SMART Design, Conduct, and Analysis in Oncology

Peter F. Thall
Department of Biostatistics
The University of Texas, M.D. Anderson Cancer Center
Houston, Texas, USA

May 21, 2014

1 Introduction

This chapter is based on my experiences as a biostatistician working with oncologists and statisticians in clinical trial design and data analysis. I will focus on sequentially adaptive decision regimes used routinely by physicians over multiple stages of a patient's therapy, known variously as treatment policies, multi-stage adaptive treatment regimes, dynamic treatment regimes (DTRs), or simply "regimes." A useful review of DTRs is given by Moodie, Richardson and Stephens (2007). Since this book is about sequential, multiple assignment, randomized trials, or SMARTs, (Murphy, 2005; Murphy et al., 2007; Lavori and Dawson, 2007), which aim to evaluate and compare DTRs in an unbiased fashion, this chapter will consist of my opinions followed by descriptions of two SMARTs conducted at M.D. Anderson Cancer Center (MDACC). The first is a completed trial in advanced prostate cancer, and the second is an ongoing trial of DTRs constructed from targeted agents for metastatic kidney cancer. The basic idea underlying both designs was to randomize each patient at enrollment and re-randomize the patient to a different treatment if and when his/her initial (frontline) treatment fails. This was motivated by the recognition that, because DTRs formalize what oncologists actually do, it is a good idea to evaluate and compare the regimes rather than only the treatments given at a particular stage of the regime, as is done conventionally. This idea originated from Randy Millikan, a clinical oncologist with whom it was my privilege to collaborate. In retrospect, the designs for these trials turned out to be both smarter and stupider than we realized when we constructed them.

In a utopian medical practice, each physician would tailor each patient's treatment based on the patient's prognostic covariates to ensure that therapeutic success is certain. A very popular idea is that this may be accomplished using genetic, protein, or other biological markers, to choose an optimal treatment for each patient. This is the fantasy called "personalized medicine" that has been promulgated widely in recent years. Of course, in the real world such perfect knowledge and perfect treatments seldom exist. Physicians know this, and to treat patients in the real world they have been practicing real personalized medicine for thousands of years. Real personalized medicine begins with recognition of the fact that a given treatment typically works in some patients and not in others. One practical way for a physician to deal with imperfect knowledge in a stochastic world is to use each patient as their own control and proceed sequentially. A general algorithm for doing this, used routinely by practicing physicians, is "Try something. If it works, give it again until the disease is cured or you can't give it any more. If it doesn't work, try something else." One may call this the "Repeat a winner and switch away from a loser (RWSL)" rule. The underlying idea is that whether a given treatment succeeds in a given patient is due to in large part to random variation, between-patient heterogeneity, and possibly treatment-prognostic covariate interactions that are not fully understood. RWSL rules typically use established conventional prognostic covariates, such as disease subtype and severity, whether the patient has been treated previously for their disease, age, and performance status.

Because physicians proceed sequentially, cancer therapy and many other areas of medical practice routinely involve multiple stages, with adaptive decisions made by the physician in each stage based on the patients latest history of treatments and outcomes. The RWSL rule is an example of a more general form of physician behavior: "Observe → act → observe → act . . . until some criterion for stopping therapy is met." Here, "act" can mean any sort of therapeutic action that a physician may take based on the patient's most recent history. A wide variety of algorithms that may be described generally by such an alternating sequence of observations and adaptive actions are used widely by physicians to treating many forms of cancer, infections, drug addiction, alcoholism, high blood pressure, or other chronic diseases. The formalism for the sequence of adaptive actions is a DTR. A regime typically takes the more specific form "Evaluate the patient's baseline covariates to diagnose the disease, make an initial treatment decision, treat the patient, evaluate the patient's outcomes and possibly updated covariates, make a second treatment decision, etc." The distinction between the physician choosing a treatment and actually using it to treat the patient is important because things may not go as planned, including patient noncompliance or drop-out, a pharmacy giving the wrong drug or dose, or a delay in treatment administration for logistical reasons.

2 Dynamic Treatment Regimes in Oncology

The main anti-cancer treatment modalities are chemotherapy (chemo), radiation therapy (RT), surgery, stem cell transplantation, and immunotherapy. Each modality carries its own risks of particular regimen-related AEs. For many cancers, a patient's regime may include two or more modalities. In general, patient outcome at any stage of cancer therapy often is complex and multivariate, typically including both desirable and undesirable treatment effects that have non-trivial probabilities of occurrence. Even for a single stage, this complicates treatment choice and outcome evaluation, and consideration of trade-offs or risk-benefit ratios between efficacy and AEs are inherent in therapeutic decision making. All of these issues are embedded in the more complex problem of multi-stage decision-making when treating a cancer patient. In oncology, actions may include choosing a treatment, modifying the dose or schedule of a treatment given in a previous stage, suspending treatment due to a regimen-related adverse event (AE), typically called "toxicity," or terminating therapy because success has been achieved, the patient is unable or unwilling to receive further treatment, or it is considered futile to continue.

Conventional oncology trial designs intentionally reduce variables by focusing on only one stage of therapy and ignoring its multi-stage structure. Unfortunately, this produces results that are of limited use to practicing oncologists. Two or more treatments given sequentially may have effects that are not obvious if the treatments given at each stage are evaluated separately. A very common practice is to compare frontline treatments in terms of overall survival while ignoring adaptive treatment decisions made by the physicians based on what is seen after frontline. Subsequent treatments may include use of consolidation chemo, salvage treatment, or modifications of dose or administration schedule to deal with interim toxicity. Ignoring such components of the actual regime easily may lead to misleading conclusions, essentially because overall effects of multi-stage regimes often are non-intuitive. A simple but important illustration is that very aggressive frontline chemo may maximize the stage 1 response rate, but if it fails then the patient patient's immune system and overall health are so compromised that any stage 2 chemo must be given at a reduced dose and so is unlikely to achieve a response. As a toy illustration, suppose that the goal is to achieve a response in one or two stages of therapy, with stage 2 given only if the stage 1 outcome is not a response. For treatments $\{a, b, c\}$, and response indicators (y_1, y_2) in the two stages, denote $\pi_{a_1,1} = \Pr(y_1 = 1|a_1 \text{ in stage 1})$ and $\pi_{a_2,2}(a_1) = \Pr(y_2{=}1 \mid y_1 = 0 \text{ with } a_1 \text{ in stage 1, } a_2 \text{ in stage 2})$. If $\pi_{a,1} = .60$ and $\pi_{b,1} = .50$, but $\pi_{c,2}(a) = .10$ and $\pi_{c,2}(b) = .50$, this implies that the 2-stage strategy (a, c) has success probability $.60 + .40 \times .10 = .64$ while the 2-stage strategy (b, c) has success probability $.50 + .50 \times .50 = .75$. So a is better than b in stage 1, but (b, c) is a better 2-stage strategy than (a, c). Unfortunately, many oncologists make the error of starting with a treatment like a.

