NOVEL CLINICAL TRIAL DESIGNS AND STATISTICAL METHODS IN THE ERA OF PRECISION MEDICINE

By

Tze Leung Lai
Michael Sklar
Nikolas Thomas Weissmueller

Technical Report No. 2020-06
June 2020

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

NOVEL CLINICAL TRIAL DESIGNS AND STATISTICAL METHODS IN THE ERA OF PRECISION MEDICINE

By

Tze Leung Lai
Michael Sklar
Stanford University

Nikolas Thomas Weissmueller
Bristol-Myers Squibb

Technical Report No. 2020-06
June 2020

This research was supported in part by National Science Foundation grant DMS 1811818.

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

Novel Clinical Trial Designs and Statistical Methods in the Era of Precision Medicine

Tze Leung Lai,[a,b*] Michael Sklar,[a*] and Nikolas Thomas Weissmueller[a,c]

[a]Department of Statistics, Stanford University, USA
[b]Center for Innovative Study Design, Stanford School of Medicine, USA
[c]Center for Observational Research and Data Science, Bristol-Myers Squibb, Redwood City, USA

Abstract

After an overview of FDA's draft guidances to industry on adaptive designs and enrichment strategies for clinical trials, we describe recent advances in adaptive confirmatory trial designs and statistical methods for their analysis. We then focus on biomarker-guided personalized therapies in the era of precision medicine, and precision-guided drug development and master protocols, and conclude with innovations in statistical science for precision medicine and regulatory submission.

Keywords: Adaptive design, confirmatory clinical trial, enrichment, master protocol, targeted therapy.

*Lai and Sklar acknowledge support of this research by the National Science Foundation under DMS-1811818.

1 Introduction

We begin by describing the background of this paper, which is related to the topics “Adaptive Multi-arm and Multi-stage Designs of Confirmatory Trials" and “Innovative and Flexible Clinical Trial Designs in the Era of Precision Medicine" in the 2019 Symposium on Biopharmaceutical Statistics in Kyoto. Section 2 focuses on the second topic, for which Section 3 describes advances in statistical methods for the analysis of these confirmatory trials for regulatory submission. The first topic, which is concerned with trial designs in the era of precision medicine, is addressed in Section 4 where we describe master protocols for precision-guided drug development and efficiency/safety testing, and the remainder of this section describes the background of master protocols.

1.1 Targeted Therapies in Oncology and FDA's Draft Guidance

Precision medicine considers the “individual variability in genes, environment, and lifestyle” of a patient to better prevent or treat illness (NIH, 2015; Garrido et al., 2018). In his State of the Union Address in January 2015, President Barack Obama launched a precision medicine initiative, that was to focus first on the improvement of cancer therapies. Experts agreed that oncology was “the clear choice,” owing to recent advances in diagnostic technology, computational capability, and scientific understanding of cancers, which remain a leading cause of morbidity and mortality worldwide (Collins and Varmus, 2015). Targeted therapies and immuno-oncology (IO) agents were among the forerunners of transformative new medicines, and have heavily utilized innovative statistical methods to meet the clinical development challenges inherent to personalized medicines (de BONO and Ashworth, 2010; Snyder et al., 2014).

Targeted therapies have established their benefit over conventional cytotoxic therapy across multiple tumors (Hodi et al., 2010; Borghaei et al., 2015; Postow et al., 2015). However, a large unmet medical need remains for most malignancies. Patients are seeking better options urgently. The comprehensive evaluation of new investigational targeted therapies in oncology, in a timely and resource efficient manner, is infeasible with conventional large randomized trials (Ersek et al., 2018, 2019). To match the right therapy with the right patients, the number of scientific questions that need to be answered during

clinical development has increased substantially. Traditionally, oncology drug development comprises a series of clinical trials where each study's objective is to establish the safety and efficacy of a single investigational therapy over the current standard of care (SOC) in a broad study population (Redman and Allegra, 2015; Berry, 2015). A targeted therapy's safety and benefit over the SOC needs to be established for a long list of considerations specific to the biomarker-defined subpopulation and pathology, including safety, therapy sequence, drug combinations, combination dosing, and the contribution of individual drug components. The reality of developing "precision medicines" is that there are fewer subjects, who are harder to find, which may jeopardize study completion and extend timelines. The costs of trials have increased with more extensive tissue sample collection, biomarker assessment and tumor imaging, more expensive comparator drugs, and the generally rising cost of medical care. Recent advances in tumor sequencing and genomics affords a more detailed understanding of the underlying biology and pathology (Lima et al., 2019). Although focusing in on molecularly defined subpopulations, this actually expands the reach of targeted therapies across tumors and lines of therapy which can be matched by specific gene signatures or biomarkers such as high expression of microsatellite instability (MSI-hi), or PD-1/PD-L1.

In 2019, 3,876 immuno-therapy compounds were in clinical development, 87% of which were oncology agents. This marks a 91% increase over the 2,030 compounds in development in 2017 (Xin et al., 2019). With additional information emerging at an increasing pace, it is expected that today's clinical protocols will require revisions tomorrow, and may need to accommodate a potential change in SOC, emerging information on safety and efficacy of similar compounds, and a better understanding of the fundamental tumor biology (Hirsch et al., 2013; Xin et al., 2019). Therefore, clinical study protocols are required that learn faster from fewer study subjects, expedite the evaluation of novel therapies, use resources judiciously, enable robust hypotheses evaluation, are operationalizable across most clinics, and afford sufficient flexibility to answer multiple research questions and respond to emerging information. Master protocols have emerged to address this challenge. The term "master protocol" refers to a single overarching design that evaluates multiple hypotheses, with the objective to improve efficiency and uniformity through standardization of proce-

dures in the development and evaluation of different interventions (Renfro and Mandrekar, 2018). In 2001, the first clinical trial to use a "master protocol" was the study B2225, an Imatinib Targeted Exploration. (McArthur et al., 2005; Park et al., 2019). However, the uptake of master protocols was slow. In 2005, STAMPEDE became only the second study to employ a master protocol design, and by 2010, there were still fewer than 10 master protocol-guided studies in the public domain. The subsequent decade from 2010 to 2019 saw a rapid growth resulting in a 10-fold increased use of master protocols in clinical studies (Park et al., 2019). A recent catalyst was the validation of the "master protocol" approach by the regulators. In 2018, the FDA issued a draft guidance for industry that advises on the use of master protocols in support of clinical development, titled "Efficient Clinical Trial Design Strategies to Expedite Development of Oncology Drugs and Biologics" (Lal, 2019). In 2019, 83 clinical trials were in the public domain that utilized master protocols.

1.2 Umbrella, Platform, and Basket Trials

The aforementioned rapid growth was also catalyzed by the successful immuno-oncology therapies targeting CTLA-4 in 2011 (ipilimumab), and PD-1 in 2014 (pembrolizumab and nivolumab); see Oiseth and Aziz (2017). Moreover, clinical development teams were faced with the need to explore quickly and efficiently a broader set of malignancies. Basket trials (59%) accounted for the largest portion of master protocols in 2019, followed by umbrella trials (22%), and platform trials (19%). The growth rate of platform trials outpaced umbrella and basket trials in the late 2010s. The majority of master protocol studies (92%) focus on oncology, and 83% enrolled adult populations (Park et al., 2019). Recently, CDER communicated that the "FDA modernizes clinical trials with master protocols" citing good practice considerations, which is expected to further encourage industry to utilize master protocols to rapidly deliver their drug pipelines; see FDA (2018). Basket trials, umbrella trials, and platform trials are implementation structures of clinical studies, and their designs are defined within a master protocol. Each trial variant provides specific flexibility in the clinical development process, and has its advantages and disadvantages (Renfro and Mandrekar, 2018; Cecchini et al., 2019). The study design and statistical consideration will need to be weighed on a case-by-case basis so that the clinical hypotheses can be an-

swered as directly as possible. Key design choices required of clinical development teams are whether to study multiple investigational drugs in one protocol, include a control arm, open multiple cohorts to test for multiple biomarkers, and whether to add or stop treatment arms during the course of the trial. Statistical analysis choices include whether to use Bayesian or frequentist methods to evaluate efficacy,how best to randomize subjects, the selection of appropriate futility and early success criteria, and what covariates to control (Renfro and Mandrekar, 2018; FDA, 2018; Renfro and Sargent, 2017; Mandrekar et al., 2015).

Basket trials investigate a single drug, or a single combination therapy, across multiple populations based on the presence of specific histology, genetic markers, prior therapies, or other demographic characteristics (FDA, 2018). They may include expansion cohorts and are especially well-suited for "signal-finding." They frequently comprise single-arm, open-label Phase I or II studies, enroll 20-50 subjects per sub-study, and use two- or multi-stage decision gates to rapidly screen multiple populations for detecting large efficacy signals (by combining multiple tumor types in one protocol) with acceptable safety profiles (Park et al., 2019; Renfro and Mandrekar, 2018). Unlike umbrella and platform trials in which the **R**ecommended **P**hase **II D**ose (RP2D) has been pre-established, a basket trial may enroll first-in-human cohorts for whom the RP2D may be established alongside any safety and efficacy signals (Cecchini et al., 2019). While often exploratory, basket trials can have registrational intent. An example is Keynote158 which studied pembrolizumab in solid tumors with high microsatellite instability (MSI-H). The simplicity of the basket protocol and its relatively small size needs to be weighed against the design's lack of control groups and limited information for sub-populations based on pooled sample analyses. Basket trial protocols may be amended to include additional tumor types and study populations (Renfro and Sargent, 2017), ineffective cohorts can be excluded in a response-adaptation approach, and new cohorts can be added, but such changes often require a protocol amendment and subsequent patient reconsenting plus retraining of study personnel. Cecchini et al. (2019) give a comprehensive review of the challenges from the perspectives of the study sponsor, regulator, investigator, and institutional review boards, and discuss the increased operational complexity and increased cost that accompany the reduction in development

time. Despite these limitations, basket protocols have become the most widely used master protocols, as they offer the smallest and fastest option, with a median study size of 205 subjects and a 22.3-month study duration (Park et al., 2019). Statistical methods which are often used to analyze these trials include frequentist sequential (LeBlanc et al., 2009; Park et al., 2019) and hierarchical Bayesian (Berry, 2006; Thall et al., 2003) methods, and the recent approaches that control the family-wise error rates for multi-arm studies (Chen et al., 2016), response adaptive randomization (Ventz et al., 2017; Lin and Bunn, 2017), calibrated Bayesian hierarchical testing and subgroup design (Chu and Yuan, 2018a,b), robust exchangeability (Neuenschwander et al., 2016), modification of Simon's two stage design to improve efficiency (Cunanan et al., 2017), and combination of frequentist and Bayesian approaches (Lin and Bunn, 2017).

