HHS Public Access

Author manuscript

Clin Trials. Author manuscript; available in PMC 2015 November 28.

Published in final edited form as:

Clin Trials. 2014 August ; 11(4): 400–407. doi:10.1177/1740774514532570.

Estimation of Optimal Dynamic Treatment Regimes

Ying-Qi Zhao, PhD and

University of Wisconsin-Madison, Department of Biostatistics and Medical Informatics, 600 Highland Ave., Madison, WI, 53705

Eric B. Laber, PhD

North Carolina State University, Department of Statistics, 2311 Stinson Dr., Raleigh, NC, 27695

Ying-Qi Zhao: yqzhao@biostat.wisc.edu

Abstract

Background—Recent advances in medical research suggest that the optimal treatment rules should be adaptive to patients over time. This has led to an increasing interest in studying dynamic treatment regimes (DTRs), a sequence of individualized treatment rules, one per stage of clinical intervention, which map present patient information to a recommended treatment. There has been a recent surge of statistical work for estimating optimal DTRs from randomized and observational studies. The purpose of this paper is to review recent methodological progress and applied issues associated with estimating optimal DTRs.

Methods—We discuss Sequential Multiple Assignment Randomized Trials (SMARTs), a clinical trial design used to study treatment sequences. We use a common estimator of an optimal DTR that applies to SMART data as a platform to discuss several practical and methodological issues.

Results—We provide a limited survey of practical issues associated with modeling SMART data. We review some existing estimators of optimal dynamic treatment regimes and discuss practical issues associated with these methods including: model building; missing data; statistical inference; and choosing an outcome when only non-responders are re-randomized. We mainly focus on the estimation and inference of DTRs using SMART data. DTRs can also be constructed from observational data, which may be easier to obtain in practice, however, care must be taken to account for potential confounding.

Keywords

Adaptive treatment strategies; Dynamic treatment regimes; Missing data; Personalized treatment; Q-learning; Sequential Multiple Assignment Randomized Trials; Outcome weighted learning; Augmented value maximization; Structural mean models

Introduction

In practice, clinical and intervention scientists adapt treatment according to the evolving health status of each patient. Dynamic treatment regimes (DTRs), also known as adaptive treatment strategies, formalize this process as a sequence of individualized treatment rules,

Correspondence to: Ying-Qi Zhao, yqzhao@biostat.wisc.edu.

one per stage of clinical intervention, which map up-to-date patient information to a recommended treatment. A DTR is said to be optimal if it maximizes the average of a desirable clinical outcome when applied to a population of interest. Note that the outcome of interest could be a measure of efficacy, side-effects, or even a composite of multiple outcomes combined into a single utility function; throughout, we assume that the outcome has been coded so that higher is better. There has been a recent surge of methodological work for estimating optimal DTRs from randomized and observational studies [1–14]. However, less attention has been given to applied issues associated with estimating optimal DTRs including model building, dealing with missing data, and choosing outcomes that define optimality. We discuss these issues within the context of data collected in a randomized clinical trial.

To review the motivation for DTRs we briefly consider an example. The treatment of advanced non-small cell lung cancer typically involves two or more lines of treatment. First-line treatments primarily consist of platinum-based doublets which include: cisplatin; gemcitabine; pemetrexed; paclitaxel; carboplatin; and vinorelbine [15]. Docetaxel, pemetrexed and erlotinib are approved second-line treatments. The question of which first-line treatment, or combination of treatments, is best depends both on patient individual characteristics and the protocol for choosing a second-line treatment as there are thought to be interactive effects between first-line and second-line treatments [15]. By considering sequences of treatments we can capture not only synergies between the first- and second-line treatments but also delayed or carry-over effects of the first-line treatment [1]. Another important consideration is the timing of the second-line therapy [16]. Figure 1 shows a schematic for treatment protocol for non-small cell lung cancer indicating where treatment choices must be made. There is interest in optimizing these treatment choices using data from a randomized clinical trial.

One type of randomized clinical trial, which provides data useful for estimating optimal DTRs, is the Sequential Multiple Assignment Randomized Trial (SMART) [17–22]. Optimal DTRs have been estimated from SMARTs for a wide range of chronic conditions including: attention deficit hyperactivity disorder [23, 24]; depression [25]; HIV infection [26, 27, 13]; schizophrenia [28]; and cigarette addiction [5]. In a SMART, subjects are randomized multiple times according to the progression of their health status. A common feature of a SMART is that the pool of available treatments depends on subject-specific characteristics. For example, in the CATIE Schizophrenia Trial [29], subjects with tardive dyskinesia could not be randomized to receive perphenazine. Another common feature of SMARTs is to first randomize subjects to a first-line therapy and subsequently re-randomize only a subset of the subjects according to their health status [30–36]. Figure 2 shows a schematic for such SMART for school-aged children with attention deficit hyperactivity disorder [37]; in this trial responders, operationalized by adequate response on the impairment rating scale [38] and individualized list of target behaviors [39], were not re-randomized. As we will show, these features present novel challenges for building high-quality and interpretable outcome models.