In some settings, SMART designs cannot be applied. For example, so-called "trimodality" therapy for esophageal cancer may or may not begin with induction chemo to debulk the disease, then continue with a combination of a different chemo plus RT (chemoradiation therapy, CRT), after which a surgeon decides whether surgery is feasible based on the patient's CRT outcomes and, if surgery is performed, one of several different procedures is chosen. In radiation therapy for esophageal cancer at MDACC, the CRT sub-regime usually is constructed from 3 possible chemos, 3 possible RT modalities, and 2 different radiation fields. If performed, surgery may be of 7 different types. In this setting, in theory there are $2 \times 3 \times 3 \times 2 \times (7{+}1) = 288$ possible regimes, although in practice only about 80 (28%) of these possibilities actually have been used. While this may seem like a good setting to conduct a SMART, algorithms for each treatment decision are well established and it is considered either unethical or infeasible to randomize among the choices at each stage. In Section 4, I will discuss two oncology trials where SMART designs were considered appropriate. To make any progress comparing the effects of the DTRs in the trimodality setting, methods that correct for bias in observational data must be applied, as was done in the following example.

In conventional evaluation of anti-cancer treatments, most published analyses focus on frontline treatments and, when evaluating overall survival or progression-free survival (PFS) time, ignore effects of salvage therapies given when frontline treatments fail. A typical example is Estey, et al. (1999), who gave results of a randomized trial of four chemotherapy combinations for acute leukemia, and concluded that there were no significant differences between the treatment arms. As done conventionally, the analyses in Estey et al. compared the frontline chemos while ignoring non-randomized salvage treatments given if the patient had disease resistant to frontline chemo or that progressed after an initial remission was achieved. Wahed and Thall (2013) re-analyzed this dataset accounting for salvage therapies and identifying 16 possible multistage regimes including both frontline and salvage. These analyses included both inverse probability of treatment weighting (IPTW)(Robins and Rotnitzky, 1992; Murphy, van der Laan and Robins, 2001; Wahed and Tsiatis, 2004) and G-estimation (Robins, 1986; Robins, Hernan, and Brumback, 2000) to correct for bias. This re-analysis estimated mean overall survival time for each regime, and reached very different conclusions. If this trial were conducted today, ideally, the design of choice would include re-randomization at salvage, i.e. it would be a SMART.

In principle, many of the ideas discussed here are applicable to studies outside oncology, such as trials of therapeutic regimes for substance abuse, behavioral disorders, or other chronic diseases. It should be kept in mind, however, that behavioral intervention trials are very different from oncology trials, essentially because cancer therapy typically is aggressive and not infrequently carries the risk of severe and possibly irreversible AEs, including regimen-related death.

3 Why Use SMART Designs?

3.1 Some Opinions on Trial Design

Since there is more to being smart about clinical trials than using SMART designs, it is worthwhile to provide a more general framework for trial design. A clinical trial is a medical experiment with human subjects. Its two purposes are to treat the patients in the trial, and to obtain useful information about treatment effects that may benefit future patients. A good design should do a reasonable job of serving both goals, despite the fact that they may be at odds with each other. To achieve this, both medical and statistical thinking must be applied carefully while constructing a trial design. This process should begin with the statistician(s) determining key elements from the physician(s), including the disease(s), trial entry criteria, treatment(s) and/or doses or treatment combinations to be evaluated, any existing standard treatment(s) that the patients would receive if they were not enrolled in the trial (which I call the "Compared to what?" question), administration schedule(s), within-patient multi-stage adaptive rules (since DTRs are ubiquitous, and should be identified), a range of anticipated accrual rates, a range of feasible sample sizes, costs, regulatory issues, and human resources. In my experience, the physician-statistician conversation may lead the physician(s) to rethink and modify some aspect of the therapeutic process, and it also may motivate the statistician(s) to develop a new design methodology.

Most clinical trials are inherently comparative, whether the protocol design is framed that way or not. This is true even for simple single-arm phase 2A "activity" trials (Gehan, 1961; Thall and Sung, 1998) in settings where no effective standard treatment exists, and the question is whether giving the experimental therapy is better than doing nothing. The real scientific goals of a clinical trial are exploration, estimation, treatment refinement, and possibly modification of future physician behavior. The use of frequentist hypothesis testing as a framework to construct trial designs often obfuscates these goals, and often leads to erroneous conclusions. Flaws with frequentist hypothesis testing include :

1) sample size computations based on numerically artificial alternatives, often with little or no attention to practical significance of the alternatives,

2) ignoring the uncertainty of estimates of key parameters used to construct a design,

3) rejection of a null hypothesis being wrongly interpreted as acceptance of a prespecified alternative hypothesis (Ratain and Karrison, 2007),

4) incorrect interpretation of p-values as probabilities of some type of error, and

5) incorrect use of p-values to quantify strength of evidence.

Useful discussions are given by Berger and Sellke (1987), McClosky (1995), and Ioannidis (2005). In contrast, Bayesian methods, such as the use of Bayes Factors (Jeffrys, 1961; Kass and Raftery, 1995), for dealing with multiple testing rely solely on the assignment of prior probabilities to models or hypotheses, and use the observed data rather than hypothetical data. A useful paper for practitioners is Westfall, Johnson, and Utts (1997). The potentially crippling effects of reliance on frequentist methods for testing multiple hypotheses are especially troubling with SMARTs, since in many settings it is very easy to generate quite a large number of regimes that should, and possibly can, be evaluated.

Statisticians often talk about “optimal designs.” Methodological research to define and derive optimal designs can be quite useful if it leads to good designs that actually can be applied. Any claim of optimality almost invariably is misleading, however, unless it is qualified by a careful explanation of the particular criterion being used to determine what is best. In the real world, no clinical trial can ever be globally optimal, because the utilities of physicians, administrators, government agencies, pharmaceutical companies, patients enrolled in the trial, and future patients are all different. The practical goal of a clinical trial is not optimality, but rather to do a good job of treating the patients and producing data of sufficient quality that, when analyzed sensibly, may benefit future patients. When designing a clinical trial, never let the perfect be the enemy of the good. The two overarching questions in constructing a clinical trial design are whether it serves the medical needs of the patients enrolled in the trial and whether it will turn out to have been worthwhile once it is completed.

