REVSTAT – Statistical Journal
Volume 9, Number 1, March 2011, 19–36

OPTIMAL DYNAMIC TREATMENT METHODS

Authors: Robin Henderson
– Mathematics & Statistics, Newcastle University, UK
Robin.Henderson@newcastle.ac.uk

Phil Ansell
– Mathematics & Statistics, Newcastle University, UK
P.S.Ansell@newcastle.ac.uk

Deyadeen Alshibani
– Mathematics & Statistics, Newcastle University, UK
Deyadeen.Alshinani@newcastle.ac.uk

Abstract:

- This paper reviews and develops methods for implementing in practice recent ideas in the field of optimal dynamic treatment allocation. Given longitudinal sequences of observational data on health status and treatment selection for a cohort of patients, the aim is to determine a regime, or decision rule, which can be used to select treatment in order to optimise some final response or outcome. The approach to this problem that has been taken in the causal inference literature is shown to be extendable to problems in the field of stochastic optimisation. New diagnostic techniques to aid in model assessment are developed, and an application in anticoagulation is presented.

Key-Words:

- *anticoagulation; bandit problems; causal inference; diagnostics; regret functions; wild bootstrap.*

AMS Subject Classification:

- 62J07, 62L07.

1. INTRODUCTION

Individualised medicine, which is one of the growing areas in health research, presents a number of statistical challenges. Without the luxury of major clinical trials, can we find methods to tailor treatment to a patient's individual circumstances, especially for those with chronic conditions? In this paper we give an overview of a selection of methods for *optimal dynamic treatment regime determination* from observational data [1], [3], [6], [11]–[13]. Our interest in the area is motivated by a collaboration in which an algorithm to determine decision rules for anticoagulation doseage is required. Anticoagulants are used to maintain blood clotting speed and reduce risk of thrombosis. They are one of the most prescribed groups of drugs in the world, being used for both treatment and prophylaxis for conditions like deep venous thrombosis, stroke, atrial fibrillation, acute myocardial infarction, prosthetic heart valves and many more. A difficulty is that there is no standard dose: the amount required varies not just between patients but also over time within patients, in response to lifestyle and dietary changes, in particular the amount of vitamin K within the body. Given a patient's current and previous values of blood clotting time, and their history of anticoagulation, can we find decision rules to provide the optimal current dose?

Three classes of methods for a general version of this problem are summarised in Section 2. We consider model formulation and estimation, and illustrate through simulations. In Section 3 we draw attention to links between recent optimal dynamic treatment methods and the longstanding stochastic scheduling research in the operational research literature. In Section 4 we propose a suite of diagnostic tests for model adequacy based on wild bootstrap residuals. In Section 5 we describe an application of the methods to the warfarin anticoagulation application which motivated our interest.

2. REGRETS, BLIPS AND REGRESSION

2.1. Modelling approaches

We assume there are K decision times, for example clinic visits. At each decision time a *state* variable is recorded, $S_1, S_2, ..., S_K$. This might be the health of a patient and can be multivariate or scalar. A decision on the *action* to be taken is then made, such as treatment allocation, leading to an action sequence $A_1, A_2, ..., A_K$. The objective is to maximise some final value Y, which may not be revealed until all K decisions have been taken, or which may accrue with time, as in the warfarin example in Section 5. Panel (a) of Figure 1 illustrates the sequence that is followed. Throughout we will assume independence between subjects and will take the standard assumption of no unmeasured confounders: all non-random elements influences action choices are captured in the observed data. We omit further technical detail on the conditions needed for valid inference.

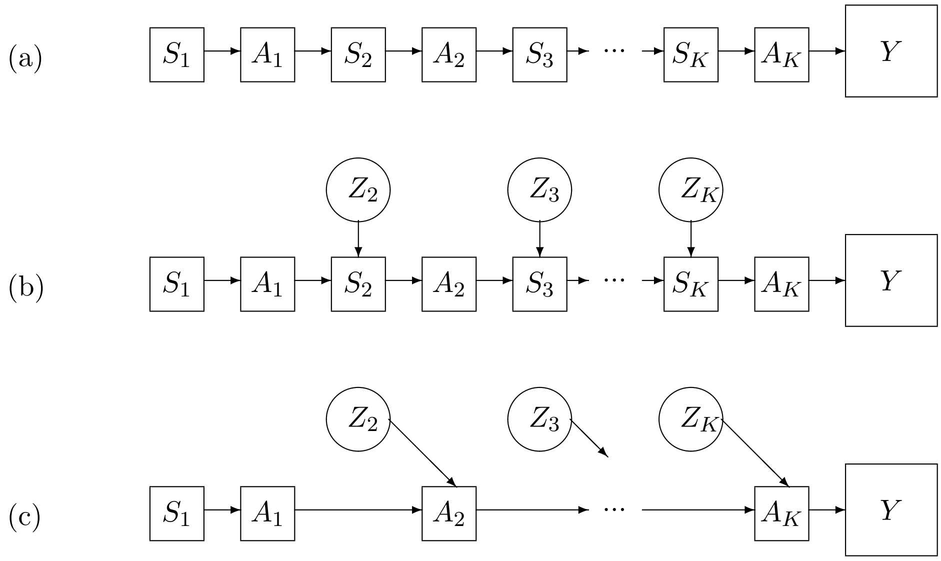

Figure 1: State, action and outcome sequence: (a) the general scenario; (b) inclusion of exogenous variables; (c) orthogonalisation.

