J. R. Statist. Soc. B (2003)
65, *Part* 2, *pp.* 331–366

Optimal dynamic treatment regimes

S. A. Murphy

University of Michigan, Ann Arbor, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, October 16th, 2002*, Professor D. Firth *in the Chair*]

Summary. A dynamic treatment regime is a list of decision rules, one per time interval, for how the level of treatment will be tailored through time to an individual's changing status. The goal of this paper is to use experimental or observational data to estimate decision regimes that result in a maximal mean response. To explicate our objective and to state the assumptions, we use the potential outcomes model. The method proposed makes smooth parametric assumptions only on quantities that are directly relevant to the goal of estimating the optimal rules. We illustrate the methodology proposed via a small simulation.

Keywords: Adaptive strategies; Causal inference; Dynamic programming; Multistage decisions

1. Introduction

Dynamic treatment regimes are individually tailored treatments that are designed to provide treatment to individuals only when and if they need the treatment. In contrast with classical treatments in which all individuals are assigned the same level and type of treatment, dynamic treatments explicitly incorporate the heterogeneity in need for treatment across individuals and the heterogeneity in need for treatment across time within an individual. In a dynamic treatment regime, decision rules for how the dosage level and type should vary with time are specified before the beginning of treatment; these rules are based on time-varying measurements of subject-specific need. The set of decision rules comprises the treatment regime.

Dynamic treatment regimes are also called adaptive strategies (Lavori and Dawson, 2000) or adaptive interventions (Collins *et al.*, 2001). When the treatment is the provision of health information designed to induce an improvement in health-related behaviours, dynamic regimes are called tailored communications (Kreuter and Strecher, 1996; Kreuter *et al.*, 2000). Dynamic treatment regimes are attractive to public policy makers because they treat only subjects who show a need for treatment, freeing public and private funds for more intensive treatment of the needy. They hold the promise of reducing non-compliance by subjects due to overtreatment or undertreatment (Lavori and Dawson, 2000; Collins *et al.*, 2001). These regimes are intended to reduce negative side-effects due to overtreatment (Bierman *et al.*, 2001). Dynamic regimes are used in tailoring health information content to provide only personally relevant information with the idea that this information will be attended to, thoughtfully processed and thus efficacious (Kreuter *et al.*, 1999).

The goal of this paper is to provide a method for estimating optimal decision rules; rules that when implemented over a time period will produce the highest mean response at the end of the

Address for correspondence: S. A. Murphy, Department of Statistics, 439 West Hall, University of Michigan, Ann Arbor, MI 48109-1092, USA.
E-mail: samurphy@umich.edu

1369–7412/03/65331

time period. The methodology proposed will use experimental or observational longitudinal data to construct estimators of the optimal decision rules. Estimating the effects of dynamic treatment regimes has been studied at length by Robins and colleagues (Robins, 1986, 1989, 1993, 1997; Murphy *et al.*, 2002; van der Laan *et al.*, 2001).

This work is motivated by the Fast Track prevention programme. This is an on-going randomized trial of a complex preventive intervention *versus* control. The intervention was designed to prevent the emergence of and to reduce the level of conduct disorders and drug use in children considered at risk owing to elevated behavioural problems in kindergarten (Bierman *et al.*, 1996; Conduct Problems Prevention Research Group, 1999a, b; McMahon *et al.*, 1996). Part of the intervention involved the implementation of a dynamic treatment regime that was designed to improve family functioning. The Fast Track team did not want to provide the highest level of home visiting to all families. It was thought that providing too many home visits might be detrimental, increasing the risk for family dependence and pejorative labelling (by self and others), and cause attrition. They decided to implement a dynamic treatment as follows. At the end of each semester, beginning with the spring semester of first grade, the family counsellor filled in a six-item questionnaire composed of questions concerning the quality of parenting and family functioning. The sum is the family functioning status. The rule for assigning the number of home visits in the following semester is

$$d_j(S_j) = 16\, I\{S_j \leqslant 8\} + 8\, I\{9 \leqslant S_j \leqslant 16\} + 4\, I\{17 \leqslant S_j\}, \qquad j = 1, 2, 3, 4,$$

where S_j is the family functioning status taken at the beginning of the jth semester, with low values indicating greater need. When this pioneering study was designed there was very little guidance in terms of how one might formulate the decision rule(s). Collins *et al.* (2001) sought to provide qualitative guidance; the goal of this paper is to provide quantitative guidance by exploring methods for estimating good rules.

The ascertainment of optimal dynamic treatment regimes belongs to the class of sequential or multistage decision problems. We consider dynamic treatment regimes in which decisions are to be made at set times; in this case the regime is a set of decision rules, with one rule per time interval. For each time interval j in $\{1, 2, \ldots, K\}$, denote the treatment decision that is to be made by a_j and denote the status (possibly a vector) at the beginning of time interval j by S_j. In general S contains predictors of the response. The jth decision rule will use the information that is available at time j and will output the treatment decision a_j. In a redesign of the Fast Track study we might want to consider a wide variety of information as part of S including information resulting from the detailed summer interviews by outside staff and severity in other domains such as the development of academic and social skills. The response at the end of time interval K is denoted by Y. So the order of occurrence is $S_1, a_1, S_2, \ldots, a_K, Y$. In general we use a bar over a variable to denote that variable and all past values of the same variable, so $\bar{a}_j = (a_1, \ldots, a_j)$.

In many applications an expert provides the multivariate distribution of $(\bar{S}_K, Y)$ indexed by the decisions $a_1, \ldots, a_K$. In this case, one traditionally uses backward induction (dynamic programming) to find decision rules resulting in a maximal mean response. These arguments are usually expressed as follows (Bather, 2000; Jordan and Bishop, 2001). Set

$$J_0(\bar{S}_K, \bar{a}_{K-1}) = \sup_{a_K}(E[Y|\bar{S}_K, \bar{a}_{K-1}, a_K]),$$

$$d_K^*(\bar{S}_K, \bar{a}_{K-1}) = \arg \sup_{a_K}(E[Y|\bar{S}_K, \bar{a}_{K-1}, a_K])$$

and then for each j

$$J_{K-j}(\bar{S}_j, \bar{a}_{j-1}) = \sup_{a_j}\{E[J_{K-j-1}(\bar{S}_{j+1}, \bar{a}_j)|\bar{S}_j, \bar{a}_{j-1}, a_j]\}, \tag{1}$$

$$d_j^*(\bar{S}_j, \bar{a}_{j-1}) = \arg\sup_{a_j}\{E[J_{K-j-1}(\bar{S}_{j+1}, \bar{a}_j)|\bar{S}_j, \bar{a}_{j-1}, a_j]\}.$$

The optimal rules are $d_1^*, \ldots, d_K^*$. It is important to recognize that the placement of $\bar{a}_{j-1}$ and a_j to the right of the '|' sign is to indicate that these conditional expectations are taken with respect to the multivariate distribution of $(S_1, S_2, \ldots, S_K, Y)$ *indexed* by the decisions $\bar{a}_j$; $\bar{a}_j$ play a role similar to parameters. For example, $E[Y|\bar{S}_K, \bar{a}_{K-1}, a_K]$ is the conditional mean of Y given $\bar{S}_K$ indexed by the sequence of the decisions $\bar{a}_K$; it is not the conditional mean of Y given $\bar{S}_K$ and 'other random variables' $\bar{a}_K$. Equation (1) is a finite time version of Bellman's equation (Bellman, 1957). The function J_{K-j} is usually called the 'optimal cost-to-go' from the present state $(\bar{S}_j, \bar{a}_{j-1})$ over the future intervals of time (Bertsekas and Tsitsiklis, 1996); we call J_{K-j} the 'optimal benefit-to-go' as we wish to maximize the mean response rather than to minimize the mean cost. Cowell *et al.* (1999) described the backward induction algorithm and also provided an alternative viewpoint using decision potentials.

For $K = 2$ the above is use of backward induction to find

$$(d_1^*, d_2^*) = \arg\sup_{d_1, d_2}\{E[E[E[Y|\bar{S}_2, a_1, a_2 = d_2(\bar{S}_2, a_1)]|S_1, a_1 = d_1(S_1)]]. \tag{2}$$

The formula for $K > 2$ is similar but long; see Cowell *et al.* (1999), chapter 8. Statisticians who are unfamiliar with dynamic programming may find the above objective function obtuse; in the next section we use Rubin's causal model to provide an alternative form for the objective function.

Bather (2000) has given a nice introduction to and discussion of these types of problem (i.e. the multivariate distribution of $(\bar{S}_K, Y)$ indexed by the decisions $\bar{a}_K$ is known or can be sampled). Although the steps in the dynamic programming algorithm are easily described, they are computationally complex owing to the alternating steps of maximizing and averaging (Bertsekas and Tsitsiklis (1996), page 3). These problems are of great interest in the engineering literature, where the decision rules are called feed-back control policies (Bertsekas and Tsitsiklis, 1996). The wide applicability of dynamic programming in finding optimal decisions combined with the computational difficulties has spawned much research. Indeed the literature for this setting is vast and spans many disciplines including management science, reinforcement learning, medical decision-making and statistics. Some recent work by statisticians in this setting includes Shachter (1986), Owens *et al.* (1997), Cowell *et al.* (1999), chapter 8, and Lauritzen and Nilsson (2001) all of whom used operations on influence diagrams to ascertain optimal sequential decision rules. Carlin *et al.* (1998) and Bielza *et al.* (2001) calculated optimal decision rules by using Monte Carlo methods to simulate from the known multivariate distribution.

Our goal is to propose methodology for estimating the optimal rules when the multivariate distribution of $(\bar{S}_K, Y)$ indexed by the decisions $\bar{a}_K$ is unknown, but experimental or observational longitudinal data are available. To do this we proceed as follows. First the counterfactual or potential outcome framework is used to provide a specification of equation (2), thus providing an alternative view of the use of dynamic programming in ascertaining optimal sequential decisions. We use this framework to formulate assumptions that justify the use of dynamic programming when only experimental or observational longitudinal data are available. Next we demonstrate that if our goal is to estimate optimal rules then it is unnecessary to estimate the full multivariate distribution of the longitudinal data. That is, in Section 3, we model this multivariate distribution with two groups of parameters that vary independently. The first group of parameters (parameters in the 'regret' functions) will be estimated and used to estimate the optimal rules and the second group of parameters (most of which are infinite dimensional) are nuisance parameters. This approach will permit us to make smoothness (i.e. parametric) assumptions on quantities that are directly relevant for estimating the optimal rules; we avoid making smoothness assumptions on other aspects of the data distribution. In Section 4 we illustrate a method that permits us to estimate the parameters in the regret functions without estimating the nuisance parameters. More importantly this method will provide a computational alternative to the interweaving maximization and averaging steps of the dynamic programming algorithm. The last section provides simulation results that illustrate the method proposed.

2. Potential outcomes and dynamic programming

Neyman (1990) introduced potential outcomes to analyse the causal effect of time-independent treatments in randomized studies. Rubin (1978) explicated Neyman's ideas and extended Neyman's work to the analysis of causal effects of time-independent treatments from observational data. Robins (1986, 1987) proposed a formal theory of causal inference that extended both Neyman's and Rubin's work to assess the direct and indirect effects of time-varying treatments from experimental and observational longitudinal studies. We use these works to specify our objective and to state the assumptions.

In the following we define the potential outcomes; these potential outcomes will be related to the observable data later. We use upper case Roman letters to denote random variables and lower case Roman letters to denote non-random variables. Since in dynamic treatment regimes we only manipulate or assign treatments, the potential outcomes are indexed only by treatments. Furthermore, we assume that a subject's outcomes are not influenced by other subjects' treatments so we index each subject's potential outcomes by only his or her treatments (see Cox (1958) and Rubin (1986) for a more complete discussion). Thus corresponding to each fixed value of the treatment vector $\bar{a}_K$ we conceptualize a potential response denoted by $Y(\bar{a}_K)$ where $Y(\bar{a}_K)$ is the response at the end of the Kth interval that a subject would have if he or she followed the treatments $\bar{a}_K$. Let $\mathcal{A}_K$ be the collection of all possible K-vectors of treatments decisions. The set of all potential responses is $\{Y(\bar{a}_K) : \bar{a}_K \text{ varying in } \mathcal{A}_K\}$. The status at the beginning of time interval j is an intermediate outcome of (past) treatments and thus the set of intermediate outcomes at time j is $\{S_j(\bar{a}_{j-1}) : \bar{a}_{j-1}$ varying in the first $j-1$ components of $\mathcal{A}_K\}$. Denote all the subject's potential outcomes by $O_{sr} = \{S_2(a_1), \ldots, S_{K-1}(\bar{a}_{K-2}), Y(\bar{a}_K); \bar{a}_K \in \mathcal{A}_K\}$.