Interactions between physicians and statisticians are only part of a complex process involving medicine, statistics, computing, ethics, regulatory issues, finances, logistics, and politics. At the institutional level, elaborate administrative processes often must be followed for protocol review that involve one or more Institutional Review Boards. A major logistical issue is that trial conduct can be complicated by interim outcome-adaptive rules that must be applied in real time. A clinical trial protocol, no matter how detailed, is an idealized representation of how the trial actually will play out. One can never know in advance precisely how new medical treatments or regimes will act in humans. For example, it may be necessary to suspend accrual and modify a design in mid-trial if unexpected AEs occur, the accrual rate is much higher or much lower than expected, or results of another ongoing or recently completed trial substantively change the original rationale for the trial design.

3.2 Some Opinions On SMART Design

SMART designs are a bold attempt to do a better job of evaluating what physicians actually do. They are motivated by two key elements. The first is the multi-stage nature of actual medical therapy. The second is the scientific goal to obtain unbiased comparisons. Combining these two elements motivates randomizing in order to compare the regimes in an unbiased fashion, and more specifically randomizing at more than one stage of the regime. Each randomization must be ethically acceptable in that, at that stage of the regime, the treatments or actions among which the patient is being randomized given their current history must be equally desirable. This criterion, applied at each stage, is the same as the usual requirement of equipoise in conventional randomized trials. If it is decided to evaluate multistage regimes rather than individual treatments by using a SMART design, it is essential to begin by determining the actual regimes that will be studied. This should include the key consideration that all regimes that are possible in the SMART design must be viable (cf. Wang, et al. 2012), that is, each regime must be a sequence of actions that the physicians actually would take.

To determine trial sample size, the first step is to elicit the anticipated accrual rate, which often may be range of values, the longest individual patient follow up time, and the maximum trial duration that the investigators planning the trial consider feasible. Simple back-of-an-envelope arithmetic then can determine a range of feasible sample sizes. This exercise may motivate either reducing the complexity of the design if a simpler feasible trial still is worthwhile, or concluding that a multi-center trial will be needed to accrue enough patients to obtain useful results. More formal sample size computation methods can be applied, including those of Feng and Wahed (2009), Dawson and Lavori (2010), and Li and Murphy (2011). The sample size computation method of Almirall et al. (2012) is quite easy to implement, and is tailored for SMARTs that aim to be pilot studies, which is likely to be the actual reality in many SMARTs. In any case, the design should be simulated on the computer, for each of a range of sample sizes under each of a reasonable set of possible scenarios, to determine the design's operating characteristics. Computer simulation results typically are extremely informative, provide a basis for calibrating design parameters, may motivate design modifications, and are an ethical necessity for complex trials. In my opinion, it is more desirable to kill computer generated patients, rather than real ones, in order to calibrate design parameters. A simple practical rule is to avoid any design that specifies a trial that never will be run because it is not feasible, or that is unlikely to be completed for one or more practical or political reasons.

Since the goals of a SMART include unbiased or approximately unbiased estimation, this provides a basis for reliably ranking the DTRs, which in turn may facilitate elimination of DTRs likely to be inferior and identification of DTRs likely to be superior. This sort of inference is very useful to the physicians or health professionals conducting the trial. A nice property of the Bayesian framework is that it allows one to compute posteriors of ranks. In a SMART, suppose that a total of m DTRs indexed by $j = 1, \cdots, m$ are studied, and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_m)$ denotes the vector of means of the final outcome. For example, if the outcome is survival time, then a larger value of μ_j corresponds to the j^{th} regime being more desirable. In this case, one may define the rank of μ_j as $R_j(\boldsymbol{\mu}) = \sum_{r=1}^{m} I(\mu_j \leq \mu_r)$, so $R_j(\boldsymbol{\mu})$ = 1 corresponds to the best regime, $R_j(\boldsymbol{\mu}) = 2$ to the second best, and so on. The posterior $p(\boldsymbol{\mu} \mid data)$ induces a posterior $p(R_1(\boldsymbol{\mu}), \cdots, R_m(\boldsymbol{\mu}) \mid data)$ on the ranks. In practice, these posteriors are computed easily using Markov chain Monte Carlo methods (Gilks, Richardson and Spiegelhgalter, 1996). A useful property of the posterior on the ranks is that, while for each $\boldsymbol{\mu}$ the vector $(R_1(\boldsymbol{\mu}), \cdots, R_m(\boldsymbol{\mu}))$ is a re-arrangement of the integers (1,...,m), the support of the marginal posterior of each $R_j(\boldsymbol{\mu})$ is not restricted to these integers but rather has support on the interval [0, m], thus quantifying posterior uncertainty about the rank of the j^{th} regime.

It is not an accident that the community of statisticians currently promoting SMART designs arose from the larger community of statisticians developing and using methods to correct for bias when analyzing observational data. It is well known that bias correction methods such as IPTW or case matching essentially attempt to construct data that are as close as possible to what would have been obtained if a randomized clinical trial (RCT) had been conducted. One rationale for randomization to compare treatments (or regimes) a and b is that it gives, in expectation, what would be achieved if there were two copies of each patient so that one could be treated with a and the other with b, with the difference $y_a - y_b$ in outcomes the so-called causal effect of a versus b, and the mean of these differences across the sample of counterfactual pairs the desired estimate. Still, randomization is not a perfect solution to the problem of obtaining unbiased comparisons simply because the world is imperfect. Those involved in the design and conduct of a SMART trial should recognize that it is hard to do real-time adaptive decision-making reasonably, much less optimally. Practical complications include patient heterogeneity, delayed outcomes, errors in recording or entering the outcomes or covariates used for interim adaptive decisions into a database, patients not treated as assigned due to error or physician decision, patient non-compliance, and informative drop-outs. All of these complications typically are associated with treatment, making conventional “intention to treat” analysis substantively misleading. The consequence of all this is that clinical trial data, from a RCT or a SMART, quite often resemble observational data. Consequently, in analyzing data from RCTs, in particular from SMARTs, it often is necessary to employ bias correction methods originally developed for analysis of observational data. An example of this is given in the following section.