Umbrella and platform trials are master protocols with exploratory or registrational intent that match biomarker-selected subgroups with subgroup-specific investigational treatments, and may include the current standard of care for the disease setting as a shared control group. They aim at identifying population subgroups that derive the most clinically meaningful benefit from an investigational therapy, and may enable a smaller, faster, and more cost-effective confirmatory phase III study. Umbrella trials are often phase II or phase II/III, have an established RP2D for each investigational therapy, and frequently include biomarker enriched cohorts (Renfro and Mandrekar, 2018). The totality of umbrella trial data enables inference on the predictive and prognostic potential of the studied biomarkers within the given disease setting (Renfro and Mandrekar, 2018). While the study of specific biomarker subsets is a key focus, the inclusion of rare populations can lead to accelerated regulatory approval to fill an unmet need but may result in long accrual and trial durations. It is possible to add or remove investigational treatments and subgroups, but the required protocol amendments can cause considerable logistic challenges for sponsors and investigators (Cecchini et al., 2019). The pre-planned and algorithmic addition or exclusion of treatments during trial conduct is what distinguishes platform trials from umbrella trials (Angus et al., 2019). Platform trials frequently include futility criteria and interim analyses, which provide guidance on whether to expand or discontinue a given investigational therapy. Platform trial cohorts often have an established RP2D for each investigational

therapy, and may be expanded directly to a registrational Phase III trial while retaining the flexibility to keep other populations in the study (Renfro and Mandrekar, 2018). Recommendations to continue or discontinue treatments are often derived by using Bayesian and Bayesian hierarchical methods (Saville and Berry, 2016; Hobbs et al., 2018). Some protocols leverage response-adaptive randomization to increase the probability that subjects are assigned to the likely superior treatment for their biomarker type, which may provide ethical and cost advantages over conventional randomization (Berry, 2006; Wen et al., 2017). Umbrella and platform trials are 2-5 fold larger and longer than the average basket trial (Park et al., 2019), and it is important to weigh the benefits of a smaller Phase I basket trial, which may be amended to provide sufficient data for accelerated approval of a novel therapy as demonstrated by Keynote-001 (Kang et al., 2017), versus the longer and more comprehensive evaluation of multiple investigational agents and subgroups. Another disadvantage co-travelling with the larger size, duration and cost of umbrella and platform trials is the potential change in the treatment landscape and SOC, which may necessitate subsequent modifications to bridge between the control and therapy arms (Lai et al., 2015; Cecchini et al., 2019; Renfro and Mandrekar, 2018).

2 Group Sequential and Adaptive Designs of Confirmatory Trials of New Treatments

As pointed out by Bartroff et al. (2013, p.77), in standard designs of clinical trials comparing a new treatment with a control (which is a standard treatment or placebo), the sample size is determined by the power at a given alternative, but it is often difficult to specify a realistic alternative in practice because of lack of information on the magnitude of the treatment effect difference before actual clinical trial data are collected. On the other hand, many trials have Data and Safety Monitoring Committees (DSMCs) who conduct periodic reviews of the trial, particularly with respect to incidence of treatment-related adverse events, hence one can use the trial data at interim analyses to estimate the effect size. This is the idea underlying group sequential trials in the late 1970s, and one such trial was the **B**eta-blocker **H**eart **A**ttack **T**rial (BHAT) that was terminated in October

1981, prior to its prescheduled end in June 1982; see Bartroff et al. (2013). BHAT, which was a multicenter, double-blind, randomized placebo-controlled trial to test the efficacy of long-term therapy with propranolol given to survivors of an acute myocardial infarction (MI), drew immediate attention to the benefits of sequential methods not because it reduced the number of patients but because it shortened a 4-year study by 8 months, with positive results for a long-awaited treatment for MI patients. The success story of BHAT paved the way for major advances in the development of group sequential methods in clinical trials and for the widespread adoption of group sequential design. Sections 3.5 and 4.2 of Bartroff et al. (2013) describe the theory developed by Lai and Shih (2004) for nearly optimal group sequential tests in exponential families to provide a definitive method amidst the plethora of group sequential stopping boundaries that were proposed in the two decades after BHAT, as reviewed in Bartroff et al. (2013).

Lai and Shih's theory is based on (a) asymptotic lower bounds for the sample sizes of group sequential tests that satisfy prescribed type I and type II error probability bounds, and (b) group sequential generalized likelihood ratio (GLR) tests with modified Haybittle-Peto boundaries that can be shown to attain these bounds. Noting that the efficiency of a group sequential test depends not only on the choice of the stopping rule but also on the test statistics, Lai and Shih use GLR statistics that have been shown to have asymptotically optimal properties for sequential testing in one-parameter exponential families and can be readily extended to multiparameter exponential families for which the type I and type II errors are evaluated at $u(\theta) = u_0$ and $u(\theta) = u_1$, respectively, where $u : \Theta \to \mathbb{R}$ is a continuously differentiable function on the natural parameter space Θ such that Kullback-Leibler information number $I(\gamma, \theta)$ is increasing in $|u(\theta) - u(\gamma)|$ for every γ; see Bartroff et al. (2013, Sections 3.7 and 4.2.4). An important consideration in this approach is the choice of the alternative θ_1 (in the one-parameter case, or u_1 in the multiparameter exponential families). To test $H_0 : \theta \leq \theta_0$, suppose the significance level is α and no more than M observations are to be taken because of funding and administrative constraints on the trial. The FSS (fixed sample size) test that rejects H_0 if $S_M \geq c_\alpha$ has maximal power at any alternative $\theta > \theta_0$. Although funding and administrative considerations often play an important role in the choice of M, justification of this choice in clinical trial protocols

is typically based on some prescribed power $1-\beta$ at an alternative $\theta(M)$ "implied" by M. The implied alternative is defined by that M and can be derived from the prescribed power $1-\beta$ at $\theta(M)$. It is used to construct the futility boundary in the modified Haybittle-Peto group sequential test (Bartroff et al., 2013, pp.81-85).

2.1 Efficient Adaptive Designs

Using Lai and Shih's theory of modified Haybittle-Peto group sequential tests, Bartroff and Lai (2008a,b) developed a new approach to adaptive design of clinical trials. In standard clinical trial designs, the sample size is determined by the power at a given alternative, but in practice, it is often difficult for investigators to specify a realistic alternative at which sample size determination can be based. Although a standard method to address this difficulty is to carry out a preliminary pilot study, the results from a small pilot study may be difficult to interpret and apply, as pointed out by Wittes and Brittain (1990), who proposed to treat the first stage of a two-stage clinical trial as an internal pilot from which the overall sample size can be re-estimated. The specific problem they considered actually dated back to Stein's (1945) two-stage procedure for testing hypothesis $H_0 : \mu_X = \mu_Y$ versus the two-sided alternative $\mu_X \neq \mu_Y$ for the means of two independent normal distributions with common, unknown variance σ^2. In its first stage, Stein's procedure samples n_0 observations from each of the two normal distributions and computes the usual unbiased estimate s_0^2 of σ^2. The second stage samples $n_1 = n_0 \vee \lfloor \left(t_{2n_0-2,\alpha/2} + t_{2n_0-2,\beta}\right)^2 2s_0^2/\delta^2 \rfloor$ observations from each population, where $\lfloor \cdot \rfloor$ denotes the greatest integer function, α is the prescribed type I error probability, $t_{\nu,\alpha}$ is the upper α-quantile of the t-distribution with ν degrees of freedom, and $1-\beta$ is the prescribed power at the alternatives satisfying $|\mu_X - \mu_Y| = \delta$. The null hypothesis $H_0 : \mu_X = \mu_Y$ is then rejected if

$$|\bar{X}_{n_1} - \bar{Y}_{n_1}| > t_{2n_0-2,\alpha/2}\sqrt{2s_0^2/n_1}.$$

Modifications of the two-stage procedure were provided by Wittes and Brittain (1990), Gould and Shih (1992), and Herson and Wittes (1993), which represent the "first generation" of adaptive designs. The second generation of adaptive designs adopts a more aggressive method to re-estimate the sample size from the estimate of δ (instead of the

nuisance parameter σ) based on the first-stage data. In particular, Fisher (1998) considers the case of normally distributed outcome variables with known common variance σ^2. Letting n be the sample size for each treatment and $0 < r < n$, he notes that after rn pairs of observations (X_i, Y_i), $S_1 = \sum_1^{rn}(X_i - Y_i) \sim N(rn\delta, 2\sigma^2 rn)$, where $\delta = \mathbb{E}(X_i - Y_i)$. Let $\gamma > 0$ and $n^* = rn + \gamma(1-r)n$ be the new total sample size for each treatment. Under $H_0 : \delta = 0$,

$$S_2 = \sum_{i=rn^*+1}^{n^*} (X_i - Y_i) \sim N(0, 2\sigma^2(1-r)\gamma n),$$

hence the test statistic $(2\sigma^2 n)^{-1/2}(S_1 + \gamma^{-1/2}S_2)$ is standard normal. Whereas Fisher uses a "variance spending" approach as $1-r$ is the remaining part of the total variance that has not been spent in the first stage, Proschan and Hunsberger (1995) use a conditional Type I error function $C(z)$ with range $[0, 1]$ to define a two-stage procedure that rejects $H_0 : \delta = 0$ in favor of $\delta > 0$ if the second-stage z-value Z_2 exceeds $\Phi^{-1}(1 - C(Z_1))$, where Z_1 is the first-stage z-value. The type I error of the two-stage test can be kept at α if $\int_{-\infty}^{\infty} C(z)\phi(z)dz = \alpha$, where ϕ and Φ are the density function and distribution function, respectively, of the standard normal distribution.

Assuming normally distributed outcomes with known variances, Jennison and Turnbull (2006 a,b) introduced adaptive group sequential tests that choose the jth group size and stopping boundary on the basis of the cumulative sample size n_{j-1} and the sample sum $S_{n_{j-1}}$ over the first $j-1$ groups, and that are optimal in the sense of minimizing a weighted average of the expected sample sizes over a collection of parameter values, subject to prescribed error probabilities at the null and a given alternative hypothesis. They showed how the corresponding optimization problem can be solved numerically by using backward induction algorithms, and that standard (non-adaptive) group sequential tests with the first stage chosen appropriately are nearly as efficient as their optimal adaptive tests. They also showed that the adaptive tests proposed in the preceding paragraph performed poorly in terms of expected sample size and power in comparison with the group sequential tests. Tsiatis and Mehta (2003) attributed this inefficiency to the use of the non-sufficient "weighted" statistic. Bartroff and Lai's (2008a,b) approach to adaptive designs, developed in the general framework of multiparameter exponential families, uses efficient generalized likelihood ratio statistics in this framework and adds a third stage to adjust for the sampling

variability of the first-stage parameter estimates that determine the second-stage sample size. The possibility of adding a third stage to improve two-stage designs dated back to Lorden (1983), who used crude upper bounds for the type I error probability that are too conservative for practical applications. Bartroff and Lai overcame this difficulty by using new methods to compute the type I error probability, and also extended the three-stage test to multiparameter and multi-arm settings, thus greatly broadening the scope of these efficient adaptive designs. Details are summarized in Chapter 8, in particular Sections 8.2 and 8.3, of Bartroff et al. (2013), where Section 8.4 gives another modification of group sequential GLR tests for adaptive choice between the superiority and non-inferiority objectives of a new treatment during interim analyses of a clinical trial to test the treatment's efficacy, as in an antimicrobial drug developed by the company of one of the coauthors of Lai et al. (2006).

