Innovative Designs of Point-of-Care Comparative Effectiveness Trials

Mei-Chiung Shih[a,*], Mintu Turakhia[b], Tze Leung Lai[c,d]

[a] *VA Palo Alto Cooperative Studies Program Coordinating Center, Mountain View, CA, USA*

[b] *Department of Medicine, Stanford University, Stanford, CA, USA*

[c] *Department of Statistics, Stanford University, Stanford, CA, USA*

[d] *Department of Health Research and Policy, Stanford University, Stanford, CA, USA*

Abstract:

One of the provisions of the health care reform legislation in 2010 was for funding pragmatic clinical trials or large observational studies for comparing the effectiveness of different approved medical treatments, involving broadly representative patient populations. After reviewing pragmatic clinical trials and the issues and challenges that have made them just a small fraction of comparative effectiveness research (CER), we focus on a recent development that uses point-of-care (POC) clinical trials to address the issue of "knowledge-action gap" in pragmatic CER trials. We give illustrative examples of POC-CER trials and describe a trial that we are currently planning to compare the effectiveness of newly approved oral anticoagulants. We also develop novel stage-wise designs of information-rich POC-CER trials under competitive budget constraints, by using recent advances in adaptive designs and other statistical methodologies.

Keywords: Adaptive design, Anticoagulants, Comparative effectiveness, Point-of-care trials, Pragmatic trials, Shared decision making

∗ Corresponding author at: VA Palo Alto Cooperative Studies Program Coordinating Center, 701B N Shoreline Boulevard, Mountain View, CA 94043, USA. *Tel*: 1-650-493-5000. *E-mail address*: mei-chiung.shih@va.gov (Mei-Chiung Shih)

1. Introduction

The past five years witness the beginning of a new era in the US health care system, following the health care reform legislation in March 2010. One of the provisions of the Patient Protection and Affordable Care Act (PPACA) is the establishment of a non-profit Patient-Centered Outcome Research Institute (PCORI) to undertake comparative effectiveness research (CER), examining the "relative health outcomes, clinical effectiveness, and appropriateness" of different medical treatments. PCORI provides funding for selected pragmatic clinical trials or large simple trials, or large-scale observational studies, involving broadly representative patient populations for CER. Observational studies are often used to provide data for CER; an example is Stukel et al. [1] that describes statistical analysis of large Medicare claims databases to compare survival rates after medical and surgical treatments for acute cardiovascular disease. The key problem with observational approaches involves 'confounding by indication', the tendency for freely choosing clinicians and patients to choose treatments with their anticipated effects in mind. Careful design of observational studies and adjustments for bias together with sensitivity analysis methods have been developed to mitigate overt biases and address uncertainties about latent biases in observational data; see [2, 3]. A more definitive way to remove these biases is to use randomization, leading to CER clinical trials. However, the cost, complexity and potential lack of impact of CER clinical trials compare unfavorably with the relative ease of observational studies. In Section 2 we give an overview of these large simple trials and the more general pragmatic trials and the issues and challenges that have made them just a small fraction of the totality of CER studies. Lai and Lavori [4] describe three methods, two of which are also reviewed in Section 2, to address these issues.

In Section 3 we focus on the remaining one of the methods, namely using point-of-

care (POC) clinical trials to close the "knowledge-action gap" described in Section 2. In particular, we review recent developments, after the publication of [4] in 2011, in both informatics and methodological advances for POC-CER trials. We also give illustrative examples of these trials. We begin Section 4 by describing one such trial planned at the Department of Veterans Affairs (VA) to compare the effectiveness of three oral anticoagulants that were approved in the US and many other countries in the last five years. Practical issues that arose during planning led us to develop a novel class of stage-wise POC-CER trials in a general framework. The stage-wise designs are information-rich and cost-effective in producing evidence-based answers to questions which evolve sequentially about the treatments. These questions not only arise endogenously during the course of the trial but also exogenously from other studies and the changing landscape of medical knowledge and practice.

As Section 4.1 shows, traditional clinical trial designs for POC-CER trials cannot handle problems of such complexity and yet require very large sample sizes and upfront commitment of a corresponding large amount of funding. Novel designs are therefore needed. Chapter 7 of [5] lists adaptive designs and "using point-of-care clinical trials to create a learning health care system" as two important innovations in clinical trial designs, and discusses their advantages and challenges. The paper [6] in this tenth anniversary issue gives an overview of the major developments and issues in adaptive designs of confirmatory trials to test new treatments in the past decade. Not only does the present paper address the other class of innovations in clinical trial designs, namely POC trials, but more importantly it also modifies some important ideas underlying the advances in adaptive designs described in [6] to resolve the difficulties and circumvent the hurdles currently facing POC-CER trials. As Section 4 shows, a major difference between the adaptive designs

of confirmatory clinical trials to test new treatments and POC trial designs is that the latter do not require blinding as they involve approved treatments and blinding may even be infeasible. The stage-wise designs developed in Section 4 capitalize on their unblinded feature to allow more efficient use of accumulated information during the course of the trial. Section 5 gives further discussion of this approach and some concluding remarks.

2. Overview of pragmatic and large simple trials for CER

2.1 Pragmatic trials as opposed to explanatory trials

About fifty years ago, Schwartz and Lellouch [7] distinguished "pragmatic trials" from clinical trials, called "explanatory trials", that aim at studying treatment effects in the presence of inter-subject variability in response. Whereas explanatory trials are exemplified by Phase I, II and III trials in the development of a new drug to build a clinical data package for regulatory approval of the drug, pragmatic trials involve approved drugs or treatments and aim at answering the question about which treatment should be used in practice. A pragmatic trial, therefore, should be conducted under "real world" conditions, in which blinding to treatment assignment is not required and clinical outcomes are measured directly rather than through surrogate endpoints that are often used to speed up explanatory trials. Hence it is also called a "naturalistic trial".