Define $\bar{S}_j = (S_1, ..., S_j)$ and $\bar{A}_j = (A_1, ..., A_j)$ to indicate the history of states and actions respectively, up to and including time j. The information available just before action j is selected is $\mathcal{F}_j = (\bar{S}_j, \bar{A}_{j-1})$ and the aim is to obtain decision *rules* $d_j(\mathcal{F}_j)$ which will maximise the expected value of Y given the information to hand. We will use $\underline{d}_j^{\text{ref}}$ to denote a known standard or reference regime, with the underscore being read as meaning all times from j to K. Similarly $\underline{d}_j^{\text{opt}}$ is the *optimal* regime, which is unknown and is the target for analysis.

Robins [14] proposed a structural nested mean model [8] approach to the problem, based on *blip functions*, which can be defined as

$$\gamma_j(a_j|\mathcal{F}_j) \;=\; E(Y \mid \mathcal{F}_j, a_j, \underline{d}_{j+1}^{\text{opt}}) - E(Y \mid \mathcal{F}_j, d_j^{\text{ref}}, \underline{d}_{j+1}^{\text{opt}}) \; .$$

Here γ_j is a function of the possible actions a_j which are available at time j, given the history $\mathcal{F}_j$ of states and actions up to that point. The blip contrasts two expectations. The first is of the final response Y given that a_j is selected at time j and under the possibly counterfactual assumption that the optimal reference regime will be followed from $j+1$ onward. The second expectation is similar except action a_j is replaced by the reference regime d_j^{ref} at time j. Robins chose the name because in each expectation the past F_j is the same, the future policy is the same, and the only difference or "blip" is between a_j and d_j^{ref} at time j.

Under Robins' approach a parametric form $\gamma_j(a_j|\mathcal{F}_j;\theta)$ is assumed for the blip function. For example we might take

$$\gamma_j(a_j|\mathcal{F}_j;\theta) \;=\; \theta_1\Big(a_j - (\theta_2 + \theta_3 S_j + \theta_4 S_{j-1})\Big)\, I(a_j \neq d_j^{\mathrm{ref}})$$

where $I(\cdot)$ is an indicator function introduced to ensure the blip is zero if the reference action is selected. Otherwise the effect of the action a_j is assumed to depend on current and previous states S_j and S_{j-1} respectively. This is a strong assumption, but an advantage of the approach is that once parameter θ is estimated it is straightforward to determine the causal effect of actions.

Murphy [13] prefers to work with *regret functions*

$$\mu_j(a_j|\mathcal{F}_j) \;=\; E(Y \mid \mathcal{F}_j, \underline{d}_j^{\mathrm{opt}}) - E(Y \mid \mathcal{F}_j, a_j, \underline{d}_{j+1}^{\mathrm{opt}}) \ .$$

These are of similar form to blip functions except they contrast the effect at time j of action a_j with the as-yet-unknown optimal rule. Thus the first expectation assumes the optimal decision is taken from j onward, whereas in the second expectation action a_j is chosen at j and then the optimal policy followed from $j+1$ onward. Regrets are non-negative since the objective is to maximise Y. They give a direct measure of the effect of choosing a sub-optimal action at time j.

Again a parametric form is assumed, for example

$$\mu_j(a_j|\mathcal{F}_j;\psi) \;=\; \psi_1\Big(a_j - (\psi_2 + \psi_3 S_j + \psi_4 S_{j-1})\Big)^2 \ .$$

This guarantees the non-negativity of the regrets and assumes the optimal action — that which has zero regret — is a linear combination of S_j and S_{j-1}. Once ψ is known the optimal action is therefore immediately obtained.

The parametric forms assumed for blips or regrets cannot be checked, since they are models for differences in counterfactuals. An alternative approach introduced independently by Almirall and colleagues [1] and Henderson and colleagues [6] attempts to incorporate parametrised regrets into a model for the actual response Y. The authors note first that the final response Y is determined by three groups of factors:

1. The initial conditions.

2. The actions selected.

3. Chance development over time.

It is straightforward to introduce initial conditions as a function of S_1 into a model for Y. The effect of actions can be modelled by regrets as above. To model chance development over time, [1] and [6] envisage a sequence of exogenous variables $Z_1, Z_2, ..., Z_K$ which influence states and Y over and above the effects of the chosen actions, as summarised in panel (b) of Figure 1. If the $\{Z_j\}$ are observed then we can model final response as

$$E[Y|\bar{S}_K, \bar{A}_K] \;=\; \beta_1(S_1) + \sum_{j=2}^{K} \beta_j^T(\bar{S}_{j-1}, \bar{A}_{j-1})\, Z_j - \sum_{j=1}^{K} \mu_j(A_j|\mathcal{F}_j) \;, \tag{2.1}$$

where β_1 is an appropriate function to capture the effect of initial conditions, and $\beta_2, \beta_3, ..., \beta_K$ are coefficients which measure the effect of the exogenous variables. In principle these can depend on the complete history of states and actions: in practice dimensionality can be managed by allowing them to depend only on recent history. The regrets μ measure the effects of actions and complete the three components. Since the $\{Z_j\}$ are unknown, Henderson *et al.* propose they be estimated by residuals from models for S_j on previous states and actions $(\bar{S}_{j-1}, \bar{A}_{j-1})$. If linear models are used then the residuals are orthogonal to the covariates. Thus, we can separate the effect of exogenous variables from the effect of earlier decisions, as displayed in panel (c) of Figure 1. See [6] for further information.