2.2 Adaptive Subgroup Selection in Confirmatory Trials

Choice of the patient subgroup to compare the new and control treatments is a natural compromise between ignoring patient heterogeneity and using stringent inclusion-exclusion criteria in the trial design and analysis. Lai et al. (2014) introduce a new adaptive design to address this problem. They first consider trials with fixed sample size, in which n patients are randomized to the new and control treatments and the responses are normally distributed, with mean μ_j for the new treatment and μ_{0j} for the control treatment if the patient falls in a pre-defined subgroup Π_j for $j = 1, \ldots, J$, and with common known variance σ^2. Let Π_J denote the entire patient population for a traditional randomized controlled trial (RCT) comparing the two treatments, and let $\Pi_1 \subset \Pi_2 \subset \cdots \subset \Pi_J$ be the J prespecified subgroups. Since there is typically little information from previous studies about the subgroup effect size $\mu_j - \mu_{0j}$ for $j \neq J$, Lai et al. (2014) begins with a standard RCT to compare the new treatment with the control over the entire population, but allows adaptive choice of the patient subgroup $\hat{I}$, in the event H_J is not rejected, to continue testing $H_i : \mu_i \leq \mu_{0i}$ with $i = \hat{I}$ so that the new treatment can be claimed to be better than control for the patient subgroup $\hat{I}$ if $H_{\hat{I}}$ is rejected. Letting $\theta_j = \mu_j - \mu_{0j}$ and

$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$, the probability of a false claim is the type I error

$$\alpha(\boldsymbol{\theta}) = \begin{cases} P_{\boldsymbol{\theta}}(\text{reject } H_J) + P_{\boldsymbol{\theta}}(\theta_{\hat{I}} \leq 0, \text{ accept } H_J \text{ and reject } H_{\hat{I}}) & \text{if } \theta_J \leq 0 \\ P_{\boldsymbol{\theta}}(\theta_{\hat{I}} \leq 0, \text{ accept } H_J \text{ and reject } H_{\hat{I}}) & \text{if } \theta_J > 0, \end{cases} \tag{1}$$

for $\boldsymbol{\theta} \in \Theta_0$. Subject to the constraint $\alpha(\boldsymbol{\theta}) \leq \alpha$, they prove the asymptotic efficiency of the procedure that randomly assigns n patients to the experimental treatment and the control, rejects H_J if $\text{GLR}_i \geq c_\alpha$ for $i = J$, and otherwise chooses the patient subgroup $\hat{I} \neq J$ with the largest value of the generalized likelihood ratio statistic $\text{GLR}_i = \{n_i n_{0i}/(n_i + n_{0i})\}(\hat{\mu}_i - \hat{\mu}_{0i})_+^2/\sigma^2$ among all subgroups $i \neq J$ and rejects $H_{\hat{I}}$ if $\text{GLR}_{\hat{I}} \geq c_\alpha$, where $\hat{\mu}_i(\hat{\mu}_{0i})$ is the mean response of patients in Π_i from the treatment (control) arm and $n_i(n_{0i})$ is the corresponding sample size. After establishing the asymptotic efficiency of the procedure in the fixed sample size case, they proceed to extend it to a 3-stage sequential design by making use of the theory of Bartroff and Lai reviewed in the preceding paragraph. They then extend the theory from the normal setting to asymptotically normal test statistics, such as the Wilcoxon rank sum statistics. These designs which allow mid-course enrichment using data collected, were motivated by the design of the DEFUSE 3 clinical trial at the Stanford Stroke Center to evaluate a new method for augmenting usual medical care with endovascular removal of the clot after a stroke, resulting in reperfusion of the area of the brain under threat, in order to salvage the damaged tissue and improve outcomes over standard medical care with intravenous tissue plasminogen activator (tPA) alone. The clinical endpoints of stroke patients are the Rankin scores, and Wilcoxon rank sum statistics are used to test for differences in Rankin scores between the new and control treatments. The DEFUSE 3 (**D**iffusion and Perfusion Imaging **E**valuation **f**or **U**nderstanding **S**troke **E**volution) trial design involves a nested sequence of $J = 6$ subsets of patients, defined by a combination of elapsed time from stroke to start of tPA and an imaging-based estimate of the size of the unsalvageable core region of the lesion. The sequence was defined by cumulating the cells in a two-way (3 volumes $\times$ 2 times) cross-tabulation as described by Lai et al. (2014, p. 195). In the upper left cell, c_{11}, which consisted of the patients with a shorter time to treatment and smallest core volume, the investigators were most confident of a positive effect, while in the lower right cell c_{23} with the longer time and largest core area, there was less confidence in the effect. The six cumulated groups, $\Pi_1, \ldots, \Pi_6$ give

rise to corresponding one-sided null hypotheses, $H_1, \ldots, H_6$ for the treatment effects in the cumulated groups.

Shortly before the final reviews of the protocol for funding were completed, four RCTs of endovascular reperfusion therapy administered to stroke patients within 6 hours after symptom onset demonstrated decisive clinical benefits. Consequently, the equipoise of the investigators shifted, making it necessary to adjust the intake criteria to exclude patients for whom the new therapy had been proven to work better than the standard treatment. The subset selection strategy became even more central to the design, since the primary question was no longer whether the treatment was effective at all, but for which patients should it be adopted as the new standard of care. Besides adapting the intake criteria to the new findings, another constraint was imposed by the NIH sponsor, which effectively limited the total randomization to 476 patients. The first interim analysis was scheduled after the 200 patients, and the second interim analysis after an additional 140 patients. DEFUSE 3 has a Data Coordinating Unit and an independent Data and Safety Monitoring Board (DSMB). Besides examining the unblinded efficacy results prepared by a designated statistician at the data coordination unit, which also provided periodic summaries on enrollment, baseline characteristics of enrolled patients, protocol violations, timeliness and completeness of data entry by clinical centers, and safety data. During interim analyses, the DSMB would also consider the unblinded safety data, comparing the safety of endovascular plus IV-tPA to that of IV-tPA alone, in terms of deaths, serious adverse events, and incidence of symptomatic intracranial hemorrhage.

In June 2017 positive results of another trial, **D**WI or CTP **A**ssessment with Clinical Mismatch in the Triage of **W**ake-Up and Late Presenting Strokes undergoing **N**eurointervention with Trevo (DAWN), which involved patients and treatments similar to those of DEFUSE 3, were announced. Enrollment in the DEFUSE 3 trial was placed on hold; an early interim analysis of the 182 patients enrolled to date was requested by the sponsor (NIH); see Albers et al. (2018) who say: “As a result of that interim analysis, the trial was halted because the prespecified efficacy boundary ($P < 0.0025$) had been exceeded." As reported by the aforementioned authors, DEFUSE 3 “was conducted at 38 US centers and terminated early for efficacy after 182 patients had undergone randomization (92 to the

endovascular therapy group and 90 to the medical-therapy group)." For the primary and secondary efficacy endpoints, the results show significant superiority of endovascular plus medical therapies. The DAWN trial "was a multicenter randomized trial with a Bayesian adaptive-enrichment design" and was "conducted by a steering committee, which was composed of independent academic investigators and statisticians, in collaboration with the sponsor, Stryker Neurovascular" (Nogueira et al., 2018). Early termination of DEFUSE 3 provides a concrete example of importance of a flexible group sequential design that can adapt not only to endogenous information from the trial but also to exogenous information from advances in precision medicine and related concurrent trials.

We conclude this section with recent regulatory developments in enrichment strategies for clinical trials and in adaptive designs of confirmatory trials of new treatments. In March 2019, the FDA released its Guidance for Industry on Enrichment Strategies for Clinical Trials to Support Determination of Effectiveness of Human Drugs and Biological Products. In November 2019, CDER and CBER of FDA released its Guidance for Industry on Adaptive Designs for Clinical Trials of Drugs and Biologics, which was an update of the 2010 CDER's Guidance for Industry on Adaptive Designs.

3 Analysis of Novel Confirmatory Trials

This section describes some advances in statistical methods for the analysis of the novel clinical trial designs of confirmatory trials in Section 2. It begins with hybrid resampling for inference on primary and secondary endpoints in Section 3.1. Section 3.2 considers statistical inference from multi-arm trials for developing and testing biomarker-guided personalized therapies.

3.1 Hybrid Resampling for Primary and Secondary Endpoints

Tsiatis et al. (1984) developed exact confidence intervals for the mean of a normal distribution with known variance following a group sequential test. Subsequently, Chuang and Lai (1998, 2000) noted that even though $\sqrt{n}(\bar{X}_n - \mu)$ is a pivot in the case of $X_i \sim N(\mu, 1)$, $\sqrt{T}(\bar{X}_T - \mu)$ is highly non-pivotal for a group sequential stopping time,

hence the need for the *exact method* of Tsiatis et al. (1984), which they generalized as follows. If $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ is indexed by a real-valued parameter θ, an exact equal-tailed confidence region can always be found by using the well-known duality between hypothesis tests and confidence regions. Suppose one would like to test the null hypothesis that θ is equal to θ_0. Let $R(\mathbf{X}, \theta_0)$ be some real-valued test statistic. Let $u_\alpha(\theta_0)$ be the α-quantile of the distribution of $R(\mathbf{X}, \theta_0)$ under the distribution F_{θ_0}. The null hypothesis is accepted if $u_\alpha(\theta_0) < R(\mathbf{X}, \theta_0) < u_{1-\alpha}(\theta_0)$. An exact equal-tailed confidence region with coverage probability $1 - 2\alpha$ consists of all θ_0 not rejected by the test and is therefore given by $\{\theta : u_\alpha(\theta) < R(\mathbf{X}, \theta) < u_{1-\alpha}(\theta)\}$. The exact method, however, applies only when there are no nuisance parameters and this assumption is rarely satisfied in practice. To address this difficulty, Chuang and Lai (1998, 2000) introduced a *hybrid resampling method* that "hybridizes" the exact method with Efron's (1987) bootstrap method to construct confidence intervals. The bootstrap method replaces the quantiles $u_\alpha(\theta)$ and $u_{1-\alpha}(\theta)$ by by the approximate quantiles u^*_α and $u^*_{1-\alpha}$ obtained in the following manner. Based on $\mathbf{X}$, construct an estimate $\hat{F}$ of $F \in \mathcal{F}$. The quantile u^*_α is defined to be α-quantile of the distribution of $R(\mathbf{X}^*, \hat{\theta})$ with $\mathbf{X}^*$ generated from $\hat{F}$ and $\hat{\theta} = \theta(\hat{F})$, yielding the confidence region $\{\theta : u^*_\alpha < R(\mathbf{X}, \theta) < u^*_{1-\alpha}\}$ with approximate coverage probability $1 - 2\alpha$. For group sequential designs, the bootstrap method breaks down because of the absence of an approximate pivot, as shown by Chuang and Lai (1998). The hybrid confidence region is based on reducing the family of distributions $\mathcal{F}$ to another family of distributions $\{\hat{F}_\theta : \theta \in \Theta\}$, which is used as the "resampling family" and in which θ is the unknown parameter of interest. Let $\hat{u}_\alpha(\theta)$ be the α-quantile of the sampling distribution of $R(\mathbf{X}, \theta)$ under the assumption that $\mathbf{X}$ has distribution $\hat{F}_\theta$. The hybrid confidence region results from applying the exact method to $\{\hat{F}_\theta : \theta \in \Theta\}$ and is given by

$$\{\theta : \hat{u}_\alpha(\theta) < R(\mathbf{X}, \theta) < \hat{u}_{1-\alpha}(\theta)\}. \tag{2}$$

The construction of (2) typically involves simulations to compute the quantiles as in the bootstrap method.