The potential outcomes model sheds light on the function to be maximized in equation (2) as follows. The mean response for regime $d_1, \ldots, d_K$ is

$$E[Y(\bar{d}_K)] = E[Y(\bar{a}_K)_{a_1=d_1(S_1),\ldots,a_K=d_K\{\bar{S}_K(\bar{a}_{K-1}),\bar{a}_{K-1}\}}] \tag{3}$$

where a bar over a variable is used to denote that variable and all past values of the same variable (e.g. $\bar{a}_K = a_1, \ldots, a_K$ and $\bar{S}_K(\bar{a}_{K-1}) = S_1, S_2(a_1), \ldots, S_K(\bar{a}_{K-1})$). The optimal rules should maximize this mean. The mean can be written as a repeated expectation; in particular for $K = 2$ we have that

$$E[Y(\bar{d}_2)] = E[E[E[Y(\bar{a}_2)|\bar{S}_2(a_1)]_{a_2=d_2\{\bar{S}_2(a_1),a_1\}}|S_1]_{a_1=d_1(S_1)}].$$

As before we may place a_1 and a_2 to the right of the '|' sign to indicate that these conditional expectations are indexed by the decisions. Then an alternative version is

$$E[E[E[Y(\bar{a}_2)|\bar{S}_2(a_1), a_1, a_2 = d_2\{\bar{S}_2(a_1), a_1\}]|S_1, a_1 = d_1(S_1)]]. \tag{4}$$

From the similarity between the above display and equation (2), we see that the dynamic programming algorithm, as expected, has the goal of finding the regime that maximizes the mean response.

Now we connect the potential outcomes with observations in a longitudinal data set and we express formula (4) in terms of these data. The observable data for a subject are $X = \{S_1, A_1, S_2, \ldots, A_K, Y\}$ where $\bar{A}_K$ is the vector of stochastic treatment decisions. $\bar{A}_K$ takes values in $\mathcal{A}_K$. We make Robins's (1997) consistency assumption, i.e. we assume that the potential outcomes are connected to the subject's data by the equalities $Y = Y(\bar{A}_K)$, $S_K = S_K(\bar{A}_{K-1})$, and so on, including $S_2 = S_2(A_1)$. In the following we use either Y or $Y(\bar{A}_K)$ to denote the observed response at the end of time interval K and either S_j or $S_j(\bar{A}_j)$ to denote the observed status at the beginning of time interval j. Thus the observable data are the pretreatment information (S_1) plus the potential outcomes corresponding to the treatment pattern $\bar{A}_K$. Most of a given subject's potential outcomes are missing; only the potential outcomes corresponding to the treatment pattern $\bar{A}_K$ can be observed.

To express formula (4) in terms of the observable data we shall need to make assumptions. To see why consider the following scenario. Suppose that, among individuals with the same past status levels and past treatment levels, the individuals with treatment A_K = high differ from the individuals with treatment A_K = low and the reason for this difference is not contained in the available data. To decide whether the decision high treatment is optimal (i.e. better than the decision low treatment) we compare the average value of Y for those with A_K = high with the average value of Y for those with A_K = low. However, any apparent difference in the two conditional means may be due to the difference in composition between the individuals with treatment A_K = high and the individuals with treatment A_K = low, i.e. A_K may not be conditionally independent of the potential outcomes O_{sr} conditional on $(\bar{S}_{K-1}, \bar{A}_{K-1})$, because there may be unmeasured 'confounders' that determine treatment and are associated with the potential outcomes. In general, assumptions about this distributional relationship must be used to identify causal effects and thus to permit causal inference (for discussion, see section 11 of Robins (1997)). We make the following independence assumption (Robins, 1997) on the relationship between the potential outcomes O_{sr} and the treatment decisions $\bar{A}_K$. We assume *no unmeasured confounders*:

$$\text{for each } j = 1, \ldots, K, A_j \text{ is independent of } O_{sr} \text{ given } \{S_1, A_1, S_2, A_2, \ldots, S_j\}.$$

This assumption is also called sequential ignorability (Robins, 2000). An alternative statement of this assumption should be possible by using the methods in Robins and Greenland (1992) or Dawid *et al.* (2001). These methods replace potential outcomes with potential experiments.

The phrase 'no unmeasured confounders' should not be misinterpreted; perhaps a better phrase would be 'no unmeasured direct confounders', as the assumption is a statement about the selection of treatment conditional on past information. Intuitively this means that an unmeasured confounder may only influence the selection of treatment through the measured past information. The no unmeasured confounders assumption would be true if the treatments are sequentially randomized. Treatments are sequentially randomized when at each time j the treatment A_j is randomized with randomization distribution allowed to depend on $\{S_1, A_1, S_2, A_2, \ldots, S_j\}$ (Robins, 1997). Lavori and Dawson (2000) and Lavori *et al.* (2000) proposed that researchers implement sequentially randomized experiments to estimate optimal decisions rules (instead of sequential randomization they used the phrase biased adaptive within-subject randomization). An additional setting in which the no unmeasured confounders assumption would be true is in computer experiments that are designed using a distribution for the decision A_j, depending only on past information. Of course, for a given observational data set, one may believe that the S_js are sufficiently rich that the no unmeasured confounders assumption holds.

Assuming no unmeasured confounders, we can write equation (3), which is a function of the multivariate distribution of the potential outcomes, as a function of the multivariate distribution of the longitudinal data:

$$E[E[\ldots E[E[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = d_K]|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]\ldots|S_1, A_1 = d_1]] \tag{5}$$

where d_j implicitly denotes $d_j(\bar{S}_j, \bar{A}_{j-1})$. This is Robins's G-computation formula (Robins, 1986, 1987, 1989, 1997; Gill and Robins, 2001). For $K = 2$ expression (5) is

$$E[Y(\bar{d}_2)] = E[E[E[Y|\bar{S}_2, A_1, A_2 = d_2(\bar{S}_2, A_1)]|S_1, A_1 = d_1(S_1)]]. \tag{6}$$

Note the subtle difference between this equation and expression (4) and the repeated expectation in equation (2); expressions (2) and (4) are functions of conditional distributions *indexed* by treatment decisions $\bar{a}_2$ whereas equation (6) is a function of conditional distributions, *conditioning* on the treatment decisions. It may appear to be patently obvious that equation (6) and the repeated expectation in equation (2) should be equal. However, from the discussion preceding the statement of the no unmeasured confounders assumption we know that this may not be true. The beauty of the no unmeasured confounders assumption combined with Robins's G-computation formula is that they provide a means by which we can say that the repeated expectations in equations (2) and (6) are equal. Robins's G-computation formula is the formula that provides the mean response to a dynamic regime in terms of the longitudinal data distribution.

A proof of Robins's G-computation formula is provided by lemma 2 in Appendix A. This proof assumes that the range of the treatment decisions $\bar{A}_K$ is countable ($\mathcal{A}_K$ is countable). Denote the conditional probability function for each A_j given $\bar{S}_j, \bar{A}_{j-1}$, by $p_j(a_j|\bar{S}_j, \bar{A}_{j-1})$. The proof assumes that the regime $\bar{d}_K$ satisfies

$$P\left[\prod_{j=1}^{K} p_j\left\{d_j(\bar{S}_j, \bar{A}_{j-1})|\bar{S}_j, \bar{A}_{j-1}\right\} > 0\right] = 1, \tag{7}$$

i.e. to prove that the repeated expectations in equations (2) and (6) are equal we not only assume no unmeasured confounders but we also make the eminently sensible assumption that treatment patterns which are consistent with the regime $\bar{d}_K$ can occur in the longitudinal data. See Gill and Robins (2001) for more general conditions and proof.

Denote the class of regimes (i.e. the vector of decision rules) satisfying equation (7) by D_P. We subscript D by the probability P since the class of regimes may vary by the distribution of the longitudinal data. For the remainder of the paper we take as our goal that of finding a regime that maximizes expression (5) over the class D_P. The assumption of no unmeasured confounders is only used to justify the use of expression (5) as an appropriate objective function for the purposes of estimating optimal regimes.

For each treatment level a_K satisfying $p_K(a_K|\bar{S}_K, \bar{A}_{K-1}) > 0$ define

$$Q_0(\bar{S}_K, \bar{A}_{K-1}, a_K) = E[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = a_K]$$

Next define

$$J_0(\bar{S}_K, \bar{A}_{K-1}) = \sup_{a_K : p_K(a_K|\bar{S}_K, \bar{A}_{K-1}) > 0} \{Q_0(\bar{S}_K, \bar{A}_{K-1}, a_K)\}.$$

For each $j = 0, \ldots, K-1$ and a_j satisfying $p_j(a_j|\bar{S}_j, \bar{A}_{j-1}) > 0$ iteratively define

$$Q_{K-j}(\bar{S}_j, \bar{A}_{j-1}, a_j) = E[J_{K-j-1}(\bar{S}_{j+1}, \bar{A}_j)|\bar{S}_j, \bar{A}_{j-1}, A_j = a_j]$$

and

$$J_{K-j}(\bar{S}_j, \bar{A}_{j-1}) = \sup_{a_j: p_j(a_j|\bar{S}_j, \bar{A}_{j-1})>0} \{Q_{K-j}(\bar{S}_j, \bar{A}_{j-1}, a_j)\}. \tag{8}$$

We call the functions $J_0, \ldots, J_{K-1}$ the optimal benefit-to-go functions. These functions differ from the displays in equation (1) in two ways: first the above optimal benefit-to-go functions are conditional on treatment decisions whereas in equation (1) they are indexed by treatment decisions and second we maximize over a restricted set of decision rules since we cannot evaluate treatment decisions that cannot occur in the longitudinal data.

Beginning with J_0 and then ascertaining each optimal benefit-to-go function forms the steps of a dynamic programming argument that ends with the maximal value equal to $J_{K-1}(S_1)$. This process is equivalent to maximizing expression (5) over the class D_P. Indeed we have the following theorem.

Theorem 1. Assume that $\mathcal{A}_K$ is countable. Assume that $E[|Y||\bar{S}_K, \bar{A}_K]$ is bounded almost surely. Then,

$$\sup_{\bar{d}_K \in D_P} (E[E[\ldots E[E[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = d_K]|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]\ldots|S_1, A_1 = d_1]]) \tag{9}$$

is equal to $E[J_{K-1}(S_1)]$.

The proof is in Appendix A.

If we assume that the supremum is achieved at a $\bar{d}_K^*$ in D_P (for example this would occur if the number of possible treatments is finite), then we can write the optimal benefit-to-go functions in terms of the potential outcomes for Y. In this case

$$J_{K-j}(\bar{S}_j, \bar{A}_{j-1}) = E[Y(\bar{A}_{j-1}, d_j^*, \ldots, d_K^*)|\bar{S}_j, \bar{A}_{j-1}] \tag{10}$$

for each $j = 1, \ldots, K$. This can be derived by following the same steps as in the proof of theorem 1 (see Appendix A). Thus the optimal benefit-to-go function represents the mean potential response conditional on the past and assuming that optimal decisions will be followed in the future.

3. The regret functions

Our goal is to estimate an optimal regime (i.e. a regime that maximizes expression (5) over the class D_P). An initial approach is to model the multivariate distribution of $(\bar{S}_K, \bar{A}_K, Y)$, say by using a parametric, semiparametric or nonparametric model, and then to apply the dynamic programming argument. By a careful choice of a semiparametric model we can avoid the dynamic programming step. Instead of parameterizing conditional mean functions or conditional distribution functions we parameterize and then estimate regret functions. At the end of this section we provide the semiparametric model that is induced by a parameterization of the regrets.

The regret functions are defined for each $j = 1, \ldots, K$ as

$$\mu_j(\bar{S}_j, \bar{A}_{j-1}, a_j) = J_{K-j}(\bar{S}_j, \bar{A}_{j-1}) - Q_{K-j}(\bar{S}_j, \bar{A}_{j-1}, a_j).$$

The μs satisfy the constraint

$$\inf_{a_j: p_j(a_j|\bar{S}_j, \bar{A}_{j-1})>0} \{\mu_j(\bar{S}_j, \bar{A}_{j-1}, a_j)\} = 0.$$

The regret function $\mu_j(\bar{S}_j, \bar{A}_{j-1}, a_j)$ provides the increase in the benefit-to-go that we forego by making decision a_j rather than the optimal decision at time j, i.e. the predictor effect is measured in terms of changes from the optimal predictor value. This is most clearly seen when the supremum in expression (9) is achieved at a $\bar{d}^*_K \in D_P$, in which case the regret is given by (using the potential outcomes model)

$$\mu_j(\bar{S}_j, \bar{A}_{j-1}, a_j) = E[Y(\bar{A}_{j-1}, d^*_j, \ldots, d^*_K)|\bar{S}_j, \bar{A}_{j-1}] - E[Y(\bar{A}_{j-1}, a_j, d^*_{j+1}, \ldots, d^*_K)|\bar{S}_j, \bar{A}_{j-1}].$$

It is clear that the backward induction argument can be based on minimizing the regrets instead of maximizing the Q-functions. Thus if we had estimates of the regrets we could then derive estimates of the optimal rules.