4 A Trial in Advanced Prostate Cancer

4.1 The Trial Design

Randy Millikan and I designed the first SMART in oncology. The goal was to evaluate four chemotherapy combinations (chemos), denoted by $\mathcal{A} = \{CVD, KA/VE, TEC, TEE\}$, for advanced prostate cancer. The trial was conducted at M.D. Anderson Cancer Center (MDACC) from December 1998 to January 2006 and enrolled 150 men. The first account of the trial design was given by Thall, Millikan, and Sung (2000). Various analyses of the trial data, including descriptions of the design and a wide variety of statistical methods for analyzing the data, are given by Thall et al. (2007), Bembom and van der Laan (2007), Millikan et al. (2008), Wang et al. (2012), and by the discussants of this last paper, Almirall, Lizotte and Murphy (2012), and Chaffee and van der Laan (2012).

Rather than simply conducting a conventional four-arm RCT, we designed the prostate cancer trial trial to evaluate multi-stage treatment regimes, each constructed using chemos from $\mathcal{A}$, that mimic the way that oncologists who treat prostate cancer behave. To do this, Randy Millikan defined a RWSL algorithm, and he chose the four chemos in $\mathcal{A}$. The algorithm was defined as follows. At enrollment, each patient's disease and prostate specific antigen (PSA) level were evaluated to obtain baseline values, and the patient was randomized fairly among the four chemos. The patient's disease and PSA level were re-evaluated at the end of each of up to four successive eight-week treatment stages. A key distinction was made between success for the chemo administered in a given stage, called "per-stage success," and overall success of the entire multi-stage regime. For a chemo used for the first time in any stage k = 1, 2, or 3, initial per-stage success ($y_k = 1$) was defined as a drop of at least 40% in PSA compared to baseline and absence of advanced of disease (AD). If this occurred, then that chemo was repeated for that patient in the next stage; if not, then that patient's chemo in the next stage was chosen by re-randomizing fairly among the three chemos not given initially to that patient. Success in stage k = 2, 3, or 4 following a success in stage $k-1$ with the same chemo was defined as a drop of at least 80% in PSA level compared to baseline and absence of AD. A maximum of two stages without per-stage success were allowed, and the patient's therapy was terminated if this occurred. Overall success was defined as two consecutive successful stages which, per the algorithm, could only be achieved using the same chemo in both stages. The oncologists in the Genitourinary Medical Oncology Department at MDACC dubbed this trial "The Scramble."

Formally, $a_1 \in \mathcal{A}$, was chosen by fair randomization for all patients. For k = 2, 3, or 4, if $y_{k-1} = 1$ then $a_k = a_{k-1}$ with probability 1, and if $y_{k-1} = 0$ then $a_k \neq a_{k-1} = a_1$, with a_k chosen by fair re-randomization among the three chemos in $\mathcal{A} - a_1$. Because a maximum of two stages with failures were allowed, each patient received 2, 3, or 4 stages of chemo. This algorithm, applied using the chemo set $\mathcal{A}$, produced 12 possible two-chemo regimes, each represented by a pair (a, b) where $a, b \in \mathcal{A}$ with $a \neq b$. In the parlance of oncology, a was the patient's frontline chemo and b was the salvage chemo given if a failed. The primary goal of the trial was to evaluate and compare the 12 possible two-chemo regimes in terms of their overall success rates. This goal was very different from that of a conventional trial,

Table 1: Possible per-stage and overall outcomes with regime (a, b) in the prostate cancer trial. S = overall success = two consecutive successful stages, and F = overall failure = S^c.

Per-stage Outcomes	Chemos	Overall Outcome	Number of stages
(y_1, y_2) = (1,1)	$a_1 = a_2 = a$	S	2
(y_1, y_2, y_3) = (0,1,1)	$a_1 = a, a_2 = a_3 = b$	S	3
(y_1, y_2, y_3, y_4) = (1,0,1,1)	$a_1 = a_2 = a, a_3 = a_4 = b$	S	4
(y_1, y_2) = (0,0)	$a_1 = a, a_2 = b$	F	2
(y_1, y_2, y_3) = (1,0,0)	$a_1 = a_2 = a, a_3 = b$	F	3
(y_1, y_2, y_3) = (0, 1,0)	$a_1 = a, a_2 = a_3 = b$	F	3
(y_1, y_2, y_3, y_4) = (1,0,1,0)	$a_1 = a_2 = a, a_3 = a_4 = b$	F	4

which would be to evaluate and compare only the chemos given initially, in stage 1. The trial was considered to be hypothesis generating, with the aim to use the results as a basis for planning a future, confirmatory phase III trial. The seven possible outcomes generated by the algorithm are summarized in Table 1.

4.2 The First Round of Analyses

For our first analysis of the data from this trial in 2007, nine years after we began the process by establishing the design and starting the trial, Dr. Millikan insisted that we use the model and method given in our initial 2000 paper. At that time, our plan was to apply more sophisticated methods, in particular to correct for bias and informative dropouts, in a later analysis to be done in collaboration with Xihong Lin. While our paper describing these initial analyses was under review at *Journal of the National Cancer Institute*, Oliver Bembom contacted Dr. Millkan and asked him to provide the trial data. Dr. Millikan complied, providing the data as requested, and a paper by Bembom and van der Laan, focusing on the importance of using inverse probability of treatment weighting (IPTW) methods for data analysis, appeared in the same issue of this journal as our paper (Bembom and van der Laan, 2007; Thall, et al. 2007).