In the next section we introduce basic concepts underpinning two common classes of DTRs estimators using data from SMARTs. Later we discuss some practical issues that arise in applying DTRs estimators, and we finish with some concluding remarks.

Estimating optimal DTRs from SMARTs

To simplify notation, we consider two-stage SMARTs with binary treatment options at each randomization. Data available from such a trial takes the form $\{(\mathrm{X_{1i}}, \mathrm{A_{1i}}, \mathrm{X_{2i}}, \mathrm{A_{2i}}, \mathrm{Y_1})\}_{i=1}^{n}$ comprising n independent and identically distributed trajectories, one for each subject. A generic trajectory (X_1, A_1, X_2, A_2, Y) is composed of $\mathrm{X_1} \in \mathrm{R}^{\mathrm{p_1}}$ which denotes baselines subject information; $A_1 \in \{-1, 1\}$ which denotes the initial (first-line) treatment; $\mathrm{X_2} \in \mathrm{R}^{\mathrm{p_2}}$ which denotes interim subject information collected during the course of the first treatment; $A_2 \in \{-1, 1\}$ which denotes the second (second-line) treatment; and $Y \in R$ which denotes an outcome coded so that higher values are better. Sample size formulae exist for sizing a SMART to compare fixed (i.e., not data-driven) treatment strategies [20, 40, 41]; see [42] for designing SMART pilots. In a trial where only "non-responders" are re-randomized, A_2 can be conceptualized as missing by design. Define $H_1 = X_1$ and $\mathrm{H_2}=(X_1^T, \mathrm{A_1}, X_2^T)^{\mathrm{T}}$ so that H_j denotes the available information before the j^{th} treatment assignment.

A DTR is a pair of functions $d = (d_1, d_2)$ where d_j is a function mapping the covariate space to the treatment space. Under d, a patient with history h_j is recommended with treatment $d_j(h_j)$. Let E^d denote expectation under the restriction that $A_1 = d_1(H_1)$ and $A_2 = d_2(H_2)$ for those re-randomized at the second stage. The optimal DTR, d^{opt}, satisfies $\mathrm{E}^{\mathrm{d^{opt}}}\mathrm{Y} \geq \mathrm{E}^{\mathrm{d}}\mathrm{Y}$ for all DTRs d. Define $Q_2(h_2, a_2) = E(Y \mid H_2 = h_2, A_2 = a_2)$ and $\mathrm{Q_1(h_1, a_1)}=\mathrm{E}(\mathrm{max_{a_2}}\ \mathrm{Q_2(H_2, a_2)}|\mathrm{H_1}=\mathrm{h_1}, \mathrm{A_1}=\mathrm{a_1})$; Q_j is called the stage-j Q-function. The stage-2 Q-function measures the quality of assigning treatment $A_2 = a_2$ to a patient presenting at the second stage with history $H_2 = h_2$. The stage-1 Q-function measures the quality of assigning treatment $A_1 = a_1$ to a subject presenting at the first stage with history $H_1 = h_1$ if he/she were treated according to the d_2^{opt} at stage-2. From dynamic programming [43] it follows that $d_j^{opt}(\mathrm{h_j})=\mathrm{argmax_{a_j}}\ \mathrm{Q_j(h_j, a_j)}$. This formulation suggests several strategies for estimating d^{opt} from SMART data. Estimators can be broadly classified into three categories: (i) regression-based methods; (ii) value maximization methods; and (iii) planning methods. Regression-based methods attempt to first estimate the Q-functions via regression models and subsequently use a plug-in estimator of d^{opt}. Regression based estimators include Q- and A-learning [1, 2, 25, 44, 45]; regret regression [6]; threshold methods [5, 4, 11]; and interactive Q-learning [46]. Regression-based estimators for censored data, discrete outcomes, continuous treatments [47, 48], and quantiles have also been developed [45, 49, 50]. We give a version of the Q-learning algorithm in the following section. Value maximization methods are based on forming an estimator of $V(d) = E^dY$ and then directly maximizing this estimator over d in some class of DTRs, say D. Value maximization methods include outcome weighted learning [9, 14], augmented value maximization [10, 12, 13], and structural mean models [7]. We give a version of outcome weighted learning in the subsequent section. Planning methods rely on systems dynamics models to simulate patient trajectories under different DTRs to find an optimum [51–53].

These models rely strongly on biological or behavioral theory, which are absent or incomplete in the settings we consider and thus we will not discuss them further.

