23 MODELING MEDICAL TREATMENT USING MARKOV DECISION PROCESSES

Andrew J. Schaefer[1,2,3], Matthew D. Bailey[1],
Steven M. Shechter[1] and Mark S. Roberts[2,3]

[1] Department of Industrial Engineering
University of Pittsburgh
Pittsburgh, PA 15261

[2] Department of Medicine
University of Pittsburgh
Pittsburgh, PA 15261

[3] Center for Research on Health Care
University of Pittsburgh
Pittsburgh, PA 15261

SUMMARY

Medical treatment decisions are often sequential and uncertain. Markov decision processes (MDPs) are an appropriate technique for modeling and solving such stochastic and dynamic decisions. This chapter gives an overview of MDP models and solution techniques. We describe MDP modeling in the context of medical treatment and discuss when MDPs are an appropriate technique. We review selected successful applications of MDPs to treatment decisions in the literature. We conclude with a discussion of the challenges and opportunities for applying MDPs to medical treatment decisions.

KEY WORDS

Markov decision processes, Stochastic dynamic programs, Optimal medical treatment, Stochastic optimal control, Medical decision making

23.1 INTRODUCTION

Medical treatment decisions must be made sequentially and in an uncertain environment. A physician determining a course of treatment must consider the patient's current health, as well as the best treatment decisions in the future. One important source of uncertainty is that different patients will respond to treatments differently. Other sources of uncertainty include the availability of scarce resources, such as cadaveric organs for transplantation, and human behavior, such as the response time for individuals to react to stroke symptoms. In current medical practice, the vast majority of these treatment decisions are made using ad hoc or heuristic strategies. However, there is a growing feeling among medical practitioners that some treatment decisions are too complicated to solve accurately using intuition alone [1, 2]. The evidence for this includes psychological experiments that indicate that short-term memory has a limited capacity to handle multiple memory constructs, and a substantial body of evidence suggesting a large variation in clinical practice [1, 3-5].

Physicians will always need to make subjective judgments about treatment strategies. However, mathematical decision models that provide insight into the nature of optimal decisions can aid treatment decisions. *Markov decision processes* (*MDPs)* (also known as stochastic dynamic programs) are an appropriate and under-utilized technique for certain types of treatment decisions. MDPs find optimal solutions to sequential and stochastic decision problems. The major advantage of MDPs is their flexibility. Although virtually every medical decision can be modeled as an MDP, the technique is most useful in classes of problems involving complex, stochastic and dynamic decisions, for which MDPs can find optimal solutions.

An MDP is similar to a Markov process (or Markov model, as it is known in the medical decision making literature), except that the decision maker must make decisions at various time epochs. The goal of an MDP is to provide an optimal *policy*, which is a decision strategy to optimize a particular criterion such as maximizing a total discounted reward. In this way, MDPs differ from other stochastic modeling techniques such as discrete-event simulation or Markov processes. Such techniques may be used to evaluate the consequences of a fully specified stochastic model, but they do not allow for the stochastic optimization of that model; they evaluate just one particular policy at a time. To evaluate exhaustively every feasible policy in this manner may be computationally prohibitive. MDPs not only provide the consequences of a policy, they guarantee that no better policy exists.

MDPs also have drawbacks. As the size of the problem increases, MDPs become harder to solve exactly. However, many techniques for finding approximate solutions to MDPs exist. This has been a fertile research area recently, but not in the context of medical treatment decisions [6, 7]. Perhaps the biggest hindrance to the broader application of MDPs is data. Obtaining quality medical data is very difficult and expensive. It is common for a large medical study to cost several million dollars. MDPs are even more data-intensive than other stochastic modeling techniques. This is because the transition probabilities governing the stochastic process, as well as the rewards, are permitted to vary according to the decision made at each decision epoch. While this flexibility is a large advantage in treatment decisions, it means that for every possible description of patient health and every possible treatment, an MDP requires enough observations to estimate accurately transition probabilities to the next epoch. In practice, this typically means that quality data covering thousands of patients is necessary for a successful and realistic MDP model. Although the use of such large patient series is not common, the increasing use of electronic medical records systems is enhancing researchers' ability to utilize large amounts of clinical data from thousands of patients [8].

In Section 23.2 we provide formal models of MDPs and discuss implementation issues such as algorithms and efficiency issues. In Section 23.3 we consider modeling issues particular to applying MDPs to health care problems. In Section 23.4 we provide a selective literature review of previous successful applications of MDPs to medical treatment problems. For each article, we describe the medical application, modeling issues and the solution technique. Finally, in Section 23.5 we provide some conclusions and discuss the future of applying MDPs to medical treatment problems.