2.2. Estimation

Moodie *et al.* [11] provide a very clear description of the estimation procedures proposed by Robins and Murphy. We provide only a brief outline here. The blips of Robins [14] can be obtained by first obtaining constructed variables which estimate at each j the response under the optimal policy:

$$H_j(\theta) \;=\; Y + \sum_{k \geq j} \Big\{ \gamma_j(d_j^{\text{opt}}|\mathcal{F}_j;\theta) - \gamma_j(A_j|\mathcal{F}_j;\theta) \Big\} \;.$$

A user-specified vector $V_j(A_j)$ of length $\dim(\theta)$ is then specified. By construction $H_j(\theta)$ is independent of $V_j(A_j)$ and so

$$0 \;=\; \sum_j H_j(\theta) \Big\{ V_j(A_j) - E\big[V_j(A_j)|\bar{S}_j, \bar{A}_{j-1}\big] \Big\}$$

is an unbiased estimating equation.

Murphy [13] takes a different approach. She defines a sum of squares involving two versions of the parameter vector ψ, say ψ and ψ^*, together with a stabilising constant c:

$$
\begin{aligned}
f_n(\psi, \psi^*, c) \;=\; & \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \Bigg(Y^i + c + \sum_{l=1, l\neq j}^{K} \mu_l(\bar{S}_l^i, \bar{A}_l^i; \psi) \\
& + \mu_j(\bar{S}_j^i, \bar{A}_j^i; \psi^*) \;-\; \hat{E}_{A_j}\Big(\mu_j(\bar{S}_j^i, \bar{A}_j^i; \psi^*) | \bar{S}_j^i, \bar{A}_{j-1}^i\Big) \Bigg)^2 .
\end{aligned}
$$

Murphy shows that consistent estimation is possible through an iterative procedure to find $(\psi, \hat{c})$ such that

$$f_n(\psi, \psi, \hat{c}) \;\leq\; f_n(\psi, \psi^*, c)$$

for all (ψ^*, c). Note that this is not the same as minimising f.

The estimation methods of Robins and Murphy are at best computationally challenging. By contrast, the approach of Almirall *et al.* [1] and Henderson *et al.* [6] is based on a model for the observed response (2.1) which means standard methods are available. Henderson *et al.* propose ordinary least squares between observed and expected responses, which is valid without any distributional assumption for responses. More efficient procedures may be possible if further assumptions are made.

2.3. Illustration

We will illustrate using a simple two-timepoint example with Normal states and binary actions as also used by Moodie *et al.* [11]. Data were generated as $S_1 \sim \mathrm{N}(450, 150^2)$, $A_1 \sim \mathrm{Bern}(0.5)$, $S_2 \sim \mathrm{N}(1.25\, S_1, 60^2)$ and $A_2 \sim \mathrm{Bern}(0.5)$. Blip functions were parametrised, leading to regrets

$$
\mu_1(a_1 | S_1; \psi) \;=\; \begin{cases} \;\;\, I(a_1 = 0)\,(\psi_{10} + \psi_{11} S_1)\,, & \psi_{10} + \psi_{11} S_1 > 0 \;, \\ -I(a_1 = 1)\,(\psi_{10} + \psi_{11} S_1)\,, & \psi_{10} + \psi_{11} S_1 < 0 \;, \end{cases}
$$

$$
\mu_2(a_1 | \bar{S}_2, A_1; \psi) \;=\; \begin{cases} \;\;\, I(a_1 = 0)\,(\psi_{20} + \psi_{21} S_1)\,, & \psi_{20} + \psi_{21} S_2 > 0 \;, \\ -I(a_1 = 1)\,(\psi_{20} + \psi_{21} S_2)\,, & \psi_{20} + \psi_{21} S_2 < 0 \;, \end{cases}
$$

and then response $Y \sim \mathrm{N}\Big(400 + 1.6\, S_1 - \mu_1(A_1 | S_1; \psi) - \mu_2(S_1 | \bar{S}_2, A_1; \psi), 60^2\Big)$.

Table 1 compares G-estimation as used by Moodie *et al.* with the regret-regression method proposed by [6]. For the latter we used ordinary least squares to fit the correctly specified model

$$E[Y | \bar{S}_2, \bar{A}_2] \;=\; \beta_0 + \beta_1 S_1 - \mu_1(A_1 | S_1; \psi) - \mu_2(S_1 | \bar{S}_2, A_1; \psi) \;.$$

The nlm routine in R was used for parameter estimation. In all simulations the algorithm converged very quickly. Both methods produce apparently unbiased estimators, as they should, with smaller standard errors under the regret-regression method.