Q-learning

In this section we assume that all subjects are re-randomized at the second stage. The Q-learning algorithm requires postulated models for the Q-functions; we consider linear working models of the form $\mathrm{Q_j}(\mathrm{H_j}, \mathrm{A_j};\theta_\mathrm{j})=H_j^T\ \alpha_\mathrm{j}+\mathrm{A_j}H_j^T\ \beta_\mathrm{j}$, where $\theta_\mathrm{j}=(\alpha_j^T, \beta_j^T)^\mathrm{T}$, $\mathrm{H_{j1}}$, $\mathrm{H_{j2}}$ are known features constructed from $\mathrm{H_j}$. Q-learning algorithm mimics the dynamic programming solution using a sequence of regressions to obtain the estimates $\widehat{\theta_j}$, and subsequently the estimated DTRs. The algorithm is summarized in the Appendix. Thus, Q-learning can be easily implemented in almost any statistical software package. Q-learning is available in packages qLearn and iqLearn of the R programming language (cran.us.r-project.org), which is freely available and callable from both SAS and SPSS. Other advantages of Q-learning include: (i) that it can be extended to accommodate discrete outcomes [45, 50], censored outcomes [49], and competing outcomes [54, 55]; (ii) measures of goodness of fit and visual diagnostics can be used to assess the quality of the fitted regression models in each stage; and (iii) the estimated Q-functions are prognostic, i.e., $max_{a_j} Q_j(h_j, a_j;\widehat{\theta_j})$, is a prediction for the outcome for a patient presenting at stage j and receiving optimal treatment thereafter.

Despite the foregoing advantages, Q-learning presents a number of challenges in practice. First, as the result of the maximization in the intermediate step (step Q2 in the appendix), modeling the stage-1 Q-function requires modeling a nonsmooth, nonmonotone transformation of the data. Even under simple generative models, it can be shown that the form of the stage-1 Q-function can be quite complex [25, 46] making it difficult to correctly specify a model. One potential remedy is to use flexible regression models, say, support vector regression or generalized additive models, to estimate the Q-functions [15, 45]; however, such models are difficult to interpret, limiting their ability to generate new scientific content. Another potential solution is to modify the Q-learning algorithm to avoid modeling after maximization [46].

A second challenge associated with Q-learning is statistical inference. Coefficients indexing the stage-1 Q-function are statistically nonregular [5, 23, 56]. A consequence is that standard methods for inference, e.g., the bootstrap or normal approximations, cannot be applied without modification. Proposed solutions include subsampling [56] and adaptive confidence intervals [23]. Both of these methods have been shown to perform well in simulations but may be conservative in small samples.

Value maximization methods

Q-learning is often called an indirect method because it estimates the optimal DTR indirectly through the estimated Q-functions. A more direct approach is to postulate an estimator of $\mathrm{V(d)} = \mathrm{E^dY}$, say $\hat{V}(\mathrm{d})$, and then estimate the optimal DTR by searching for $\hat{d}=\mathrm{argmax}_{\mathrm{d}\in\mathrm{D}}\ \hat{V}(\mathrm{d})$, where D is a prespecified class of DTRs. Estimators of this form are

called value maximization methods or policy search methods and have received a great deal of attention recently [7, 9, 10, 12–14]. A potential advantage of value maximization methods is that, because they need not rely on models for the Q-function, they may be more robust to model specification. Conversely, fewer assumptions about the trajectory distribution may lead to estimators with higher variability.

Methods for estimating V(d), include: inverse probability-weighting [9, 14]; augmented inverse probability-weighting [10, 13]; and marginal structural mean models [7]. Marginal structural mean models are most effective with low-dimensional histories and a small class of potential regimes D. Inverse probability-weighting and augmented inverse probability-weighting estimators can be applied with high-dimensional histories and very large classes of regimes, however, they are nonsmooth functions of the observed data making the search for the optimal regime within D computationally challenging. Both [9] and [10] connected the problem of maximizing inverse probability-weighting and augmented inverse probability-weighting estimators of V(d) with weighted classification problems and were thereby able to leverage existing classification algorithms to approximately compute $\mathrm{argmax}_{\mathrm{d}\in\mathrm{D}}\,\hat{V}(\mathrm{d})$. We briefly review a simple value maximization algorithm.