23.2 FUNDAMENTALS OF MDP METHODOLOGY

Markov decision processes, or stochastic dynamic programs, are a general framework for modeling dynamic systems under uncertainty. Under mild separability assumptions, discrete-time MDPs can be applied to a variety of systems where decisions are made sequentially to optimize a stated performance criterion. An MDP binds previous, current, and future system decisions through the proper definition of system states, defined as variables that contain the relevant information for making future decisions. The system model evolves in the following manner: The condition or state of the system is observed (or partially observed), an action is taken, a reward is received (or cost incurred), and the system transitions to a new state according to a known probability distribution. The state variables must be defined so that given the current state of the system the future transitions and rewards are independent of the past. This is the standard assumption of a Markov process. MDPs are typically used to model dynamic systems; therefore the decisions are assumed to occur sequentially. However, static decisions can also be modeled using MDPs when the problem's decisions or reward structure are separable: then a one-time decision can be optimized by decomposing it into a sequence of sub-decisions.

23.2.1 Finite-horizon MDPs

We now introduce the fundamentals of MDP methodology. For more complete coverage we refer the reader to Puterman, Bertsekas, or Bellman [9-11]. Following the notation of Puterman, the basic model of a finite-horizon, discrete-time MDP is defined by $(S, A, p_t(\cdot|\cdot,\cdot), r_t(\cdot,\cdot), N)$, where S is the set of defined states and for every state $s \in S$, A is the set of all feasible actions or decisions and A_s are those actions available at state *s*. The system progresses to state s' from state *s* when action $a \in A_s$ is chosen at decision epoch *t*, $(t = 1, \ldots, N)$, with known probability transition $p_t(s' | s, a)$. When action $a \in A_s$ is chosen from state *s* at decision epoch *t*, a reward $r_t(s,a)$ is received. We define a policy $\pi = \{d_1, d_2, \ldots, d_{N-1}\}$ as a sequence of decision rules, where a decision rule is a mapping from states to actions, so that $d_t(s) \in A_s$. The application of a policy π induces a probability distribution over the states at various stages, where the state of the system after *t* transitions is X_t and the action chosen, Y_t, is a function of this state. The objective is to compute the policy that maximizes a given criterion in expectation.

Three commonly used criteria (when beginning in state *s*) are: the total expected reward,

$$v_N^\pi(s) = E_s^\pi \left\{ \sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right\};$$

the total discounted expected reward,

$$v_N^\pi(s) = E_s^\pi \left\{ \sum_{t=1}^{N-1} \lambda^{t-1} r_t(X_t, Y_t) + \lambda^{N-1} r_N(X_N) \right\},$$

for $0 \le \lambda < 1$; and the average reward per stage,

$$v_N^\pi(s) = E_s^\pi \left\{ \frac{1}{N} \left(\sum_{t=1}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right) \right\}.$$

For a finite *N*, the optimal policy for both the average reward per stage and the total reward criterion are equivalent. For the infinite-horizon case, which will be discussed shortly, there is a distinction.

We will present the fundamentals of the total discounted expected reward criterion. Under the standard assumptions of a basic MDP model, where S and A are finite and the rewards are bounded, i.e. $|r_t(s,a)| \le M < \infty$ for every state-action pair (s,a), and $t \le N$, then $v_N^{\pi}(s)$ exists and is bounded. We seek a policy π^* such that $v_N^{\pi}(s) \le v_N^{\pi^*}(s)$ for every $s \in S$. As a result of the principal of optimality [11], the separability of the MDP decisions and rewards can be exploited to decompose this N-period problem into a sequence of $N-1$ single-stage problems, by recursively solving backward from stage $N-1$ to 1:

$v_N(s) = r_N(s)$ for every $s \in S$, and

$$v_t(s) = \max_{a \in A_s} \left\{ r_t(s,a) + \sum_{s' \in S} \lambda p(s' \mid s,a) v_{t+1}(s') \right\}$$

for $t = N-1, \ldots, 1$ and $s \in S$.

Here $v_t(s)$ is the total discounted-expected reward of the $N-t$ stage problem beginning in state s at stage t or as a single-stage problem with terminal rewards $v_{t+1}(s')$, which are known at the time $v_t(s)$ is computed. This is the true computational benefit of MDPs, the ability to reduce a problem into manageable subproblems and still attain the optimal solution. The optimal policy is defined to be the sequence of decision rules, mapping states to the actions that maximize the above recursion, i.e.

$$d_t(s) = \arg\max_{a \in A_s} \left\{ r_t(s,a) + \sum_{s' \in S} \lambda p(s' \mid s,a) v_{t+1}(s') \right\}$$

for $t = N-1, \ldots, 1$ and $s \in S$.