Since an exact method for constructing confidence regions is based on inverting a test, such a method is implicitly or explicitly linked to an ordering of the sample space of the test statistic used. The ordering defines the p-value of the test as the probability (under

the null hypothesis) of more extreme values (under the ordering) of the test statistic than that observed in the sample. Under a total ordering $\leq$ of the sample space of (T, S_T), Lai and Li (2006) call (t, s) a qth quantile if $P\{(T, S_T) \leq (t, s)\} = q$, which generalizes Rosner and Tsiatis' exact method for randomly stopped sums S_T of of independent normal random variables with unknown mean μ. For the general setting where a stochastic process $\mathbf{X}_u$, in which u denotes either discrete or continuous time, is observed up to a stopping time T, Lai and Li (2006) define $\mathbf{x} = \{\mathbf{x}_u : u \leq t\}$ to be a qth quantile if

$$P\{\mathbf{X} \leq \mathbf{x}\} \geq q, \quad P\{\mathbf{X} \geq \mathbf{x}\} \geq 1 - q, \tag{3}$$

under a total ordering $\leq$ for the sample space of $\mathbf{X} = \{\mathbf{X}_u : u \leq T\}$. For applications to confidence intervals of a real parameter θ, the choice of the total ordering should be targeted toward the objective of interval estimation. Let $\{U_r : r \leq T\}$ be real-valued statistics based on the observed process $\{\mathbf{X}_s : s \leq T\}$. For example, let U_r be an estimate of θ based on $\{\mathbf{X}_s : s \leq r\}$. A total ordering on the sample space of $\mathbf{X}$ can be defined via $\{U_r : r \leq T\}$ as follows:

$$\mathbf{X} \geq \mathbf{x} \text{ if and only if } U_{T \wedge t} \geq u_{T \wedge t}, \tag{4}$$

in which $\{u_r : r \leq t\}$ is defined from $\mathbf{x} = \{\mathbf{x}_r : r \leq t\}$ in the same way as $\{U_r : r \leq T\}$ is defined from $\mathbf{X}$ and which has the attractive feature that the probability mechanism generating $\mathbf{X}_t$ needs only to be specified up to the stopping time T in order to define the quantile. Bartroff et al. (2013, p.164) remark that if $U_r = \sqrt{r}(\bar{X}_r - \mu_0)$ then the Lai-Li ordering is equivalent to Siegmund's ordering and also to the Rosner-Tsiatis ordering, but "the original Rosner-Tsiatis ordering requires $n_1, \ldots, n_k$ (or the stochastic mechanism generating them to be completely specified" and has difficulties "described in the last paragraph of Sect. 7.1.3 if this is not the case."

Bartroff et al. (2013, Sections 7.4 and 7.5) describe how this ordering can be applied to implement resampling for secondary endpoints together with applications to time-sequential trials which involve interim analyses at calendar time t_j $(1 \leq j \leq k)$, with $0 < t_1 < \cdots < t_k = t^*$ (the prescribed duration of the trial), and which have time to failure as the primary endpoint; Lai et al. (2009) have also extended this approach to inference on secondary endpoints in adaptive or time-sequential trials.

3.2 Statistical Inference from Multi-Arm Trials for Developing and Testing Biomarker-Guided Personalized Therapies

Lai et al. (2013) first elucidate the objectives underlying the design and analysis of these multi-arm trials that attempt to select the best of k treatments for each biomarker-classified subgroup of cancer patients in Phase II studies, with objectives that include (a) treating accrued patients with the best (yet unknown) available treatment, (b) developing a biomarker-guided treatment strategy for future patients, and (c) demonstrating that the strategy developed indeed has statistically significantly better treatment effect than some predetermined threshold. The group sequential design therefore uses an outcome-adaptive randomization rule, which updates the randomization probabilities at interim analyses and uses GLR statistics and modified Haybittle-Peto rules to include early elimination of inferior treatments from a biomarker class. It is shown by Lai et al. (2013) to provide substantial improvements, besides being much easier to implement, over the Bayesian outcome-adaptive randomization design used in the BATTLE (**B**iomarker-integrated **A**pproaches of **T**argeted **T**herapy for **L**ung Cancer **E**limination) trial of personalized therapies for non-small cell lung cancer. An April 2010 editorial in *Nature Reviews in Medicine* points out that BATTLE design, which "allows researchers to avoid being locked into a single, static protocol of the trial" that requires large sample sizes for multiple comparisons of several treatments across different biomarker classes, 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 by Lai et al. (2013, pp.651-653, 662), targeted therapies that target the cancer cells (while leaving healthy cells unharmed) and the "right" patient population (that has the genetic or other markets for the sensitivity to the treatment) have great promise in cancer treatments but also challenges in designing clinical trials for drug development and regulatory approval. One challenge is to identify the biomarkers that are predictive of response and another is to develop a biomarker classifier that can identify patients who are sensitive to the treatments. We can address these challenges by using recent advances in contextual multi-arm bandit theory, which we summarize below.

The K-arm bandit problem, introduced by Robbins (1952) for the case $K = 2$, is prototypical in the area of stochastic adaptive control that addresses the dilemma between

"exploration" (to generate information about the unknown system parameters needed for efficient system control) and "exploitation" (to set the system inputs that attempt to maximize the expected rewards from the outputs.) Robbins considered the problem of which of K populations to sample from sequentially in order to maximize the expected sum $E\left(\sum_{i=1}^{N} Y_i\right)$. Let $\mathcal{F}_t$ be the history (or more formally, the σ-algebra of events) up the time t. An allocation rule $\phi = (\phi_1, \ldots, \phi_N)$ is said to be "adaptive" if $\{\phi_t = k\} \in \mathcal{F}_{t-1}$ for $k = 1, \ldots, K$. Suppose Y_t has density function f_{θ_k} when $\phi_t = k$, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. Let μ_k be the mean of the kth population, which is assumed to be finite. Then

$$E_{\boldsymbol{\theta}}\left(\sum_{t=1}^{N} Y_t\right) = \sum_{t=1}^{N}\sum_{k=1}^{K} E_{\boldsymbol{\theta}}\{E_{\boldsymbol{\theta}}(Y_t I_{\{\phi_t=k\}}|\mathcal{F}_{t-1})\} = \sum_{k=1}^{K} \mu(\theta_k) E_{\boldsymbol{\theta}} T_N(k), \tag{5}$$

where $T_N(k) = \sum_{t=1}^{N} I_{\{\phi_t=k\}}$ is the total sample size from population k. If the population k^* with the largest mean were known, then obviously one should sample from it to receive expected reward $N\mu_{k^*}$, where $\mu_{k^*} = \max_{1\le k\le K} \mu_k$. Hence maximizing the expected sum $E_{\boldsymbol{\theta}}(\sum_{t=1}^{N} Y_t)$ is equivalent to minimizing the regret, or shortfall from $N\mu_{k^*}$:

$$R_N(\boldsymbol{\theta}) = N\mu_{k^*} - E_{\boldsymbol{\theta}}\left(\sum_{t=1}^{N} Y_t\right) = \sum_{k:\mu(\theta_k)<\mu_{k^*}} \{\mu_{k^*} - \mu(\theta_k)\} E_{\boldsymbol{\theta}} T_N(k), \tag{6}$$

in which the second equality follows from (5) and shows that the regret is a weighted sum of expected sample sizes from inferior populations. Making use of this representation in terms of expected sample sizes, Lai and Robbins (1985) derive an the asymptotic lower bound, as $N \to \infty$, for the regret $R_N(\boldsymbol{\theta})$ of uniformly good adaptive allocation rules:

$$R_N(\boldsymbol{\theta}) \ge (1 + o(1)) \sum_{k:\mu(\theta_k)<\mu(\theta^*)} \frac{\mu(\theta^*) - \mu(\theta_k)}{I(\theta_k, \theta^*)} \log N, \tag{7}$$

where $\theta^* = \theta_{k^*}$ and $I(\boldsymbol{\theta}, \lambda) = E_\theta\{\log(f_\theta(Y)/f_\lambda(Y))\}$ is the Kullback-Leibler information number; an adaptive allocation rule is called "uniformly good" if $R_N(\boldsymbol{\theta}) = o(N^a)$ for all $a > 0$ and $\boldsymbol{\theta}$. They show that the asymptotic lower bound (7) can be attained by the "upper confidence bound" (UCB) rule that samples from the population (arm) with the largest upper confidence bound, which incorporates uncertainty in the sample mean by the numbers of observations sampled from the arm (i.e., width of a one-sided confidence interval.)