Assume that binary treatments are equally randomized at each stage. Then, under mild regularity conditions, it can be shown that $\mathrm{V}(\mathrm{d})=4E(Y1_{A_1=d_1(H_1)}1_{A_2=d_2(H_2)})$ where 1_z equals one if z is true and zero otherwise [14]. The inverse probability weighted estimator is based on the foregoing expression for V(d) and is given by

$$\hat{V}(\mathrm{d})=4n^{-1}\sum_{i=1}^{n}Y_i\,1_{A_{1i}=d_1(H_{1i})}1_{A_{2i}=d_2(H_{2i})}.$$

For illustration, assume D is the space of all linear decision rules. Then, for an $\mathrm{d}\in\mathrm{D}$, we may associate a vector $\beta=(\beta_1^T,\beta_2^T)^{\mathrm{T}}$ so that $\mathrm{d_j}(\mathrm{h_j})=\mathrm{sign}(h_j^T\beta_\mathrm{j})$ where sign(x)=1 if x>0 and sign(x)=−1 if x<0, and write

$$\hat{V}(d)=\hat{V}(\beta)=4n^{-1}\sum_{i=1}^{n}Y_i\,1_{A_{1i}H_{1i}^T\beta_1>0}1_{A_{2i}H_{2i}^T\beta_2>0}.$$

An estimator of the optimal DTR is obtained by solving for $\tilde{\beta}=argmax_{\beta_2}\hat{V}(\beta)$. However, the indicator functions make this a mixed integer linear program, which is known to be computationally burdensome. Approaches to finding $\tilde{\beta}$ include employing a stochastic search algorithm, for example simulated annealing or a generic algorithm [13], or using a concave surrogate for the indicator functions [14]. Depending on the optimization method, additional constraints on 9 may be required to ensure a unique solution.

Value maximization methods are appealing because they avoid strong and potentially incorrect assumptions about the outcome distribution. Furthermore, the class of regimes D can be restricted to only include regimes which are logistically feasible, parsimonious, interpretable, or otherwise desirable. Drawbacks of value maximization methods include: computational complexity; the lack of a prognostic model; the potential lack of a

scientifically meaningful estimand; and, as mentioned previously, potentially higher variability.

Additional practical considerations

In addition to the issues raised in the foregoing section, there are a number of important practical considerations associated with estimating optimal DTRs from SMART data. Here, we provide an overview of those that we have found to be most common.

Missing data

SMARTs, like any clinical trial, are prone to missing data. Dealing with missing data in SMARTs is complicated by the sequential design and the fact that treatment randomizations depend on evolving subject status. For example, in a trial where only responders are re-randomized at the second stage, a subject that is lost to follow-up during the first stage will be missing: second stage history which contains his/her responder status; second stage treatment; and outcome. Whether the second stage treatment is truly missing or missing by design depends on the subject's unobserved responder status. Another complication is that the timing and number of clinic visits may be dependent on patient outcomes [29]; thus, a natural approach is to use multiple imputation and sequentially impute missing data as needed. For example, if clinic visits are dependent on patient status, one would first impute patient status, then, conditional on the imputed status, one would subsequently impute the next visit time, etc. Shortreed et al. provide a sequential multiple imputation strategy for SMARTs that can be used with existing multiple imputation software [57].

Both regression-based and value-maximization methods for estimating optimal DTRs can be extended for use with multiply imputed datasets by either aggregation or concatenation. In the aggregation approach one first estimates the optimal DTR separately for each imputed data set and then take a majority vote. For example, if $\hat{d}^{(1)}, \ldots, \hat{d}^{(M)}$ are M stage-j decision rules estimated across M multiply imputed datasets, then the aggregated rule is $\hat{d}_j^A(h_j) = sign\left(\sum_{m=1}^{M} \hat{d}_j^{(m)}(h_j)\right)$. Note that in the case of linear decision rules, the aggregated decision rule is equivalent to simply averaging the coefficients indexing the decision rules to form a single linear decision rule. Alternatively, concatenation involves stacking the M imputed datasets on top of each other to form a single large dataset and then estimating a single decision rule. This is advantageous if fitting the model on each imputed dataset is computationally expensive or if it is desired that the final decision be sparse (the average of sparse vectors need not be sparse). However, concatenation may preclude estimation of between-imputation variance estimation used in standard multiple imputation variance formulas [58].

Choosing an outcome in responder trials

SMARTs where only a subset of subjects, which we generically term 'non-responders,' are re-randomized at the second stage are common, especially in cancer clinical trials. We assume responders are followed until the end of the study.