Large simple trials, which attracted much attention in the 1980s beginning with [8], are basically large pragmatic trials that aim at answering important health care questions, or confirming conclusions from meta-analyses of small trials, or identifying small but still worthwhile improvements in treatment outcomes for common diseases. One such trial was ISIS (International Studies of Infarct Survival), an RCT of IV atenolol versus placebo following myocardial infarction (MI) which involved 16,000 subjects and showed 15% reduction in mortality by day 7 [9]. A subsequent study involved 58,050 subjects from 1086

hospitals and used a $2 \times 2 \times 2$ factorial design to test oral captopril, oral mononitrate, an IV magnesium sulphate in an immediate post-MI period. It found significant reduction in mortality for captopril, but not for the other two treatments [10].

Another large pragmatic trial was ALLHAT (the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial), a randomized, double-blind, multi-center clinical trial sponsored by the National Heart Lung and Blood Institute in conjunction with the VA. It recruited more than 42,000 patients from 623 primary care clinics and its aim was to determine if the combined incidence of fatal coronary heart disease and non-fatal myocardial infarction differs between diuretic (chlorthalidone) treatment and each of three alternative antihypertensive pharmacologic treatments: a calcium antagonist (amlodipine), and ACE inhibitor (lisinopril), and an alpha adrenergic blocker (doxazosin). A lipid-lowering subtrial ($\geq$ 10,000 patients) was designed to determine whether lowering cholesterol with an HMG Co-A reductase inhibitor (pravastatin), in comparison with usual care, reduced mortality in a moderately hypercholesterolemic subset of participants. ALLHAT was the largest antihypertensive trial ever conducted, and the second largest lipid-lowering trial. It recruited many patients over age 65, women, African-Americans and patients with diabetes. The study was conducted between 1994 and 2002 largely in community practice settings. In ALLHAT, hypertensive patients were randomly assigned to receive one of four drugs in a double-blind design, and a limited choice of second-step agents were provided for patients not controlled on first-line medication. Patients were followed every three months for the first year and every four months thereafter for an average of six years of follow-up. This landmark study cost over $100 million, the final results were presented in 2002 [11], and [12] anticipated the results of this pragmatic trial would translate into clinical practice:

> Physicians and policymakers can be confident that the reported outcomes in

> this study are likely to predict results that will be observed across a wide range of practice settings. This eliminates a common obstacle to physician implementation of clinical research findings.

Yet, six years later, *The New York Times* article under the headline THE MINIMAL IMPACT OF A BIG HYPERTENSION STUDY on November 28, 2008 quoted C. Furberg, chair of ALLHAT, as saying "The impact was disappointing." The reasons cited for this "blunted impact" include the difficulty of persuading doctors to change, scientific disagreement about the government's interpretation of the results, and heavy marketing by pharmaceutical companies of their own drugs, paying speakers to "publicly interpret the Allhat results in ways that made their products look better."

*2.2 Equipoise-stratified randomization for CER and the STAR*D trial*

The STAR*D (Sequenced Alternatives to Relieve Depression) trial was a multi-site, prospective, randomized, multi-step clinical trial of outpatients with nonpsychotic major depressive disorder [13]. It compared seven treatment options in patients who did not attain a satisfactory response with citalopram, a selective serotonin reuptake inhibitor antidepressant. After receiving citalopram participants without sufficient symptomatic benefit were eligible for randomization among four switch options (sertraline, bupropion, venlafaxine, cognitive therapy) and three citalopram augment options (bupropion, buspirone, cognitive therapy). It was clear to the study designers that few patients would be willing to be randomized among all seven options, so other design options were considered. One possibility was to randomize patients between two overall strategies: "switch" or "augment", allowing physician choice to determine which specific option would be implemented. This design was not adopted. Instead, it ascertained before randomization the set of options that the patient-clinician dyad considered to be equally reasonable, given the patient's preferences,

and his or her state after a trial of citalopram. This set of options characterizes the patient's Equipoise Stratum (ES). A total of 1429 patients were randomized under this scheme. The largest ES were the "Medication Switch Only" group, allowing randomization among the three medications (40%) and the "Medication Augmentation Only", allowing randomization between two options (29%). The "Any Augmentation" (10%) and "Any Switch" (7%) were the next largest, and only 5% of patients were randomized among options that contrasted a switch and augment condition. The randomization roughly sorted patients (and their clinicians) into two groups: those who obtained partial benefit from citalopram and therefore were interested in augmentation, and those who obtained no benefit and were interested only in switching. Thus, ES enabled the study to "self-design" in assigning patients to the parts of the experiment that were relevant to current practice and to patient preference. Lai and Lavori [4] mention the preceding ES randomization approach as one of the innovative design methods for pragmatic trials comparing the effectiveness of multiple treatments.

2.3 Sequential multiple-assignment randomization trials (SMART)

Another innovative design method for pragmatic trials described in [4] is sequential multiple-assignment randomization (SMAR) for dynamic treatment strategies for the management of patients whose disease has a chronic dimension, neither acutely fatal nor completely curable. These treatment strategies are usually cobbled together from disparate sources of information. For example, the choice of drugs for the initial treatment of newly diagnosed hypertension or bipolar manic disorder may be well-informed by carefully controlled trials conducted as part of the registration process or comparative effectiveness studies. However, the choice of a second-line drug if the first drug fails to bring the hypertension or mania under control may not have such a strong evidence base, and as the

patient experiences multiple failures of disease control, the basis for making such decisions may thin out completely. Furthermore, the best way to start treating the disease may depend on the downstream options. If a highly effective but risky treatment can be deployed successfully as a salvage treatment after failure of a less effective but non-toxic treatment, then it may produce an overall better long-term benefit to use it in that way. But if the progression of disease makes the riskier treatment less effective in the salvage role, the conclusion may be reversed. The SMAR design has been proposed and used in clinical trials to study dynamic treatment strategies [14, 15, 16, 17, 18, 19, 20]. It offers the ability to distinguish between short-term ("myopic") outcome differences (on rates of remission, by initial induction option) and long-term survival differences.