We directly model the regrets; this model may be nonparametric, semiparametric or parametric. Estimation is discussed in the next section. One possible class of parametric models is based on a known ‘link’ function $f(u)$; these functions provide the link between the regret and the decision rule. The minimal value of each f should be achieved at $u = 0$ and be equal to 0 ($f(0) = 0$); so each f is a non-negative function. For a positive scale parameter $\eta_j(\bar{s}_j, \bar{a}_{j-1})$ set

$$\mu_j(\bar{s}_j, \bar{a}_j) = \eta_j(\bar{s}_j, \bar{a}_{j-1})\, f\{a_j - d_j(\bar{s}_j, \bar{a}_{j-1})\}. \tag{11}$$

The constraints on the link function imply that $d_j(\bar{s}_j, \bar{a}_{j-1})$ is the optimal decision rule based on past information $(\bar{s}_j, \bar{a}_{j-1})$. We form parsimonious parametric models for the optimal decision $d_j(\bar{s}_j, \bar{a}_{j-1})$ and for the scale parameter $\eta_j(\bar{s}_j, \bar{a}_{j-1})$. Parsimony can be important; in many settings a simple rule is easier to implement than a complicated rule.

The shape of the link f determines how the modelled regret will change as a treatment decision deviates from the optimal. To provide flexibility in the rate at which the modelled regret changes as a treatment decision deviates from the optimal, we model and estimate a multiplicative unknown scale parameter $\eta_j(\bar{s}_j, \bar{a}_{j-1})$. Large values of $\eta_j(\bar{s}_j, \bar{a}_{j-1})$ imply that a small difference in the treatment decision from the optimal produces a large regret and vice versa. Suppose that the possible decisions are real values (e.g. doses of a drug); then we might believe that the regret will have a ‘U’-shape so that for doses smaller than the optimal the subject receives insufficient treatment, yet for doses larger than the optimal the subject suffers toxicity or side-effects and thus the subject does not benefit as much as if the optimal dose is delivered. In this case we might use $f(u) = u^2$. Alternatively we might believe that the particular treatment does not cause toxicity or side-effects; in this case the link might be positive quadratic for $u < 0$ and equal to 0 thereafter. In the future it will be important to model the link rather than to assume it known.

To provide a more flexible regret we can combine two links; for example we might base a model for the regret on

$$\mu_j(\bar{s}_j, \bar{a}_j) = \eta_j(\bar{s}_j, \bar{a}_{j-1})\, f\{a_j - d_j(\bar{s}_j, \bar{a}_{j-1})\} + \eta'_j(\bar{s}_j, \bar{a}_{j-1})\, f'\{a_j - d_j(\bar{s}_j, \bar{a}_{j-1})\}$$

where $f(u) = u^2\, I\{u \leqslant 0\}$ and $f'(u) = u^2\, I\{u \geqslant 0\}$. We can allow the link function to change by interval and/or by past information. This is particularly relevant if the type of decision may change by interval and/or if the type of decision differs by most recent status and most recent past decision.

If in an interval two different decisions must be made, then we may make the decisions sequentially, forming two ‘intervals’ from the one and using a different link for each type of decision. For example suppose that in each interval educational staff must make two decisions; first the staff member must decide whether the child is to receive special education or not. If the child is to receive special education then the staff member must recommend a certain number of minutes of special education per day; otherwise the staff member must recommend a number of tutoring sessions per week. First we break each interval into two intervals corresponding to the two decisions; then we might use equation (11) for a given link f and $a_j \in \{0, 1\}$ denoting special education by 1 and no special education by 0. Then for possibly different link functions f' and f'' the regret for the second decision would be

$$\begin{aligned}\mu_{j+1}(\bar{s}_{j+1}, \bar{a}_{j+1}) = I\{a_j = 1\}\, &\eta_{j+1}(\bar{s}_{j+1}, \bar{a}_j)\, f'\{a_{j+1} - d_{j+1}(\bar{s}_{j+1}, \bar{a}_j)\} \\ &+ I\{a_j = 0\}\, \eta_{j+1}(\bar{s}_{j+1}, \bar{a}_j)\, f''\{a_{j+1} - d_{j+1}(\bar{s}_{j+1}, \bar{a}_j)\}\end{aligned}$$

where the link f' is used to parameterize the regret when we are choosing between minutes of special education and the link f'' is used to parameterize the regret when we are deciding the frequency of tutoring sessions. In the next section we provide a method to estimate the regret functions.

We can write the mean of Y given $(\bar{S}_K, \bar{A}_K)$ in terms of the regrets

$$E[Y|\bar{S}_K, \bar{A}_K] = \mu_0 + \sum_{j=1}^{K} \phi_j(\bar{S}_j, \bar{A}_{j-1}) - \sum_{j=1}^{K} \mu_j(\bar{S}_j, \bar{A}_j) \tag{12}$$

where $\mu_0 = E[J_{K-1}(S_1)]$ and the ϕ_js are defined so that the right-hand side is equal to the left-hand side; $\phi_j(\bar{S}_j, \bar{A}_{j-1}) = J_{K-j}(\bar{S}_j, \bar{A}_{j-1}) - Q_{K-j+1}(\bar{S}_{j-1}, \bar{A}_{j-1})$ for $j = 1, \ldots, K$. Note that $E[\phi_j(\bar{S}_j, \bar{A}_{j-1})|\bar{S}_{j-1}, \bar{A}_{j-1}]$ is 0. As mentioned previously, parameterizing the regrets induces a semiparametric model for the longitudinal data X. When Y is continuous this model is given by

$$g\left[y - \mu_0 - \sum_{j=1}^{K} \{\phi_j(\bar{s}_j, \bar{a}_{j-1}) - \mu_j(\bar{s}_j, \bar{a}_j)\}|\bar{s}_K, \bar{a}_K\right] \prod_{j=1}^{K} p_j(a_j|\bar{s}_j, \bar{a}_{j-1}) \prod_{j=1}^{K} f_j(s_j|\bar{s}_{j-1}, \bar{a}_{j-1})$$

where

(a) $g(\cdot|\bar{s}_K, \bar{a}_K)$ is the mean 0 conditional density of Y given $(\bar{S}_K, \bar{A}_K)$ and must belong to the class of mean 0 densities for fixed values of $(\bar{s}_K, \bar{a}_K)$,

(b) $p_j(a|\bar{s}_j, \bar{a}_{j-1})$ denotes the conditional probability function for each A_j given $(\bar{S}_j, \bar{A}_{j-1})$ at times $j = 1, \ldots, K$ and each p_j must belong to the class of probability functions in a for fixed values of $(\bar{s}_j, \bar{a}_{j-1})$,

(c) $f_j(s|\bar{s}_{j-1}, \bar{a}_{j-1})$ is the conditional density of S_j given $(\bar{S}_{j-1}, \bar{A}_{j-1})$ at times $j = 1, \ldots, K$, and each f_j must belong to the class of probability densities in s for fixed values of $(\bar{s}_{j-1}, \bar{a}_{j-1})$,

(d) $\mu_j(\bar{s}_j, \bar{a}_j)$ is the jth regret $J_{K-j}(\bar{s}_j, \bar{a}_{j-1}) - Q_{K-j}(\bar{s}_j, \bar{a}_j)$ and each μ_j must belong to the class of functions satisfying $\inf_{a_j : p_j(a_j|\bar{s}_j, \bar{a}_{j-1}) > 0}\{\mu_j(\bar{s}_j, \bar{a}_j)\} = 0$,

(e) $\phi_j(\bar{s}_j, \bar{a}_{j-1})$ is $J_{K-j}(\bar{s}_j, \bar{a}_{j-1}) - Q_{K-j+1}(\bar{s}_{j-1}, \bar{a}_{j-1})$ and must belong to the class of functions satisfying $\int \phi_j(\bar{s}_j, \bar{a}_{j-1})\, f_j(s_j|\bar{s}_{j-1}, \bar{a}_{j-1})\; \mathrm{d}s_j = 0$ and

(f) μ_0 is $E[J_{K-1}(S_1)]$ and takes values on the real line. μ_0 is the mean response to the optimal decision regime.

As discussed above we are primarily interested in the parameters composing the regrets μ_j; the unknown functions g, p_js, f_js, ϕ_js and scalar μ_0 are nuisance parameters.

3.1. Advantages of modelling the regrets

Modelling the regrets has several nice conceptual and practical properties. First, we parameterize the optimal rules, i.e. we impose parsimony on aspects of the multivariate distribution of $(\bar{S}_K, \bar{A}_K, Y)$ that are of direct relevance for our goal. In particular, this approach combined with the estimation method to follow permits the straightforward use of statistical methods such as hypothesis testing and model selection. Thus we can test whether particular features of the past information are needed in the optimal rule.

A second nice property is that the constraints on the form of the optimal decision are made explicit. To highlight this second property we contrast the direct parametric modelling of the regrets with parameterizing both $Q_0(\bar{S}_K, \bar{a}_K) = E[Y|\bar{S}_K, \bar{A}_K = \bar{a}_K]$ and the density of S_j given $(\bar{S}_{j-1}, \bar{A}_{j-1})$ for each j and then by alternating supremum and averaging steps forming estimates of all the Q-functions. This alternative approach leads to implicit constraints on the form of the regret functions $\mu_1, \ldots, \mu_{K-1}$ and the $Q_1, \ldots, Q_{K-1}$-functions. Implicit constraints on $Q_1, \ldots, Q_{K-1}$ occur because these are joint functions of Q_0 and the conditional distributions of the S_js.

The finite dimensional parametric models for Q_0 and the distribution of S_j given $(\bar{S}_{j-1}, \bar{A}_{j-1})$ also constrain the form of the regret as follows. According to the dynamic programming algorithm we would first maximize Q_0 over a_K, yielding $J_0(\bar{S}_K, \bar{a}_{K-1})$. Thus the parameterization of Q_0 determines the form of J_0. From equation (12) we have that

$$J_0(\bar{S}_K, \bar{a}_{K-1}) = \mu_0 + \sum_{j=1}^{K} \phi_j(\bar{S}_j, \bar{a}_{j-1}) - \sum_{j=1}^{K-1} \mu_j(\bar{S}_j, \bar{a}_j)$$

(recall that $Q_0 = E[Y|\bar{S}_K, \bar{A}_K = \bar{a}_K]$). Thus the parameterization of Q_0 determines the shape of this function, particularly in a_{K-1} via the term $\phi_K(\bar{S}_K, \bar{a}_{K-1}) - \mu_{K-1}(\bar{S}_{K-1}, \bar{a}_{K-1})$. At the same time, the parametric model for the distribution of S_K given $(\bar{S}_{K-1}, \bar{A}_{K-1})$ constrains the shape of $\phi_K(\bar{s}_K, \bar{a}_{K-1})$ in a_{K-1} since $E[\phi_K(\bar{S}_K, \bar{A}_{K-1})|\bar{S}_{K-1}, \bar{A}_{K-1})] = 0$. These two constraints lead to a limited set of forms for the regret function. This is particularly a problem when the parametric models are non-linear. This situation is similar to that highlighted by Robins and Wasserman (1997) who, in testing for effects of treatment decisions, found that the hypothesis of no treatment effect may be excluded by the explicit restrictions imposed by parametric models on other, less scientifically interesting, parts of the multivariate data distribution. Because the constraints are implicit it is difficult in any given situation to check how much and in what way they constrain the regret function.