Let R be an indicator of response, which takes the value 1 if the subject is a responder and 0 otherwise, then R is contained in H_2 for all subjects. We assume that R is assessed at the end of a fixed time-period, e.g., six-months from baseline. Those deemed non-responders are immediately re-randomized. For responders, Y is collected prior to the assignment of A_2 and hence is part of H_2; indeed, H_2 will contain different information for responders and non-responders. Thus, $Q_2(H_2, A_2) = Y\ R + E\ (Y \mid H_2, R = 0, A_2)$. In Q-learning one would thus estimate the stage-2 Q-function by regressing Y on H_2 and A_2 only using the non-responder data. Let $Q_2(H_2, a_2; \widehat{\theta}_2)$ denote this estimator. The second step of the Q-learning algorithm, Q_2, is to compute $\hat{Y} = RY + (1 - R) max_{a_2} Q_2(H_2, a_2; \widehat{\theta}_2)$. However, note that $\hat{Y}$ will typically be more variable for responders than non-responders since non-responder data has been projected onto H_2. This can complicate building high-quality models of the stage-1 Q-function. One approach to alleviate this problem is to regress Y on information collected prior to classification of responder status. Let $\widetilde{H_2}$ denote information collected prior to responder classification and let $\tilde{Y}$ be an estimator of $\mathrm{E}(\mathrm{Y} | \widetilde{H_2})$ built using only responder data. Then, $\hat{Y} * = R\tilde{Y} + (1 - R) max_{a_2} Q_2(H_2, a_2; \widehat{\theta}_2)$ is used in place of $\hat{Y}$ in the first stage regression of Q-learning. Note that this does not affect the validity of the Q-learning algorithm since H_1 and A_1 are contained in $\widetilde{H_2}$ so that $\mathrm{E}(\mathrm{YR} | \mathrm{H}_1, \mathrm{A}_1) = \mathrm{E}\left\{\mathrm{E}(\mathrm{YR} | \widetilde{H_2}, \mathrm{R}) | \mathrm{H}_1, \mathrm{A}_1\right\}$.

Conclusions

We have tried to provide a limited survey of practical issues associated with estimation of optimal DTRs from SMART data. While estimation of and inference for DTRs is a rapidly growing area of statistics methodological research, it is equally important to address more practical issues associated with modeling SMART data. Given the rapidly growing interest in estimating optimal DTRs from SMARTs we believe a crucial open issue is the development of valid sample size formulae for testing data-driven DTRs.

Acknowledgments

This work was supported by National Institute of Health [P01 CA142538]. Funding for this conference was made possible (in part) by 2 R13 CA132565-06 from the National Cancer Institute. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

Appendix

Q-learning algorithm is as follows:

Q1 Find $\widehat{\theta}_2 = argmin_{\theta_2} \sum_{i=1}^{n} (Y_i - Q_2(H_{2i}, A_{2i}; \theta_2))^2$;

Q2 Define $\hat{Y} = max_{a_2} Q_2\left(H_2, a_2; \widehat{\theta}_2\right)$;

Q3 Find $\widehat{\theta}_1 = argmin_{\theta_1} \sum_{i=1}^{n} \left(\widehat{Y}_i - Q_1(H_{1i}, A_{1i}; \theta_1)\right)^2$;

then $\hat{d}_j^Q(h_j) = argmax_{a_j} Q_j\left(h_j, a_j; \widehat{\theta}_J\right)$. Q-learning, in the simple case considered, requires fitting two linear regressions (steps Q1 and Q3) and making linear predictions for each subject (step Q2).