Thall et al [19] describe a two-stage randomized trial of a total of 12 different strategies of first-line and second-line treatments for androgen-independent prostate cancer. After randomization to one of four treatments, patients who responded were continued on the initial treatment, and those who did not were randomized among the other three treatments not assigned initially. The intent of this Phase II SMART design was to select a candidate treatment for evaluation in a Phase III trial. The treatment with the best initial success rate was the combination TEC (weekly paclitaxel, estramjustine, and carboplatin), and the authors proposed that it be taken forward to Phase III testing. Dynamic treatment trials suffer from more frequent missing longitudinal data and dropouts or non-compliance. Additional care is needed to analyze the data and [21, 22] describe methods to do this.

3. Embedded experiments via POC: A new era for CER

3.1 PPACA and closing the "implementation gap" of comparative trials via POC

Chapter 7 of [5] describes POC as an approach to melding (a) a randomized trial,

which "remain(s) the gold standard for determining a treatment's safety and efficacy" but has "high costs and extended timelines" and problematic integration of its results into clinical care, and (b) an observational study that is "less expensive and produce(s) quicker results" which are not reliable enough to guide clinical care. The essence of POC is to use "randomization to remove selection bias in an observational study." Its long-term goal is "to create a true learning health care system." It introduces a new relationship between clinical care and research to address the current difficulty in which "there are too many research questions, too few investigators, and too little funding" in clinical effectiveness research. Since "clinical care dollars, being spent in any case, can generate the data," a health care system that can learn from its data and experiences can "make taking care of patients a whole lot quicker, more effective, and probably cheaper."

Thus POC can be regarded as an innovation in pragmatic trials to bridge the gap, such as in the ALLHAT trial of Section 2.1, between knowledge generated from a comparative clinical trial and actions taken in clinical care, which is called the "implementation gap" in discussions of CER and evidence-based medicine. Lai and Lavori [4] describe a POC trial as an "embedded clinical trial" that can bring the benefits of knowledge generated from the trial to "improve health care without having to mount a separate implementation strategy." They suggest to use it to improve experimental design for CER which, as mentioned in Section 1, plays an important role in the new US health care system following the the health care reform act, PPACA, in 2010, and the 2009 economic stimulus package that allocated over one billion dollars to CER studies. The high cost of medical care has led to an urgent interest in weeding out costly, ineffective medical care, and POC-CER introduced in [23] is a timely innovation to study comparative effectiveness via pragmatic clinical trials. Two years before the publication of [23], Luce et al [24] argued for transformational

change in randomized clinical trials as they are "the most rigorous method of generating comparative effectiveness evidence and will necessarily occupy a central role in an expanded national CER agenda," but "as currently designed and conducted, are ill-suited to meet the evidentiary needs implicit in the IOM definition of CER."

3.2 POCs at the VA

As pointed out in [23], planning for a trial comparing two standard regimens (sliding scale versus weight based) of insulin administration for hospitalized diabetic (hyperglycemia) patients at the VA led the authors to the development of POC-CER. The VA already had data in an electronic medical record (EMR) that includes electronic ordering of medications and protocols for both of these insulin regimens. Review of EMR data at the VA Boston Healthcare System demonstrated that each of these two regimens is used with approximately equal frequency and discussions with treating clinicians indicated that choice of the administration method is based on personal preference and not on patient-specific determinants. There were no published data comparing the effectiveness or the adverse effects of the sliding scale or a weight-based insulin protocol in treating patients with hyperglycemia.

The trial, which is currently ongoing, is a multi-site, open-label, randomized clinical trial comparing sliding scale regular insulin to a weight-based regimen for control of hyperglycemia in non-ICU hospitalized patients. The primary outcome is length of hospital stay up to 30 days, which has important cost implications and may be shortened if diabetic control can be made more efficient. Secondary outcomes include degree of glycemic control during the hospitalization and readmission within 30 days of discharge for glycemic control. The goals of the trial are to answer the clinical question and to test the feasibility of POC trials. All non-ICU patients who require in-hospital sliding scale or weight-based

insulin therapy are eligible for the trial. VA's computerized patient record system (CPRS) is used to present the possibility of randomization to the clinician at the point of care of an eligible patient. The decision to obtain informed consent from the patient is made by the clinician at the time of an insulin order. Patients who provide consent are randomized through CPRS to one of the two insulin regimens, and followed until 30 days after randomization. Data on outcomes and covariates are collected directly from CPRS. The details of implementing the clinical trial processes in CPRS, including the randomization procedure, can be found in [25]. The summary on statistical methods in [23, p.183] says:

> Adaptive randomization will assign up to 3000 patients, preferentially to the currently 'winning' strategy, and all care will proceed according to usual practices. Based on a Bayesian stopping rule, the study has acceptable frequentist operating characteristics (Type I error 6%, power 86%) against a 12% reduction of median length of stay from 5 to 4.4 days.

Details of the adaptive randomization, given in [23, pp.188-190], are summarized in Section 5.1 which also gives a discussion of the Bayesian method and provides a simpler alternative. The trial modifies the assignment probability to either regimen of insulin administration each time the study accrues a batch of additional patients, with batch sizes of at least 100.

This trial is a pilot study of POC implementation in clinical trials at the VA, whose information systems (e.g., CPRS) are key to the efficiency and scalability of POC research. Chapter 7 of [5] mentions that informatics is already available at the VA to position it for POC research:

> Because VA patients' electronic records are available at any VA facility nationwide, additional opportunities to participate in trials present themselves. . ..

> Mining the EHR (electronic health record) data allows that patients to be identified and facilitates the patient's engagement in the trial. . .. Another potential benefit of point-of-care trials would be to create a culture change in the way clinicians and patients think about treatment trials. If doctors and patients want . . . to provide and receive evidence-based medicine, then they need to be part of the evidence-gathering process (facilitated by the informatics).

