial design in cancer drug development. J Clin Oncol. 2000; 18:684. [PubMed: 10653884]
30. Ellenberg S, Eisenberger M. An efficient design for Phase III studies of combination chemotherapies. Cancer Treat Rep. 1985; 69(10):1147–1154. [PubMed: 4042093]
31. Fedorov, VV. Theory of optimal experiments. Studden, WJ.; Klimko, EM., editors. Academic Press; New York: 1972. translated from the RussianProbability and mathematical statistics, No 12
32. Fiore L, Brophy M, D'Avolio L, Conrad C, O'Neil G, Sabin T, Kaufman J, Hermos J, Swartz S, Liang M, Gaziano M, Lawler E, Ferguson R, Lew R, Doras G, Lavori P. A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. Clin Trials. 2011; 8:183–195.10.1177/1740774511398368 [PubMed: 21478329]
33. Food and Drug Administration Center for Drug Evaluation and Research. Guidelines for industry: adaptive design clinical trials for drugs and biologics. Rockville, MD: 2010. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf

34. Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: Design issues. J Natl Cancer Inst. 2010; 102(3):152–160.10.1093/jnci/djp477 [PubMed: 20075367]
35. Gu M, Lai TL. Repeated significance testing with censored rank statistics in interim analysis of clinical trials. Stat Sin. 1998; 8(2):411–428.
36. Habermann TM, Weller EA, Morrison VA, Gascoyne RD, Cassileth PA, Cohn JB, Dakhil SR, Woda B, Fisher RI, Peterson BA, Horning SJ. Rituximab-CHOP versus CHOP alone or with maintenance rituximab in older patients with diffuse large B-cell lymphoma. J Clin Oncol. 2006; 24(19):3121–3127.10.1200/JCO.2005.05.1003 [PubMed: 16754935]
37. Haines LM, Perevozskaya I, Rosenberger WF. Bayesian optimal designs for Phase I clinical trials. Biometrics. 2003; 59(3):591–600.10.1111/1541-0420.00069 [PubMed: 14601760]
38. Huang X, Ning J, Li Y, Estey E, Issa JP, Berry DA. Using short-term response information to facilitate adaptive randomization for survival clinical trials. Stat Med. 2009; 28(12):1680–1689.10.1002/sim.3578 [PubMed: 19326367]
39. Hunsberger S, Rubinstein L, Dancey J, Korn E. Dose escalation trial designs based on a molecularly targeted endpoint. Stat Med. 2005; 24(14):2171–2181.10.1002/sim.2102 [PubMed: 15909289]
40. Inoue LYT, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. Biometrics. 2002; 58(4):823–831.10.1111/j.0006-341X.2002.00823.x [PubMed: 12495136]
41. Ivanova A. A new dose-finding design for bivariate outcomes. Biometrics. 2003; 59(4):1001–1007.10.1111/j.0006-341X.2003.00115.x [PubMed: 14969479]
42. Ivanova A, Wang K. Bivariate isotonic design for dose-finding with ordered groups. Stat Med. 2006; 25(12):2018–2026.10.1002/sim.2312 [PubMed: 16220476]
43. Karrison T, Huo D, Chappell R. A group sequential, response-adaptive design for randomized clinical trials. Control Clin Trials. 2003; 24(5):506–522.10.1016/S0197-2456(03)00092-8 [PubMed: 14500050]
44. Kay-Lambkin FJ, Baker AL, McKetin R, Lee N. Stepping through treatment: Reflections on an adaptive treatment strategy among methamphetamine users with depression. Drug Alcohol Rev. 2010; 29(5):475–482.10.1111/j.1465-3362.2010.00203.x [PubMed: 20887570]
45. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev, Drug Discov. 2004; 3:711–716.10.1038/nrd1470 [PubMed: 15286737]
46. Korn E. Nontoxicity endpoints in phase I trial designs for targeted, non-cytotoxic agents. J Natl Cancer Inst. 2004; 96(13):977–978.10.1093/jnci/djh208 [PubMed: 15240771]
47. Lai TL, Li W. Confidence intervals in group sequential trials with random group sizes and applications to survival analysis. Biometrika. 2006; 93(3):641–654.
48. Lai TL, Shih MC. Power, sample size and adaptation considerations in the design of group sequential clinical trials. Biometrika. 2004; 91(3):507–528.
49. Lai TL, Shih MC, Su Z. Tests and confidence intervals for secondary endpoints in sequential clinical trials. Biometrika. 2009; 96:903–915.10.1093/biomet/asp063
50. Lavori P, Dawson R. Developing and comparing treatment strategies: An annotated portfolio of designs. Psychopharmacol Bull. 1998; 34(1):13–18. [PubMed: 9564193]
51. Lavori P, Dawson R. A design for testing clinical strategies: biased adaptive within-subject randomization. J R Stat Soc Ser A Stat. 2000; 163(1):29–38.
52. Lavori PW, Dawson R. Dynamic treatment regimes: Practical design considerations. Clin Trials. 2004; 1(1):9–20. [PubMed: 16281458]
53. Lavori PW, Dawson R. Adaptive treatment strategies in chronic disease. Annu Rev Med. 2008; 59:443–453.10.1146/annurev.med.59.062606.122232 [PubMed: 17914924]
54. Lavori P, Dawson R, Rush A. Flexible treatment strategies in chronic disease: Clinical and research implications. Biol Psychiatry. 2000; 48(6):605–614. [PubMed: 11018231]
55. Lavori P, Rush A, Wisniewski S, Alpert J, Fava M, Kupfer D, Nierenberg A, Quitkin F, Sackeim H, Thase M, Trivedi M. Strengthening clinical effectiveness trials: Equipoise-stratified randomization. Biol Psychiatry. 2001; 50(10):792–801. [PubMed: 11720698]
56. Ledford H. Clinical drug tests adapted for speed. Nature. 2010; 464(7293):1258.10.1038/4641258a [PubMed: 20428134]