References

1. Murphy SA. Optimal dynamic treatment regimes. J Roy Stat Soc B. 2003; 65:331–366.
2. Robins, JM. Optimal structural nested models for optimal sequential decisions. In: Lin, DY.; Heagerty, PJ., editors. Proc Second Seattle Symp on Biostatistics. Springer; 2004. p. 189-326.
3. Murphy SA. A generalization error for q-learning. J Mach Learn Res. 2005; 6:1073–1097. [PubMed: 16763665]
4. Moodie EEM, Richardson TS. Estimating optimal dynamic regimes: Correcting bias under the null. Scand Stat Theory Appl. 2009; 37:126–146. [PubMed: 20526433]
5. Chakraborty B, Murphy S, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes. Stat Methods Med Res. 2010; 19:317–343. [PubMed: 19608604]
6. Henderson R, Ansell P, Alshibani D. Regret-regression for optimal dynamic treatment regimes. Biometrics. 2010; 66:1192–1201. [PubMed: 20002404]
7. Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content. Int J Biostat. 2010; 6:1–47.
8. Zhao Y, Zeng D, Socinski MA, et al. Reinforcement learning strategies for Clin Trials in nonsmall cell lung cancer. Biometrics. 2011; 67:1422–1433. [PubMed: 21385164]
9. Zhao YQ, Zeng D, Rush AJ, et al. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc. 2012; 107:1106–1118. [PubMed: 23630406]
10. Zhang B, Tsiatis AA, Laber EB, et al. A robust method for estimating optimal treatment regimes. Biometrics. 2012; 68:1010–1018. [PubMed: 22550953]
11. Goldberg Y, Song R, Kosorok MR. Adaptive q-learning. From Probability to Statistics and Back. High-Dimensional Models and Processes. 2012; 150
12. Zhang B, Tsiatis AA, Davidian M, et al. Estimating optimal treatment regimes from a classification perspective. Stat. 2012; 1:103–114. [PubMed: 23645940]
13. Zhang B, Tsiatis A, Laber E, et al. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika. 2013 To appear.
14. Zhao, YQ.; Zeng, D.; Laber, E., et al. Report. University of Wisconsin-Madison; US: Jan. 2014 New statistical learning methods for estimating optimal dynamic treatment regimes.
15. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. Stat Med. 2009; 28:3294–3315. [PubMed: 19750510]
16. Fidias P, Dakhil S, Lyss A, et al. Phase iii study of immediate versus delayed docetaxel after induction therapy with gemcitabine plus carboplatin in advanced non-small-cell lung cancer: updated report with survival. J Clin Oncol. 2007; 25(18 suppl):388s.
17. Lavori PW, Dawson R. A design for testing clinical strategies: biased adaptive within-subject randomization. J Roy Stat Soc A. 2000; 163:29–38.
18. Lavori PW, Dawson R. Dynamic treatment regimes: practical design considerations. Clin Trials. 2004; 1:9–20. [PubMed: 16281458]
19. Dawson R, Lavori P. Placebo-free designs for evaluating new mental health treatments: the use of adaptive treatment strategies. Stat Med. 2004; 23:3249–3262. [PubMed: 15490427]
20. Murphy SA. An experimental design for the development of adaptive treatment strategies. Stat Med. 2005; 24:1455–1481. [PubMed: 15586395]
21. Murphy SA, Oslin DW, Rush AJ, et al. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. Neuropsychopharmacol. 2007; 32:257–262.
22. Nahum-Shani I, Qian M, Almiral D, et al. Experimental design and primary data analysis methods for comparing adaptive interventions. Psychol Methods. 2012; 17:457–477. [PubMed: 23025433]

23. Laber, EB.; Qian, M.; Lizotte, D., et al. Report, Stat Dept. Univ. of Michigan; 2011. Statistical inference in dynamic treatment regimes. Report no. 506
24. Nahum-Shani I, Qian M, Almiral D, et al. Q-learning: A data analysis method for constructing adaptive interventions. Psychol Methods. 2012; 17:478–494. [PubMed: 23025434]
25. Schulte, PJ.; Tsiatis, AA.; Laber, EB., et al. Stat Sci. 2013. Q- and a-learning methods for estimating optimal dynamic treatment regimes. In Press
26. Moodie EEM, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics. 2007; 63:447–455. [PubMed: 17688497]
27. Cain LE, Robins JM, Lanoy E, et al. When to start treatment? a systematic approach to the comparison of dynamic regimes using observational data. Int J Biostat. 2010; 6(2) Article 18.
28. Shortreed SM, Laber E, Lizotte DJ, et al. Informing sequential clinical decision-making through reinforcement learning: an empirical study. Mach Learn. 2011; 84:109–136. [PubMed: 21799585]
29. Stroup TS, McEvoy JP, Swartz MS, et al. The national institute of mental health clinical antipsychotic trials of intervention effectiveness (catie) project. Schizophrenia Bull. 2003; 29:15–31.
30. Joss R, Alberto P, Bleher E, et al. Combined-modality treatment of small-cell lung cancer: Randomized comparison of three induction chemotherapies followed by maintenance chemotherapy with or without radiotherapy to the chest. Ann Oncol. 1994; 5:921–928. [PubMed: 7696164]
31. Stone RM, Berg DT, George SL, et al. Granulocyte– macrophage colony-stimulating factor after initial chemotherapy for elderly patients with primary acute myelogenous leukemia. N Engl J Med. 1995; 332:1671–1677. [PubMed: 7760868]
32. Tummarello D, Mari D, Graziano F, et al. Randomized, controlled phase iii study of cyclophosphamide, doxorubicin, and vincristine with etoposide (cav-e) or teniposide (cav-t), followed by recombinant interferon-α maintenance therapy or observation, in small cell lung carcinoma patients with complete responses. Cancer. 1997; 80:2222–2229. [PubMed: 9404698]
33. Matthay KK, Villablanca JG, Seeger RC, et al. Treatment of high-risk neuroblastoma with intensive chemotherapy, radiotherapy, autologous bone marrow transplantation, and 13-cis-retinoic acid. N Engl J Med. 1999; 341:1165–1173. [PubMed: 10519894]
34. Habermann TM, Weller EA, Morrison VA, et al. Rituximab-chop versus chop alone or with maintenance rituximab in older patients with diffuse large b-cell lymphoma. J Clin Oncol. 2006; 24(19):3121–3127. [PubMed: 16754935]
35. Auyeung SF, Long Q, Royster EB, et al. Sequential multiple assignment randomized trial design of neurobehavioral treatment for patients with metastatic malignant melanoma undergoing high-dose interferon-alpha therapy. Clin Trials. 2009; 6:480–490. [PubMed: 19786415]
36. Mateos MV, Oriol A, Martinez-Lopez J, et al. Bortezomib, melphalan, and prednisone versus bortezomib, thalidomide, and prednisone as induction therapy followed by maintenance treatment with bortezomib and thalidomide versus bortezomib and prednisone in elderly patients with untreated multiple myeloma: a randomised trial. Lancet Oncol. 2010; 11:934–941. [PubMed: 20739218]
37. William, E.; Pelham, J.; Fabiano, G.; Waxmonsky, J., et al. Adaptive pharmacological and behavioral treatments for children with adhd: Sequencing, combining, and escalating doses. Institute of Educational Sciences' Third Annual Research Conference; Washington, DC. 2008.
38. Fabiano GA, Pelham WE, Waschbusch DA, et al. A practical measure of impairment: Psychometric properties of the impairment rating scale in samples of children with attention deficit hyperactivity disorder and two school-based samples. J Clin Child Adolesc Psychol. 2006; 35:369–385. [PubMed: 16836475]
39. Pelham WE, Hoza B, Pillow DR, et al. Effects of methylphenidate and expectancy on children with ADHD: Behavior, academic performance, and attributions in a summer treatment program and regular classroom setting. J Consult Clin Psychol. 2002; 70:320–335. [PubMed: 11952190]
40. Li Z, Murphy SA. Sample size formulae for two-stage randomized trials with survival outcomes. Biometrika. 2011; 98:503–518. [PubMed: 22363091]
41. Feng W, Wahed AS. A supremum log rank test for comparing adaptive treatment strategies and corresponding sample size formula. Biometrika. 2008; 95:695–707.