Table 1: Summary of simulation results based on Moodie *et al.* scenario. One thousand repetitions at sample size $n = 500$.

True ψ	G-estimation*		Regret-regression	
	Mean	SE	Mean	SE
250.0	250.01	17.20	250.20	11.39
−1.0	−1.00	0.04	−1.00	0.03
720.0	720.30	24.05	719.85	10.82
−2.0	−2.00	0.04	−2.00	0.02

* These results are taken from Moodie *et al.* (2007), who used the doubly robust form of G-estimation: their equation (2), which is the most efficient of the methods they considered.

Table 2 investigates how estimated parameters translate into decision regime performance. One thousand repetitions at sample size $n = 500$ were generated. After each repetition a further 10 000 observations were generated using each of four different decision rules: the gold standard of always choosing the optimal decision; equally likely randomised decisions; and following the estimated decision rules obtained from the first stage data by G-estimation of the regret functions and by the regret-regression procedure. Column $\bar{Y}$ gives the mean achieved response for each procedure, and column "Err" gives the overall percentage of times a suboptimal decision was made, pooled over both decision times. Columns 'Cut 1' and 'Cut 2' summarise the estimated cutpoints at each decision time, with the true values given in the gold standard row. Again we see that both G-estimation and regret-regression perform well, with again less variability when regret-regression is used.

Table 2: Further summary of simulation results based on Moodie *et al.* scenario. See text for explanation.

	$\bar{Y}$	SE	Err	Cut 1	SE	Cut 2	SE
Gold	1120.1	2.4	0.0	250.0		360.0	
Random	780.0	3.5	50.0				
Regrets (G-est.)	1119.6	2.8	0.6	249.9	9.9	359.5	12.7
Regret-regression	1120.0	2.5	0.3	250.5	6.3	359.9	2.6

3. REGRET-REGRESSION FOR A TWO-ARM BANDIT PROBLEM

The methods summarised above were developed with the aim of causal inference from observational data. In this section we argue that they can also be applied to problems from the stochastic optimisation literature. We will illustrate using the classic two-arm bandit problem.

At time j the state value S_j is a 2-vector (M_{0j}, M_{1j}), where $M_{0j} \in \{0, 1\}$ is the value of *arm zero* and $M_{1j} \in \{0, 1\}$ of *arm one*. The action A_j is to choose one of the arms. Response Y is then incremented by a reward which depends on the current value of the chosen arm. In our example the rewards are 6 or 4 for the two values of arm zero, and 8 or 3 for the two values of arm one. If arm zero is selected then M_{0j} is updated for time $j+1$ according to a Markov chain but M_{1j} remains at its previous value. The opposite happens if arm one is selected: M_{1j} is updated but M_{0j} is unchanged. In our example the transition matrices are

$$P_0 = \begin{pmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{pmatrix} \qquad \text{and} \qquad P_1 = \begin{pmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{pmatrix}.$$

This is a special case of the so-called multi-armed bandit problem. A single resource is available to process a collection of competing projects (arms) over an infinite horizon. At each decision time $j = 0, 1, ...,$ a decision must be taken as to which arm will be selected for processing. If arm k is chosen at time j then a discounted reward of

$$\lambda^j R_k(M_{kj})$$

is gained, where $\lambda \in [0, 1)$ is a discount rate, $R_k(\cdot)$ is a reward function and M_{kj} is the value of a Markov chain modelling the evolution of arm k at time j. After a unit of time dedicated to project k, it changes state according to a Markov law of motion P_k. The states of the other arms remain unchanged.

The objective is to find a policy for allocating arms for processing that maximises the total expected discounted reward over an infinite horizon. In principle, for particular problems the use of dynamic programming and the application of Bellman's principle of optimality [2] would allow these classical problems to be solved. However, as the size of the problem increases, the computational difficulties become intractable. Additionally, no insight into the structure of the optimal policy is obtained. An alternative method of solution, based around *forwards induction*, was introduced by Gittins and Jones [5]. They defined a dynamic allocation index (DAI) as

$$G_k(x_k) \;=\; \sup_{\tau > 0} \frac{E\left[\sum_{t=0}^{\tau-1} \lambda^t R(M_t) | M_0 = x_k\right]}{1 - E\left[\lambda^\tau\right]} \quad ,$$

where the bandit is initially in state x_k and τ is a positively valued stopping time defined on the process. The Gittins Index policy is the one that selects the arm with the current largest DAI. Such policies, since Whittle [17], are now referred to as *Gittins Index policies*. There are a number of methods for calculating the Gittins index including direct calculation, calibration methods, linear programming and special purpose algorithms; see [4] for more details. The "largest to smallest algorithm" [15] was implemented for the illustration here.