The VA Cooperative Studies Program's Clinical Trial Center at MAVERIC (Massachusetts Veterans Epidemiology Research and Information Center) has been developing state-of-the-art informatics tools not only for POC trials but also for the broader goal of creating a learning health care system within the VA.

The experience gained from planning and conducting this POC trial on insulin administration regimens has already led to other POC trials being planned at the VA. One such trial that has received funding aimed at determining whether chlorthalidone is more effective than hydrochlorothiazide at preventing cardiovascular outcomes in veterans over age 65 with hypertension (ClinicalTrials.gov Identifier: NCT02185417). Both medications are thiazide-type diuretics that have been used for more than 50 years and are considered first-line treatment for hypertension. Patients currently prescribed hydrochlorothiazide will be randomized to either continue taking hydrochlorothiazide or to receive chlorthalidone. All patient care, including the study drug, will continue to be managed by the primary care provider. Study operations will be conducted centrally and patient data will be collected passively through VA and non-VA databases. The primary outcome is time to a major cardiovascular event, defined as a composite outcome comprised of the first occurrence after randomization of any of the following: stroke, myocardial infarction, urgent coronary revascularization because of unstable angina, hospitalization for acute congestive heart

failure, non-cancer death. Secondary outcomes include time to event for each component of the composite primary outcome and additional cardiovascular events. The study will enroll up to 13,500 veterans over 3 years and follow them on average for 3 years, resulting in a total study duration of 4.5 years.

4. Novel stage-wise designs of information-rich POC-CER trials under competitive budget constraints

This section begins with our recent experience in planning a POC trial that compares the effectiveness of newly approved target-specific oral anticoagulants (TSOACs) in stroke prevention in patients with atrial fibrillation. The scientific background is described in Section 4.1. Issues with the large sample size of the design led us to consider modifications, which are described in Section 4.2. The first modification we tried was changing the original design in Section 4.1 to a group sequential design, which can use information acquired during the course of the trial to substantially reduce the actual sample size. We then tried also to enhance the scope of the trial by using recent advances in adaptive designs [26] that can incorporate mid-course modification of sample size and early stopping for futility, group sequential partial likelihood testing, adaptive randomization that allows elimination and addition of treatments, and personalized treatment choice based on baseline patient characteristics. Applying these advances led us to settle down on a much more flexible design that could even adapt to new scenarios which might emerge during the course of the trial. Then funding and drug adherence issues, which are described in Section 4.4, arose and led us to refine the adaptive design further into a stage-wise POC trial. Section 4.3 describes a general CER framework in which a stage-wise POC trial proceeds in stages, publishes results on the endpoints to be addressed for the stage, makes a go/no go decision, and adapts the design for the next stage to the information collected so far. Although this

may appear to be basically a group sequential or adaptive design, an important innovation in the stage-wise POC trial is that it takes advantage of the unblinded nature of the trial to break it into stages whose findings on the stage-wise endpoints can be made pubic after each stage. Funding for the trial can therefore also be broken into stages. This is particularly valuable in view of the total cost of the typically large POC trial. The funding agency can also adapt to the stage-wise results in deciding how much more to fund, and these results may also suggest alternative funding agencies. In Section 4.4 we return to the POC-CER trial of TSOACs to illustrate this idea.

4.1 CER of TSOACs: Study objectives and initial design of a POC trial

Atrial fibrillation and atrial flutter (AF, collectively) are abnormal heart rhythms of the upper chamber of the heart. AF increases the risk of stroke. The loss of coordinated electromechanical atrial activity predisposes to impaired atrial emptying, stasis of blood, and a prothrombotic state [27]. These factors cause blood clots to form in the heart and embolize systemically to cause stroke or other organ failure [28, 29], which is the major cause of AF-related morbidity and mortality. Among AF patients, the annual incidence of stroke is 4.5%. The estimated annual direct and indirect cost of stroke in the US is $57.9 billion. Until recently, warfarin has remained the mainstay of anticoagulation therapy. Safety and effectiveness of warfarin is directly attributable to quality of anticoagulation, as determined by the time within the therapeutic INR range of 2.0-3.0 (TTR). In the past five years, four new TSOACs have received approval in the US and many other countries for patients with AF [30, 31, 32, 33]. These fixed-dose agents, which inhibit thrombin (Factor II) or Factor Xa, are administered once or twice daily and unlike warfarin, do not require routine laboratory testing. Although the primary outcome of time to stroke or systemic embolism was nearly identical in the pivotal trials for these agents, there were

considerable differences in inclusion and exclusion criteria. For example, patients in the trial [31] required prior stroke or higher stroke risk (CHADS2 score ≥ 3) for enrollment. All three pivotal trials showed at least noninferiority compared to warfarin for prevention of stroke and systemic embolism. All studies were superior to warfarin for reduction in intracranial hemorrhage and fatal bleeding. Despite these findings, there exists considerable uncertainty as to optimal first-line strategy in veteran and non-veteran patients, and this has led to reluctance and restraint in national VA policy to direct anticoagulation. There are also increasing concerns about the effectiveness and safety of the new TSOACs outside of the trials. This motivated planning a POC-CER trial of the TSOACs at the VA, including all patients with AF and with an indication for anticoagulation as potential study participants, who can be identified through the VA's EMR either at the entry of new patients or at outpatient visits of patients already on oral anticoagulation.