New applications and advances in information technology and biomedicine in the new millenium have led to the development of *contextual multi-arm bandits*, also called bandits with side information or covariates, while the classical multi-arm bandits reviewed above are often referred to as "context-free" bandits. Personalized marketing (e.g., Amazon) uses web sites to track a customer's purchasing records and thereby to maket products that are individualized for the customer. Recommender systems select items such as movies (e.g., Netflix) and news (e.g., Yahoo) for users based on the users' and items' features(covariates). Whereas classical K-arm bandits reviewed above aim at choosing ϕ_i sequentially so that $E_{\boldsymbol{\theta}}(\sum_{i=1}^N Y_i)$ is as close as possible to $N \max_{1\le k\le K} \mu_k$, contextual bandits basically replace $N\mu_k$ by $\sum_{i=1}^N \mu_k(\boldsymbol{x_i})$, where $\boldsymbol{x_i}$ is the covariate of the ith subject, noting that analogous to (5),

$$E_{\boldsymbol{\theta}}(Y_i) = \sum_{k=1}^K E_{\boldsymbol{\theta}}\{E_{\boldsymbol{\theta}}(Y_i I_{\{\phi_i=k\}}|\boldsymbol{x_i}, \mathcal{F}_{t-1})\} = \sum_{k=1}^K E_{\boldsymbol{\theta}}(\mu_k(\boldsymbol{x_i}) I_{\{\phi_i=k\}}). \tag{8}$$

Assuming $\boldsymbol{x_i}$ to be i.i.d. with distribution G, we can define $g^*(x) = \arg\max_{1\le k\le K} \mu(\theta_k, x)$, $\theta^*(x) = \theta_{k^*(x)}$ and the regret

$$\begin{aligned} R_N(\boldsymbol{\theta}, B) &= N \int_B \mu(\theta^*(\boldsymbol{x}), \boldsymbol{x}) dG(\boldsymbol{x}) - \sum_{i=1}^N \sum_{k=1}^K \int_B \mu(\theta_k, \boldsymbol{x}) E_{\boldsymbol{\theta}}(I_{\{\phi_i=k\}}) dG(\boldsymbol{x}) \\ &= \sum_{k=1}^K \int_B \{\mu(\theta^*(\boldsymbol{x}), \boldsymbol{x}) - \mu(\theta_k, \boldsymbol{x})\} E_{\boldsymbol{\theta}} T_N(k, \boldsymbol{x}) dG(\boldsymbol{x}) \end{aligned} \tag{9}$$

for Borel subsets B of the support of G, where $T_N(k, B) = \sum_{i=1}^N I_{\{\phi_i=k, \boldsymbol{x_i}\in B\}}$, noting that the measure $E_{\boldsymbol{\theta}} T_N(k, \cdot)$ is absolutely continuous with respect to G, hence $E_{\boldsymbol{\theta}} T_N(k, \boldsymbol{x})$ in (9) is its Radom-Nikodym derivative with respect to G. For contextual bandits, an arm that is inferior at $\boldsymbol{x}$ may be the best at $\boldsymbol{x}'$. Therefore the uncertainty in the sample mean reward at $\boldsymbol{x}_t$ does not need to be immediately accounted for, and adaptive randomization (rather than UCB rule) can yield an asymptotically optimal policy.

To achieve the objectives (a), (b) and (c) in the first paragraph of this subsection, Lai et al.(Lai et al., 2013, pp.654-655) use contextual bandit theory which we illustrate below with $J = 3$ groups of patients and $K = 3$ treatments, assuming normally distributed responses with mean μ_{jk} and known variance 1 for patients in group j receiving treatment k. Using Bartroff and Lai's adaptive design (2008a,b) reviewed in Section 2.1, let n_i denote the total

sample size up to the time of the ith interim analysis, n_{ij} denote the total sample size from group j in those n_i patients, and let n_{ijk} be the total sample size from biomarker class j receiving treatment k up to the ith interim analysis. Because it is unlikely for patients to consent to being assigned to a seemingly inferior treatment, randomization in a double blind setting (in which the patient and the physician both do not know whether treatment or control is assigned) is needed for informed consent. Contextual bandit theory suggests assigning the highest randomization probability between interim analyses i and $i+1$ to $\hat{k}_j^{(i)} = \arg\max_k \hat{\mu}_{jk}$ (which is the MLE of $k_j^* = \arg\max_k \mu_{jk}$) and eliminating treatment k from the set of $\mathcal{K}_{ij}$ of surviving treatments at the ith interim analysis if the GLR statistic $l_j^i(k, \hat{k}_j^{(i)})$ exceeds $5\delta_{ij}$, where $\delta_{ij} \to 0$ but $\sqrt{n_{ij}}\delta_{ij} \to \infty$, with a randomization scheme in which

$$\pi_{jk}^{(i)} = \left(1 - \varepsilon|\mathcal{K}_{ij} \setminus \mathcal{H}_{ij}|\right) / |\mathcal{H}_{ij}|, \tag{10}$$

in which $|A|$ denotes the cardinality of a finite set A and $\mathcal{H}_{ij} = \{k \in \mathbb{X}_{ij} : |\hat{\mu}_{jk}^{(i)} - \hat{\mu}_{j,\hat{k}_j^{(i)}}| \leq \delta_{ij}\}$. Equal randomization (with randomization probability $1/K$) for the K treatments is used up to the first interim analysis. In context-free multi-arm bandit theory, this corresponds to the ε-greedy algorithm which has been shown by Auer et al. (2002) to provide an alternative to the UCB rule for attaining the asymptotic lower bound for the regret. Lai et al. (2013) introduce a subset selection method for selecting a subset of treatments at the end of the trial to be used for future patients, with an overall probability guarantee of $1-\alpha$ to contain the best treatment for each biomarker class, and such that the expected size of the selected subset is as small as possible in some sense. They also develop a group sequential GLR test with prescribed type I error to demonstrate that the developed treatment strategy improves the mean treatment effect of SOC by a given margin.

4 Precision Medicine and Master Protocols

This section begins with an overview of precision-guided drug development in Section 4.1 and master protocols to collect data for this goal and for testing the new treatment developed. Section 4.2 describes the FRACTION, ADVISE, and Checkmate-848 master protocols at Bristol-Meyers Squibb (BMS). Sections 4.3 and 4.4 present an example of adaptive

basket trial design, and discuss new opportunities for statistical science in precision-guided drug development and regulatory approval.

4.1 Precision-Guided Drug Development and Basket Protocols

Janet Woodcock, director of FDA's Center for Drug Evaluation and Research (CDER), published in 2017 a seminal paper on master protocols of "mechanism-based precision medicine trials," affordable in cost, time, and sample size, to study multiple therapies, multiple diseases, or both; see Woodcock and LaVange (2017). Table 2 of the paper lists six such trials to illustrate the concept: (i) B2225, a Phase II basket trial, (ii) BRAF V600, an early Phase II basket trial, (iii) NCI-Match, a Phase I followed by Phase II umbrella trial, (iv) BATTLE-1, a Phase II umbrella trial, (v) I-SPY 2, a Phase II platform trial, and (vi) Lung-MAP, a Phase II-III trial with a master protocol to study 4 molecular targets for NSCLC initially, to be trimmed to 3 targets for the PHASE III confirmatory trial. We have discussed the BATTLE (respectively, I-SPY) trials for therapies to treat NSCLC (respectively, breast cancer) in Section 3.2. For NCI-Match, a treatment is given across multiple tumors sharing a common biomarker; see Conley and Doroshow (2014) and Do et al. (2015). Hyman et al. (2015) describe the BRAF V600 basket trial, after noting that (a) BRAF V600 mutations occur in almost 50% of cutaneous melanomas and result in constitutive activation of downstream signaling through the MAPK (mitogen-activated protein kinase) pathway, based on previous studies reported by Davies et al. (2002) and Curtin et al. (2005); (b) Vemurafenib, a selective oral inhibitor of BRAF v600 kinase produced by Roche-Genentech, has been shown to improve survival of patients with BRAF V600E mutation-positive metastic melanoma according to Chapman et al. (2011); and (c) efforts by the Cancer Genome Atlas and other initiatives have identified BRAF V600 mutations in non-melanoma cancers (De Roock et al., 2010; Van Cutsem et al., 2011; Weinstein et al., 2013; Kris et al., 2014). They point out that "the large number of tumor types, low frequency of BRAF V600 mutations, and the variety of some of the (non-melanoma) cancers make disease specific studies difficult (unaffordable) to conduct." Hyman et al. (2015) therefore use six "baskets" (NSCLC, ovarian, colorectal, and breast cancers, multiple myeloma, cholangiocarncinoma) plus a seventh ("all-others") basket which "permitted

enrollment of patients with any other BRAF V600 mutation-positive cancer" in their Phase II basket trial of Vemurafenib. The Phase II trial uses Simon's two-stage design "for all tumor-specific cohorts in order to minimize the number of patients treated if vemurafenib was deemed ineffective for a specific tumor type." The primary efficacy endpoint was response rate at week 8. "Kaplan-Meier methods were used to estimate progression-free and overall survival. No adjustments were made for multiple hypothesis testing that could result in positive findings."

In the BRAF V600 trial, 122 adults received at least one dose of Vemurafenib (20 for NSCLC, 37 for colorectal cancer, 5 for multiple myeloma, 8 for cholangocarcinoma, 18 for ECD or LCH, 34 for breast, ovarian, and "other" cancers), and 89% of these patients had at least one previous line of therapy. Vemurafenib showed (a) "efficacy in BRAF V600 mutation-positive NSCLC" compared to standard second-line docetexal in molecularly unselected patients, and (b) for ECD or LCH "which are closely related orphan diseases with no approved therapies," the response rate was 43% and none of the patients had disease progression while receiving therapy, despite a median treatment duration of 5.9 months. Hyman et al. (2015) point out that "one challenge in interpreting the results of basket studies is drawing inferences from small numbers of patients." Following up on this point, Berry (2015) discusses other challenges for inference from basket trials. In particular, he points out that even though patients have the same biomarker, different tumor sites and tumor types may have different response rates and simply pooling trial results across tumor types may mislead interpretation. On the other hand, different tumors may have similar response rates and hierarchical Bayesian modeling can help borrow information across these types to compensate for the small sample sizes.

We include here another basket trial led by our Stanford colleague, Dr. Shivaani Kummar. She collaborated with investigators at Loxo Oncology in South San Francisco, and other investigators at UCLA, USC, Harvard, Cornell, Vanderbilt, MD Anderson, and Sloan Kettering, to design and conduct a basket trial involving seven specified cancer types and an eighth basket ("other cancers") to evaluate the efficacy and safety of larotretinib, a highly selective TRK inhibitor produced by Loxo Oncology in South San Francisco, for adults and children who had TRK fusion-positive cancers. A total of 55 patients were enrolled into one

of three protocols and treated with larotretinib: a Phase I study involving adults, a Phase I-II study involving adults and children, and a Phase II study involving adolescents and adults with TRK fusion-positive tumors. The Phase II study uses the recommended dose of the drug twice daily. The dose-escalation Phase I study and the Phase I portion of the Phase I-II study do not require the subjects to have TRK fusions although the combined analysis only includes "patients with prospectively identified TRK fusions." The primary endpoint for the combined analysis was the overall response assessed by an independent radiology committee. Secondary endpoints include duration of response, progression-free survival, and safety. At the data-cutoff date 7/17/2017, the overall response rate was 75%, and 7 of the patients had complete response while 34 had partial response; see Drilon et al. (2018). In the accompanying editorial of that issue in *NEJM*, André (2018) says that "this study is an illustration of what is likely to be the future of drug development in rare genomic entities" and that according to the Magnitude of Clinical Benefit Scale for single-arm trials recently developed by the European Society of Medical Oncology, "studies that show rates of objective response of more than 60% and a median progression-free survival of more than 6 months, as the study conducted by Drilon et al. does, are considered to have the highest magnitude of clinical benefit" in line with the pathway for single-arm trials of treatments of rare diseases with well-established natural histories to receive approval from regulatory agencies. André (2018) also mentions that the study by Drilon et al. "did not find any difference in efficacy among the 12 tumor histotypes (including those in the all-other basket)," proving a successful "trans-tumor approach" in the case of TRK fusions with larotrectinib, but that "some basket trials have not shown evidence of trans-tumor efficacy of targeted therapies, notably BRAF inhibitors." He points out the importance of developing "statistical tools to support a claim that a drug works across tumor types" and to provide "a more in-depth understanding of the failure of some targets in a trans-tumor approach."