A rather different reaction was given by Armstrong, et al. (2007a), who actually waited for us to publish our results before writing their letter. Armstrong et al. (2007a) cited results of the so-celled "TAX327" study (Tannock, et al., 2004), which concluded that docetaxel + prednisone was superior to mitoxantrone + prednisone in terms of overall survival for treating men with advanced prostate cancer. Armstrong et al. (2007a) criticized us for not including a "docetaxel single-agent comparator," described our therapeutic approach as "aggressive and toxic," and provided several other interesting opinions, including criticism of our use of change in PSA to characterize outcome along with AD. To respond to these criticisms, we proceeded empirically by using estimated effects of docetaxel + prednisone on survival. We first obtained the fitted survival time regression model derived in the analysis of the TAX327 study data given by Armstrong, et al. (2007b). Using the actual covariates of the patients in The Scramble with this fitted prognostic model, we computed covariate-specific estimates of how long each of our patients would have been expected to survive if he had received docetaxel + prednisone every 3 weeks, as in the superior arm of the TAX327 study. We then compared the resulting predicted survival curve associated with this hypothetical treatment to the Kaplan-Meier estimate based on the actual survival time data of our patients. The two curves are given in Figure 1, which is reproduced from Millikan, Logothetis and Thall (2008). This figure appears to indicate that, amazingly, the patients in The Scramble survived much longer than they would have survived if they had been treated with docetaxel + prednisone. This survival time comparison was graphical, and furthermore it was informal in that we made no correction for potential bias due to between-trial effects or other possible confounding variables, nor did we perform a comparative test of hypothesis. Although survival time was not the primary endpoint of The Scramble, it seems that the trial's RWSL algorithm, applied with the four chemos noted earlier, provided greatly improved survival for the patients enrolled in The Scramble compared to what would have been obtained with docetaxel + prednisone if they had been enrolled in that arm of the TAX327 study. A possible alternative explanation is that the oncologists and supportive care at MDACC were superior compared to the corresponding elements of the TAX327 study, although this seems dubious given the high level of communication between oncologists at large medial centers. This graphical comparison, while crude, appears to provide an empirical illustration of how a SMART design can benefit the patients enrolled in the trial, at least in the setting of treatment for advanced prostate cancer. A more formal comparison would use the combined data from both trials and apply some form of IPTW, G-estimation, matching, or other bias correction method. To my knowledge, such a formal comparative analysis of these two particular data sets has not yet been carried out.

4.3 The Second Round of Analyses

The first analyses of The Scramble, described above, were based on the 150 eligible patients who were randomized. However, 47 (31%) of these patients did not complete the multistage regime as specified by the protocol algorithm. Both Bembom and van der Laan (2007) and Thall et al. (2007) classified these 47 patients as dropouts, assumed that dropout was noninformative, and carried out a complete case analysis. Given this background, Xihong Lin, Lu Wang, and Andrea Rotnitzky ("The Harvard Gang") and I decided to re-analyze the data, this time accounting for the possibility that these dropouts were informative. To start, Dr. Lin asked me to ask Dr. Millikan the specific reason for each dropout. My subsequent conversation with Dr. Millikan turned out to be quite important, since it led to a process that actually introduced new information into the data set and motivated the use of a utility function. This was because it turned out that, for 35 of the 47 patients whom we had considered to be dropouts, in fact their therapies were stopped by their attending physicians due to either progressive disease (PD) or severe toxicity. This adaptive decision rule was not included in the multi-stage algorithm given in the protocol because, as Dr. Millikan explained to me, it is such standard clinical practice that it did not seem worth formalizing. After a very heated and highly productive discussion, we agreed to account for these adaptive decision rules by defining the per-stage outcome in terms of both efficacy and toxicity. This yielded a much more accurate and more informative bivariate ordinal

Table 2: Elicited utilities of the seven actual possible per-stage outcomes in the prostate cancer trial. Toxicity = 0 for no severe toxicity, 1 for toxicity that precludes further therapy but allows efficacy to be evaluated, and 2 for toxicity precluding further therapy and not allowing efficacy to be evaluated. Efficacy = 0 for favorable response, 1 for no favorable response but no PD, 2 for PD, and 3 for inevaluable efficacy due to PD.

		Per-Stage Efficacy			
		0	1	2	3
Per-Stage Toxicity	0	1.0	0.5	0.1	–
	1	0.8	0.3	0	–
	2	–	–	–	0

per-stage outcome $y_k = (y_{k,T}, y_{k,E})$, for k = 1,2,3,4. We defined $y_{k,T} = 0$ if there was no toxicity, 1 if toxicity occurred at a level severe enough to preclude further therapy but allow efficacy to be evaluated in that stage, and 2 if toxicity occurred at a level severe enough to preclude further therapy and not allow efficacy to be evaluated in that stage. For efficacy, we defined $y_{k,E} = 0$ if a favorable per-protocol response was observed, 1 if a favorable per-protocol response was not observed but PD did not occur, 2 if PD occurred, and 3 if efficacy was inevaluable due to severe toxicity. Dr. Millikan revisited the data and went through the painstaking process of determining y_k for each stage of each patient. In fact, only seven of the 12 possible combinations $(y_{k,T}, y_{k,E})$ could occur, and I elicited joint utilities $U(y_k)$ for each of these seven possibilities from Dr. Millikan. These are given in Table 2.

The results of our analyses of this extended version of the data set from The Scramble are described by Wang et al. (2012), which includes detailed accounts of the newly extended DTRs, which we dubbed "viable switch rules," IPTW methods, inferences for different numerical versions of the utility function, descriptions of counterfactual outcomes, and an analysis of the 12 remaining informative dropouts. While Wang et al. (2012) called $U(y_k)$ a "scoring function" in order to avoid potentially controversial connotations of the word "utility," in fact $U(y_k)$ is a utility function, and it is very similar to those used by Thall et al. (2011, Table 1) and Thall and Nguyen (2012, Table 1) for sequentially adaptive dose-finding. One notable event in this process was that, after I inveigled Dr. Millikan to travel to Crystal City to give an invited talk at an ENAR meeting, he and Dr. Rotnitzky met in person and were able to discuss several key scientific issues. While initially it was our intention to analyze only the per-stage outcomes, Dr. Rotnitzky insisted that we also perform a survival analysis of the regimes, which gave a much more complete picture of the actual viable regime effects. In their discussion of this paper, Almirall, Lizotte and Murphy (2012) explained how The Scramble fit into the world of SMARTs, and provided a very useful sensitivity analysis of the utility functions. Chaffee and van der Laan (CVDL, 2012), in their discussion, reiterated the well-known fact that, because the regimes in a SMART are specified by design, when there are few dropouts IPTW can improve efficiency by weighting for the conditional probabilities of dropout. Of course, this is why Wang et al. (2012) used IPTW. CVDL also argued that targeted maximum likelihood ought to be used for SMARTs with high rates of ignorable dropouts. While assuming ignorability of dropouts in any trial unavoidably is controversial, it seems pretty clear that, in most cancer trials, because each patient's life is at stake it is difficult to believe that any dropout is truly ignorable.