In the above solution and model we assumed that the decision horizon was a finite N. Often there is no defined horizon or the number of stages is so large that it may be approximated by an infinite horizon. In these instances we utilize the techniques discussed in the next section.

23.2.2 Infinite-horizon MDPs

Infinite-horizon models require an infinite amount of data. Therefore, it is typically assumed that data are time-homogeneous or changing so slowly that homogeneity is a reasonable assumption. As a result, the state of an infinite-horizon MDP must be carefully defined to ensure that the system transitions are stationary. If the data are naturally time-dependent, the time-homogeneity assumption can be satisfied by properly augmenting the state definition with the time at which a system transition occurs. The presence of stationary system transitions allows for the use of several elegant solution techniques and easily characterized optimal policies. We replace the above finite-horizon criteria with an infinite-horizon variant by taking the limit of each measure as N goes to infinity. Unlike the total expected reward and discounted expected reward criteria, the analysis and solution methodologies for the average reward criterion depend on the structure of the underlying Markov processes [12]. Again we focus on the problem of maximizing a stream of discounted rewards, which is assured to converge as a result of the bounded rewards assumption.

One of the key insights into infinite-horizon MDPs is that as a result of the assumptions of an infinite horizon, time-homogeneity, and Markov property, under a stationary policy π, i.e. $d_t(s) = d(s)$ for all $s \in S$ and $t = 1, 2, \ldots$, the expected reward vector is also stationary

$$v_t^{\pi}(s) = v^{\pi}(s) \text{ for } t = 1, 2, \ldots \text{ and } s \in S,$$

and $v^{\pi}(s)$ is the unique solution to the set of equations:

$$v^{\pi}(s) = r\,(s, d(s)) + \sum_{s' \in S} \lambda p(s' \mid s, d(s))\, v^{\pi}(s') \text{ for } s \in S. \qquad (1)$$

It is well known that a stationary policy is optimal for these MDPs. In addition, the optimal vector v^* is the solution of the following equations, known as Bellman's equations:

$$v^*(s) = \max_{a \in A_s} \left\{ r\,(s, a) + \sum_{s' \in S} \lambda p(s' \mid s, a)\, v^*(s') \right\} \text{ for } s \in S.$$

Given any initial bounded vector $v_\circ$, it can be shown that the following sequence converges to a solution of Bellman's equations:

$$v_k(s) = \max_{a \in A_s} \left\{ r\,(s, a) + \sum_{s' \in S} \lambda p(s' \mid s, a)\, v_{k-1}^*(s') \right\} \text{ for } s \in S \text{ and } k = 1, 2, \ldots. \qquad (2)$$

However, this solution procedure, known as value iteration, may require an infinite number of iterations [13]. As a result, another technique, policy iteration, is typically used to search over the finite space of policies [11, 14].

In policy iteration, we begin with a policy π°, evaluate that policy by solving the set of linear equations in (1) to find $v^{\pi^{\circ}}$, use this value to choose the actions that maximize the equations in (2) to perform a policy improvement step, and determine the next policy π^{1}. This process is continued until identical policies are found in subsequent iterations. Each iteration results in a policy with an improved optimal reward vector and therefore, for an MDP with finite state and action spaces, policy iteration will terminate with the optimal policy in a finite number of steps. There are several variants of the above techniques; however, the most successful solution methodologies will typically exploit the natural structure of a particular problem instance.

23.2.3 Partially observed MDPs

The above finite- and infinite-horizon MDPs fall into a broader class of MDPs that assume perfect state information – in other words, an exact description of the system. However, often such precision is either too strong an assumption or is not plausible within the model. For example, the state of an MDP could be results from a series of medical tests. These results may supply a better idea of the true state of the patient, but are subject to the error of the tests. Extensions of MDPs, called partially observed Markov decision processes (POMDPs), have been developed to deal with imperfect information [15, 16]. In these models it is assumed that uncertainty exists in the transitions of the system itself and in our knowledge of which state the system truly occupies. Therefore, the objective is to find an optimal policy based on the observations of the system and the previous decision rules applied. It is possible to replace the partially observed state with a sufficient statistic that can be interpreted as a likelihood estimation of the true state of the system given the observations seen. In this manner, the model can be transformed to one with perfect information using the sufficient statistic as the state definition [17]. However, this conversion results in computationally intractable models for systems with even moderately sized underlying true state spaces. As a result, heuristics or approximation techniques must be employed to effectively generate solutions to realistic problem instances.