BioPharma Dive, a company in Washington, D.C. that provides news and analysis of clinical trials, drug discovery, and development, FDA regulations and approvals, for biotech and biopharmaceutical corporations, has a 2019 article sponsored by Paraxel, a global provider of biopharmaceutical services headquarted in Waltham, MA, highlighting

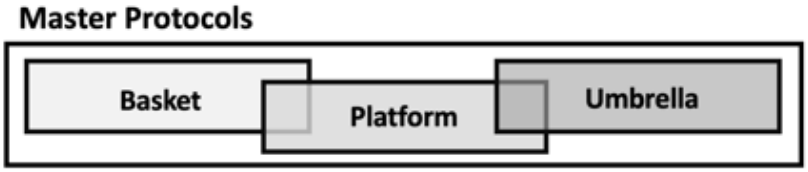
that “in the past five years, we’ve seen a sharp increase in the number of trials designed with a precision medicine approach," and that “in 2018 about one of every four trials approved by the FDA was a precision medicine therapy;" see (BioPharma Dive, 2019). Moreover, “developing these medicines requires changes to traditional clinical trial designs, as well as the use of innovative testing procedures that result in new types of data," and “the FDA has taken proactive steps to modernize the regulatory framework" that "prioritizes novel clinical trials and real-world data solutions to provide robust evidence of safety and efficacy at early stages." The February 12, 2020, news item of BioPharma Dive is about Merck’s positive results for its cancer drug Keytruda, when combined with chemotherapy, in breast cancer patients on whom a certain amount of tumor and immune cells express a protein that make Keytruda truly effective for this difficult-to-treat form of breast cancer called “triple negative." The news item the following day is that the FDA granted BMS’s CAR-T treatment (called liso-cel) of a type of lymphoma priority reviews, setting up a decision by August 17, 2020; see BioPharma Dive (2020*a*,*b*). Liso-cel was originally developed by the biotech company Juno Therapeutics before its acquisition by Celgene in 2018. In Jan 2019, BMS announced its $74 billion acquisition of Celgene and completed the acquisition in November that year after regulatory approval by all the government agencies required by the merger agreement.

Since 2016, Stanford University has held an annual drug discovery symposium, focusing on precision-guided drug discovery and development. We briefly describe here the work of Brian Kobilka, one of the founding conference organizers and the director of Kobilka Institue of Innovative Drug Discovery (KIDD) at The Chinese University of Hong Kong, Shenzhen, and his former mentor and Nobel Prize co-winner Robert Lefkowitz. In a series of seminal papers from 1981 to 1984 published by Lefkowitz and his postdoctoral fellows at the Howard Hughes Medical Institute and Departments of Medicine and Biochemistry at Duke University, the β_2-subtypes of the pharmacologically important β-andrenergic receptor (βAR) were purified to homogeneity and demonstrated to retain binding activity. Dixon, Sigal, and Strader of Merck Research Laboratories subsequently collaborated with Lefkowitz, Kobilka and others on their team at Duke to derive an amino-acid sequence of peptides which indicated significant amino-acid homology with bovine rhodopsin and were

able to find a genomic intronless clone in 1986. In his Dec. 2012 Nobel Lecture, Lefkowitz highlights the importance of the discovery, saying: "Today we know that GPCRs (G protein coupled receptors), also known as seven transmembrane receptors, represent by far the largest, most versatile and most ubiquitous of the several families of plasma membrane receptors ... Moreover, these receptors are the targets for drugs accounting for more than half of all prescription drug sales in the world (Pierce et al., 2002)." Kobilka highlights in his Nobel lecture his efforts to understand the structural basis of β_2AR using advances in X-ray crystallography and later in electron microscopy to study the crystal structure of β_2AR. He concludes his Nobel lecture by saying: "While the stories outlined in this lecture have advanced the field, much work remains to be done before we can fully understand and pharmacologically control signaling by these fascinating membrane proteins." This work is continued at the Kobilka Institute of Innovative Drug Discovery and by his and other groups at Stanford, Lefkowitz's group at Duke, and other groups in other centers in academia and industry, in North America, Asia, and Europe.

4.2 FRACTION, ADVISE, and Checkmate-848 Trials at BMS

Bristol-Myers Squibb had initiated in 2016 a Phase II program comprising multiple tumor-specific platform trials, collectively known as FRACTION (**F**ast **R**eal-time **A**ssessment of **C**ombination **T**herapies in **I**mmuno- **O**ncology). Motivated by an objective response rate (ORR) of 58% for the combination of nivolumab and ipilimumab, and ORRs of 44% and 19% for the monotherapies in untreated metastatic melanoma (Larkin et al., 2015), BMS hypothesized that nivolumab as an adjuvant to the next-generation IO agents may provide clinically meaningful benefit in a broader set of malignancies. However, with the plethora of new IO agents in early development, conventional independent Phase II designs were considered insufficient to study the various IO therapy permutations. Therefore, FRACTION aimed to "reduce the time and number of patients needed to identify potentially beneficial regimens that merit advancing to Phase II/III registrational trials" (Simonsen et al., 2018). FRACTION comprises three Phase II platform trials that focus on advanced NSCLC, gastric cancer (GC), and renal cell carcinoma (RCC). Each tumor-specific FRACTION study is governed by a master protocol and subprotocols. Efficacious combination treatments that



	Basket	Platform	Umbrella
Investigational therapy	Single (or multiple)	Single or multiple	Multiple
Populations	Multiple	Single or multiple	Single (or multiple)
Other		Add/drop sub-studies	
Median study size*	n = 205	n = 892	n = 346
Median study duration*	22.3 months	58.9 months	60.9 months
Prevalence*	59%	19%	22%
* Park et al. 2019			
Example studies	BRAF V600 CheckMate-848	I-SPY 2 FRACTION	LUNG-MAP ADVISE

Figure 1: *Examples of Master Protocols. A "master protocol" is the overarching term and equally applies to basket, umbrella, and platform trials .*

meet prespecified early decision criteria enroll further subjects, while futile treatment arms are terminated early. The decision to continue or stop treatment is based on the number of objective responses. The primary endpoints are ORR, median duration of response (DOR) and progression-free survival (PFS), according to the Response Evaluation Criteria In Solid Tumors (RECIST) v1.1 criteria at week 24 of each treatment regimen. Similar to I-SPY 2, FRACTION uses adaptive randomization informed by a subject's pretreatment biomarker status. Specifically, the NSCLC FRACTION uses Simon-Fleming 4-stage designs, while the GC and RCC FRACTIONS use Simon's 2-stage designs; see Simonsen et al. (2018) who suggest that single-stage designs or Bayesian continuous monitoring (Thall and Simon, 1994) could also be used. A limitation of the FRACTION design is that successful therapies require a confirmatory Phase II/III study for registrational approval, as early decisions are based on small sample sizes using surrogate measures of survival. The program is ongoing, and it remains to be determined if the FRACTION approach can substantially accelerate the development of new IO combinations.

In collaboration with multiple academic and industry partners, BMS recently initiated an adaptive Phase I, randomized, multi-center, open-label trial, named ADVISE, which prospectively evaluates the feasibility and utility of real-time biomarker profiles in selecting IO combination therapies, and which is still ongoing. The investigational combina-

tions in ADVISE (**AD**apti**V**e Biomarker Trial That **I**nform**S** the **E**volution of combination immuno-oncology therapies) use nivolumab (anti-PD1) as a backbone therapy and add a single next-generation IO agent chosen from the following list: lirimumab (anti-KIR), relatlimab (anti-LAG3), cabiralizumab (anti-CSF-1R), ipilimumab (anti-CTLA-4), BMS-986156 (anti-GITR), BMS-986205 (IDO1), and stereotactic radiation therapy (SBRT), hence a seven-arm trial; see Luke et al. (2018); Ott et al. (2017); Heinzerling et al. (2016); Siu et al. (2017); Pardoll (2012); Ascierto et al. (2017); Bell et al. (2016). Prior clinical studies have shown that the nivolumab-based IO regimens in ADVISE are pharmacologically active and tolerable. It is open to subjects across a broad range of pre-treated solid tumors: advanced melanoma, NSCLC, renal-cell carcinoma (RCC), urothelial bladder cancer (UBC), squamous cell carcinoma of the head and neck (SCCHN), and gastric or gastroesophageal junction carcinoma (GEC). Fifty subjects are randomized to one of the seven study arms, based on a pre-defined decision algorithm that considers the subject's immune profile of the tumor microenvironment. Each patient's immunohistochemical (IHC) phenotype signature, as evaluated by baseline biopsy, is matched with the target profile for a given molecular mechanism of action of the investigational therapy. The IHC signature measures low, medium, or high expressions of LAG-3, CSF-1R, GITR, IDO, NKp46, and FOCP3. Where biomarker levels cannot clearly recommend a specific combination therapy, patients are randomized to any of the treatment arms. For example, if all markers are expressed at moderate levels, the patient is randomized. While efficacy is the primary endpoint, the secondary endpoints are safety and the change of histopathologic features and biomarker expression patterns from baseline (Luke et al., 2018).

Keynote 158 was the first tissue-agnostic immuno-oncology study to receive accelerated approval from the FDA. It evaluated pembrolizumab across multiple solid tumors expressing genetic markers for high microsatellite instability (MSI-H), or mismatch repair deficiency (dMMR). Objective response rate (ORR) data for the 149 randomized patients supported the accelerated approval of pembrolizumab for MSI-H or dMMR solid tumors (Ott et al., 2017; Pardoll, 2012). Additional promising tissue-agnostic biomarkers have emerged since then. For example, tumors with a high mutational burden (TMB-H) may contain numerous neo-antigens, and are expected to be more immunogenic than tumors

with comparatively low mutational burden (Heinzerling et al., 2016). TMB measures the number of mutations carried by tumor cells and median TMB has been shown to correlate with ORR; see Bell et al. (2016) and Ascierto et al. (2017). CheckMate-848 is a randomized, open-label, Phase II basket trial of nivolumab monotherapy, or nivolumab in combination with ipilimumab, in advanced or metastatic TMB-H solid tumors (Siu et al., 2017). The study enrolls subjects with a refractory, metastatic, or unresectable histologically solid malignant tumor with TMB-H, who are refractory to standard local therapies, or for whom no standard treatment is available. Excluded are subjects with melanoma, NSCLC, renal cell carcinoma or hematological malignancy at primary site of disease. The primary endpoint is ORR in participants with a high expression of tissue- and blood-based TMB. Duration of response (DOR), Time to Response (TTR), Clinical Benefit Rate (CBR), progression-free survival (PFS), and overall survival (OS) are secondary endpoints. A potential limitation is that nivolumab monotherapy may not be the standard of care across the included tumor types at the scheduled end of CheckMate-848.