An additional nice property resulting from direct models for the regrets is that parameterization of the regrets does not place constraints on other aspects of the multivariate distribution of $(\bar{S}_K, \bar{A}_K, Y)$ when $E[Y|\bar{S}_K, \bar{A}_K]$ is unbounded. For continuous Y with unbounded support, this can be seen from the likelihood provided earlier. Since Y has unbounded support there is no *a priori* restriction on the value of $E[Y|\bar{S}_K, \bar{A}_K]$. Thus, as can be seen from the likelihood, μ_j and p_j are variation independent of the nuisance parameters μ_0, g, ϕ_j and f_j, for $j = 1, \ldots, K$. We are interested in parameters composing μ_j and will use models for p_j; parametric models for the μ_js and p_js do not constrain the possible values of the nuisance parameters and all modelling assumptions are explicit, i.e. we know that we are not accidently making implicit, perhaps untenable, assumptions about nuisance parts of the multivariate distribution. To highlight this property consider an alternative approach of directly parameterizing each of the Q-functions. Note that

$$Q_1(\bar{S}_{K-1}, \bar{a}_{K-1}) = E[\inf_{a_K}\{Q_0(\bar{S}_K, \bar{a}_K)|\bar{S}_{K-1}, \bar{A}_{K-1} = \bar{a}_{K-1}\}].$$

Thus parametric models for Q_1 and Q_0 constrain the conditional distribution of S_K given $(\bar{S}_{K-1}, \bar{A}_{K-1})$. This is a well-known problem regarding the compatibility of marginal and conditional models (Q_0 plays the role of the conditional mean and Q_1 plays the role of the marginal mean). For a now classic example, see Hougaard (1986), who showed that, if we assume that the proportional hazards model from survival analysis holds for both marginal and conditional models, then subject to natural conditions the mixing density (here the density of S_K given $(\bar{S}_{K-1}, \bar{A}_{K-1})$) must be a positive stable distribution with infinite mean. Thus the restrictions on the conditional distribution of S_K given $(\bar{S}_{K-1}, \bar{A}_{K-1})$ can be quite surprising.

Bertsekas and Tsitsiklis (1996), working in a similar setting to that in Section 1 (i.e. the distribution of $\bar{S}_K$ and Y, indexed by the treatment decisions, is known or can be simulated from), illustrated the approximation of the Q-functions by neural network architectures such as splines, wavelets and classical neural networks. Brockwell and Kadane (2001) used a discretization method combined with the use of 'features' summarizing the past history to approximate the Q-functions. These approaches share the first property with the method proposed here, i.e. they impose parsimony on aspects of the multivariate distribution of $(\bar{S}_K, \bar{A}_K, Y)$ that are of direct relevance for the goal of estimating an optimal regime. The architectures discussed by Bertsekas and Tsitsiklis (1996) can be considered nonparametric modelling methods when S can assume only a few values and the possible decisions are small; in this case this method shares the third property of avoiding implicit constraints on other aspects of the multivariate distribution. However, in many cases S may assume many values or the set of possible decisions is not small; then owing to the curse of dimensionality this approach must be considered parametric and thus implicit restrictions are placed on the conditional density of each S_j given $\bar{S}_{j-1}$.

A fourth nice property of this modelling approach is that it directly leads to a simple estimator of the mean response to the optimal dynamic regime. This follows from equation (12) which implies that

$$E[J_{K-1}(S_1)] = \sum_{j=1}^{K} E[\mu_j(\bar{S}_j, \bar{A}_j)] + E[Y].$$

Under the assumption of no unmeasured confounders, $E[J_{K-1}(S_1)]$ is the mean response to an optimal dynamic regime. Thus, given estimators of the regrets and a sample ($\bar{S}_{Ki}$, $\bar{A}_{Ki}$, Y_i, $i = 1, \ldots, n$), an estimator of the mean response to an optimal dynamic regime is

$$n^{-1} \sum_{i=1}^{n} \left\{ \sum_{j=1}^{K} \hat{\mu}_j(\bar{S}_{ji}, \bar{A}_{ji}) + Y_i \right\}.$$

All four properties are analogues of properties that Robins (1986, 1987, 1989, 1997) established for structural nested mean models. Structural nested mean models are models for effects relative to the predictor value of 0, whereas the models presented here are for effects relative to the optimal predictor value. Additionally, the likelihood here is similar in form to the likelihood for Robins's structural nested mean models.

4. Estimation

Estimation procedures can be based on the following least squares characterization of the regret functions. Denote the true regret functions with a subscript of 0, i.e. $\mu_{01}, \ldots, \mu_{0K}$. Also we assume that the support of each conditional probability function $p_j(a|\bar{s}_j, \bar{a}_{j-1})$ for given

$(\bar{s}_j, \bar{a}_{j-1})$ is known. Thus the infimum in the definition of the regret function is over a known set of decisions.

Theorem 2. Assume that both Y and each component of $\bar{\mu}_{0K}$ are square integrable. Then given a vector of square integrable functions $\bar{\mu}_K$, where each μ_j satisfies both $\inf_{a: p_j(a|\bar{S}_j, \bar{A}_{j-1})>0}\{\mu_j(\bar{S}_j, \bar{A}_{j-1}, a)\} = 0$ and

$$E\left[Y + \sum_{l=1}^{K} \mu_l(\bar{S}_l, \bar{A}_l) - \sum_a \mu_j(\bar{S}_j, \bar{A}_{j-1}, a)\, p_j(a|\bar{S}_j, \bar{A}_{j-1})\right]^2$$
$$\leqslant E\left[Y + \sum_{l=1, l\neq j}^{K} \mu_l(\bar{S}_l, \bar{A}_l) + m_j(\bar{S}_j, \bar{A}_j) - \sum_a m_j(\bar{S}_j, \bar{A}_{j-1}, a)\, p_j(a|\bar{S}_j, \bar{A}_{j-1})\right]^2 \qquad (13)$$

for all square integrable m_j, $j = 1, \ldots, K$, we have that $\bar{\mu}_K$ is almost surely equal to $\bar{\mu}_{0K}$. We can replace Y by $Y + c$ for a scalar c and/or we can replace the $\Sigma_{l=1}^{K}$ by $\Sigma_{l\geqslant j}^{K}$ and the sum $\Sigma_{l=1, l\neq j}^{K}$ by $\Sigma_{l>j}^{K}$ and in both cases the same result holds.

See Appendix A for a proof of this result.

There are various ways to base estimation of the regret functions on the above least squares characterization. For simplicity suppose that the conditional probability functions $\bar{p}_K$ are known. First we formulate a model for the regrets with a p-dimensional unknown parameter β, say $\mu_j(\bar{s}_j, \bar{a}_j; \beta)$, so that $\inf_a\{\mu_j(\bar{s}_j, \bar{a}_{j-1}, a; \beta)\} = 0$. Then we replace the expectation in inequality (13) by an average over the data with the aim of finding a $\hat{\beta}$ for which

$$\mathbb{P}_n\left\{Y + \sum_{l=1}^{K} \mu_l(\bar{S}_l, \bar{A}_l; \hat{\beta}) - \sum_a \mu_j(\bar{S}_j, \bar{A}_{j-1}, a; \hat{\beta})\; p_j(a|\bar{S}_j, \bar{A}_{j-1})\right\}^2$$
$$\leqslant \mathbb{P}_n\left\{Y + \sum_{l=1, l\neq j}^{K} \mu_l(\bar{S}_l, \bar{A}_l; \hat{\beta}) + m_j(\bar{S}_j, \bar{A}_j) - \sum_a m_j(\bar{S}_j, \bar{A}_{j-1}, a)\; p_j(a|\bar{S}_j, \bar{A}_{j-1})\right\}^2$$

for chosen m_j, $j = 1, \ldots, K$ (for a function f of the ith subject's data X_i, $\mathbb{P}_n\{f(X)\}$ is defined as $(1/n)\Sigma_{i=1}^{n} f(X_i)$, assuming that the sample observations are independent draws from a distribution). In the simulations, we use the model for the regrets as the m_j.

In general the conditional probability functions $\bar{p}_K$ for the longitudinal data will be unknown (however, we continue to assume that the support of each conditional probability function $p_j(a|\bar{s}_j, \bar{a}_{j-1})$ for given $(\bar{s}_j, \bar{a}_{j-1})$ is known). These densities can be estimated by postulating a model and then using maximum likelihood, i.e., given the model $p_j(a|\bar{s}_j, \bar{a}_{j-1}; \alpha)$, $j = 1, \ldots, K$, with unknown parameter α, maximize the log-likelihood

$$\mathbb{P}_n\left[\sum_{j=1}^{K} \log\{p_j(A_j|\bar{S}_j, \bar{A}_{j-1}; \alpha)\}\right]$$

to find $\hat{\alpha}_n$. In this paper we assume that the support of each p_j does not vary by the unknown parameter α. To estimate the parameters in the regret functions, we search for a $(\hat{\beta}_n, \hat{c}_n)$ for which

$$\sum_{j=1}^{K} \mathbb{P}_n\left\{Y + \hat{c}_n + \sum_{l=1}^{K} \mu_l(\bar{S}_l, \bar{A}_l; \hat{\beta}_n) - \sum_a \mu_j(\bar{S}_j, \bar{A}_{j-1}, a; \hat{\beta}_n)\; p_j(a|\bar{S}_j, \bar{A}_{j-1}; \hat{\alpha}_n)\right\}^2$$

$$\leqslant \sum_{j=1}^{K} \mathbb{P}_n \left\{ Y + c + \sum_{l=1, l \neq j}^{K} \mu_l(\bar{S}_l, \bar{A}_l; \hat{\beta}_n) + \mu_j(\bar{S}_j, \bar{A}_j; \beta) \right.$$
$$\left. - \sum_a \mu_j(\bar{S}_j, \bar{A}_{j-1}, a; \beta)\, p_j(a|\bar{S}_j, \bar{A}_{j-1}; \hat{\alpha}_n) \right\}^2 \qquad (14)$$

for all β and c. (The inclusion of the unknown scalar c does not change consistency but greatly improves the stability of the algorithm.) One way to implement the search computationally is to start with an initial value of $\hat{\beta}_n$, say $\hat{\beta}_{n1}$, substitute $\hat{\beta}_{n1}$ for $\hat{\beta}_n$ in inequality (14), minimize over (β, c), to obtain $(\hat{\beta}_{n2}, \hat{c})$, discard $\hat{c}$, replace $\hat{\beta}_n$ in inequality (14) with $\hat{\beta}_{n2}$ and iterate the minimization process until convergence. The least squares criterion uses the conditional variability of the decision levels in the data to estimate the optimal decisions. Intuitively this can be seen by acting as if the above least squares function (14) is smooth in β and differentiating with respect to β and c. The A_js minus their conditional expectations given the past play an analogous role to covariates in a linear regression. This means, quite naturally, that, if given the past information A_j is deterministic, this method cannot lead to estimators of the decision rules.

When the modelled regrets and optimal decisions are smooth functions of β, simple Taylor series arguments can be used to derive an estimator of the asymptotic variance of $\hat{\beta}_n$. The formula is provided in Appendix A. An estimator of the mean response μ_0 to an optimal dynamic regime is

$$\hat{\mu}_0 = \mathbb{P}_n \left\{ \sum_{j=1}^{K} \mu_j(\bar{S}_j, \bar{A}_j; \hat{\beta}) + Y \right\}.$$

The formula for the asymptotic variance $\hat{\mu}_0$ is also in Appendix A.

4.1. Further comments

The least squares characterization in inequality (13) and/or (14) leads to relatively simple computations in contrast with the very natural approach of first estimating $Q_0(\bar{S}_K, \bar{a}_K) = E[Y|\bar{S}_K, \bar{A}_K = \bar{a}_K]$ and the density of S_j given $(\bar{S}_{j-1}, \bar{A}_{j-1})$ for each j and second carrying out the interwoven supremum and expectation steps composing the dynamic programming algorithm. See Bertsekas and Tsitsiklis (1996), page 3, for comments on the computational issues. Furthermore, as discussed at the end of Section 2, to avoid imposing implicit constraints on the decision rule, we would want to make the models for $Q_0(\bar{S}_K, \bar{a}_K)$ and the conditional densities of S_j given the past nonparametric. Yet if the dimension or number of values that S can assume and/or the number of decisions is large the curse of dimensionality will result in highly variable estimators.

Note that the above estimation method depends on a correct parameterization of the conditional probability functions $\bar{p}_K$. Thus we are assuming an overall model for the multivariate distribution of the data that stipulates a parametric model for the regrets $\bar{\mu}_K$ and a parametric form for the conditional probability functions $\bar{p}_K$ (all other parts are nonparametric).

In general, estimators based directly on the least squares characterization in inequality (13) and/or (14) will not lead to efficient estimators of β. There are several options to remedy this situation. If the dimensionality of the problem is small so that we avoid the curse of dimensionality (e.g. K is small and the status measure can assume only a few values and the possible treatment levels are few) then we can use nonparametric models to estimate each component of the multivariate distribution and then use dynamic programming. If this method is feasible then we do not need to assume a parametric model for the conditional probability functions. Alternatively we can adjust the above estimator to achieve double robustness (and a 'local' type of efficiency; Robins (2000)). Conceptually this is a straightforward adaptation of Robins's (2000) work on structural nested mean models. A third option is to change the model from a model in which most parts of the multivariate distribution are left unspecified (i.e. nonparametric) to a model in which the entire multivariate distribution is parametric; then we may use the likelihood to construct maximum likelihood estimators of the regret functions.