For the special two-arm two-value case described at the opening of this section, the Gittins Index policy under almost no discounting ($\lambda = 0.9999$) is to choose arms 1,0,1,1 for states $(0{,}0), (0{,}1), (1{,}0)$ and (1,1) respectively. Note that a play-the-winner rule which optimises current reward would be 1,0,1,0 in the same order. The difference is at state (1,1) where the rewards on offer are (4,3). The Gittins policy of choosing arm one acknowledges future expectation — the possibility that the arm one reward value could change from 3 to 8 — whereas the play-the-winner rule is myopic and takes the higher immediate reward of 4 on offer from arm zero.

The Gittins policy is derived under an assumption that the process continues indefinitely and the optimal policy is stationary. We can use the regret-regression method to examine optimal dynamic policies for fixed length horizons. We simply simulate the process with actions chosen randomly and then fit a linear model incorporating regrets and residuals from dummy variables to describe the values. After each action the model includes residuals associated with eight dummy variables: one for each state/action combination. We choose optimal actions by working from the final timepoint and changing the action to ensure regrets are positive, starting with a working guess at which actions are optimal. Since linear models are used, this is a trivial task even when large samples are used to smooth out the noise generated by the Markov chains.

Table 3 illustrates for $K = 5$, showing the optimal action for each state S_j ($j = 1, 2, ..., 5$) for this problem, along with the regrets for choosing a suboptimal action. It is interesting to compare the optimal choices with the Gittins policy. In states (0,0) and (1,0) they are the same: choose action $A_j = 1$ and hence take reward 8 units. State (1,1) has Gittins and optimal actions the same at $A_j = 1$ until time $j = 5$ at which final time the higher short-term reward under action $A_j = 0$ should be taken. State (0,1) also has a change in optimal action near the end, but this time at the penultimate decision stage. When in this state at earlier times, the optimal dynamic policy is to choose action $A_j = 1$ whereas the stationary Gittins policy is to choose $A_j = 0$.

For reference we give the mean reward under four decision regimes:

Regime	Mean Y
Random, prob 0.5	25.2
Play-the-winner	26.0
Gittins	27.1
Optimal dynamic	27.8

Table 3: Optimal actions and regrets for two-arm bandit problem with horizon $K = 5$. The first reward is obtained if $A_j = 0$, the second if $A_j = 1$. See text for other parameter values.

State S_j	Rewards	Optimal action				
		$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
(0,0)	6 or 8	1	1	1	1	1
(0,1)	6 or 3	1	1	1	0	0
(1,0)	4 or 8	1	1	1	1	1
(1,1)	4 or 3	1	1	1	1	0

State S_j	Rewards	Regret				
		$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
(0,0)	6 or 8	0.32	0.32	0.32	0.80	2.00
(0,1)	6 or 3	0.32	0.32	0.32	0.40	3.00
(1,0)	4 or 8	0.88	0.88	0.88	1.60	4.00
(1,1)	4 or 3	0.88	0.88	0.88	0.40	1.00

4. DIAGNOSTICS

We return to the general problem of Section 2 and focus on the regret-regression approach based on (2.1). An advantage of this approach is that we model the actual responses and hence can obtain residuals between observed and fitted values. Plots of residuals against covariates, fitted values, selected actions or estimated regrets can be used for diagnostic assessment and model comparisons. However, we have made no assumptions on response Y other than independence and our model (2.1) for the mean. In particular we have not assumed homogeneity of variance, which implies that whilst there should be no trends in the means of plots of residuals there may well be systematic patterns in the scatter, even for a correctly specified model. Further, standard bootstrap methods can be problematic when observations are independent but not identically distributed.

We propose to test for trend in residual plots using the wild bootstrap or conditional multiplier method [7], [10]. Suppose we have variables $\{D_i\}$ $(i = 1, 2, ..., n)$ which are independent with zero mean and finite but not necessarily equal variance. Suppose further that $T_0 = n^{-1/2} \sum_{i=1}^{n} D_i$ converges in distribution to some variable D. Let $\{\xi_i\}$ $(i = 1, 2, ..., n)$ be independent and identically distributed with zero mean and unit variance. Then $T_1 = n^{-1/2} \sum_{i=1}^{n} \xi_i D_i$ also converges in distribution to D. The wild bootstrap resampling method is to generate N independent copies of $\{\xi_i\}$ and use the resulting N copies of T_1 as an empirical estimator of the distribution of T_0. Note that all original variables D_i contribute exactly once to each T_1: there is no omission or duplication as in the standard bootstrap.

A complication is that residuals $R_i = Y_i - E[Y_i|\bar{S}_{iK}, \bar{A}_{iK}]$ are not independent. Our proposal is to base a test statistic on a contrast: $T_0 = n^{-1/2}\sum_{i=1}^{n} c_i R_i$ where $\sum_{i=1}^{n} c_i = 0$. We then obtain N resamples of $T_1 = n^{-1/2}\sum_{i=1}^{n} \xi_i c_i R_i$ and we compare the observed T_0 with the empirical distribution of T_1 to obtain a test of trend. The detail is as follows:

1. Order the residuals against a chosen covariate (or the fitted value).
2. Divide the residuals into six equally sized groups 1 to 6 corresponding to lowest sixth to highest sixth covariate values (with minor adjustments below if the six groups cannot be equal).
3. Select a contrast set from the following:

Test		Contrast coefficients c					
		1	2	3	4	5	6
T_1	Trend	1	1	1	−1	−1	−1
T_2	Curvature	1	1	−2	−2	1	1
T_3	Lower tail	1	−1	0	0	0	0
T_4	Upper tail	0	0	0	0	1	−1

4. Compute T_0 with the chosen contrasts for the six groups. Compute N wild bootstrap versions as described above using standard Normal $\{\xi_i\}$ and obtain an empirical p-value as the proportion of resampled test statistics which are more extreme than T_0.