4.3 Design of CK2 Inhibitor Basket Trial

This basket trial is an ongoing project of the SPARK Translational Research Program at Stanford led by Drs. Kevin Grimes and Teresa Purzner, in collaboration with Senhwa Biosciences at San Diego and Taipei, which produces silmitasertib (CX-4945), a potent and selective inhibitor of CK2 (casein kinase 2) that is part of the hedgehog signaling pathway which the neurosurgeon Dr. Purzner uses to treat pediatric patients with recurrent medulloblastoma (MB), a rare malignant pediatric brain tumor. Senhwa Biosciences and the Pediatric Brain Tumor Consortium (www.pbtc.org) signed a clinical trial agreement in June 2018 to study CX-4945. It is believed that CX-4945 may be efficacious for other cancers associated with the hedgehog signaling pathway. Senhwa Biosciences announced that FDA approved its IND application of CX-4945 for basal-cell carcinoma (BCC) in Nov 2018, and that it had begun recruiting adult patients from six clinical centers in the US. Moreover, in Jan 2017, it announced that the FDA had granted orphan drug designation to CX-4945 for the treatment of cholangiocarcinoma. Thus, a basket trial design is used for CX-4945, involving three baskets: BCC for adults, MB for children, and all other cancers

from the clinical centers in the study.

4.4 New Opprtunities for Statistical Science

Woodcock and LaVange (2017) point out new opportunities for statistical science in the design and analysis of master protocols; "With multiple questions to address under a single protocol, usually in an area of unmet need, and an extensive infrastructure in place to handle data flow, master protocols are a natural environment for considering innovative trial designs. The flexibility to allow promising new therapies to enter and poor-performing therapies to discontinue usually requires some form of adaptive design, but the level of complexity of those adaptations can vary according to the objectives of the master protocol." Section 4.3 summarizing our recent work at the SPARK Translational Research Program illustrates the opportunities for adaptive design noted by Woodcock and LaVange, especially how the design should adapt to "the objectives of the master protocol" during the course of the trial. They also point out that "two types of innovation are hallmarks of master protocols: the use of a trial network with infrastructure in place to streamline trial logistics, improve data quality, and facilitate data collection and sharing; and the use of a common protocol that incorporates innovative statistical approaches to study design and data analysis, enabling a broader set of objectives to be met more effectively than would be possible in independent trials," such as the PBTC that will run the MB basket of CX-4945. Recent advances in hidden Markov models and MCMC schemes that we are developing for cryo-EM analysis at Stanford is another example of new opportunities for statistical science in drug discovery. This will be coupled with innovative designs for regulatory submission. It is an exciting interdisciplinary team effort, merging statistical science with other sciences and engineering.

References

Albers, G. W., Marks, M. P., Kemp, S., Christensen, S., Tsai, J. P., Ortega-Gutierrez, S., McTaggart, R. A., Torbey, M. T., Kim-Tenser, M., Leslie-Mazwi, T. et al. (2018),

'Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging', *New England Journal of Medicine* **378**(8), 708–718.

André, F. (2018), 'Developing anticancer drugs in orphan molecular entities—a paradigm under construction', *New England Journal of Medicine* .

Angus, D. C., Alexander, B. M., Berry, S., Buxton, M., Lewis, R., Paoloni, M., Webb, S. A., Arnold, S., Barker, A., Berry, D. et al. (2019), 'Adaptive platform trials: definition, design, conduct and reporting considerations', *Nature Reviews Drug Discovery* **18**(10), 797.

Ascierto, P., Bono, P., Bhatia, S., Melero, I., Nyakas, M., Svane, I., Larkin, J., Gomez-Roca, C., Schadendorf, D., Dummer, R. et al. (2017), 'LBA18 efficacy of BMS-986016, a monoclonal antibody that targets lymphocyte activation gene-3, in combination with nivolumab in pts with melanoma who progressed during prior anti–PD-1/PD-L1 therapy (mel prior io) in all-comer and biomarker-enriched populations', *Annals of Oncology* **28**(suppl_5).

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002), 'Finite-time analysis of the multiarmed bandit problem', *Machine Learning* **47**(2-3), 235–256.

Bartroff, J., Lai, T. L. and Shih, M.-C. (2013), *Sequential Experimentation in Clinical Trials: Design and Analysis*, Springer.

Bell, R. B., Leidner, R. S., Crittenden, M. R., Curti, B. D., Feng, Z., Montler, R., Gough, M. J., Fox, B. A., Weinberg, A. D. and Urba, W. J. (2016), 'OX40 signaling in head and neck squamous cell carcinoma: overcoming immunosuppression in the tumor microenvironment', *Oral Oncology* **52**, 1–10.

Berry, D. (2006), 'Bayesian clinical trials', *Nature Reviews Drug Discovery* **5**(1), 27–36.

Berry, D. (2015), 'The Brave New World of clinical cancer research: adaptive biomarker-driven trials integrating clinical practice with clinical research', *Molecular Oncology* **9**(5), 951–959.

BioPharma Dive (2019), 'Drug development innovations that work: Precision medicine (Part 3 in a series)', https://www.biopharmadive.com/spons/drug-development-innovations-that-work-precision-medicine-part-3-in-a-ser/556187/.

BioPharma Dive (2020*a*), 'Bristol-Myers finds FDA receptive to speedy review of key cell therapy', https://www.biopharmadive.com/news/bristol-myers-liso-cel-car-t-speedy-fda-review/572273/.

BioPharma Dive (2020*b*), 'Merck builds case for Keytruda use in breast cancer', https://www.biopharmadive.com/news/merck-keytruda-breast-cancer-roche-tecentriq/572198/.

Borghaei, H., Paz-Ares, L., Horn, L., Spigel, D. R., Steins, M., Ready, N. E., Chow, L. Q., Vokes, E. E., Felip, E., Holgado, E. et al. (2015), 'Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer', *New England Journal of Medicine* **373**(17), 1627–1639.

Cecchini, M., Rubin, E. H., Blumenthal, G. M., Ayalew, K., Burris, H. A., Russell-Einhorn, M., Dillon, H., Lyerly, H. K., Reaman, G. H., Boerner, S. et al. (2019), 'Challenges with novel clinical trial designs: master protocols', *Clinical Cancer Research* **25**(7), 2049–2057.

Chapman, P. B., Hauschild, A., Robert, C., Haanen, J. B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M. et al. (2011), 'Improved survival with vemurafenib in melanoma with BRAF V600E mutation', *New England Journal of Medicine* **364**(26), 2507–2516.

Chen, C., Li, X., Yuan, S., Antonijevic, Z., Kalamegham, R. and Beckman, R. A. (2016), 'Statistical design and considerations of a phase 3 basket trial for simultaneous investigation of multiple tumor types in one study', *Statistics in Biopharmaceutical Research* **8**(3), 248–257.

Chuang, C.-S. and Lai, T. L. (1998), 'Resampling methods for confidence intervals in group sequential trials', *Biometrika* **85**(2), 317–332.

Chuang, C.-S. and Lai, T. L. (2000), 'Hybrid resampling methods for confidence intervals', *Statistica Sinica* pp. 1–33.

Collins, F. S. and Varmus, H. (2015), 'A new initiative on precision medicine', *New England Journal of Medicine* **372**(9), 793–795.

Conley, B. A. and Doroshow, J. H. (2014), Molecular analysis for therapy choice: NCI MATCH, *in* 'Seminars in oncology', Vol. 41, p. 297.

Cunanan, K. M., Iasonos, A., Shen, R., Begg, C. B. and Gönen, M. (2017), 'An efficient basket trial design', *Statistics in Medicine* **36**(10), 1568–1579.

Curtin, J. A., Fridlyand, J., Kageshita, T., Patel, H. N., Busam, K. J., Kutzner, H., Cho, K.-H., Aiba, S., Bröcker, E.-B., LeBoit, P. E. et al. (2005), 'Distinct sets of genetic alterations in melanoma', *New England Journal of Medicine* **353**(20), 2135–2147.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W. et al. (2002), 'Mutations of the braf gene in human cancer', *Nature* **417**(6892), 949–954.

de BONO, J. S. and Ashworth, A. (2010), 'Translating cancer research into targeted therapeutics', *Nature* **467**(7315), 543–549.

De Roock, W., Claes, B., Bernasconi, D., De Schutter, J., Biesmans, B., Fountzilas, G., Kalogeras, K. T., Kotoula, V., Papamichael, D., Laurent-Puig, P. et al. (2010), 'Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis', *The Lancet Oncology* **11**(8), 753–762.

Do, K., O'Sullivan Coyne, G. and Chen, A. P. (2015), 'An overview of the NCI precision medicine trials—NCI MATCH and MPACT', *Chinese Clinical Oncology* **4**(3).

Drilon, A., Laetsch, T. W., Kummar, S., DuBois, S. G., Lassen, U. N., Demetri, G. D., Nathenson, M., Doebele, R. C., Farago, A. F., Pappo, A. S. et al. (2018), 'Efficacy of larotrectinib in TRK fusion–positive cancers in adults and children', *New England Journal of Medicine* **378**(8), 731–739.

Ersek, J. L., Black, L. J., Thompson, M. A. and Kim, E. S. (2018), 'Implementing precision medicine programs and clinical trials in the community-based oncology practice: barriers and best practices', *American Society of Clinical Oncology Educational Book* **38**, 188–196.

Ersek, J. L., Graff, S. L., Arena, F. P., Denduluri, N. and Kim, E. S. (2019), 'Critical aspects of a sustainable clinical research program in the community-based oncology practice', *American Society of Clinical Oncology Educational Book* **39**, 176–184.

FDA (2018), 'Master protocols: Efficient clinical trial design strategies to expedite development of oncology drugs and biologics. guidance for industry', *Federal Register* .

Fisher, L. D. (1998), 'Self-designing clinical trials', *Statistics in Medicine* **17**(14), 1551–1562.

Garrido, P., Aldaz, A., Vera, R., Calleja, M., de Alava, E., Martin, M., Matias-Guiu, X. and Palacios, J. (2018), 'Proposal for the creation of a national strategy for precision medicine in cancer: a position statement of SEOM, SEAP, and SEFH', *Clinical and Translational Oncology* **20**(4), 443–447.

Gould, L. and Shih, W. J. (1992), 'Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance', *Communications in Statistics-Theory and Methods* **21**(10), 2833–2853.

Heinzerling, L., Ott, P. A., Hodi, F. S., Husain, A. N., Tajmir-Riahi, A., Tawbi, H., Pauschinger, M., Gajewski, T. F., Lipson, E. J. and Luke, J. J. (2016), 'Cardiotoxicity associated with ctla4 and pd1 blocking immunotherapy', *Journal for Immunotherapy of Cancer* **4**(1), 50.

Herson, J. and Wittes, J. (1993), 'The use of interim analysis for sample size adjustment', *Drug Information Journal* **27**(3), 753–760.

Hirsch, B. R., Califf, R. M., Cheng, S. K., Tasneem, A., Horton, J., Chiswell, K., Schulman, K. A., Dilts, D. M. and Abernethy, A. P. (2013), 'Characteristics of oncology clinical trials: insights from a systematic analysis of clinicaltrials. gov', *JAMA Internal Medicine* **173**(11), 972–979.

Hobbs, B. P., Chen, N. and Lee, J. J. (2018), 'Controlled multi-arm platform design using predictive probability', *Statistical Methods in Medical Research* **27**(1), 65–78.