23.2.4 Semi-Markov decision processes

The above discussion focused on models where the time between decision epochs is fixed and has no effect on the rewards of the system. However, in health care and other applications, decisions may occur over continuous time intervals, such as when varying treatments can be administered. The time between these transitions may depend on the action selected or may occur randomly. In these instances, an extension of MDPs called semi-Markov decision processes (SMDPs) can be employed. These models allow system transitions to occur in continuous time and allow for the inclusion of a probability distribution over the amount of time spent in a state. Through problem transformations and redefinitions, techniques and solution algorithms analogous to those of discrete-time MDPs have been developed for this class of problems [18, 19].

23.3 MODELING ISSUES

23.3.1 Benefit of MDP modeling over traditional decision modeling in health care

For simple medical treatment decisions, a decision tree can be utilized to discover the best course of action. A terminal node of a decision tree usually represents the expected utility (such as life expectancy or quality-adjusted life years) of a patient whose health progression follows that branch of the tree. The path to that terminal node may be complex, and the calculation of that value, requires knowing how the patient may transition between various health states from the initial decision point until death. Modeling these transitions in a standard tree requires a large number of nodes representing multiple time periods in the model, resulting in a tree explosion [20]: the situation is even more complex if the decision can be made at various times, which requires the use of *embedded* decision nodes, making the analysis and interpretation of standard trees almost impossible. As the complexity of the problem increases, the standard decision tree becomes impractical.

Markov models are popular in medical decision making because they can handle some of the difficulty described above. They allow for a simpler representation of the future states and possible transitions that may occur until the patient dies. Solutions to Markov models are obtained via matrix algebra, cohort simulations, or Monte Carlo simulations. Markov models have their limitations, however, because they are not well suited to handle the situation in which decisions may be made at multiple time points. This deficiency of traditional Markov models is precisely the advantage of using Markov *decision* processes for treatment decisions.

Rather than evaluating a decision tree based on a one-time decision (as is often the case in traditional decision trees and Markov models), MDPs allow the "do-nothing" option in each time period and consider the "do-something" option at any later decision epoch [21]. For example, organ transplantation can be modeled as an MDP in which the action each time a donor organ becomes available is to either accept the organ or reject it and wait for a better one. The MDP methodology is especially beneficial because it offers the flexibility of choosing possibly different actions across multiple time periods according to the patient's state. For example, a doctor treating an

HIV-infected patient using highly active anti-retroviral therapy (HAART) may consider different doses and different combinations of drugs at different times during the course of treatment. The action chosen depends on the patient's state, which could include side effects, level of CD4 cells and viral RNA, signs of drug resistance, and degree of adherence to the regimen. Just about any situation where one wants to optimize a process over multiple time periods can be modeled using an MDP. As discussed above, though, exact solutions for large-scale problems may be computationally infeasible and one may need to resort to approximate heuristics.

23.3.2 Issues in modeling disease treatment decisions

Many MDP applications in health care must address the same important modeling issues. For example, MDPs that attempt to optimize a treatment plan or surgery time for a disease require a model of how a patient's health evolves both before and after an intervention. In the case of the optimal time to transplant a liver from a living donor, it is important to develop both a good natural history model of how a patient's health changes in the absence of a transplant and a post-transplant survival model that determines when a patient dies. The natural history model is used to determine transition probabilities between health states from one period to the next if the patient chooses to wait another day for the transplant. In MDP terminology, the survival model determines a terminal reward – the expected remaining life of the patient after receiving a new liver – when the transplant action is chosen.

Another modeling issue in health care MDPs is determining the rewards associated with actions. Optimal disease treatments are usually concerned with maximizing both total life years and quality of life. The quality-adjusted life year (QALY) is a popular measure in the medical literature that blends these two goals [22, 23]. This approach considers a patient's utility for various health states and multiplies the length of life under these health states by the utility weight. One can assess these utilities in various ways including the standard gamble, the time-tradeoff, and the visual rating scale [24]. When quality adjustment is used, the decision to wait another day for treatment or surgery can have very different payoffs for different patients. As Ahn and Hornberger note, for example, some kidney patients may not mind dialysis as much as others and hence would be willing to wait longer for a better donor match [25].