57. Lee J, Gu X, Liu S. Bayesian adaptive randomization designs for targeted agent development. Clin Trials. 2010; 7:574–583. [PubMed: 20667935]
58. Liu A, Li Q, Yu K, Yuan V. A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. Clin Trials. 2010; 7(5):537–545. [PubMed: 20685769]
59. LoRusso PM, Boerner SA, Seymour L. An overview of the optimal planning, design, and conduct of phase I studies of new therapeutics. Clin Cancer Res. 2010; 16(6):1710–1718.10.1158/1078-0432.CCR-09-1993 [PubMed: 20215546]
60. Luce BR, Kramer JM, Goodman SN, Connor JT, Tunis S, Whicher D, Schwartz JS. Rethinking randomized clinical trials for comparative effectiveness research: The need for transformational change. Ann Intern Med. 2009; 151(3):206–W45. [PubMed: 19567619]
61. Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2002; 58(1):48–57.10.1111/j.0006-341X.2002.00048.x [PubMed: 11890326]
62. Ma B, Britten C, Siu L. Clinical trial designs for targeted agents. Hematol/oncol Clin North Am. 2002; 16(5):1287–1305.
63. Maitournam A, Simon R. On the efficiency of targeted clinical trials. Stat Med. 2005; 24(3):329–339.10.1002/sim.1975 [PubMed: 15551403]
64. Mandrekar SJ, Sargent DJ. Predictive biomarker validation in practice: Lessons from real trials. Clin Trials. 2010; 7 in press.
65. McKay J, Lynch K, Shepard D, Pettinati H. The effectiveness of telephone-based continuing care for alcohol and cocaine dependence—24-month outcomes. Arch Gen Psychiatry. 2005; 62(2):199–207. [PubMed: 15699297]
66. Murphy S. An experimental design for the development of adaptive treatment strategies. Stat Med. 2005; 24(10):1455–1481.10.1002/sim.2022 [PubMed: 15586395]
67. Murphy, S.; McKay, J. Adaptive treatment strategies: An emerging approach for improving treatment effectiveness; Clinical Science, the Newsletter of the Society for the Science of Clinical Psychology, Section III of the Division of Clinical Psychology of the American Psychological Association. 2004. p. 7-13.Winter-Spring Issuehttp://sites.google.com/site/sscpwebsite/newsletters-1
68. Murphy S, van der Laan M, Robins J. Marginal mean models for dynamic regimes. J Am Stat Assoc. 2001; 96(456):1410–1423.10.1198/016214501753382327 [PubMed: 20019887]
69. Murphy SA, Oslin DW, Rush AJ, Zhu J. MCATS. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. Neuropsychopharmacology. 2007; 32(2):257–262.10.1038/sj.npp.1301241 [PubMed: 17091129]
70. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for Phase 1 clinical trials in cancer. Biometrics. 1990; 46(1):33–48.10.2307/2531628 [PubMed: 2350571]
71. O'Quigley J, Hughes MD, Fenton T. Dose-finding designs for HIV studies. Biometrics. 2001; 57(4):1018–1029.10.1111/j.0006-341X.2001.01018.x [PubMed: 11764240]
72. Ratain MJ, Sargent DJ. Optimising the design of Phase II oncology trials: The importance of randomisation. Eur J Cancer. 2009; 45(2 Sp Iss SI):275–280.10.1016/j.ejca.2008.10.029 [PubMed: 19059773]
73. Rubin DB. On the limitations of comparative effectiveness research. Stat Med. 2010; 29(19):1991–1995.10.1002/sim.3960 [PubMed: 20683890]
74. Rubinstein L, Crowley J, Ivy P, Leblanc M, Sargent D. Randomized Phase II designs. Clin Cancer Res. 2009; 15(6):1883–1890.10.1158/1078-0432.CCR-08-2031 [PubMed: 19276275]
75. Rush A, Fava M, Wisniewski S, Lavori P, Trivedi M, Sackeim H, Thase M, Nierenberg A, Quitkin F, Kashner T, Kupfer D, Rosenbaum J, Alpert J, Stewart J, McGrath P, Biggs M, Shores-Wilson K, Lebowitz B, Ritz L, Niederehe G. STAR D Investigators Group. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. Control Clin Trials. 2004; 25(1):119–142.10.1016/S0197-2456(03)00112-0 [PubMed: 15061154]
76. Simon RM. Optimal two-stage designs for Phase II clinical trials. Control Clin Trials. 1989; 10:1–10. [PubMed: 2702835]