At the time when the trial was first planned, only three TSOACs had been approved: dabigatran, rivaroxaban, and apixaban. It was designed to mimic clinical practice as closely as possible in order to inform evidence-based choices at the VA. The primary objective of the trial is to determine the comparative effectiveness of dabigatran, rivaroxaban and apixaban in veterans with VA-managed anticoagulation, and the primary effectiveness outcome is the composite of stroke and systemic embolism, intracranial hemorrhage and death. The secondary objectives are to determine the comparative safety, with fatal bleeding as the primary safety outcome, and cost effectiveness of the three anticoagulants. The sample size calculation in the trial design was based on the pairwise comparisons among the TSOACs, with study participants randomized to one of the three TSOACs using a permuted block randomization scheme that is stratified by site. Based on results from the pivotal studies and the VA's EMR, the two-year primary event rate was anticipated to be 15% for dabiga-

tran and apixaban and 13% for rivaroxaban. The corresponding number of patients needed to treat per year to prevent one primary event is about 100, which is lower than the typical cutoff of 100-150, and therefore this 2% difference is considered clinically meaningful. To have 80% power to detect a 2% difference in the two-year event rate (15% vs 13%, hazard ratio 0.86) between two treatment groups at a significance level 5%/3 (due to Bonferroni correction for 3 tests), 1763 events are needed. Assuming 3 years of recruitment and adding 1.5 years of follow-up after the last randomized participant (total study duration 4.5 years, median follow-up 3 years), the number of patients needed is 4381 per group. Rounding this number up to 4400 led to a total sample size of 4400×3=13,200 patients to be enrolled over 3 years.

4.2 Group sequential modification, adaptive randomization and an innovative POC design

The sample size determined above for the pragmatic trial followed the usual framework for sample size calculations in the trial design of these trials. However, the assumptions on the event rates and effect sizes made at the planning stage may be found to differ substantially from the observed data during the course of the trial. A related example is the TASTE (Thrombus Aspiration during ST-segment Elevation) myocardial infarction trial [34]. On the basis of mortality data from patients with ST-segment elevation myocardial infarction who underwent percutaneous coronary intervention (PCI) in Sweden between 2008 and 2009, the trial design assumed that the one-month mortality with PCI would be 6.3%. Under this assumption it was calculated that 456 events would have to occur for the trial to have 80% power to detect a hazard ratio of at least 1.3 with PCI as compared with PCI plus thrombus aspiration, leading to a sample size of 5000. When enrollment approached 5000 patients, the 30-day mortality (estimated without knowledge of treatment assignments) was observed to be substantially lower (2.9%) in the study cohort. This led

the Steering Committee of the trial to amend the protocol by increasing the sample size to 7138 patients and to adopt a group sequential design for which interim analysis was conducted by a data monitoring committee. This experience of TASTE suggests that the protocol of the trial should allow the possibility of mid-course modification of the sample size or adaptive design. Clearly there are limits on the maximum sample size because of time and resource constraints. An advantage of a group sequential or adaptive design is that it allows early stopping for futility if an interim analysis shows little chance of a significant result when the trial is continued to its maximum sample size.

A possible reason for the relatively small treatment differences among the anticoagulants is the heterogeneity of the patient population. These differences may be magnified in certain patient classes. Lai, Liao and Kim [26] have recently introduced a novel group sequential design to develop and test biomarker-guided personalized therapies involving approved cancer treatments. The design can enhance substantially the trial's findings by fulfilling multiple objectives, which include (a) treating accrued patients in the trial with the best (yet unknown) available treatment, (b) developing a treatment strategy for future patients, and (c) demonstrating that the strategy developed indeed has a better treatment effect than that of the standard of care, or that of any of the approved treatments. In a group sequential trial, sequential decisions are made only at times of interim analysis. Equal randomization is applied to the treatments up to the first interim analysis. Then an adaptive randomization scheme is used, assigning the highest randomization probability to the leading treatment in each biomarker class. In addition, generalized likelihood ratio statistics are used for early elimination of significantly inferior treatments from a biomarker class, with the elimination threshold so chosen that there is a guaranteed probability of $1-\alpha$ that the best treatment for each biomarker class is not eliminated, where α corresponds to

the type I error. To accomplish this, [26] uses subset selection ideas from the selection and ranking literature, in which selecting a subset of treatments, with a guaranteed probability of at least $1 - \alpha$ that it contains the best treatment.

For the stroke prevention trial in AF patients, in place of the biomarkers for cancer patients considered in [26], there are two important patient characteristics that may affect treatment choice. The first is presence or absence of renal dysfunction, as some of the TSOACs are eliminated mainly through the renal system and may therefore accumulate more and be less safe in patients with chronic kidney disease. The second is risk of stroke: low to moderate risk ($1 \leq$ CHADS2 score ≤ 2) versus moderate to high risk (CHADS2 score ≥ 3). In this connection, it should be noted that the pivotal trial of rivaroxaban only enrolled patients with moderate to high risk but the anticoagulant's labeling does not have this restriction. The adaptive randomization scheme in [26], which allows arm elimination that is tantamount to assigning zero randomization probability to an eliminated arm, can also be modified to allow inclusion of a new arm in the trial after its initiation. At the time when the POC-CER trial of TSOACs was planned, there was a fourth TSOAC, edoxaban, for which a pivotal trial was near completion. Anticipating that edoxaban might have gained approval at some time of some interim analysis of the trial, the adaptive randomization scheme in [26] could be modified to assign to the additional treatment the same randomization probability as that of the leading treatment until the next interim analysis, when this treatment would have been administered to a reasonable number of patients to assess its efficacy in comparison with other treatments. This anticipation turned out to be a reality and edoxaban received FDA approval after completion of the pivotal trial [33]. It is an example of the questions and issues referred to in the last paragraph of Section 1, that can arise exogenously from other studies.

4.3 Budget constraints and a novel stage-wise design

A major concern of the VA Cooperative Studies Program with launching a large pragmatic trial of the scope described in Section 4.1 is the cost. Even though the trial is embedded into treatment of the patients when it is POC, there are still high costs of data collection and analysis of the large information-rich trial ("informatics costs") and the "opportunity cost" of forgoing possibly more valuable trials that compete for the same (or part of the) large pool of clinical centers, doctors and patients of the trial. However, it is virtually impossible for a funding agency to determine the relative values of the outcomes of competing trials before these large pragmatic trials that are still being considered for funding have any results or even accrual and execution data. A rational way of allocating resources in this case is to put budget constraints on several promising candidates and proceed with the trials under these constraints to generate some data that can help estimate their relative values, so that the "winners" will receive continued funding and the "losers" are suspended. The I-SPY2 trial reviewed in Section 2.3 of [6] has a similar philosophy although it is for Phase II testing of new breast cancer therapies.