An important area of research in medical treatment decisions concerns the correct way to discount future health consequences. A ubiquitous model to handle this is the discounted-utility (DU) model in which the same discount rate (appropriately compounded) is applied to all future outcomes [26]. In this way, outcomes that occur earlier are preferred to equally valued outcomes that occur later. Over the last couple of decades, however, there has been much research questioning the normative aspects of the DU model [27]. Some of this research demonstrates preference reversals as the time until an event draws nearer, which is inconsistent with DU theory. For example, one study showed that one month before birth, many women wanted to avoid using anesthesia, but during labor they often changed their mind and preferred the anesthesia [28]. Such reversals can be handled by an alternative discounting model – hyperbolic discounting [29]. Other observed phenomena that are inconsistent with traditional DU models include sign effects (where gains are discounted more than losses), magnitude effects (where small outcomes are discounted more than larger ones), and preferences for improving sequences over worsening sequences [27].

A common and recommended practice in cost-effectiveness analyses is to use the same discount rate for both monetary and health outcomes [30]. However, people usually do not discount these two types of outcomes in the same way [31]. Rather, people often demonstrate higher discount rates for health than for money, and, moreover, do not demonstrate a correlation between discount rates in these areas [31]. This suggests that we must pay careful attention to the valuation and discounting of outcomes in an MDP.

23.3.3 Appropriateness of MDPs

Under mild assumptions about the reward functions, any discrete-time sequential decision under uncertainty can be modeled as an MDP. However, data limitations and computational effort may impose limits on one's ability to solve large-scale MDP models in health care. MDP models differ from other models used for treatment decisions. A discrete-event simulation estimates the behavior of a system under uncertainty but is generally unable to make optimal decisions within the simulation. An exception is optimization via simulation, in which parameters governing the simulation are optimized by estimating gradients [32]. In contrast, an MDP allows decisions to be embedded within a Markov process. Rather than an estimate of system behavior, an MDP implicitly considers all possible decision rules or policies and produces the one that behaves the best under a given optimality criteria.

23.3.4 State definition

Selecting the appropriate level of descriptive detail contained in the states of an MDP model is extremely important. From a modeling perspective, the more detailed the information contained in the states the better, since this detail provides a greater distinction among patients. However, increasing the state space makes the model more difficult to solve. Furthermore, data limitations may make a large state space undesirable. For instance, there may be state-action pairs (*s,a*) for which few or no clinical observations occurred. This is typically the case in health care models. States can either be functions of physiological measures (e.g. laboratory values, heart rate, CD4 counts) or can be defined based on subjective judgments such as survival probability.

When insufficient data exist to derive a transition probability distribution or estimated rewards for a set of state-action pairs, two main modeling approaches can be used. One method is to aggregate states judiciously and/or actions to accumulate enough observations for sufficient estimates. For this approach it is important that the aggregated states and/or actions can be justified clinically, since the model cannot distinguish among different patients in the same state. The other approach is to use empirical models of clinical phenomena to estimate the effect of one state-action pair by considering similar state-action pairs for which sufficient data exist. For instance, a statistical model such as a regression model might be able to estimate the effects of a particular state-action pair by considering the results of all states with the same action. This approach may be more successful in estimating rewards than transition probabilities.

23.4 APPLICATIONS OF MDPs TO MEDICAL TREATMENT DECISIONS

We summarize previous successful applications of MDPs to medical treatment decisions. Despite the appropriateness of MDPs for medical treatment decisions, the fact that relatively few such applications exist illustrates the difficulties in developing successful applications.

Epidemic Control Lefèvre developed a continuous-time MDP formulation to address the problem of controlling an epidemic in a closed population of *N* people [33]. The state of the system was described by the number of people infected, and the rest of the population was considered susceptible. Transition probabilities depended on the rate of infection from some external causes, the internal rate of disease transfer from those infected to the uninfected, and the rate at which the infected recovered from the disease. At any point in time, the decision-maker could choose two parameter levels: 1) the amount of the population to quarantine, and 2) the amount of medical treatment to apply to the infected population. Utilizing these definitions, the model minimized the total expected discounted cost over an infinite horizon where the costs incorporated the social cost of people being infected, the cost of quarantining, and the cost of administering medical treatment to those infected. Rather than use real data to solve an instance of the problem, Lefèvre developed the structure of the optimal policy according to the form of the various input parameters. In order to do this, he used a technique that allows one to convert a continuous-time MDP into an equivalent discrete-time MDP [19, 34].

Drug Infusion Hu et al. considered the problem of choosing an appropriate drug infusion plan for the administration of anesthesia [35]. The main decision in this problem was the level at which to set the drug infusion rate to reach a target concentration. Too much anesthesia can cause problems with blood pressure, heart rate, or recovery from the anesthetic state, but too little anesthesia can make the patient more aware of the painful operation. They modeled the problem as a POMDP, which in its pure form was computationally unsolvable. Fast heuristics were necessary for this problem since the maintenance of drug concentrations at target levels is very time sensitive.