Simulation results (not shown) indicate that all of the tests have the correct size for correctly specified models and that none uniformly dominates for power. We propose that all four be adopted in practice and in addition we recommend a fifth test based on the extremum of the cumulative residuals:

$$T_5 = \max_j \left\{ \sum_{i=1}^{j} R_i \right\}.$$

5. APPLICATION

Rosthøj *et al.* [16] and Henderson *et al.* [6] describe analyses of data on warfarin treatment of patients on long term anticoagulation. There are 303 patients with 14 clinic visits each. At each visit the International Normalised Ratio (INR) of blood clotting time was recorded, along with the *change* in prescribed dose of anticoagulant. If INR is too high then patients have risk of severe bleeding, whereas if INR is too low then there is risk of thrombosis. The aim therefore is to adjust dose to maintain as closely as possible INR within a target range, which can depend on underlying condition of the patient.

The response variable Y used in previous analysis is the overall percentage time the patient spent in range (PTR), which was to be maximised. The state variable S_j used by [16] and [6] is a standardised version of the INR, defined to be zero if the patient has INR in range, and otherwise the scaled distance to the nearest target boundary, with scaling by the population standard deviation. This ensures comparability between patients with different conditions and target intervals. The action variable A_j is the change in dose, in mg Warfarin. The first four visits are considered as a stabilisation period, and since there is no information after the final visit this is not used for the analyses to come. Thus $K=9$ and the data for analysis consist of states $S_1, S_2, ..., S_9$ and actions $A_1, A_2, ..., A_9$ for the 303 patients. Henderson *et al.* [6] also worked with a discretised state S_j^* given by

$$S_j^* \;=\; \begin{cases} 1\,, & S_j \le -0.3 \qquad \text{(very low)}\,, \\ 2\,, & -0.3 < S_j < 0 \quad \text{(low)}\,, \\ 3\,, & S_j = 0 \qquad\quad\;\; \text{(in range)}\,, \\ 4\,, & 0 < S_j < 0.55 \quad \text{(high)}\,, \\ 5\,, & S_j \ge 0.55 \qquad\; \text{(very high)}\,. \end{cases}$$

Rosthøj *et al.* used the methods of [13] and were able to fit only one very simple regret model:

$$\mu_j(a_j|\mathcal{F}_j) \;=\; \begin{cases} I(a_j \neq 0)\big(5.84 + 1.59\, a_j^2\big)\,, & S_j = 0\,, \\ 0.24\big(a_j + 2.01\, S_j\big)^2\,, & S_j \neq 0\,. \end{cases} \tag{5.1}$$

Here the optimal decision by construction is to leave dose unchanged if INR is within range, and is otherwise to change in proportion to state. For high states the dose should be increased so as to reduce clotting time, and the opposite for low states. The regret for a suboptimal decision increases quadratically as the dose change moves away from optimal. The model is overly simple and not claimed to be realistic, but Rosthøj *et al.* were unable to obtain convergence of either the G-estimation or iterative methods (see Section 2) for more realistic models.

Henderson *et al.* used the regret-regression approach based on (2.1) and had no difficulty in fitting more realistic models. Their final selection assumed that the regret function depended on the current discretised state S_j^* and the previous standardised state S_{j-1}. For category s of S_j^* the model is:

$$\mu_j(a_j|\mathcal{F}_j, S_j^* = s; \psi) \;=\; \psi_{s1} f(A_j - \psi_{s2} - \psi_{s3} S_{j-1})\;, \tag{5.2}$$

where $f(u) = u$ if $u \ge 0$ and $f(u) = u^2$ otherwise. Parameter estimates and bootstrap standard errors from 100 resamples are given in the upper part of Table 4.

Table 4: Parameter estimates and bootstrap standard errors for anticoagulation model example. Upper section: $\lambda = 1$ and analysis as [6]. Lower section: $\lambda = 0.3$.

$S_j^* = s$	ψ_{s1}	SE	ψ_{s2}	SE	ψ_{s3}	SE
−2	0.67	0.32	2.15	0.27	−1.11	0.38
−1	0.38	0.11	2.74	0.18	−1.57	0.67
0	0.97	0.36	−0.14	0.32	−1.12	0.74
1	2.38	0.27	−2.33	0.26	−0.98	0.27
2	2.83	0.79	−3.00	0.44	0.25	0.21
1	0.28	0.17	1.86	0.33	−1.05	0.58
2	0.12	0.11	3.00	0.43	−1.54	0.81
3	0.23	0.27	−0.10	0.17	−1.21	0.75
4	1.24	0.37	−1.57	0.42	−0.27	0.43
5	1.47	0.60	−1.98	0.69	0.39	0.39

To illustrate our diagnostic test suggestion, we will consider residuals from the two fitted models plotted against the regret following the first considered visit time. Figure 2 shows that the residuals from model (5.1) are more variable than those from model (5.2), with perhaps more evidence of trend in the early and later segments. To investigate, we applied the five wild bootstrap tests of Section 4. p-values from 200 wild bootstrap samples are given in Table 5. They confirm the early and late trends for model (5.1) are significant and the model is not fully adequate, but there are no significant trends for model (5.2).