It is useful to summarize what we actually learned from this 14-year process. First, The Scramble illustrates the fact that one cannot design an experiment optimally until after it already has been carried out. The second point is that, because dropouts and other deviations from protocol designs occur quite commonly, analysis of data from a well-designed randomized trial often is very similar to analysis of observational data. Third, my recommendation to a medical statistician who wants to do a good job of study design or data analysis is that they should talk to their doctor. This often is not one conversation, but rather is a process than may play out over many months or years, during which the statistician must learn about specific medical practices and the physician must learn about statistical methods. Not infrequently, this process results in improvements of both paradigms. Fourth, what initially appear to be dropouts in a clinical trial actually may be patients whose treatment was stopped by a physician applying an adaptive decision rule as part of routine medical practice. Such a rule, once recognized and made well-defined, should be operationalized as part of the DTR in the statistical paradigm. Fifth, if you want to do a better job analyzing a complicated data set that you do not fully understand, find people who are smarter than you and listen carefully to what they say. This is why I enlisted The Harvard Gang for this challenging project. Finally, lest we become enamored with the beauty and clarity of new statistical methods and the scientific process, it should be kept in mind that prostate cancer has not yet been cured. Perhaps it is time for statisticians to devote more attention to treatment discovery and refinement, rather than only study design and data analysis.

5 A Trial in Metastatic Kidney Cancer

Given the successful completion of the prostate cancer trial, the GU Oncology Department at MDACC embraced the idea of evaluating DTRs, rather than only individual treatments. This motivated a SMART in metastatic kidney cancer, called Sequential Two-agent Assessment in Renal Cell Carcinoma Therapy, "START." This was activated in 2010, and Nizar Tannir is the trial's PI. A detailed account of the design is given in Thall, et al. (2007). This trial was motivated by the lack any truly effective treatment for metastatic renal cell cancer, which has a median PFS of nine months.

The START trial evaluates six two-stage DTRs constructed from the three targeted agents pazopanib (p), bevacizumab (b), and everolimus (e). Both p and b are VEGF pathway inhibitors, which means that they are designed to block the tumor's blood supply, while e inhibits the rapamycin (mTOR) pathway, which is a central regulator of cell metabolism, growth, proliferation, and survival. Initially, each patient is randomized among {p, b, e} using the Pocock-Simon (1975) method to balance on two prognostic covariates, an indicator of whether the patient received prior cytokine or vaccine treatment, and a three-level risk category variable. The stage 1 outcome is time to overall treatment failure, measured from the date of randomization to the date of disease progression (worsening compared to baseline at first randomization), discontinuation of protocol treatment for any reason, or death. The stage 1 outcome is the final failure if a patient drops out or dies. Only patients whose stage 1 outcome is disease progression receive a stage 2 treatment, which is chosen by re-randomizing the patient fairly between the two agents not received initially by that patient. The six two-stage regimes are thus $\mathcal{D} = \{(p,b),\ (p,e),\ (b,e),\ (b,p),\ (e,b),\ (e,p)\}$. Denote the stage 1 time to progression by t_1 and observed failure time by t_1^o with $\delta_1 = I(t_1^o = t_1)$, and define (t_2^o, t_2, δ_2) similarly for stage 2. The per-stage outcomes are $y_k = (t_k^o, \delta_k)$, for $k = 1, 2$, and overall time-to-failure is $y = t_1^o + \delta_1 t_2^o$. The goal is to estimate $\mu_d = \mathrm{E}(y \mid d)$ for each $d \in \mathcal{D}$ and select the best regime on that basis. An important point is that, if the patient's therapy ends with discontinuation or death at t_1^o in stage 1, then the stage 1 agent contributes to the estimates of μ_d for two regimes. For example if a patient is randomized to p and drops out or dies, then $\delta_1 = 0$ and t_1^o contributes to the estimates of both $\mu_{(p,b)}$ and $\mu_{(p,a)}$.

An important property of this design is that each agent appears as either the first or second element of four of the six regimes. This is attractive for a pharmaceutical company making a given agent a since, for example, if a more conventional three-arm RCT were conducted based on stage 1 only, then a would be given to only 33% of the patients, rather than 66%. More generally, with DTRs of this form, for each individual agent the number of opportunities to be part of a winning sequence is larger than the corresponding number if the agent is considered alone. A key point is that accounting for the six regimes allows the possibility that, for example, the effect on t_2 of b given after p may differ from the effect of b given after e. A possible ethical question is why the START design apparently does not include an established standard treatment comparator arm. At the time the trial was begun, based on conventional trials pazopanib was an established standard for frontline and everolimus an established standard for second line treatment for these patients. Thus, the strategy (p, e) may be considered a (frontline, second line) "control" arm.

There were several practical complications to deal with in constructing the START design. These included interval censoring of progression times and possible delay in starting the stage 2 therapy. A specialized computer program necessarily was required to simulate the design and establish its operating characteristics. There were several very time consuming iterations of this process due to successive requirements and advice from various parties involved. The first design (February, 2006) had a maximum of $N = 240$ patients, studied 12 strategies constructed from 4 agents, and assumed an accrual rate of 12 patients per month. To respond to criticisms and suggestions from individuals at the Cancer Therapy Evaluation Program (CTEP)of the National Cancer Institute (NCI), we excluded two agents from the stage 1 pool, yielding 8 regimes, assumed an accrual rate of 9 patients per month, and re-designed and re-simulated the trial (April, 2006). Subsequently, we were informed by the regulators at CTEP/NCI that they were no longer interested in our trial. In January, 2007, after individuals at several pharmaceutical companies expressed an interest, the trial was designed a third time, now with the current 6-regime structure of START, but with different agents and $N = 360$, assuming an accrual rate of 12 patients per month. At the behest of Christopher Logothetis, Chair of the GU Oncology Department at MDACC, I made the arduous journey from Houston to Chicago and presented this latest version of the design at the annual meeting of the Kidney Cancer Association (October, 2007). Following the advice of several oncologists at this meeting, I decreased N to 240 and added the the following Bayesian interim weeding rule. Denoting μ^{max} = maximum of the six regime mean failure times, a strategy d will be stopped if $\Pr\{\mu_d < \mu^{max} - 3 \ months \mid data\} > .90$, applied when 120 patients are fully evaluated (May, 2007). This rule, which does not appear in Thall et al. (2007), does appear in the trial protocol, and may be called a between-regime "drop the losers" rule. Between establishing this design in 2007 and trial activation in 2010, extensive negotiations with various pharmaceutical companies resulted in the three agents $\{p, b, e\}$ that actually are being studied in the START trial. Thus, the process from first design to trial activation included multiple design modifications and took four years.