77. Simon RM. An agenda for *clinical trials*: Clinical trials in the genomic era. Clin Trials. 2004; 1(5): 468–470. http://ctj.sagepub.com/content/1/5/468.short, http://ctj.sagepub.com/content/1/5/468.full.pdf+html. 10.1191/1740774504cn046xx [PubMed: 16279285]

78. Simon RM. Development and validation of therapeutically relevant multi-gene biomarker classifiers. J Natl Cancer Inst. 2005; 97(12):866–867. http://jnci.oxfordjournals.org/content/97/12/866.full.pdf+html. 10.1093/jnci/dji168 [PubMed: 15956642]

79. Simon RM. Development and validation of biomarker classifiers for treatment selection. J Stat Plan Inference. 2008; 138(2):308–320.10.1016/j.jspi.2007.06.010 [PubMed: 19190712]

80. Simon RM. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Person Med. 2010; 7(1):33–47.10.2217/PME.09.49

81. Storer B. Design and analysis of phase I clinical trials. Biometrics. 1989; 45:925–937. [PubMed: 2790129]

82. TenHave T, Coyne J, Salzer M, Katz I. Research to improve the quality of care for depression: alternatives to the simple randomized clinical trial. Gen Hosp Psych. 2003; 25(2):115–123.10.1016/S0163-8343(02)00275-X

83. Thall PF, Cook JD. Dose-finding based on efficacy-toxicity trade-offs. Biometrics. 2004; 60(3): 684–693.10.1111/j.0006-341X.2004.00218.x [PubMed: 15339291]