5. Simulation results and further discussion

The goal of the simulations is to demonstrate that the estimation methodology proposed here is promising and deserves further investigation. We simulate data from the following prototypical sequentially randomized experiment. This type of experiment would be used to estimate optimal decision rules, rules that maximize academic achievement Y at the end of $K = 10$ intervals. In each interval there are two decisions. The first decision is whether the child should receive special education (1, yes; 0, no). This decision is binary. If the child is recommended for special education then a level 1 or greater special education per day is to be assigned. If the child is not to receive special education then an amount of tutoring per week, 0 or greater, is assigned. Each decision should be based on the child's status as assessed at prior time intervals and at the beginning of the present time interval $(S_1, \ldots, S_K)$. In the simulation of the sequentially randomized experiment, the treatment assignment probabilities for the decision about whether a child should receive special education are uniform on $\{0, 1\}$. The treatment amount assignment probabilities for special education are uniform on 1, 2, 3 and for reading tutoring are uniform on 0, 1, 2, 3. Although these treatment assignment probabilities do not depend on the past, in a sequentially randomized experiment we can allow the treatment assignment probabilities to depend on the past statuses and treatments.

We divide each interval into two subintervals; in the first subinterval the 'yes–no' special education decision is made and in the second subinterval the amount of the appropriate treatment is to be decided. Thus the effective number of intervals is $2K = 20$. The remaining data are simulated as follows. The conditional density of Y, $g(\cdot|\bar{s}_2, \bar{a}_2)$, is a normal, mean 0, variance 0.64 density. The marginal density of S_1 and f_1 is normal with mean $\text{mean}_1 = 0.5$ and variance 0.01. The conditional density of each S_j given $(\bar{S}_{j-1}, \bar{A}_{j-1})$, $j > 2$ and odd, is normal, mean $\text{mean}_j = 0.5 + 0.2S_{j-1} - 0.07A_{j-1}A_{j-2} - 0.01A_{j-1}(1 - A_{j-2})$, where A_{j-2} assumes values 1 or 0 according to whether special education is assigned. The conditional variance of S_j is set to 0.01. For j even, S_j is set equal to S_{j-1}. Except for the regret functions the terms in the telescoping form of the conditional mean of Y are the ϕ_js and μ_0. These are set as $\phi_j(\bar{s}_j, \bar{a}_{j-1}) = -5(s_j - \text{mean}_j)$ for j odd, and $\mu_0 = 30$. Recall that μ_0 is the optimal mean response. There are no ϕ_j-terms for j even.

In simulation 1 the true regret for j odd (decide to assign special education or not) is given by $6(a_j - I\{s_j > 5/9\})^2$ and the true regret for j even is given by $1.5a_{j-1}(a_j - 2s_j)^2 + 1.5(1 - a_{j-1})(a_j - 5.5s_j)^2$ (the treatment amounts are allowed to be continuous dosage levels). This specification of the regrets means that the simulated mean $E[Y|\bar{A}_{20} = \bar{a}_{20}, \bar{S}_{20} = \bar{s}_{20}]$ is given by

$$30 - 5 \sum_{j=1, j \text{ odd}}^{20} (s_j - \text{mean}_j) - \sum_{j=1, j \text{ odd}}^{20} 6(a_j - I\{s_j > 5/9\})^2$$
$$- \sum_{j=2, j \text{ even}}^{20} 1.5a_{j-1}(a_j - 2s_j)^2 + 1.5(1 - a_{j-1})(a_j - 5.5s_j)^2.$$

The regrets are 0 at the optimal decision; thus, for j odd, $d_j^* = I\{s_j > 5/9\}$ and, for j even, $d_j^* = 2a_{j-1}s_j + 5.5(1 - a_{j-1})s_j$.

In the analysis of each simulated sequentially randomized experiment, we fit a quadratic link ($f(u) = u^2$). In the odd time intervals the fitted regret for the yes–no special education decision is

$$\mu_j(\bar{s}_j, \bar{a}_j) = \beta_1 \left[a_j - \frac{\exp\{30(s_j - \beta_2)\}}{1 + \exp\{30(s_j - \beta_2)\}} \right]^2 \tag{15}$$

for $a_j \in \{0, 1\}$ (1, yes; 0, no), i.e. we approximate the non-smooth $I\{s_j > \beta_2\}$ by the smooth function $\exp\{30(s_j - \beta_2)\}/[1 + \exp\{30(s_j - \beta_2)\}]$. The choice of 30 is arbitrary; in the future the pros and cons of other choices should be considered. In the even time intervals the fitted regret for the amount of treatment (special education or tutoring) is

$$\mu_{j,2}(\bar{s}_j, \bar{a}_j) = \beta_4 a_{j-1}\{a_j - (\beta_3 + \beta_5 s_j)\}^2 + \beta_7(1 - a_{j-1})\{a_j - (\beta_6 + \beta_8 s_j)\}^2. \tag{16}$$

So the fitted decision functions are smooth: $d_j = \exp\{30(s_j - \beta_2)\}/[1 + \exp\{30(s_j - \beta_2)\}]$ for j odd and $d_j = a_{j-1}(\beta_3 + \beta_5 s_j) + (1 - a_{j-1})(\beta_6 + \beta_8 s_j)$ for j even.

We conducted a number of simulations; here three are discussed. First for illustration we provide, in Table 1, estimates of the βs for simulation 1. It is not surprising that β_1 is poorly estimated since we are fitting only an approximation of the optimal decision rule in the odd time intervals.

Scientific interest is not directly concerned with the accurate estimation of the decision regime but rather is most concerned with the ability of the estimated regime to produce an optimal response. Thus to compare the simulation results we provide box plots and a table that evaluate the mean response to the estimated decision regimes. For each data set in a simulation, we estimate first the βs and then the corresponding decision rule. The estimated rules comprise the estimated regime. Then for each of these estimated decision regimes we calculate the mean response under the estimated treatment regime by Monte Carlo methods. To be more precise, we generate 10 000 observations (complying with the estimated decision regime) and form the mean response. Thus we have a mean response corresponding to each estimated treatment regime in the simulation. Since our simulations are of size 1000 data sets, we have 1000 (Monte-Carlo-estimated) mean responses per simulation: Table 2 and Fig. 1. The first simulation is as above (labelled 'simulation 1'). From Table 2, we see that on average (across estimations of the regimes) the estimated regimes in simulation 1 lead to a mean response of 29.27; recall

Table 1. Estimates of parameters in the decision regime: simulation 1†

True β	*Average $\hat{\beta}$*	*Standard error*	*Average estimated standard error*
6.0	6.89	0.210	0.210
0.56	0.56	0.002	0.002
0.0	0.05	0.184	0.181
1.5	1.50	0.125	0.125
2.0	2.01	0.255	0.247
0.0	0.06	0.128	0.125
1.5	1.48	0.078	0.083
5.5	5.54	0.358	0.358

†Simulations of 1000 data sets of size 1000.

Table 2. Descriptive statistics for the mean response to each of 1000 estimated treatment regimes†

Statistic	*Results for the following simulations:*		
	Simulation 1	*Simulation 2*	*Simulation 3*
Mean	29.27	29.54	28.22
Median	29.27	29.54	28.29
Standard deviation	0.19	0.16	0.47

†The mean response is evaluated by using 10 000 Monte Carlo repetitions. The optimal mean response is 30.

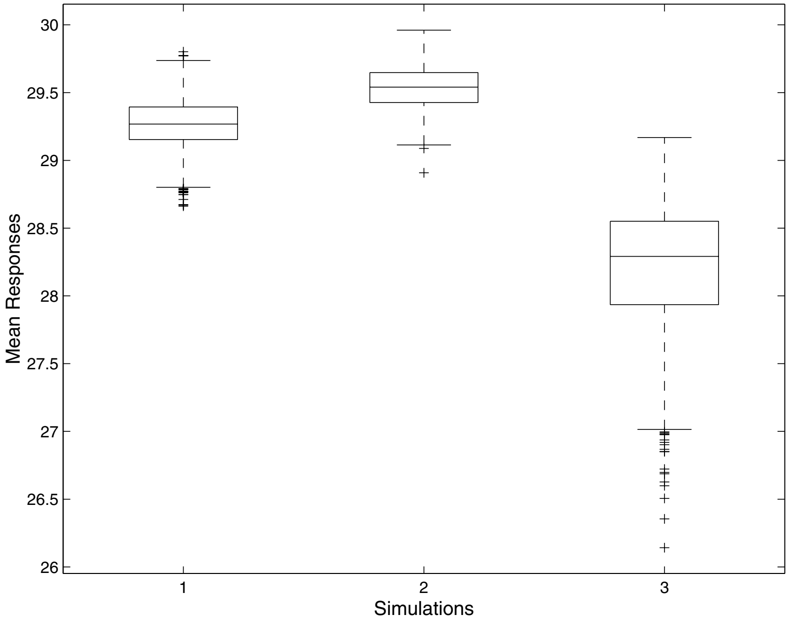

Fig. 1. Mean responses for estimated optimal regimes

that the optimal mean response is 30. This less-than-optimal performance is attributable to the fact that we are approximating the discontinuous decision rule in the odd time intervals with a smooth function. As a comparison, consider the following simple non-dynamic regime. In the simulations the mean of S_j at each time j is between 0.4 and 0.5, indicating that the 'average child' would not have been assigned to special education $(0.5 < 5/9)$. In addition the amount of academic tutoring that is optimally assigned for a child with $S_j = 0.45$ is $5.5 \times 0.45 = 2.475$. So in our simple non-dynamic regime all children are assigned 2.475 units of academic tutoring at each time interval. The use of this simple regime results in an average mean response of 15.98. Thus the use of the estimated optimal rule compared with this simple rule results in an average increase of $29.27 - 15.98 = 13.29$ in the response.

Table 2 and Fig. 1 contain results for two further simulations. These two simulations differ from simulation 1 only in the form of their true (i.e. simulated) regrets. In both cases the true link function is non-quadratic (hence the fitted quadratic link is misspecified). Simulation 2 uses data simulated with the link

$$f(u) = \begin{cases} u^2 & \text{if } u^2 \geqslant 0.83, \\ 0 & \text{otherwise} \end{cases}$$

in the odd-numbered intervals and links

$$f(u) = \begin{cases} u^2 & \text{if } u^2 \geqslant 3.33, \\ 0 & \text{otherwise} \end{cases}$$

for both amounts of special education and tutoring respectively in the even-numbered intervals. From Table 2, under simulation 2, we see that, even though the link function is misspecified, the mean response to the estimated optimal regimes is closer to the optimal value of 30 than in simulation 1 (where the link function is correctly specified). This is not surprising since in simulation 2 the true regrets are 0 for a range of treatment levels; thus the estimated rules only need to provide a treatment level within a range to produce nearly optimal results.

In simulation 3, we consider the opposite situation to that in simulation 2. In simulation 3 the regrets are 'peaked' at the optimal; here the link is

$$f(u) = \begin{cases} |u| & \text{if } u^2 < 1.5, \\ u^2 - 1.5 + \surd 1.5 & \text{otherwise} \end{cases}$$

in the odd-numbered intervals and links

$$f(u) = \begin{cases} |u| & \text{if } u^2 < 2.5, \\ u^2 - 2.5 + \surd 2.5 & \text{otherwise} \end{cases}$$

for both amounts of special education and tutoring respectively in the even-numbered intervals. As might be expected a misspecification of the link (we fit a quadratic link) results in the poorest mean response of the three simulations. Additionally the variability in mean response across the 1000 estimated optimal regimes more than doubles. Here the 'peakedness' of the regret implies that the estimated rules must provide treatment levels that are very close to the optimal treatment levels to produce a nearly optimal response. The poor performance and increase in variability are highlighted in Fig. 1 by the box plots of the mean responses.

In general our simulations indicate that the estimation procedure can be sensitive to misspecifications of the link function and the use of a smooth decision rule to approximate a discrete-valued decision rule; thus in future we plan to explore the usefulness of more flexible link functions and accurate parameterizations of the rules.