Important issues in the START trial are the rationale for the maximum sample size of 240, and how the trial might have been designed using conventional tests of hypotheses. The main criteria for choosing sample size were feasibility and the ability to obtain reasonable correct selection rates under an array of different scenarios specified in terms of $E(t_1)$ and $E(t_2)$. Based on $N = 240$ patients, assuming that 20% of patients will discontinue therapy in stage 1, 32 patients are expected to receive both stages of each strategy. If, instead, a hypothesis test based approach were taken, the 15 pairs of regimes might be compared using a two-sided test with null median failure time 15.7 months and power .80 to detect a targeted value of 22 months, a 40% improvement, controlling overall type I error rate at .05. This would require a maximum of 611 patients for each pair of strategies, thus 1833 patients total. The expected maximum trial duration would be slightly over 13 years. The START design also replaces the conventional approach in oncology of doing three single arm trials in what here we call stage 1, and doing three more single arm trials of the agents as second line therapy for patients who progress in stage 1. In this regard, a useful way to think of the total sample size of 240 is to compare the START trial to these six conventional single-arm phase II trials, each of size 40. The data from these six trials would be of very limited use because the failure to randomize would provide data of little use for for unbiased comparisons. Moreover, conducting six single-arm trials would fail to account for the joint effects of two agents given sequentially. Another alternative approach would be to conduct two trials, each with 120 patients randomized among $\{p, b, e\}$. One trial would compare these as frontline agents, and the second trial would compare them as second line agents. With this approach, as with six single-arm trials, the benefits of linking each (frontline, salvage) pair would be lost. This is an essential advantage of START, since the effects of pairs of agents given in sequence are not intuitively clear based in how each agent may behave in one stage of therapy, as either frontline or salvage. Logistically, the additional effort of randomizing 240 patients sequentially in a single trial among the six strategies is minimal. Moreover, administratively, it is much easier to organize one trial rather than two or six trials.

6 Discussion

DTRs reflect actual physician behavior. Compared to conventional trials that focus on one stage of therapy, SMARTs reflect this behavior by providing a basis for unbiased estimation of multi-stage strategy effects. This, in turn, provides a relaible basis for identifying strategies likely to be either superior or inferior. These are practical goals that are more concordant with how therapeutic advances actually are achieved, compared to testing hypotheses.

In most applications, designing a SMART is more challenging and time-consuming than constructing a conventional RCT. This is because DTRs are inherently more complex than single-stage treatments, statistical modeling of the sequences of treatments and outcomes is required, the properties of the trial must be validated by computer simulation, and computer software must be developed for this purpose. In contrast, conduct of a trial to evaluate and compare DTRs actually is very similar to that of a conventional trial that includes within-patient adaptive rules.

The idea of designing clinical trials to study multi-stage treatment sequences is just catching on. SMART trials are rare in oncology. When confronted by simple explanations of why focusing on only one stage of therapy in a clinical trial can lead to very misleading conclusions, many physicians simply are unwilling to be convinced, and more than a few react angrily. From a purely logical viewpoint this may seem strange, since so much of medical practice, and indeed most human behavior, involves sequences of adaptive decisions. The explanation seems to lie in the fact that, while human beings act sequentially, actually planning more than one step ahead can be very difficult and non-intuitive. Moreover, for clinical trialists, the implication of SMARTs and the theory underlying DTRs is that "Everything you know is wrong," which can be very unsettling. Still, it is quite encouraging that the physicians in the GU Oncology Department at MDACC, and many others in the medical community, understand the advantages of SMARTs and have begun to use them to design and conduct actual trials.

Acknowledgements
This research was supported by NIH/NCI grant RO1 CA 83932. Figure 1 originally appeared in the Journal of the National Cancer Institute, and is reproduced here with the permission of Oxford University Press.

Bibliography

1. Almirall D, Compton SN, Gunlicks-Stoessel M, Duan N, Murphy SA. (2012). Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Statistics in Medicine, 31(17), 1887-1902

2. Almirall D, Lizotte D, Murphy SA. Comment on "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer" by Wang, Rotnitzky, Lin, Millikan, and Thall. Journal of the American Statistical Association, 107, 509-512, 2012.

3. Armstrong AJ , Garrett-Mayer ES, Eisneberger M. Comment on 'Adaptive therapy for androgen independent prostate cancer: A randomized selection trial including four regimens' by Thall et al., J National Cancer Institute. 100:681-682, 2008.

4. Armstrong AJ , Garrett-Mayer ES , Yang YC , de WR , Tannock IF , Eisenberger M. A contemporary prognostic nomogram for men with hormone-refractory metastatic prostate cancer: a TAX327 study analysis . Clin Cancer Res. 13 (21): 6396 6403, 2007.

5. Bembom O, van der Laan, M. Statistical methods for analyzing sequentially randomized trials. J Natl Cancer Inst 99: 1577 82, 2007.

6. Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P values and evidence. Journal of the American Statistical Association 82(397) : 112-122, 1987.

7. Chaffee P, van der Laan M. Comment on "Evaluation of Viable Dynamic Treatment Regimes in a Sequentially Randomized Trial of Advanced Prostate Cancer" by Wang, Rotnitzky, Lin, Millikan, and Thall. Journal of the American Statistical Association, 107, 513-517, 2012.

8. Dawson R, Lavori PW. Sample size calculations for evaluating treatment policies in multi-stage design. Clinical Trials. 7:643652, 2010.

9. Estey EH, Thall PF, Pierce S., Cortes, J., Beran, M., Kantarjian, H., Keating, M.J., Andreeff, M. and Freireich, E. Randomized phase II study of Fludarabine +Cytosine Arabinoside+ Idarubicin +/- All Trans Retinoic Acid +/- Granulocyte-colony stimulating factor in poor prognosis newly diagnosed acute myeloid leukemia and myelodysplastic syndrome. Blood, 93:2478-2484, 1999

10. Feng W, Wahed A. Sample size for two-stage studies with maintenance therapy. Statist. Med., 28, 20282041, 2009

11. Gehan EA. The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. Journal of Chronic Diseases. 13:346-353 1961.

12. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC: Boca Raton, 1996.

13. Ioannidis JPA. Why most published research findings are false.PLoS Medicine. 2:0696-0701, 2005.

14. Jeffreys H. Theory of Probability, 3rd ed. Oxford Classic Texts in the Physical Sciences. 1961. Oxford Univ. Press, Oxford.

15. Kass R, Raftery A. Bayes factors. Journal of the American Statistical Association 90, 773-795, 1995.

16. Lavori PW, Dawson R. Improving the efficiency of estimation in randomized trials of adaptive treatment strategies. Clinical Trials. 4(4):297308, 2007.