Blinding is one of the major costs of confirmatory clinical trials in drug development, and removing blinding leads to large cost savings for POC trials of comparative effectiveness of approved treatments. There is also another important advantage of unblinding that POC-CER studies can capitalize on to circumvent the difficulties mentioned in the preceding paragraph. Because the data are unblinded, it is convenient to analyze the data continuously or at prespecified times of interim analysis of the data accumulated during the course of the trial, with multifaceted analysis targeted towards to multiple objectives listed in the penultimate paragraph of Section 4.2. More importantly, we can break the trial into stages, with specific goals and research outputs at each stage. Although the over-

arching objectives of the entire trial remain the same, the stages can provide data from the stage-wise design for the funding agency to assess not only the progress of the stages towards these overarching objectives, which may still take considerable time and resources to complete, but also the value added to clinical practice from the trial's stage-wise results that have been funded. Basically the stage-wise design functions as a group sequential design with the adaptive features of Section 4.2 insofar as the entire trial and its overarching objectives are concerned, but also separates itself into stages with their own specific questions that can be addressed at the end of a stage. Each stage can therefore also receive additional funding from other agencies that are interested in the stage's specific questions. An illustrative example is given in the next section.

4.4 Short versus sustained anticoagulation clinic management of TSOAC

We now return to the POC trial of comparative effectiveness of TSOACs described in Sections 4.1 and 4.2. Besides FDA's approval of edoxaban described in the last paragraph of Section 4.2, another external event that required our modification of the original trial design was the *Guidance for the Oversight and Monitoring* of TSOACs, issued by VA's PBM (Pharmacy and Benefits Management) Office in February 2014, suggesting that all VA specialized anticoagulation clinics should manage new initiatives of TSOACs for a minimum of three months (http://www.pbm.va.gov). In addition, the American Heart Association, American College of Cardiology, and Heart Rhythm Society also pointed out in their new AF management guidelines the importance of increasing patients' adherence to anticoagulation therapy. Because of the short half-life of TSOACs compared to warfarin, safety and effectiveness of TSOACs may be sensitive to even small deviations in adherence. Data from the VA Health Care System in VA users with AF prescribed dabigatran between October 1, 2010, and September 30, 2012 showed that in 5,376 patients

(age 71±10 years; CHADS2VASC 3.2±1) with a median follow-up of 244 days, only 72% of patients were noted to be adherent; the same cohort also showed that lower adherence was associated with increased risk for combined all-cause mortality and stroke [35]. Therefore, although TSOACs are typically more convenient for patients and were designed to obviate the need for patient monitoring, adherence remains an issue and there are opportunities to use sustained clinic management of TSOACs to improve patient adherence [35, 36]. The anticoagulation clinic management interventions include lab monitoring together with patient and caregiver education at baseline and periodic follow-up, and monitoring patients for adherence, refilling medications on time, and tolerance of medication.

A principal barrier to sustained clinic management is financial reimbursement. Currently anticoagulation clinic and care providers receive reimbursement for management of warfarin, but there is no national coverage determination for TSOAC management. Therefore VA's PBM and Anticoagulation Workgroup have recently expressed that a new priority is to clarify whether sustained anticoagulation clinic management of TSOACs can improve adherence. We can modify the group sequential/adaptive design of Section 4.2 into a stage-wise design, in which the first stage has this high-priority question as its main finding. The basic idea is to add an extra layer of randomization to the sites, which are registered anticoagulation clinics, so that participating sites are randomized to short-term (3-month) versus sustained (12-month) clinic management for all patients prescribed a TSOAC. Adherence is measured by the proportion of patients on TSOACs managed by that site with appropriate adherence, which is defined as a medication possession ratio of $\geq 80\%$. At the end of the first stage, which corresponds to the first interim analysis of the overall group sequential design, a major statistical analysis to be reported is whether sustained clinic management significantly improves the short-term management at initiation.

If that is the case, the finding is expected to promote changes in reimbursement policy across payers and lead to widespread adoption of sustained clinic management. The future stages of the group sequential design will use sustained clinic management and combined the results with those under sustained clinic management in the first stage. If no significant improvement is found, then short-term clinic management will be used instead and combined with the results under short-term management in the first stage.

5. Discussion

5.1. Integrating advances in statistical methodologies into POC-CER

POC-CER has inherent complexities, which can result in prohibitively large sample sizes if one treats them by standard methods. For example, it typically involves multiple sites (either for accrual or to increase the generalizability of findings) and a relatively permissive implementation to accommodate local site variations in the delivery of care. A case in point is in extending the original VA POC trial that compares two different insulin dosing regimens from the first site to the others because of the variety of ways that different VA health care centers deliver diabetic care. The prescribing physician can be a trainee working under a chief resident during a short-term rotation at one site, or can be a hospitalist with a permanent position at another site. Moreover, patient populations differ across sites, in race, social class, income, and rate of co-morbid conditions, even within the VA system. Readers of the trials literature often question whether the average effects studied in clinical trials accurately reflect the possible heterogeneity of true effects across subpopulations. One way to deal with such concerns is to divide the sites into classes and study CER locally within each class. This, however, would substantially increase the already large sample size. Another way is to use a random effects model to account for site

effects. The endpoint in that study is length of stay (LOS) at hospital up to 30 days, which is a time to event endpoint. The adaptive randomization scheme referred to in Section 3.2 is based on the Bayesian posterior probability that one regimen has a shorter median LOS than the other given the accumulated data, assuming exponentially distributed LOS truncated at 30 days and a conjugate prior (inverse Gamma) distribution for the median LOS. Including random effects would make the computation of the posterior probability substantially more complicated although it can be carried out by Markov chain Monte Carlo (MCMC).