42. Almirall D, Compton SN, Gunlicks-Stoessel M, et al. Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. Stat Med. 2012; 31:1887–902. [PubMed: 22438190]
43. Bellman, R. Dynamic Programming. Princeton Univeristy Press; Princeton: 1957.
44. Blatt, D.; Murphy, SA.; Zhu, J. Report. University of Michigan; US: 2004. A-learning for approximate planning.
45. Moodie EEM, Dean N, Sun YR. Q-learning: Flexible learning about useful utilities. Stat Biosciences. Sep 12.2013 Epub ahead of print. doi: 10.1007/s12561-013-9103-z
46. Laber, E.; Linn, K.; Stefanski, L. Report. North Carolina State University; US: Jan. 2014 Interactive q-learning.
47. Rich, B.; Moodie, EEM.; Stephen, DA. Report. Mcgill University; CA: 2014. Adaptive individualized dosing in pharmacological studies: Generating candidate dynamic dosing strategies for warfarin treatment.
48. Laber, EB.; Zhao, YQ. Report. North Carolina State University; US: Jan. 2014 Tree-based methods for optimal treatment allocation.
49. Goldberg Y, Kosorok MR. Q-learning with censored data. Ann Stat. 2012; 40:529–560. [PubMed: 22754029]
50. Linn, K.; Laber, E.; Stefanski, L. Report. North Carolina State University; US: Jan. 2014 Interactive q-learning for probabilities and quantiles.
51. Rivera DE, Pew MD, Collins LM. Using engineering control principles to inform the design of adaptive interventions: A conceptual introduction. Drug Alcohol Depend. 2007; 88:S31–S40. [PubMed: 17169503]
52. Navarro-Barrientos JE, Rivera DE, Collins LM. A dynamical systems model for understanding behavioral interventions for weight loss. Adv Social Comput Lect Notes Comput Sci. 2010; 6007:170–179.
53. Navarro-Barrientos JE, Rivera DE, Collins LM. A dynamical model for describing behavioural interventions for weight loss and body composition change. Math Comput Modell Dyn Syst. 2011; 17:183–203.
54. Lizotte DJ, Bowling M, Murphy SA. Linear fitted-q iteration with multiple reward functions. J Mach Learn Res. 2012; 13:3253–3295. [PubMed: 23741197]
55. Laber E, Lizotte D, Ferguson B. Set-valued dynamic treatment regimes for competing outcomes. Biometrics. Jan 8.2014 Epub ahead of print. doi: 10.1111/biom12132
56. Chakraborty B, Laber E, Zhao YQ. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. Biometrics. 2013; 69:714–723. [PubMed: 23845276]
57. Shortreed, S.; Laber, E.; Pineau, J., et al. Technical Report. 2010. Imputations methods for the clinical antipsychotic trials of intervention and effectiveness study. Technical Report SOCS-TR-2010.8
58. Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2. Chichester: Wiley; 2002.