84. Thall P, Russell K. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in phase I/II clinical trials. Biometrics. 1998; 54(1):251–264. [PubMed: 9544520]

85. Thall P, Simon R, Ellenberg S, Shrager R. Optimal 2-stage designs for clinical trials with binary response. Stat Med. 1988; 7(5):571–579. [PubMed: 3387716]

86. Thall P, Millikan R, Sung H. Evaluating multiple treatment courses in clinical trials. Stat Med. 2000; 19(8):1011–1028. [PubMed: 10790677]

87. Thall PF, Logothetis C, Pagliaro LC, Wen S, Brown MA, Williams D, Millikan RE. Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. J Natl Cancer Inst. 2007; 99(21):1613–1622.10.1093/jnci/djm189 [PubMed: 17971530]

88. Thall PF, Nguyen HQ, Estey EH. Patient-specific finding based on bivariate outcomes and covariates. Biometrics. 2008; 64(4):1126–1136.10.1111/j.1541-0420.2008.01009.x [PubMed: 18355387]

89. Thompson W. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika. 1933; 25(Part 3/4):285–294.

90. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B, Traficante N, Fereday S, Hung JA, Chiew YE, Haviv L, Gertig D, de-Fazio A, Bowtel DDL. Australian Ovarian Cancer Study Group. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res. 2008; 14(16):5198–5208.10.1158/1078-0432.CCR-08-0196 [PubMed: 18698038]

91. Tunis SR, Benner J, McClellan M. Comparative effectiveness research: Policy context, methods development and research infrastructure. Stat Med. 2010; 29(19):1963–1976.10.1002/sim.3818 [PubMed: 20564311]

92. Unutzer J, Katon W, Williams J, Callahan C, Harpole L, Hunkeler E, Hoffing M, Arean P, Hegel M, Schoenbaum M, Oishi S, Langston C. Improving primary care for depression in late life—The design of a multicenter randomized trial. Med Care. 2001; 39(8):785–799. [PubMed: 11468498]

93. Vickers AJ, Ballen V, Scher HI. Setting the bar in phase II trials: The use of historical data for determining “go/no go” decision for definitive phase III testing. Clin Cancer Res. 2007; 13(3): 972–976.10.1158/1078-0432.CCR-06-0909 [PubMed: 17277252]

94. Von Hoff D, Turner J. Response rates, duration of response, and dose response effects in phase I studies of antineoplastics. Invest New Drugs. 1991; 9:115–122. [PubMed: 1827432]

95. Wahed AS. Inference for two-stage adaptive treatment strategies using mixture distributions. J R Stat Soc, Ser C, Appl. 2010; 59(Part 1):1–18.

96. Wahed AS, Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2004; 60(1): 124–133.10.1111/j.0006-341X.2004.00160.x [PubMed: 15032782]

97. Whitehead J, Brunier H. Bayesian decision procedures for dose determining experiments. Stat Med. 1995; 14(9–10):885–893. [PubMed: 7569508]

98. Wittes J, Brittain E. The role of internal pilot studies in increasing the efficiency of clinical trials. Stat Med. 1990; 9(1–2):65–72. [PubMed: 2345839]

99. Wolbers M, Helterbrand JD. Two-stage randomization designs in drug development. Stat Med. 2008; 27(21):4161–4174.10.1002/sim.3309 [PubMed: 18570274]

100. Yin G, Yuan Y. Bayesian dose finding in oncology for drug combinations by copula regression. J R Stat Soc, Ser C, Appl Stat. 2009; 58(2):211–224.10.1111/j.1467-9876.2009.00649.x

101. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Stat Med. 2009; 28(26):3294–3315.10.1002/sim.3720 [PubMed: 19750510]

102. Zhou X, Liu S, Kim ES, Herbst RS, Lee JL. Bayesian adaptive design for targeted therapy development in lung cancer—A step toward personalized medicine. Clin Trials. 2008; 5(3):181–193.10.1007/s12561-011-9042-5 [PubMed: 18559407]