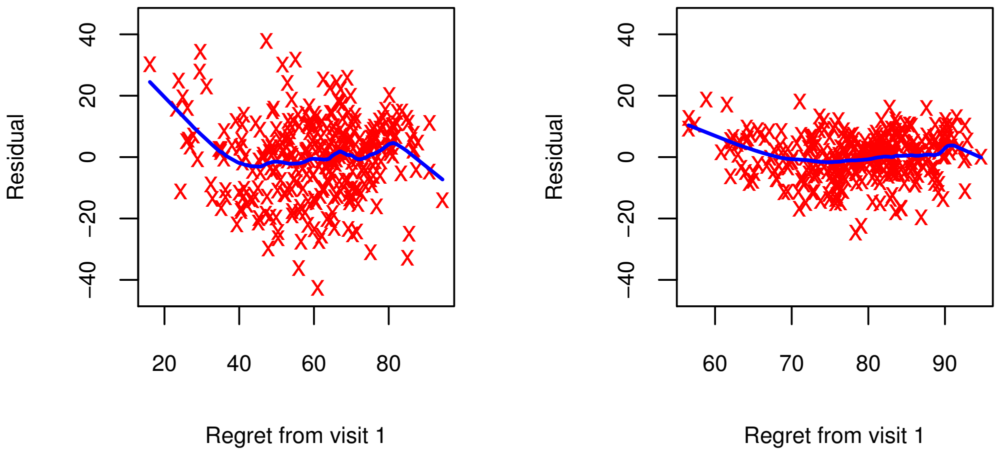

Figure 2: Warfarin residuals against regret at time 1. Left plot: model (5.1); right plot: model (5.2). The solid line is a smooth through the data.

Table 5: Wild bootstrap p-values for residuals in Figure 2.

Model	Test				
	T_1	T_2	T_3	T_4	T_5
(5.1)	0.830	0.000	0.593	0.018	0.055
(5.2)	0.753	0.611	0.136	0.870	0.816

We summarise now a new analysis of the warfarin data with a revised response variable. As well as a decision on the dose to be taken, at each clinic visit there is also a recommendation as to the timing of the next visit. Generally overly frequent visits are discouraged. Letting $N(\tau)$ be the number of visits in follow-up time τ we propose a new response

$$Y = Y(\tau) = \lambda\, PTR(\tau) + (1-\lambda)\frac{\tau}{N(\tau)}\ , \qquad 0 \le \lambda \le 1\ , \tag{5.3}$$

which weights together percentage time in range and average time between visits. Overly frequent visits thus reduce the response. For the warfarin data $N(\tau)$ is fixed at nine visits of interest but the time τ taken varies considerably between patients.

We will use model (5.2) for analysis. Choosing $\lambda = 1$ gives the previous results. To explore, we also analysed for a variety of other values for λ. To illustrate, the lower part of Table 4 gives parameter estimates at $\lambda = 0.3$ together with bootstrap standard errors obtained from 100 resamples. The general trend against s is the same as for $\lambda = 1$ but since the response is on a different scale it is hard to make a direct comparison. Instead, in Figure 3 we show the estimated optimal actions at a variety of combinations of current and previous state. The crosses indicate the values obtained when $\lambda = 1$ and the other points indicate values at a sequence of decreasing λ. As expected, increase in dose is indicated when INR is low, and decrease when INR is high, with previous INR moderating the action. Generally there is little effect of λ except at high INR, where the recommendation would be to reduce dose by a smaller amount if timing of visits is of interest. The rationale is that large dose changes are usually followed by quick return visits to monitor the effect. If this is to be discouraged then more modest changes are recommended. There lack of effect of λ at the low values of INR reflects the asymmetry in risk: very low values of INR need immediate strong action.