The model of exponentially distributed survival times for the two dosing regimens is a special case of the proportional hazards model, under which the hazard ratio is time-invariant and is widely used to compare the two survival distributions. Although it is conventional and innocuous to assume the exponential model at the planning stage (when there is insufficient information on the actual survival distributions) for determining the sample size and study duration to attain some prescribed power, data collected during the course of the trial may show substantial departures from the simple model assumed. Hence a more convincing and robust analysis is to use the logrank statistic that is the efficient score statistic of the proportional hazards model, in which random effects can also be readily incorporated [37]. As explained in [6] and [38], the Bayesian approach does not have type I error guarantees even after "frequentist twists" of the type used in [23] and [39]. Section 4.1 of [6] points out that mainstream statistical methods have well-established theories, efficiency properties, and implementation details/software and that many of the advanced techniques already provide powerful tools for developing efficient and flexible adaptive designs that also have valid type I errors or confidence levels. Although Bayesian methods are part of mainstream statistics, "so are parametric, semiparametric, and nonparametric

(empirical) likelihood methods." The more complex case of time-to-event endpoints is discussed in [38] which also reviews semiparametric/nonparametric and Bayesian (MCMC-based) survival analysis in mainstream statistics. The data-driven adaptive designs of POC-CER trials that we propose to get around the prohibitively large sample sizes under standard trial designs can use these advances in statistical methodologies together with some refinements and modifications, as shown in Section 4.2 and in the closely related paper [40].

LOS is often used as an outcome measure for pragmatic trials. Besides the VA insulin study, another well-known example is the COMPANION (Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure) trial [41]. The trial involved a total of 1520 patients who had advanced heart failure due to ischemic or non-ischemic cardiomyopathies and a QRS interval of at least 120 msec. These patients were randomly assigned in a 1:2:2 ratio to receive optimal pharmacologic therapy alone or in combination with cardiac-resynchronization therapy with either a pacemaker or a pacemaker-defibrillator. The primary endpoint was time to death from or hospitalization for any cause. The main finding was that the hazard ratio for the risk of this combined endpoint was 0.81 for the pacemaker group (P=0.015, logrank test) and 0.80 for the pacemaker-defibrillator group (P=0.010, logrank test). In a subsequent analysis of these data, Anand et al. [42] noted that an assessment of the true reduction in hospitalization risk should also take into account varying follow-up times, leading to the consideration of average number of hospitalization days per patient-year of follow-up, average length of stay per hospital admission, and recurrence of hospitalization. Moreover, because comparison of hospitalization risks between treatment groups "must consider the competing risk of death and varying follow-up times," [42] used the nonparametric analysis of recurrent events and the competing risk of mortality

introduced by Ghosh and Lin [43].

5.2. Shared decision making in pragmatic trials

The short versus sustained anticoagulation clinic management of TSOACs in Section 4.4 only represents one of the exogenous questions concerning improving patient-centered outcomes of anticoagulation treatments. Another closely related question is shared decision making. Seaburg et al. [44] pointed out last year the importance of Shared Decision Making (SDM) in treating AF patients:

> AF accounts for more than a third of all hospitalization for cardiac rhythm disturbances. Hospitalization for AF has risen dramatically over the past 20 years and projected to rise as population ages. ... Studies have demonstrated significant gaps in AF patients' knowledge about their condition, as well as knowledge of the risks and benefits of the treatment they are currently taking for their AF despite their disease being treated for years. ... The Institute of Medicine included patient-centered care as 1 of 6 key quality domains. One of the most important attributes of patient-centered care is active patient participation in the decision-making process. SDM, described as the pinnacle of patient-centered care, is characterized by patient and clinician partnership, ... joint deliberation considering the pros and cons of each option, and agreement about which treatment to implement.

They mentioned development of decision aids as an important direction for SDM because these aids provide "statistical information based on current evidence" to "help clinicians and their patients deliberate jointly." Decision aids for AF treatment "often use validated scoring systems to assist in the estimation of risk," with CHADS2 and CHA2DS2-VASc

scores as well-known examples in AF treatment. Beginning in 2014, Circulation has a new series on SDM, and it is anticipated that new advances and decision aids may appear in the near future that can be added to the stage-wise design, for their testing and possible inclusion in the pragmatic trial. Ting, Brito and Montori [45] and Lin and Fagerlin [46] discuss how to measure the quality of SDM.

Acknowledgment

This research was supported in part by the Clinical and Translational Science Award 1UL1 RR025744 for the Stanford Center for Clinical and Translational Education and Research (Spectrum) from the National Center for Research Resources, National Institutes of Health, and by the National Science Foundation (T.L. Lai) and the US Department of Veterans Affairs Cooperative Studies Program (M.-C. Shih).

References

[1] Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vemeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management of AMI survival using propensity score and instrumental variable methods. J Amer Med Assoc. 2007;297:278–285.

[2] Rubin DB. Estimating causal effects of treatment in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.

[3] Rubin DB. Bayesian inference for causal effects: the role of randomization. Ann Stat. 1978;6:34–58.

[4] Lai TL, Lavori PW. Innovative clinical trial designs: Toward a 21st-century health care system. Statistics in Biosciences. 2011;3:145–168.

[5] Institute of Medicine. Public Engagement and Clinical Trials: New Models and Disruptive Technologies. Washington DC: National Academics Press; 2012.

[6] Lai TL, Lavori PW, Tsang KW. Adaptive design of confirmatory trials: Advances and challenges. Contemp Clin Trials. 2015;this issue.

[7] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chron Dis. 1967;20(8):637–648.

[8] Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials. Stat Med. 1984;3(4):409–422.

[9] ISIS-1 (First International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous atenolol among 16027 cases of suspected acute myocardial infarction. Lancet. 1986;2:57–66.

[10] ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. Lancet. 1995;345:669–685.

[11] ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). J Amer Med Assoc. 2002;288(23):2891–2997.