One of the main difficulties in this problem arose from the inability to directly observe patient parameters such as anesthesia concentration in the blood and the clearance rate of the drug. This lead to two main issues in the model: 1) the best way to estimate the prior and posterior distributions for these parameters (i.e., whether to use a continuous or discrete distribution), and 2) how much to emphasize active versus passive gathering of information (i.e., how much cost should be incurred now to obtain useful information that can be used more effectively later). The authors developed their own discretization technique for estimating the parameter distribution. This technique has most of the advantages of using continuous and discrete distributions without incurring high computational costs. They applied six approximation methods to determine suboptimal though useful treatment strategies. Three of these treatment strategies emphasized active gathering of information, and the other three strategies emphasized passive gathering. Based on their results, they planned on implementing one of the passive gathering policies into the STANPUMP program at Stanford Medical School, which administers intravenous anesthetics.

Kidney Transplantation Ahn and Hornberger described a model of kidney transplantation that allowed patients to accept or reject an offered kidney based on the quality of the organ [25]. For a potential kidney, they estimated the one-year graft survival of that kidney in a certain patient. For that patient, they also determined the one-year graft survival acceptance threshold that maximized his or her quality-adjusted life expectancy (QALE). The QALE was based on patient-specific ratings for being in different health states. Rather than solve the problem explicitly as an MDP, the authors restricted their search to threshold policies, thereby reducing the problem to finding the optimal threshold level. If the expected one-year graft survival for the kidney-patient pair exceeded the threshold, the patient accepted the transplant; otherwise the patient rejected it. Their model was further simplified by having just five states: 1) alive on dialysis waiting for a transplant, 2) not eligible for transplantation, 3) received a functioning renal transplant, 4) transplant failed, and 5) death. They assumed that patients transitioned between the different states according to a Markov chain with probabilities based on published graft and patient survival rates in the United States.

Spherocytosis Treatment Magni et al. used an MDP approach to decide on therapy for mild hereditary spherocytosis, a disease that causes the chronic destruction of red blood cells [21]. For patients with a mild form of this disease, the main medical treatments considered were prophylactic splenectomy and/or cholecystectomy or no surgery at all. The state of the patient was described through the severity of gallstones and the presence of or years since removal of the spleen. The authors considered gallstone natural history, risk of surgical mortality, and natural causes of death in deriving transition probabilities. They estimated these probabilities and quality-of-life utilities based on published mortality tables and previous studies. They assumed that decisions were made every year with the overall objective of maximizing the patient's quality-adjusted life years. The optimal solution to the MDP model resulted in the following strategy: If a six-year old patient does not have gallstones, then as long as she does not develop gallstones, wait until she is fifteen and then perform splenectomy surgery. If gallstones do appear before the age of fifteen, then both cholecystectomy and splenectomy are suggested.

Treatment of Ischemic Heart Disease Hauskrecht and Fraser applied a POMDP formulation to the problem of treating patients with ischemic heart disease (IHD) [36]. IHD results from the heart not receiving adequate oxygen and is usually caused when the coronary arteries narrow. For patients with this disease, physicians must choose among various diagnostic procedures (such as an angiogram or one of many varieties of stress test), which may be followed by a therapeutic intervention such as medication, surgery (such as angioplasty or bypass surgery), or nothing at all. The state of the patient was described by a variety of variables including the level of coronary heart disease, ischemia level, history of coronary artery bypass grafting, history of percutanerous transluminal coronary angioplasty, and stress test results. The uncertainty of the patient health state arises from the inability to know exactly the level of coronary artery occlusion or the homodynamic impact of that occlusion on myocardial ischemia. Some variables, such as level of chest pain, are directly observable.

Hauskrecht and Fraser framed their POMDP as an infinite-horizon discounted model that seeks a treatment strategy that minimizes total lifetime costs (where the costs incorporate duration of life, quality of life, and monetary costs). To solve their model they used heuristic procedures along with methods that take advantage of special problem structure. They validated their model by devising treatments for ten case patients and then having a cardiologist evaluate their model's treatment strategy. Almost all of the model's recommendations were deemed clinically reasonable, though the experiment also revealed areas for model improvement. Overall, their POMDP formulation was very effective and efficient in generating good treatment strategies for IHD.