5.1. Discussion

Causal Bayesian networks provide a natural alternative to the method presented here. For example, we construct a tentative causal Bayesian network (possibly containing unobserved latent variables) that incorporates the assumption that the associated conditional distributions would be the same even if the decisions are set by outside intervention. We 'learn' the structure of the causal Bayesian network from the data, i.e. we postulate multivariate models for the group of observed and unobserved variables, estimate parameters and assess the fit. This is an area of intense research. An overview of 'learning' the structure of the network with references is provided by Cowell *et al.* (1999), chapter 11. See also Heckerman (1998). Then, given a particular multivariate model with estimated parameters, we follow the dynamic programming arguments. The dynamic programming steps are computationally difficult; Cooper (1990) showed that probabilistic inference using these types of networks can be NP hard (problems that cannot necessarily be solved in polynomial time on a sequential computer). Cowell *et al.* (1999), chapter 11, discussed the use of local computations and decision potentials designed to reduce the computational burden. It would be most interesting to assess whether the methodology proposed here is competitive. Note that these methods do not directly parameterize the optimal decision rules; furthermore the learning process assumes that there is good scientific information about the distribution of all unobserved variables and about the relationship between the unobserved variables and the observed variables.

Another interesting alternative to the method presented here is the use of Robins's structural nested mean model (Robins, 1986, 1987, 1989, 1997). However, to form the mean response to a dynamic regime we must (as is the case with Bayesian networks) model the distribution of each status S_j as outcomes of past statuses and past treatment.

Lavori and Dawson (2000) proposed the use of an approximate Bayesian bootstrap to impute the values of all potential outcomes. After all potential outcomes have been imputed, one would calculate the mean response corresponding to each of a variety of dynamic treatment regimes and compare these to find the best dynamic regime. Unfortunately, in the cases that we envision, the number of time intervals combined with the continuity of the status will preclude the use of the nonparametric approximate Bayesian bootstrap.

This work raises some interesting issues. An important practical problem is the appropriate design of sequentially randomized trials to be used for estimating an optimal decision regime. From a statistical standpoint this is a difficult area because practical and ethical considerations might limit the variability in the treatment levels, yet variability in the treatment levels is crucial for high quality estimates. We need to understand better the consequences of low variability in the observed treatment decisions ($\bar{A}_j$) given the past information. Clearly one consequence of low variability and/or a small number of possible values of A_j will be less precise estimation of the rules. A second type of problem that can arise is that in the experiment or observational study expensive information or information that is difficult to collect may have been used in the selection of the treatment. For practical applicability the rules should not depend on this information. For example, in the Fast Track study, staff may have used information from detailed summer interviews to assign the treatment; however, in future, summer interviews may not be available. So the goal is to find the rules that optimally use a specified subset of the past information. The work by van der Laan *et al.* (2001) should be useful in developing methodology for this problem. A third issue is that we assume that decisions are made in discrete time; practically, this means that, in any time interval, treatment decisions are made on all or most subjects. However, when the timings of the treatment decisions are so variable across subjects that the chance that a decision occurs in any time interval becomes exceedingly small it is better to move to a continuous time framework. It would be very useful to generalize this work to continuous time. Another useful generalization would allow for a time-varying response. We could choose to estimate a regime that would optimize a one-dimensional summary of the time-varying response such as a weighted average of the response with earlier responses contributing less to the average than do later responses. This is similar to minimizing discounted expectations in infinite horizon, sequential decision problems (see Bather (2000), chapter 9, or section 2.1.2 of Bertsekas and Tsitsiklis (1996)). When the one-dimensional summary can be expressed as an average, the work presented here should generalize without much alteration.

Because of the similarity in the likelihood between this method and Robins's nested structural mean model, many of Robins's results should generalize to this setting. In particular, when the available data are observational, the assumption of no unmeasured confounders is suspect. In this setting Robins has developed sensitivity analyses for the nested structural mean model (Robins *et al.* (1999), sections 8.1b and 8.2b). It would be most interesting to develop analyses that examine how the rules would change as we allow for deviations from the no unmeasured confounders assumption.

An additional interesting area was described in Section 1: suppose that it is feasible to simulate from the known multivariate distribution. It is unclear whether the methodology presented here can provide the basis for better computational algorithms that provide approximate optimal regimes. This could be of great utility as the traditional use of a Bayesian network is computationally difficult.

Acknowledgements

The research of the author is partially supported by National Institute for Drug Abuse grant 2P50 DA 10075, National Institute for Drug Abuse career scientist award K02 DA 15674 and National Science Foundation grant DMS 9802885.

Appendix A

Lemma 1. Let (Z_1, Z_2, Z_3) be a random vector where (Z_1, Z_2) are scalars, Z_3 is an m-vector and Z_2 is countably discrete. Assume that $E|Z_1|$ is finite. Denote a version of the regular conditional probability density of Z_2 given Z_3 by $p(z_2|z_3)$. Assume that $d : R^m \to R$ is measurable and that $P[p\{d(Z_3)|Z_3\} > 0] = 1$. Then any two versions of $E[Z_1|Z_3, Z_2 = d(Z_3)]$ are almost surely equal on the σ-field generated by Z_3.

Proof. By definition of the conditional expectation of Z_1 given Z_3 on the set $\{Z_2 = d(Z_3)\}$, we have that, for any B, a set in the σ-field generated by Z_3,

$$\int_{B\cap\{w:Z_2(w)=d\{Z_3(w)\}\}} Z_1\,\mathrm{d}P = \int_{B\cap\{w:Z_2(w)=d\{Z_3(w)\}\}} E[Z_1|Z_3, Z_2 = d(Z_3)]\,\mathrm{d}P$$

where $E[Z_1|Z_3, Z_2 = d(Z_3)]$ belongs to the σ-field generated by Z_3 intersected with the set $\{w : Z_2(w) = d\{Z_3(w)\}\}$. Suppose that we have two versions of $E[Z_1|Z_3, Z_2 = d(Z_3)]$; on $\{Z_2 = d(Z_3)\}$ these can be written as $h_1(Z_3)$ and $h_2(Z_3)$. By definition,

$$0 = \int_{B\cap\{w:Z_2(w)=d\{Z_3(w)\}\}} h_1(Z_3) - h_2(Z_3)\,\mathrm{d}P.$$

But this is the same as (by the definition of p)

$$0 = \int_B \{h_1(Z_3) - h_2(Z_3)\}p\{d(Z_3)|Z_3\}\,\mathrm{d}P.$$

Since $P[p\{d(Z_3)|Z_3\} > 0] = 1$, $P\{h_1(Z_3) = h_2(Z_3)\} = 1$ (not just on the set $\{w : Z_2(w) = d\{Z_3(w)\}\}$).

Lemma 2. Assume that the range of the treatment decisions $\bar{A}_K$ is countable and that $E|Y|$ is finite. Assume that no unmeasured confounders hold. Assume that the regime $\bar{d}_K$ is measurable and satisfies equation (7), where $p_j(a_j|\bar{S}_j, \bar{A}_{j-1})$ is a regular conditional density for the conditional distribution of A_j given $\bar{S}_j$ and $\bar{A}_{j-1}$. Then the repeated expectation (5)

$$E[E[\ldots E[E[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = d_K]|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]\ldots|S_1, A_1 = d_1]]$$

is well defined and is equal to $E[Y(\bar{d}_K)]$.

Proof. For the repeated expectation to be well defined it must be shown to assume a unique (only one!) value. As discussed by Gill and Robins (2001), difficulties occur because the value of each conditional expectation is arbitrary when the conditioning set has probability 0. Assumption (7) and lemma 1 will allow us to show uniqueness. Because each d_j is measurable and by the definition of conditional expectations, $E[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = d_K]$, $E[E[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = d_K]|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]$, etc. are defined almost surely P. Using lemma 1 and assumption (7) (for example in the first case equate Y with Z_1, $(\bar{S}_K, \bar{A}_{K-1})$ with Z_3 and A_K with Z_2) we see that the conditional expectation is uniquely defined almost surely P. Thus expression (5) is well defined.

To see that expession (5) is equal to $E[Y(\bar{d}_K)]$, first note that $E[Y|\bar{S}_K, \bar{A}_{K-1}, A_K = d_K]$ is equal to $E[Y(\bar{A}_{K-1}, d_K)|\bar{S}_K, \bar{A}_{K-1}, A_K = d_K]$ almost surely by definition of Y. But O_{sr} is independent of A_K conditionally on $(\bar{S}_K, \bar{A}_{K-1})$; thus the above conditional expectation is equal to $E[Y(\bar{A}_{K-1}, d_K)|\bar{S}_K, \bar{A}_{K-1}]$ almost surely. Thus expression (5) is equal to

$$E[E[\ldots E[E[Y(\bar{A}_{K-1}, d_K)|\bar{S}_K, \bar{A}_{K-1}]|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]\ldots|S_1, A_1 = d_1]].$$

Next the repeated expectation, $E[E[Y(\bar{A}_{K-1}, d_K)|\bar{S}_K, \bar{A}_{K-1}]|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]$ is equal to $E[Y(\bar{A}_{K-1}, d_K)|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]$ almost surely. Now we repeat the arguments made before, i.e. first use the definition of Y; then, since O_{sr} is independent of A_{K-1} conditionally on $(\bar{S}_{K-1}, \bar{A}_{K-2})$, the above conditional expectation becomes $E[Y(\bar{A}_{K-2}, d_{K-1}, d_K)|\bar{S}_{K-1}, \bar{A}_{K-2}]$. The repeated use of these arguments proves that expression (5) is indeed equal to $E[Y(\bar{d}_K)]$.

Lemma 3. Let Z_1 be a random variable, Z_3 a random m-vector and Z_2 a random variable with range $\mathcal{A}$ where $\mathcal{A}$ is countable. $E[|Z_1||Z_2, Z_3]$ is bounded almost everywhere. Denote a version of the regular conditional density of Z_2 given Z_3 by $p(z_2|z_3)$. Then given $\varepsilon > 0$ there is a measurable function d such that

$$E[Z_1|Z_3, Z_2 = d(Z_3)] \geqslant \sup_{a:p(a|Z_3)>0} (E[Z_1|Z_3, Z_2 = a]) - \varepsilon$$

almost everywhere P.

Proof. Without loss of generality assume that $\mathcal{A}$ is the positive integers. Consider the set of positive integers j for which

$$E[Z_1|Z_3 = z_3, Z_2 = j] \geqslant \sup_{i:p(i|z_3)>0} (E[Z_1|Z_3 = z_3, Z_2 = i]) - \varepsilon$$

and $p(j|z_3) > 0$. Let $d(z_3)$ be the minimum integer in this set. Then we have

$$\{z_3 : d(z_3) > j\} = \{z_3 : \max_{i \leqslant j, p(i|z_3)>0} (E[Z_1|Z_3 = z_3, Z_2 = i]) < \sup_{a:p(a|z_3)>0} (E[Z_1|Z_3 = z_3, Z_2 = a]) - \varepsilon.$$

This set is measurable for all values of j; thus d is measurable.

A.1. Proof of theorem 1

First expression (5) is no greater than

$$E[E[\ldots E[J_0(\bar{S}_K, \bar{A}_{K-1})|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]\ldots|S_1, A_1 = d_1]].$$

This is because $p_K\{d_K(\bar{S}_K, \bar{A}_{K-1})|\bar{S}_K, \bar{A}_{K-1}\} > 0$ almost surely and d_K must take values in $\{a_K : p_K(a_K|\bar{S}_K, \bar{A}_{K-1}) > 0\}$. Let $\varepsilon > 0$. Furthermore use lemma 3 to see that expression (9) is at least as large as

$$\sup_{\bar{d}_{K-1} \in D_P} \{E[E[\ldots E[J_0(\bar{S}_K, \bar{A}_{K-1})|\bar{S}_{K-1}, \bar{A}_{K-2}, A_{K-1} = d_{K-1}]\ldots|S_1, A_1 = d_1]]\} - \varepsilon$$

because D_P includes all measurable rules d_K depending on $\bar{S}_K$ and $\bar{A}_{K-1}$ with $p_K\{d_K(\bar{S}_K, \bar{A}_{K-1})|\bar{S}_K, \bar{A}_{K-1}\} > 0$ almost surely. Thus expression (9) is within ε of the first term in the former display.

Evaluating the repeated expectation, the first term of the previous display is

$$\sup_{\bar{d}_{K-1} \in D_P} \{E[E[\ldots E[Q_1(\bar{S}_{K-1}, \bar{A}_{K-2}, d_{K-1})|\bar{S}_{K-2}, \bar{A}_{K-3}, A_{K-2} = d_{K-2}]\ldots|S_1, A_1 = d_1]]\}. \tag{17}$$

Similarly we note that expression (17) is within ε of

$$\sup_{\bar{d}_{K-2} \in D_P} \{E[E[\ldots E[J_1(\bar{S}_{K-1}, \bar{A}_{K-2})|\bar{S}_{K-2}, \bar{A}_{K-3}, A_{K-2} = d_{K-2}]\ldots|S_1, A_1 = d_1]]\}$$

and thus this display is within 2ε of expression (9). Finish the proof by repeating this argument and recognizing that ε is arbitrary.