[12] Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: Increasing the value of clinical research for decision making in clinical and health policy. J Amer Med Assoc. 2003;290(12):1624–1632.

[13] Rush A, Fava M, Wisniewski S, Lavori P, Trivedi M, Sackeim H, et al. Sequenced treatment alternatives to relieve depression (STAR*D): rantionale and design. Contr Clin Trial. 2004;25:119–142.

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

[15] Murphy S, van der Laan M, Robins J. Marginal mean models for dynamic regimes. J Amer Statist Assoc. 2001;96(456):1410–1423.

[16] Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2002;58(1):48–57.

[17] Lavori PW, Dawson R. Dynamic treatment regimes: Practical design considerations. Clin Trials. 2004;1(1):9–20.

[18] Murphy S. An experimental design for the development of adaptive treatment strategies. Stat Med. 2005;24(10):1455–1481.

[19] Thall PF, C L, Pagliaro LC, Wen S, Brown MA, Williams D, et al. Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. J Nat Cancer Inst. 2007;99(21):1613–1622.

[20] Wolbers M, Helterbrand JD. Two-stage randomization designs in drug development. Stat Med. 2008;27(21):4161–4174.

[21] Bembom O, van der Laan MJ. Statistical methods for analyzing sequentially randomized trials. J Nat Cancer Inst. 2007;99(21):1577–1582.

[22] Wang L, Rotnitzky A, Lin X, Millikan RE, Thall PF. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. J Am Stat Assoc. 2012;107(498):493–508.

[23] Fiore LD, Brophy M, Ferguson RE, D'Avolio L, Hermos JA, Lew RA, et al. A point-of-care clinical trial comparing insulin administered using a sliding scale versus a weight-based regimen. Clin Trials. 2011;8(2):183–195.

[24] Luce BR, Kramer JM, Goodman SN, Connor JT, Tunis S, Whicher D, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. Ann Intern Med. 2009;151(3):206–209.

[25] D'Avolio L, Ferguson R, Goryachev S, Woods P, Sabin T, O'Neil J, et al. Implementation of the Department of Veterans Affairs' first point-of-care clinical trial. J Amer Med Informat Assoc. 2011;19:e170–e176.

[26] Lai TL, Liao OYW, Kim DW. Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. Contemp Clin Trials. 2013;36(2):651–663.

[27] Rockson S, Albers G. Comparing the guidelines: Anticoagulation therapy to optimize stroke prevention in patients with atrial fibrillation. J Am Coll Cardiol. 2004;43(6):929–935.

[28] Wolf P, Benjamin E, Belanger A, Kannel W, Levy D, D'Agostino R. Secular

trends in the prevalence of atrial fibrillation: The Framingham study. Am Heart J. 1996;131(4):790–795.

[29] Hart R, Sherman D, Easton J, Cairns J. Prevention of stroke in patients with nonvalvular atrial fibrillation. Neurology. 1998;51(3):674–681.

[30] Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, et al. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med. 2009;361(12):1139–1151.

[31] Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med. 2011;365(10):883–891.

[32] Granger CB, Alexander JH, McMurray JJV, Lopes RD, Hylek EM, Hanna M, et al. Apixaban versus warfarin in patients with atrial fibrillation. N Engl J Med. 2011;365(11):981–992.

[33] Guigliano RP, Ruff CT, Braunwald E, Murphy SA, Wiviott SD, Halperin JL, et al. Edoxaban versus warfarin in patients with atrial fibrillation. N Engl J Med. 2013;369(22):2093–2104.

[34] Frobert O, Lagerqvist B, Olivecrona GK, Omerovic E, Gudnason T, Maeng M, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med. 2013;369(17):1587–1597.

[35] Shore S, Carey EP, Turakhia MP, Jackevicius CA, Cunningham F, Pilote L, et al. Adherence to dabigatran therapy and longitudinal patient outcomes: Insights from the Veterans Health Administration. Am Heart J. 2014;167:810–817.

[36] Shore S, Ho PM, Lambert-Kerzner A, Glorioso TJ, Carey EP, Cunningham F, et al. Site-level variation in and practices associated with dabigatran adherence. J Amer Med Assoc. 2015;313(14):1–8.

[37] Pankratz VS, de Andrade M, Themeau. Random-effects Cox proporitonal hazards model: general variance components methods for time-to-eent data. Genet Epidemiol. 2005;28(2):97–109.

[38] He P, Lai TL, Zheng S. Design of clinical trials with failure-time endpoints and interim analysis: An update after fifteen years. Contemp Clin Trials. 2015;this issue.

[39] Berry DA. Bayesian clinical trials. Nature Rev Durg Disc. 2006;5:27–36.

[40] Shih MC, Lavori PW. Sequential methods for comparative effectiveness experiments: Point of care clinical trials. Statistica Sinica. 2013;23(4):1775–1791.

[41] Bristow MR, Saxon LA, Boehmer J, Krueger S, Kass DA, De Marco T, et al. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. N Engl J Med. 2004;350(21):2140–2150.

[42] Anand IS, Carson P, Galle E, Song R, Boehmer J, Ghali JK, et al. Cardiac resynchronization therapy reduces the risk of hospitalization in patients with advanced heart failure: results from the Comparison of Medical Therapy, Pacing and Defibrillation in Heart Failure (COMPANION) trial. Circulation. 2009;119(7):969–977.

[43] Ghosh D, Lin DY. Nonparametric analysis of recurrent events and death. Biometrics. 2000;56(2):554–562.

[44] Seaburg L, Hess EP, Coylewright M, Ting HH, McLeod CJ, Montori VM. Shared

decision making in atrial fibrillation: where we are and where we should be going. Circulation. 2014;129(6):704–710.

[45] Ting HH, Brito JP, Montori VM. Shared decision making: science and action. Circ Cardiovasc Qual Outcomes. 2014;7(2):323–327.

[46] Lin GA, Fagerlin A. Shared decision making: state of the science. Circ Cardiovasc Qual Outcomes. 2014;7(2):328–334.