Breast Cancer Screening and Treatment Ivy used a POMDP to develop a cost benefit analysis of mammogram frequency and treatment options for breast cancer [37]. The goal was to minimize the total expected cost over a patient's lifetime, where costs were based on the patient's condition, exams, and treatment options. The model consisted of three states: no disease, non-invasive breast cancer, and invasive breast cancer. It was assumed that all patients started in the no-disease state, transitioned to the non-invasive state after a random number of years (according to a geometric distribution based on age) and then transitioned to the invasive stage after another random number of years (the model was flexible enough to relax the assumptions that all non-invasive cancers became invasive or that one must enter the non-invasive state before reaching the invasive state). The part of the model that was partially observable was the patient's condition. Two types of exams – clinical breast exams (CBE) and mammograms – could be performed to get information about the patient's state. At the beginning of each time period, the decision-maker must choose whether to perform a CBE alone or a CBE with a mammogram. If a mammogram was performed and the results were abnormal then the decision-maker could choose either a lumpectomy or a mastectomy. If the mammogram was normal the decision-maker could choose to cease treatment. Using estimates from the literature on costs, test specificity, test sensitivity, and disease progression rates, Ivy solved the dynamic program and characterized optimal decision regions based on the perceived probabilities of the different states of breast cancer.

Liver Transplantation Alagoz et al. presented an MDP model for deciding the optimal time to perform a living-donor liver transplantation [38]. In these types of transplants, the friend or relative of a patient agrees to donate a portion of her liver, and the livers of both the donor and the patient regenerate to a normal size. The goal of the model was to determine when to perform the surgery in order to maximize the expected life years of the patient. The model considered the daily decision of whether or not to transplant. If a transplant was performed, the reward was the expected remaining life years post-transplant, and this was based on survival-analysis estimates [39]. If no transplant was performed, then the patient died in the next day with some probability or transitioned to another health state and increased her life by one day. The transition to other health states was governed by a natural history model of pre-transplant survival [39]. Alagoz et al. used the policy iteration algorithm to solve the MDP and generated an optimal stationary policy to transplant or wait at least another day as a function of the liver quality and the patient health at the start of the day [38].

23.5 CONCLUSIONS

MDPs are a powerful and appropriate technique for medical treatment decisions. MDPs provide optimal policies to stochastic and dynamic decisions. Examples of such decisions naturally arise in finding optimal disease treatment plans. Despite a wealth of potential applications, there have been very few successful applications of MDPs in the medical arena. This is due to several factors, particularly heavy data requirements and computational limitations. However, several recent trends appear to help ameliorate these limitations. First, the medical community is rapidly developing a more quantitative understanding of disease progression and the effects of treatment options. Additionally, the operations research/management science community is improving the solution methodology for MDPs, particularly approximate solutions of MDPs. Also, computing capacity continues to become cheaper. Finally, more hospitals are using electronic medical record systems to gather large amounts of patient data. This confluence of factors will open the door for the increased application of MDPs to medical treatment problems.

Acknowledgments

This work was supported in part by National Science Foundation grant DMI-0223084.

References

[1] Morris, A.H. (2000). Developing and implementing computerized protocols for standardization of clinical decisions. *Annals of Internal Medicine*, 132, 373-83.

[2] Tversky, A. and D. Kahneman (1982). Availability: a heuristic for judging frequency and probability. In *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic and A. Tversky, (Eds.), Cambridge University Press, New York.

[3] Pilote, L., R.M. Califf, S. Sapp, D.P. Miller, D.B. Mark, W.D. Weaver, J.M. Gore, P.W. Armstrong, E.M. Ohman and E.J. Topol for the GUSTO-1 Investigators (1995). Regional variation across the United States in the management of acute myocardial infarction. *New England Journal of Medicine*, 333, 565-572.

[4] Nattinger, A.B., M.S. Gottlieb, J. Veum, D. Yahnke and J.S. Goodwin (1992). Geographic variation in the use of breast-conserving treatment for breast cancer. *New England Journal of Medicine*, 326, 1102-7.

[5] Wennberg, J. and A. Gittelsohn (1973). Small area variations in health care delivery. *Science*, 182, 1102-1108.

[6] Van Roy, B. (2002). Neuro-dynamic programming: Overview and recent trends. In *Handbook of Markov Decision Processes: Methods and Applications*, E. Feinberg and A. Schwartz, (Eds.), Kluwer Academic Press, Boston, MA.

[7] de Farias, D.P. and B. Van Roy (2003). The linear programming approach to approximate dynamic programming. *Operations Research* 51, 850-856.