Hodi, F. S., O'Day, S. J., McDermott, D. F., Weber, R. W., Sosman, J. A., Haanen, J. B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J. C. et al. (2010), 'Improved survival with ipilimumab in patients with metastatic melanoma', *New England Journal of Medicine* **363**(8), 711–723.

Hyman, D. M., Puzanov, I., Subbiah, V., Faris, J. E., Chau, I., Blay, J.-Y., Wolf, J., Raje, N. S., Diamond, E. L., Hollebecque, A. et al. (2015), 'Vemurafenib in multiple nonmelanoma cancers with BRAF V600 mutations', *New England Journal of Medicine* **373**(8), 726–736.

Kang, S. P., Gergich, K., Lubiniecki, G. M., de Alwis, D. P., Chen, C., Tice, M. A. and Rubin, E. H. (2017), 'Pembrolizumab KEYNOTE-001: an adaptive study leading to accelerated approval for two indications and a companion diagnostic', *Annals of Oncology* **28**(6), 1388–1398.

Kobilka, B. (2013), 'The structural basis of g-protein-coupled receptor signaling (nobel lecture)', *Angewandte Chemie International Edition* **52**(25), 6380–6388.

Kris, M. G., Johnson, B. E., Berry, L. D., Kwiatkowski, D. J., Iafrate, A. J., Wistuba, I. I., Varella-Garcia, M., Franklin, W. A., Aronson, S. L., Su, P.-F. et al. (2014), 'Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs', *JAMA* **311**(19), 1998–2006.

Lai, T. L., Lavori, P. W. and Liao, O. Y.-W. (2014), 'Adaptive choice of patient subgroup for comparing two treatments', *Contemporary Clinical Trials* **39**(2), 191–200.

Lai, T. L., Lavori, P. W. and Tsang, K. W. (2015), 'Adaptive design of confirmatory trials: Advances and challenges', *Contemporary Clinical Trials* **45**, 93–102.

Lai, T. L., Liao, O. Y.-W. and Kim, D. W. (2013), 'Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research', *Contemporary Clinical Trials* **36**(2), 651–663.

Lai, T. L. and Robbins, H. (1985), 'Asymptotically efficient adaptive allocation rules', *Advances in Applied Mathematics* **6**(1), 4–22.

Lai, T. L., Shih, M.-C. and Zhu, G. (2006), 'Modified Haybittle Peto group sequential designs for testing superiority and non-inferiority hypotheses in clinical trials', *Statistics in Medicine* **25**(7), 1149–1167.

Lal, R. (2019), 'FDA modernizes clinical trials with master protocols', `https://www.fda.gov/drugs/cder-small-business-industry-assistance-sbia/fda-modernizes-clinical-trials-master-protocols-february-26-2019-issue`.

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J. J., Cowey, C. L., Lao, C. D., Schadendorf, D., Dummer, R., Smylie, M., Rutkowski, P. et al. (2015), 'Combined nivolumab and ipilimumab or monotherapy in untreated melanoma', *New England Journal of Medicine* **373**(1), 23–34.

LeBlanc, M., Rankin, C. and Crowley, J. (2009), 'Multiple histology Phase II trials', *Clinical Cancer Research* **15**(13), 4256–4262.

Lefkowitz, R. J. (2013), 'A brief history of g-protein coupled receptors (nobel lecture)', *Angewandte Chemie International Edition* **52**(25), 6366–6378.

Lima, Z. S., Ghadamzadeh, M., Arashloo, F. T., Amjad, G., Ebadi, M. R. and Younesi, L. (2019), 'Recent advances of therapeutic targets based on the molecular signature in breast cancer: genetic mutations and implications for current treatment paradigms', *Journal of Hematology & Oncology* **12**(1), 38.

Lin, J. and Bunn, V. (2017), 'Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials', *Contemporary Clinical Trials* **54**, 48–59.

Lorden, G. (1983), 'Asymptotic efficiency of three-stage hypothesis tests', *The Annals of Statistics* **11**(1), 129–140.

Luke, J. J., Azad, N. S., Edwards, R., Huang, S.-M. A., Comprelli, A., Monga, M., Reilly, T. P. and Hodi, F. S. (2018), 'Phase I, open-label, adaptive biomarker trial that informs

the evolution of combination immuno-oncology therapies, a precision IO approach to personalized medicine.', *Journal of Clinical Oncology* .

Mandrekar, S. J., Dahlberg, S. E. and Simon, R. (2015), 'Improving clinical trial efficiency: thinking outside the box', *American Society of Clinical Oncology Educational Book* **35**(1), e141–e147.

McArthur, G. A., Demetri, G. D., Van Oosterom, A., Heinrich, M. C., Debiec-Rychter, M., Corless, C. L., Nikolova, Z., Dimitrijevic, S. and Fletcher, J. A. (2005), 'Molecular and clinical analysis of locally advanced dermatofibrosarcoma protuberans treated with imatinib: Imatinib target exploration consortium study B2225', *Journal of clinical oncology* **23**(4), 866–873.

Neuenschwander, B., Wandel, S., Roychoudhury, S. and Bailey, S. (2016), 'Robust exchangeability designs for early phase clinical trials with multiple strata', *Pharmaceutical Statistics* **15**(2), 123–134.

NIH (2015), 'What is precision medicine?', `https://ghr.nlm.nih.gov/primer/precisionmedicine/definition`.

Nogueira, R. G., Jadhav, A. P., Haussen, D. C., Bonafe, A., Budzik, R. F., Bhuva, P., Yavagal, D. R., Ribo, M., Cognard, C., Hanel, R. A. et al. (2018), 'Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct', *New England Journal of Medicine* **378**(1), 11–21.

Oiseth, S. J. and Aziz, M. S. (2017), 'Cancer immunotherapy: a brief review of the history, possibilities, and challenges ahead', *Journal of Cancer Metastasis and Treatment* **3**(10), 250–61.

Ott, P. A., Hodi, F. S., Kaufman, H. L., Wigginton, J. M. and Wolchok, J. D. (2017), 'Combination immunotherapy: a road map', *Journal for Immunotherapy of Cancer* **5**(1), 16.

Pardoll, D. M. (2012), 'The blockade of immune checkpoints in cancer immunotherapy', *Nature Reviews Cancer* **12**(4), 252–264.

Park, J. J., Siden, E., Zoratti, M. J., Dron, L., Harari, O., Singer, J., Lester, R. T., Thorlund, K. and Mills, E. J. (2019), 'Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols', *Trials* **20**(1), 1–10.

Pierce, K. L., Premont, R. T. and Lefkowitz, R. J. (2002), 'Seven-transmembrane receptors', *Nature Reviews Molecular Cell Biology* **3**(9), 639–650.

Postow, M. A., Callahan, M. K. and Wolchok, J. D. (2015), 'Immune checkpoint blockade in cancer therapy', *Journal of clinical oncology* **33**(17), 1974.

Proschan, M. A. and Hunsberger, S. A. (1995), 'Designed extension of studies based on conditional power', *Biometrics* pp. 1315–1324.

Redman, M. W. and Allegra, C. J. (2015), The master protocol concept, *in* 'Seminars in oncology', Vol. 42, Elsevier, pp. 724–730.

Renfro, L. and Mandrekar, S. J. (2018), 'Definitions and statistical properties of master protocols for personalized medicine in oncology', *Journal of biopharmaceutical statistics* **28**(2), 217–228.

Renfro, L. and Sargent, D. (2017), 'Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples', *Annals of Oncology* **28**(1), 34–43.

Robbins, H. (1952), 'Some aspects of the sequential design of experiments', *Bulletin of the American Mathematical Society* **58**(5), 527–535.

Saville, B. R. and Berry, S. (2016), 'Efficiencies of platform clinical trials: a vision of the future', *Clinical Trials* **13**(3), 358–366.

Simonsen, K. L., Fracasso, P. M., Bernstein, S. H., Wind-Rotolo, M., Gupta, M., Comprelli, A., Reilly, T. P. and Cassidy, J. (2018), 'The fast real-time assessment of combination therapies in immuno-oncology (FRACTION) program: innovative, high-throughput clinical screening of immunotherapies', *European Journal of Cancer* **103**, 259–266.

Siu, L. L., Gelmon, K., Chu, Q., Pachynski, R., Alese, O., Basciano, P., Walker, J., Mitra, P., Zhu, L., Phillips, P. et al. (2017), 'Abstract ct116: BMS-986205, an optimized indoleamine 2, 3-dioxygenase 1 (ido1) inhibitor, is well tolerated with potent pharmacodynamic (pd) activity, alone and in combination with nivolumab (nivo) in advanced cancers in a Phase I/IIA trial'.

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S. et al. (2014), 'Genetic basis for clinical response to ctla-4 blockade in melanoma', *New England Journal of Medicine* **371**(23), 2189–2199.

Thall, P. F. and Simon, R. (1994), 'Practical Bayesian guidelines for Phase IIB clinical trials', *Biometrics* pp. 337–349.

Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H. and Benjamin, R. S. (2003), 'Hierarchical bayesian approaches to Phase II trials in diseases with multiple subtypes', *Statistics in Medicine* **22**(5), 763–780.

Tsiatis, A. A. and Mehta, C. (2003), 'On the inefficiency of the adaptive design for monitoring clinical trials', *Biometrika* **90**(2), 367–378.

Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. (1984), 'Exact confidence intervals following a group sequential test', *Biometrics* pp. 797–803.

Van Cutsem, E., Kohne, C.-H., Láng, I., Folprecht, G., Nowacki, M. P., Cascinu, S., Shchepotin, I., Maurel, J., Cunningham, D., Tejpar, S. et al. (2011), 'Cetuximab plus irinotecan, fluorouracil, and leucovorin as first-line treatment for metastatic colorectal cancer: updated analysis of overall survival according to tumor KRAS and BRAF mutation status', *Journal of Clinical Oncology* **29**(15), 2011–2019.

Ventz, S., Barry, W. T., Parmigiani, G. and Trippa, L. (2017), 'Bayesian response-adaptive designs for basket trials', *Biometrics* **73**(3), 905–915.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R. et al. (2013), 'The cancer genome atlas pan-cancer analysis project', *Nature Genetics* **45**(10), 1113.

Wen, S., Ning, J., Collins, S. and Berry, D. (2017), 'A response-adaptive design of initial therapy for emergency department patients with heart failure', *Contemporary Clinical Trials* **52**, 46–53.

Wittes, J. and Brittain, E. (1990), 'The role of internal pilot studies in increasing the efficiency of clinical trials', *Statistics in Medicine* **9**(1-2), 65–72.

Woodcock, J. and LaVange, L. M. (2017), 'Master protocols to study multiple therapies, multiple diseases, or both', *New England Journal of Medicine* **377**(1), 62–70.

Xin, Y. J., Hubbard-Lucey, V. M. and Tang, J. (2019), 'Immuno-oncology drug development goes global', *Nature reviews. Drug discovery* **18**(12), 899.