17. McCloskey DN. The insignificance of statistical significance. Scientific American 272(4) : 104-105, 1995.

18. Millikan R, Logothetis C, Thall PF. Response to comments on 'Adaptive therapy for androgen independent prostate cancer: A randomized selection trial including four regimens' by P.F. Thall et al., J National Cancer Institute. 100(9):682-683, 2008.

19. Murphy SA. An experimental design for the development of adaptive treatment strategies. Statistics in Medicine. 24(10):14551481, 2005

20. Murphy SA, Collins LM, Rush AJ. Customizing treatment to the patient: Adaptive treatment strategies. Drug and Alcohol Dependence. 88:S1S3, 2007.

21. Murphy SA, van der Laan M, Robins JM, and CPPRG. Marginal mean models for dynamic treatment regimes. Journal of the American Statistical Association, 96, 14101424, 2001.

22. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31:102-115,1975

23. Ratain MJ, Karrison TG. Testing the wrong hypothesis in phase II oncology trials: There is a better alternative. Clinical Cancer Research. 13:781-782, 2007.

24. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods application to control of the healthy survivor effect. Math. Modeling, 7:13931512, 1986.

25. Robins, J. M., Hernan, M. A. and Brumback, B. Marginal structural models and causal inference in epidemiology. Epidemiology, 11, 550560, 2000.

26. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In AIDS Epidemiology, Methodological Issues (eds N. Jewell, K. Dietz and V. Farewell), pp. 297331, 1992. Boston: Birkhuser.

27. Tannock IF, de Wit R, Berry WR, et al. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. New England Journal of Medicine. 351(15): 1502 1512, 2004.

28. Thall PF, Logothetis C, Pagliaro L, Wen S, Brown MA, Williams D, Millikan R. Adaptive therapy for androgen independent prostate cancer: A randomized selection trial including four regimens. J National Cancer Institute. 99:1613-1622, 2007

29. Thall PF, Millikan R, Sung, H-G. Evaluating multiple treatment courses in clinical trials. Statistics in Medicine, 19: 1011-1028, 2000.

30. Thall PF, Nguyen HQ. Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. J Biopharmaceutical Statistics 22:785-801, 2012.

31. Thall PF, Sung H-G. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. Statistics in Medicine, 17:1563-1580, 1998.

32. Thall PF, Sung H-G, Estey EH. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. J American Statistical Assoc, 97:29-39, 2002.

33. Thall PF, Szabo A, Nguyen HQ, Amlie-Lefond CM, Zaidat OO. Optimizing the concentration and bolus of a drug delivered by continuous infusion. Biometrics. 67:1638-1646, 2011

34. Thall PF, Wooten LH, Logothetis CJ, Millikan R, Tannir NM. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Statistics in Medicine. 26:4687-4702, 2007.

35. Li Z, Murphy SA. Sample size formulae for two-stage randomized trials with survival outcomes. Biometrika. 98:503-518, 2011.

36. Wahed AS, Thall PF. Evaluating joint effects of induction-salvage treatment regimes on overall survival in acute leukemia. J Royal Statistical Society, Series C (Applied Statistics). 62:67-83, 2013.

37. Wahed AS, Tsiatis AA. Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomization designs in clinical trials. Biometrics, 60, 124133, 2004.

38. Wang L, Rotnitzky A, Lin X, Millikan R, Thall PF. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. J American Statistical Assoc. 107:493-508, 2012. (Rejoinder to comments, pages 518-520)

39. Westfall PH, Johnson WO, Utts JM. A Bayesian perspective on the Bonferroni adjustment. Biometrika 84:419-427,

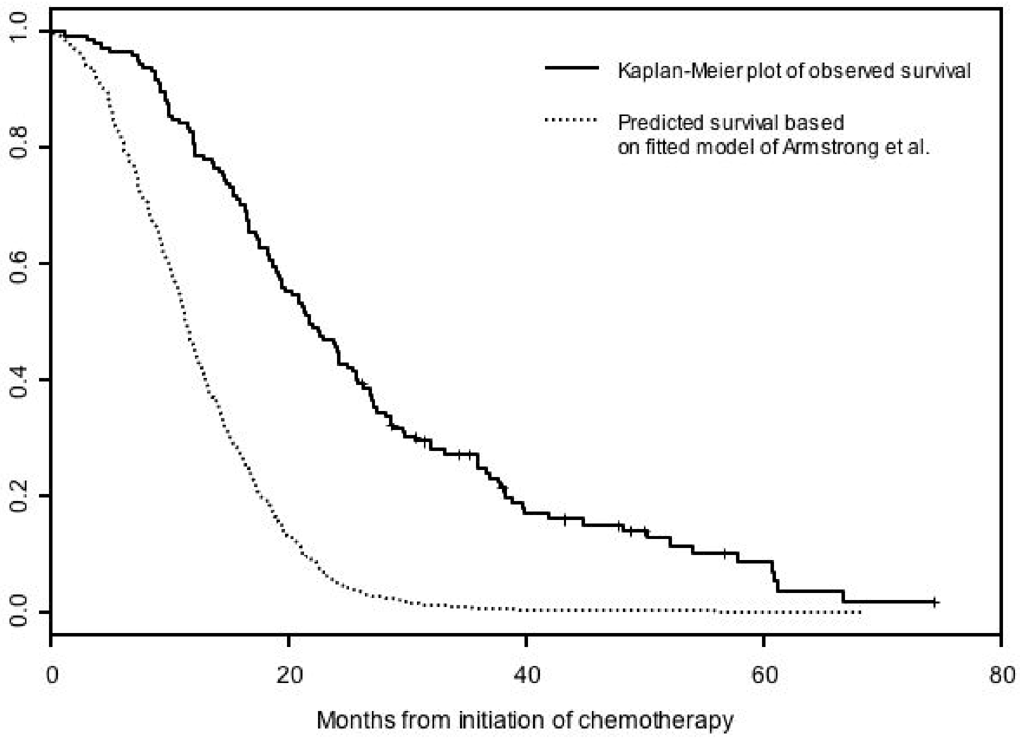

Figure 1: The solid line is the Kaplan-Meier plot computed using the actual survival time data from the trial reported by Thall et al. (2006). The dotted line was computed using the covariates of the patients from this trial and the parameter estimates from the fitted survival model given by Armstrong, et al. (2007b) using data from the TAX327 trial. The dotted line represents a hypothetical survival time distribution that might have been obtained if the patients in the trial reported by Thall et al. (2006) had be treated with docetaxel + prednisone.