A.2. Proof of theorem 2

We only need inequality (13) to hold for $\bar{m}_K = \bar{\mu}_{0K}$. Note that

$$E[\mu_K(\bar{S}_K, \bar{A}_K)|\bar{S}_K, \bar{A}_{K-1}] = \sum_a \mu_K(\bar{S}_K, \bar{A}_{K-1}, a)\ p_K(a|\bar{S}_K, \bar{A}_{K-1}).$$

Consider the Kth inequality

$$E\left[Y + \sum_{l=1}^{K} \mu_l(\bar{S}_l, \bar{A}_l) - E[\mu_K(\bar{S}_K, \bar{A}_K)|\bar{S}_K, \bar{A}_{K-1}]\right]^2 \leqslant E\left[Y + \sum_{l=1}^{K-1} \mu_l(\bar{S}_l, \bar{A}_l) + \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_{0K}(\bar{S}_K, \bar{A}_K)|\bar{S}_K, \bar{A}_{K-1}]\right]^2.$$

Combining terms we have

$$2\,E\left[Y + \sum_{l=1}^{K-1} \mu_l(\bar{S}_l, \bar{A}_l) + \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]\right] \{\mu_K(\bar{S}_K, \bar{A}_K) - \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_K - \mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]\} + E[\mu_K(\bar{S}_K, \bar{A}_K) - \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_K - \mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]]^2 \leqslant 0.$$

Since the second part of the product in the first term has conditional mean 0, the first term simplifies to yield

$$2\,E[Y + \mu_{0K}(\bar{S}_K, \bar{A}_K)][\mu_K(\bar{S}_K, \bar{A}_K) - \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_K - \mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]] + E[\mu_K(\bar{S}_K, \bar{A}_K) - \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_K - \mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]]^2 \leqslant 0.$$

Recall that $\mu_{0K}(\bar{S}_K, \bar{A}_K)$ is equal to $-E[Y|\bar{S}_K, \bar{A}_K]$ plus a term that is constant in A_K; thus the first term is identically 0 and we have

$$E[\mu_K(\bar{S}_K, \bar{A}_K) - \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_K - \mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]]^2 \leqslant 0,$$

implying that the quantity inside the expectation is almost surely equal to 0. Thus,

$$E[[\mu_K(\bar{S}_K, \bar{A}_K) - \mu_{0K}(\bar{S}_K, \bar{A}_K) - E[\mu_K - \mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]]^2|\bar{S}_K, \bar{A}_{K-1}] = 0 \qquad \text{almost surely,}$$

i.e.

$$\sum_a [\mu_K(\bar{S}_K, \bar{A}_{K-1}, a) - \mu_{0K}(\bar{S}_K, \bar{A}_{K-1}, a) - E[\mu_K - \mu_{0K}|\bar{S}_K, \bar{A}_{K-1}]]^2\ p_K(a|\bar{S}_K, \bar{A}_{K-1}) = 0$$

almost surely. Fix a sample point in this set of probability 1. Then for each a with $p_K(a|\bar{S}_K, \bar{A}_{K-1}) > 0$ we have

$$\mu_{0K}(\bar{S}_K, \bar{A}_{K-1}, a) - \mu_K(\bar{S}_K, \bar{A}_{K-1}, a) = E[\mu_{0K} - \mu_K|\bar{S}_K, \bar{A}_{K-1}].$$

Recall that the supremum of $\mu_{0K}(\bar{S}_K, \bar{A}_{K-1}, a)$ over such a is 0 and by assumption the same holds for μ_K. Thus $E[\mu_{0K} - \mu_K|\bar{S}_K, \bar{A}_{K-1}] = 0$ and

$$\mu_{0K}(\bar{S}_K, \bar{A}_{K-1}, a) - \mu_K(\bar{S}_K, \bar{A}_{K-1}, a) = 0.$$

We have that $\mu_{0K} = \mu_K$ with probability 1.

Next we consider inequality (13) for $j = K - 1$ and $\bar{m}_{K-1} = \bar{\mu}_{0,K-1}$. The proof is virtually identical with the above. Using the result just shown that $\mu_{0K} = \mu_K$, we have

$$E\left[Y+\sum_{l=1}^{K-1}\mu_l(\bar S_l,\bar A_l)+\mu_{0K}(\bar S_K,\bar A_K)-E[\mu_{K-1}|\bar S_{K-1},\bar A_{K-2}]\right]^2$$
$$\leqslant E\left[Y+\sum_{l=1}^{K-2}\mu_l(\bar S_l,\bar A_l)+\sum_{l=K-1}^{K}\mu_{0l}(\bar S_l,\bar A_l)-E[\mu_{0,K-1}|\bar S_{K-1},\bar A_{K-2}]\right]^2.$$

Combining terms as before,

$$\begin{aligned}2\,E&\left[Y+\sum_{l=1}^{K-2}\mu_l(\bar S_l,\bar A_l)+\sum_{l=K-1}^{K}\mu_{0l}(\bar S_l,\bar A_l)-E[\mu_{0,K-1}|\bar S_{K-1},\bar A_{K-2}]\right]\{\mu_{K-1}(\bar S_{K-1},\bar A_{K-1})\\&-\mu_{0,K-1}(\bar S_{K-1},\bar A_{K-1})-E[\mu_{K-1}-\mu_{0,K-1}|\bar S_{K-1},\bar A_{K-2}]\}+E[\mu_{K-1}(\bar S_{K-1},\bar A_{K-1})\\&-\mu_{0,K-1}(\bar S_{K-1},\bar A_{K-1})-E[\mu_{K-1}-\mu_{0,K-1}|\bar S_{K-1},\bar A_{K-2}]]^2\leqslant 0.\end{aligned}$$

Since the second part of the product in the first term has conditional mean 0, the first term simplifies to yield

$$\begin{aligned}2\,E&[Y+\mu_{0,K-1}(\bar S_{K-1},\bar A_{K-1})+\mu_{0K}(\bar S_K,\bar A_K)]\{\mu_{K-1}(\bar S_{K-1},\bar A_{K-1})-\mu_{0,K-1}(\bar S_{K-1},\bar A_{K-1})\\&-E[\mu_{K-1}-\mu_{0,K-1}|\bar S_{K-1},\bar A_{K-2}]\}+E[\mu_{K-1}(\bar S_{K-1},\bar A_{K-1})-\mu_{0,K-1}(\bar S_{K-1},\bar A_{K-1})\\&-E[\mu_{K-1}-\mu_{0,K-1}|\bar S_{K-1},\bar A_{K-2}]]^2\leqslant 0.\end{aligned}$$

Since $\mu_{0,K-1}(\bar S_{K-1},\bar A_{K-1})$ is equal to $-E[Y+\mu_{0,K}(\bar S_K,\bar A_K)|\bar S_{K-1},\bar A_{K-1}]$ plus a term constant in A_{K-1} the first term is 0. We have

$$E[\mu_{K-1}(\bar S_{K-1},\bar A_{K-1})-\mu_{0,K-1}(\bar S_{K-1},\bar A_{K-1})-E[\mu_{K-1}-\mu_{0,K-1}|\bar S_{K-1},\bar A_{K-2}]]^2=0,$$

implying that the quantity inside the expectation is almost surely equal to 0. As in the arguments for $j=K$ this implies that $\mu_{0,K-1}=\mu_{K-1}$ with probability 1.

Continuing in this fashion we see that $\bar\mu_{0K}=\bar\mu_K$ with probability 1. Also it is easy to see that we can replace the $\Sigma_{l=1}^K$ by $\Sigma_{l\geqslant j}^K$ and the sum $\Sigma_{l=1 l\neq j}^K$ by $\Sigma_{l>j}^K$ and the same result holds.

A.3. Variance of $\hat\beta$

$\hat\beta$ can be used when the regrets and optimal decisions are smooth functions of β. For a p column vector, say V, denote $VV^{\mathrm T}$ by $V^{\otimes 2}$. Also denote the first derivative with respect to α of $p_j(A_j|\bar S_j,\bar A_{j-1};\alpha)$ by $\dot p_j(A_j|\bar S_j,\bar A_{j-1};\alpha)$ and set

$$S_\alpha(\bar S_K,\bar A_K;\alpha)=\sum_{j=1}^K \dot p_j(A_j|\bar S_j,\bar A_{j-1};\alpha)/p_j(A_j|\bar S_j,\bar A_{j-1};\alpha).$$

Denote the second derivative with respect to α of $\mathbb P_n[\Sigma_{j=1}^K\log\{p_j(A_j|\bar S_j,\bar A_{j-1};\alpha)\}]$ and evaluated at $\alpha=\hat\alpha_n$ by $-\hat I_{\alpha\alpha}$. Denote the first derivative of $\mu_j(\bar S_j,\bar A_j;\beta)$ with respect to β by $\dot\mu_j(\bar S_j,\bar A_j;\beta)$. Set $\hat c(\beta,\alpha)$ equal to

$$\frac{-1}{K}\sum_{j=1}^K\mathbb P_n\left\{Y+\sum_{l=1}^K\mu_l(\bar S_l,\bar A_l;\beta)-\sum_{a_j}\mu_j(\bar S_j,\bar A_{j-1},a_j;\beta)\,p_j(a_j|\bar S_j,\bar A_{j-1};\alpha)\right\},$$

set $S_\beta(Y,\bar S_K,\bar A_K;\hat\beta_n,\hat\alpha_n)$ to

$$\sum_{j=1}^K\left\{Y+\hat c(\hat\beta_n,\hat\alpha_n)+\sum_{l=1}^K\mu_l(\bar S_l,\bar A_l;\hat\beta_n)-\sum_{a_j}\mu_j(\bar S_j,\bar A_{j-1},a_j;\hat\beta_n)\,p_j(a_j|\bar S_j,\bar A_{j-1};\hat\alpha_n)\right\}$$

$$\times\left\{ \frac{\partial}{\partial\beta}\hat{c}(\beta,\hat{\alpha}_n)\Big|_{\beta=\hat{\beta}_n} + \dot{\mu}_j(\bar{S}_j,\bar{A}_j;\hat{\beta}_n) - \sum_{a_j}\dot{\mu}_j(\bar{S}_j,\bar{A}_{j-1},a_j;\hat{\beta}_n)\ p_j(a_j|\bar{S}_j,\bar{A}_{j-1};\hat{\alpha}_n)\right\}$$

and set $-\hat{I}_{\beta\alpha}$ equal to

$$\begin{aligned}&\sum_{j=1}^{K}\mathbb{P}_n\left\{\dot{\mu}_j(\bar{S}_j,\bar{A}_j;\hat{\beta}_n) - \sum_{a_j}\dot{\mu}_j(\bar{S}_j,\bar{A}_{j-1},a_j;\hat{\beta}_n)\ p_j(a_j|\bar{S}_j,\bar{A}_{j-1};\hat{\alpha}_n)\right\}\\ &\quad\times\left\{\frac{\partial}{\partial\alpha}\hat{c}(\hat{\beta}_n,\alpha)\Big|_{\alpha=\hat{\alpha}_n} - \sum_{a_j}\mu_j(\bar{S}_j,\bar{A}_{j-1},a_j;\hat{\beta}_n)\ \dot{p}_j(a_j|\bar{S}_j,\bar{A}_{j-1};\hat{\alpha}_n)\right\}^{\mathrm{T}}\\ &\quad-\sum_{j=1}^{K}\mathbb{P}_n\left\{Y+\hat{c}(\hat{\beta}_n,\hat{\alpha}_n)+\sum_{l=1}^{K}\mu_l(\bar{S}_l,\bar{A}_l;\hat{\beta}_n) - \sum_{a_j}\mu_j(\bar{S}_j,\bar{A}_{j-1},a_j;\hat{\beta}_n)\ p_j(a_j|\bar{S}_j,\bar{A}_{j-1};\hat{\alpha}_n)\right\}\\ &\quad\times\sum_{a_j}\dot{\mu}_j(\bar{S}_j,\bar{A}_{j-1},a_j;\hat{\beta}_n)\ \dot{p}_j(a_j|\bar{S}_j,\bar{A}_{j-1};\hat{\alpha}_n)^{\mathrm{T}}.\end{aligned}$$