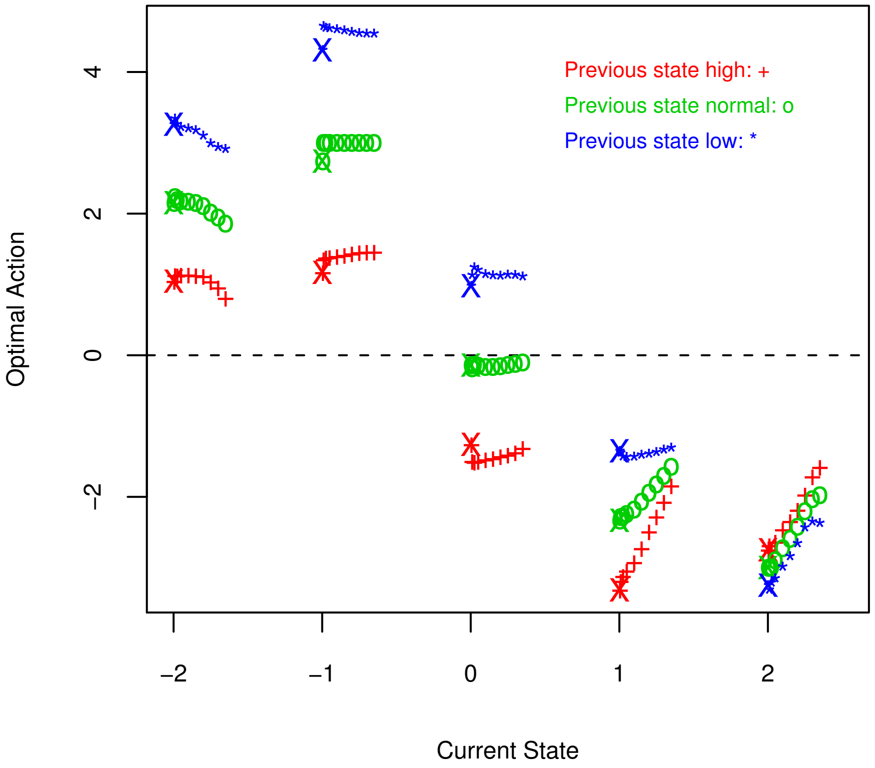

Figure 3: Effect of changing λ in response 5.3. The crosses mark optimal actions when $\lambda = 1$. The other points show how the optimal action changes as λ varies through $\{0.995, 0.99, 0.95, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3\}$, moved to the right for display purposes.

6. DISCUSSION

We have presented an overview of the structural nested mean model approach to optimal dynamic treatment regime determination, with focus on blip models [14], regret models [13], and regret-regression models [1], [6]. Although there has been growing discussion in the literature on causal inference for dynamic treatment regimes, the area is still very much underdeveloped and there are few genuine applications in realistic problems. One issue is the computational challenge faced for reasonable sized data sets. Another is the assumption of balanced data, in the sense of common clinic or visit times. Methods which allow irregular timing of visits are needed. In this case the definition of regrets and blips is problematic. The counting process approach may be fruitful [9] but much further research is needed. Nonetheless we see great promise in this type of approach.

ACKNOWLEDGMENTS

We are grateful to the organisers of the meeting *Statistical Modelling: Challenges in Health, 2010*, for the invitation to present this work.

REFERENCES

[1] Almirall, D.; Ten Have, T. and Murphy, S.A. (2010). Structural nested mean models for assessing time-varying effect moderation, *Biometrics*, to appear.

[2] Bellman, R.E. (1957). *Dynamic Programming*, Princeton University Press, Princeton.

[3] Chakraborty, and Murphy, S.A. (2009). Inference for nonregular parameters in optimal dynamic treatment regimes, *Statistical Methods in Medical Research*, **19**, 317–343.

[4] Gittins, J.C. (1989). *Multi-Armed Bandit Indices*, Wiley, Chichester.

[5] Gittins, J.C. and Jones, D.M. (1974). A dynamic allocation index for the sequential design of experiments, *Progress in Statistics, European Meeting of Statisticians*, **1**, 241–266.

[6] Henderson, R.; Ansell, P. and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes, *Biometrics*, to appear.

[7] Lin, D.Y.; Fleming, T.R. and Wei, L.J. (1994). Confidence bands for survival curves under the proportional hazards model, *Biometrika*, **81**, 73–81.

[8] Lok, J.; Gill, R.; van der Vaart, A. and Robins, J.M. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models, *Statistica Neerlandica*, **58**, 1–25.

[9] Lok, J. (2008). Statistical modeling of causal effects in continuous time, *Annals of Statistics*, **36**, 1464–1507.

[10] Martinussen, T. and Scheike, T.H. (2006). *Dynamic Regression Models for Survival Data*, Spring-Verlag, New York.

[11] Moodie, E.; Richardson, T.S. and Stephens, D. (2007). Demystifying optimal dynamic treatment regimes, *Biometrics*, **63**, 447–455.

[12] Moodie, E.E.M. and Richardson, T.S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null, *Scandinavian Journal of Statistics*, **37**, 126–146.

[13] Murphy, S. (2003). Optimal dynamic treatment regimes (with discussion), *Journal of the Royal Statistical Society Series B*, **65**, 331–366.

[14] Robins, J.M. (2004). Optimal structured nested models for optimal sequential decisions. In "Proceedings of the Second Seattle Symposium on Biostatistics" (D.Y. Lin and P.J. Heagerty, Eds.), Springer, New York, 189–326.

[15] Robinson, D.R. (1982). Algorithms for evaluating the dynamic allocation index, *Operations Research Letters*, **1**, 72–74.

[16] Rosthøj, S.; Fullwood, C.; Henderson, R. and Stewart, S. (2006). Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach, *Statistics in Medicine*, **25**, 4197–4215.

[17] Whittle, P. (1980). Multi-armed bandits and the Gittins index, *Journal of the Royal Statistical Society Series B*, **42**, 143–149.