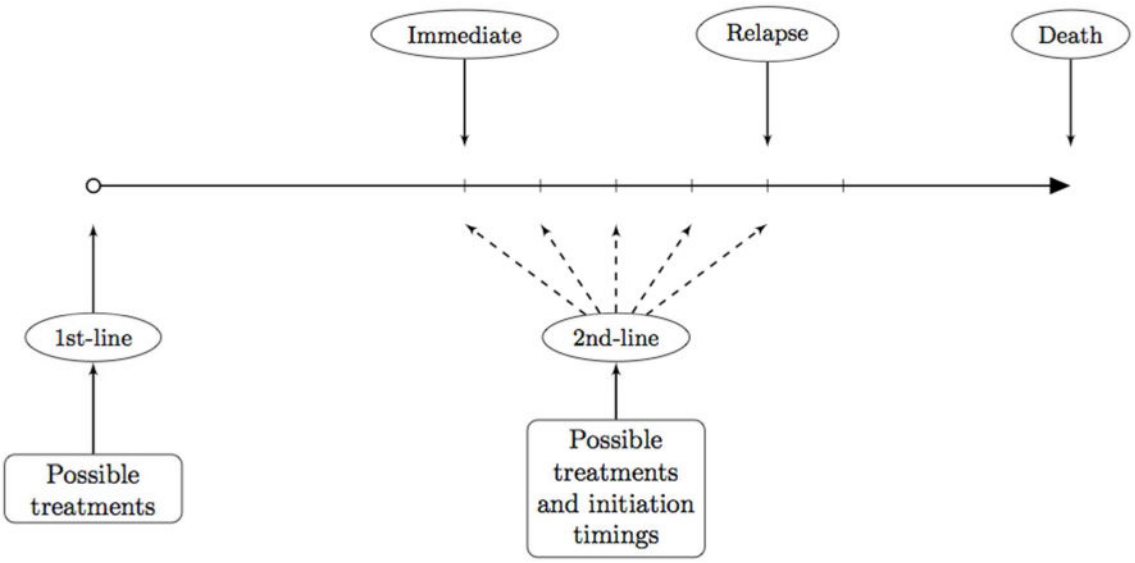

Figure 1.

Non-small Cell Lung Cancer. There are multiple possible 1st-line treatments to start with. As a follow-up, multiple possible 2nd –line treatments exist. In addition, there are many choices for when to initiate the 2nd-line treatment. For example, it can be initiated once the 1st-line treatment is finished (immediate); it can be initiated when the disease progresses (progression); or any time inbetween.

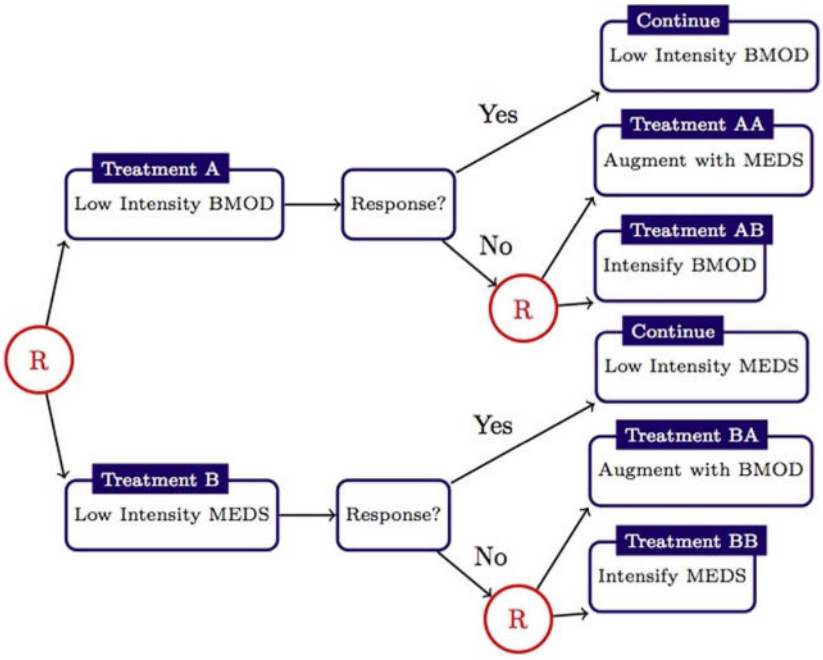

Figure 2.

Schematic describing the Adaptive Pharmacological and Behavioral Treatments for Children with ADHD SMART [W. Pelham (PI)]; randomizations, denoted by a circled letter 'R,' were with equal probability. Responder status is based on subject Impairment Rating Scale [38] and Individualized List of Target Behaviors [39]. see [22] for additional details.