Lastly set $-\hat{I}_{\beta\beta}$ equal to

$$\begin{aligned}&\sum_{j=1}^{K}\mathbb{P}_n\left\{\frac{\partial}{\partial\beta}\hat{c}(\beta,\hat{\alpha}_n)\Big|_{\beta=\hat{\beta}_n} + \sum_{l=1}^{K}\dot{\mu}_l(\bar{S}_l,\bar{A}_l;\hat{\beta}_n) - \sum_{a_j}\dot{\mu}_j(\bar{S}_j,\bar{A}_{j-1},a_j;\hat{\beta}_n)\ p_j(a_j|\bar{S}_j,\bar{A}_{j-1};\hat{\alpha}_n)\right\}\\ &\quad\times\left\{\dot{\mu}_j(\bar{S}_j,\bar{A}_j;\hat{\beta}_n) - \sum_{a_j}\dot{\mu}_j(\bar{S}_j,\bar{A}_{j-1},a_j;\hat{\beta}_n)\ p_j(a_j|\bar{S}_j,\bar{A}_{j-1};\hat{\alpha}_n)\right\}.\end{aligned}$$

For a sample of n subjects we estimate the asymptotic variance of $\hat{\beta}$ by

$$(1/n)\hat{I} = (1/n)\hat{I}_{\beta\beta}^{-1}\ \mathbb{P}_n\{S_\beta(Y,\bar{S}_K,\bar{A}_K;\hat{\beta}_n,\hat{\alpha}_n) + \hat{I}_{\beta\alpha}\hat{I}_{\alpha\alpha}^{-1}\ S_\alpha(\bar{S}_K,\bar{A}_K;\hat{\alpha}_n)\}^{\otimes 2}(\hat{I}_{\beta\beta}^{-1})^{\mathrm{T}}.$$

A.4. Variance of $\hat{\mu}_0$

An estimator of the asymptotic variance of $\hat{\mu}_0$ is given by

$$(1/n)\mathbb{P}_n\left[\sum_{j=1}^{K}\mu_j(\bar{S}_j,\bar{A}_j;\hat{\beta}_n) + Y + \hat{I}_{0\beta}\hat{I}^{-1}\{S_\beta(Y,\bar{S}_K,\bar{A}_K;\hat{\beta}_n,\hat{\alpha}_n) + \hat{I}_{\beta\alpha}\hat{I}_{\alpha\alpha}^{-1}\ S_\alpha(\bar{S}_K,\bar{A}_K;\hat{\alpha}_n)\}\right]^2$$

where

$$\hat{I}_{0\beta} = \mathbb{P}_n\left\{\sum_{j=1}^{K}\dot{\mu}_j(\bar{S}_j,\bar{A}_j;\hat{\beta}_n)\right\}^{\mathrm{T}}.$$

References

Bather, J. (2000) *Decision Theory: an Introduction to Dynamic Programming and Sequential Decisions*. Chichester: Wiley.

Bellman, R. (1957) *Dynamic Programming*. Princeton: Princeton University Press.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996) *Neuro-dynamic Programming*. Belmont: Athena Scientific.

Bielza, C., Müller, P. and Insua, D. R. (2001) Decision analysis by augmented probability simulation. Unpublished. M. D. Anderson Cancer Center, Houston.

Bierman, K. L., Nix, R., Maples, J. J., Murphy, S. A. and Conduct Problems Prevention Research Group (2001) Evaluating the use of clinical judgment in the context of an adaptive intervention design: the Fast Track prevention program. Unpublished. Pennsylvania State University, University Park.

Brockwell, A. E. and Kadane, J. B. (2001) A gridding method for sequential analysis problems. Unpublished.

Carlin, B. P., Kadane, J. B. and Gelfand, A. E. (1998) Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, **54**, 964–975.

Collins, L. M., Murphy, S. A. and Bierman, K. A. (2001) Design and evaluation of adaptive preventive interventions. Unpublished.

Conduct Problems Prevention Research Group (1999a) Initial impact of the Fast Track prevention trial for conduct problems: I, the high-risk sample. *J. Consult. Clin. Psychol.*, **67**, 631–647.

Conduct Problems Prevention Research Group (1999b) Initial impact of the Fast Track prevention trial for conduct problems: II, classroom effects. *J. Consult. Clin. Psychol.*, **67**, 648–657.

Cooper, G. F. (1990) The computational complexity of probabilistic inference using belief networks. *Artif. Intell.*, **42**, 393–405.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L. and Spiegelhalter, D. J. (1999) *Probabilistic Networks and Expert Systems*. New York: Springer.

Cox, D. R. (1958) *The Design and Planning of Experiments*. London: Chapman and Hall.

Dawid, A. P., Didelez, V. and Murphy, S. A. (2001) On the conditions underlying the estimability of causal effects from observational data. Unpublished. University College London, London.

Gill, R. D. and Robins, J. M. (2001) Causal inference for complex longitudinal data: the continuous case. *Ann. Statist.*, to be published.

Heckerman, D. (1998) A tutorial on learning with Bayesian networks. In *Learning in Graphical Models* (ed. M. I. Jordan), pp. 301–354. Dordrecht: Kluwer.

Hougaard, P. (1986) A class of multivariate failure time distributions. *Biometrika*, **73**, 671–678.

Jordan, M. I. and Bishop, C. M. (2001) *An Introduction to Probabilistic Graphical Models*. To be published.

Kreuter, M., Farrell, D., Olevitch, L. and Brennan, L. (2000) *Tailoring Health Messages, Customizing Communication with Computer Technology*. Hillsdale: Erlbaum.

Kreuter, M. W. and Strecher, V. J. (1996) Do tailored behavior change messages enhance the effectiveness of health risk appraisals?: results from a randomized trial. *Hlth Educ. Res.*, **11**, 97–105.

Kreuter, M. W., Strecher, V. J. and Glassman, B. (1999) One size does not fit all: the case for tailoring print materials. *Ann. Behav. Med.*, **21**, 276–283.

van der Laan, M. J., Murphy, S. A. and Robins, J. M. (2001) Analyzing dynamic regimes using structural nested mean models. Unpublished.

Lauritzen, S. L. and Nilsson, D. (2001) Representing and solving decision problems with limited information. *Mangmnt Sci.*, **47**, 1235–1251.

Lavori, P. W. and Dawson, R. (2000) A design for testing clinical strategies: biased adaptive within-subject randomization. *J. R. Statist. Soc.* A, **163**, 29–38.

Lavori, P. W., Dawson, R. and Rush, A. J. (2000) Flexible treatment strategies in chronic disease: clinical and research implications. *Biol. Psychiat.*, **48**, 605–614.

McMahon, R. J., Slough, N. and Conduct Problems Prevention Research Group (1996) Family-based intervention in the Fast Track Program. In *Preventing Childhood Disorders, Substance Abuse, and Delinquency* (eds R. De V. Peters and R. J. McMahon), pp. 65–89. Newbury Park: Sage.

Murphy, S. A., van der Laan, M. J., Robins, J. M. and Conduct Problems Prevention Research Group (2002) Marginal mean models for dynamic regimes. *J. Am. Statist. Ass.*, **96**, 1410–1423.

Neyman, J. (1990) On the application of probability theory to agricultural experiments (Engl. transl. by D. M. Dabrowska and T. P. Speed). *Statist. Sci.*, **5**, 465–480.

Owens, D. K., Shachter, R. D. and Nease, R. F. (1997) Representation and analysis of medical decision problems with influence diagrams. *Med. Decsn Mak.*, **17**, 241–262.

Robins, J. M. (1986) A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Comput. Math. Applic.*, **7**, 1393–1512.

Robins, J. M. (1987) Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect". *Comput. Math. Applic.*, **14**, 923–945.

Robins, J. M. (1989) The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: a Focus on AIDS* (eds L. Sechrest, H. Freeman and A. Mulley), pp. 113–159. US Public Health Service.

Robins, J. M. (1993) Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *Proc. Biopharm. Sect. Am. Statist. Ass.*, 24–33.

Robins, J. M. (1997) Causal inference from complex longitudinal data. *Lect. Notes Statist.*, **120**, 69–117.

Robins, J. M. (2000) Robust estimation in sequentially ignorable missing data and causal inference models. *Proc. Bayesian Statist. Sci. Sect. Am. Statist. Ass.*, 6–10.

Robins, J. M. and Greenland, S. (1992) Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3**, 143–155.

Robins, J. M., Rotnitzky, A. and Scharfstein, D. O. (1999) Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiolgy: the Environment and Clinical Trials* (eds M. E. Halloran and D. Berry), pp. 1–92. New York: Springer.

Robins, J. M. and Wasserman, L. (1997) Estimation of effects of sequential treatments by reparameterizing

directed acyclic graphs. In *Proc. 13th Conf. Uncertainty in Artificial Intelligence* (eds D. Geiger and P. Shenoy), pp. 409–442. San Francisco: Morgan Kaufmann.
Rubin, D. B. (1978) Bayesian inference for causal effects: the role of randomization. *Ann. Statist.*, **6**, 34–58.
——(1986) Which ifs have causal answers. *J. Am. Statist. Ass.*, **81**, 961–962.
Shachter, R. D. (1986) Evaluating influence diagrams. *Ops Res.*, **34**, 871–882.

Discussion on the paper by Murphy

Elja Arjas (*University of Helsinki*)

If I am right the burning issue in social and educational programmes, of which the Fast Track study is an example, is not how to determine an individually optimal dynamic treatment regime but, rather, how a limited total resource should be optimally shared between those who need help. Treatments cannot then be assigned purely on the basis of individual status and treatment histories. The situation is different in clinical trials, where the availability of drugs is not a problem but where they often have unwanted side-effects. The simple logic 'more is better' does not then necessarily apply.

My second comment deals with the concept of *potential outcomes* and the *no unmeasured confounders* postulate. I understand that it is tempting to postulate the existence of individual potential outcomes that are indexed by a list of treatments received, the main advantage being that the appealing concept of an 'individual causal effect' associated with one regime *versus* another can then be defined directly as the contrast between the corresponding potential outcomes. But I have great difficulty in understanding what the postulate of no unmeasured confounders, as formulated in the paper, would mean in a concrete study.

To illustrate this point, consider again the Fast Track study. It is mentioned in Section 5.1 that

> 'staff may have used information from detailed summer interviews to assign treatment; however, in future, summer interviews may not be available'.

It is obvious that decision rules which are to be used later cannot be based on information which then will not be available. But if the summer interviews were actually determinants of how the treatments were assigned in the original study, but are no longer available when the data are analysed, the resulting statistical inference can be seriously confounded. How does this problem relate to potential outcomes? I think that it would be more natural to formulate the no unmeasured confounders assumption by referring, instead of to potential outcomes, to *potential confounders*. Somewhat more formally, we would say that $\{U_1, U_2, \ldots, U_j\}$ are potential confounders at time j if the prediction of the response Y, given the observed past $\{S_1, A_1, S_2, A_2, \ldots, S_j\}$, would change if the conditioning would also involve known values of $\{U_1, U_2, \ldots, U_j\}$. A natural way to formulate the no unmeasured confounders assumption is now to require that for each j, given the observed past $\{S_1, A_1, S_2, A_2, \ldots, S_j\}$, A_j is chosen independently of all such potential confounders $\{U_1, U_2, \ldots, U_j\}$. Of course, such an assumption can never be verified from the data if the potential confounders have not been measured. But at least this alternative formulation would lead the analyst to contemplate the possible existence of factors whose values are unknown but that nevertheless might have influenced the treatment assignments that were made when the original study was carried out.

My third comment concerns the methods of statistical inference. Frankly, the many estimation methods, ranging from maximum likelihood to least squares based on nonparametric frequency estimates, left me in a state of considerable confusion. Would it not be more logical, and simpler, to start from the likelihood expression (cf. the formula below equation (12) in the paper)

$$\prod_{j=1}^{K} f_j(S_j|\bar{S}_{j-1}, \bar{A}_{j-1}) \prod_{j=1}^{K} p_j(A_j|\bar{S}_j, \bar{A}_{j-1})\, g(Y|\bar{S}_K, \bar{A}_K).$$

Here we can see likelihood contributions coming, in an alternating fashion, for each j, first from observing a new value for the status variable S_j and then from recording the corresponding treatment assignment A_j, and ending after K steps with the contribution of the observed response Y given the entire status and treatment history. Under the above-mentioned version of the no unobserved confounders postulate the middle term in this likelihood does not depend on the potential confounder variables $U_1, U_2, \ldots, U_K$. Using statistical terminology in a somewhat liberal manner we include here parameters involved in the definition of the functions f_j and g among such potential confounders. But then, in likelihood inference (including Bayesian), the middle term will only have the role of a proportionality constant, and therefore the inference regarding the functions f_j and g is unaffected by what particular distributions p_j were used