[8] Tierney, W.M., J.M. Overhage and C.J. McDonald (1995). Toward electronic medical records that improve care. *Annals of Internal Medicine*, 122, 725-726.

[9] Puterman, M.L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York.

[10] Bertsekas, D.P. (2001). *Dynamic Programming and Optimal Control*. Athena Scientific Press, Belmont, MA.

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

[12] Arapostathis, A., V. Borkar, E. Fernandez-Gaucherand, M.K. Ghosh and S.I. Marcus (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31, 282-344.

[13] Shapley, L.S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39, 1095-1100.

[14] Howard, R.A. (1960). *Dynamic Programming and Markov Processes*. Technology Press of Massachusetts Institute of Technology, Cambridge, MA.

[15] Lovejoy, W.S. (1991). A survey of algorithmic methods for partially observed Markov decision problems. *Annals of Operations Research*, 28, 47-66.

[16] White, C.C. and W.T. Scherer (1989). Solution procedures for partially observed Markov decision processes. *Operations Research*, 37, 791-797.

[17] Streibel, C.T. (1965). Sufficient statistics in the optimal control of stochastic systems. *Journal of Mathematical Analysis and Applications*, 12, 576-592.

[18] Jewell, W.S. (1963). Markov-renewal programming I: Formulation, finite return models; Markov-renewal programming II, infinite return models, example. *Operations Research*, 11, 938-971.

[19] Serfozo, R. (1979). An equivalence between continuous and discrete time Markov decision processes. *Operations Research*, 27, 616-620.

[20] Roberts, M.S. and F.A. Sonnenberg (2000). Decision modeling techniques. In *Decision Making in Health Care*, F. A. Sonnenberg and G. Chapman, (Eds.), Cambridge University Press, Cambridge, UK.

[21] Magni, P., S. Quaglini, M. Marchetti and G. Barosi (2000). Deciding when to intervene: a Markov decision process approach. *International Journal of Medical Informatics*, 60, 237-253.

[22] Torrance, G.W. (1976). Social preferences for health states: an empirical evaluate of three measurement techniques. *Socio-Economic Planning Sciences*, 10, 129-136.

[23] Torrance, G.W., D.H. Feeny, W.J. Furlong, R.D. Barr, Y. Zhang and Q. Wang (1996). Multiattribute utility function for a comprehensive

health status classification system. Health Utilities Index Mark 2. *Medical Care*, 34, 702-722.

[24] Drummond, M.F., B. O'Brien, G.W. Stoddart and G.W. Torrance (1997). *Methods for the Economic Evaluation of Health Care Programmes.* Oxford University Press, Oxford.

[25] Ahn, J.H. and J.C. Hornberger (1996). Involving patients in the cadaveric kidney transplant allocation process: A decision-theoretic perspective. *Management Science*, 42, 629-641.

[26] Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies*, 4, 155-161.

[27] Frederick, S., G. Loewenstein and T. O'Donoghue (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, XL, 351-401.

[28] Christensen-Szalanski, J.J. (1984). Discount functions and the measurement of patients' values. Women's decisions during childbirth. *Medical Decision Making*, 4, 47-58.

[29] Kirby, K.N. and N.N. Markovic (1995). Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts. *Organizational Behavior and Human Decision Processes*, 64, 22-30.

[30] Gold, M.R., J. Siegel, L. Russell and M. Weinstein, Eds. (1996). *Cost-Effectiveness in Health and Medicine*. Oxford University Press, New York.

[31] Chapman, G.B. (2003). Time discounting of health outcomes. In *Time and Decision: Economic and Psychological Perspectives on Intertemporal Choice*, G. Loewenstein, D. Read and R. F. Baumeister, (Eds.), Russell Sage Foundation, New York.

[32] Pflug, G. and U. Dieter (1992). *Simulation and Optimization: Proceedings of the International Workshop on Computationally Intensive Methods in Simulation and Optimization, held at the International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria, August 23-25,1990*. Springer-Verlag, Berlin.

[33] Lefevre, C. (1981). Optimal control of a birth and death epidemic process. *Operations Research*, 29, 971-982.

[34] Lippman, S. (1973). Applying a new technique in the optimization of exponential systems. *Operations Research*, 23, 687-710.

[35] Hu, C., W.S. Lovejoy and S.L. Shafer (1993). Comparison of some suboptimal control policies in medical drug therapy. *Operations Research*, 44, 696-709.

[36] Hauskrecht, M. and H. Fraser (2000). Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine*, 18, 221-244.

[37] Ivy, J.S. (2002). A maintenance model for breast cancer detection and treatment. Submitted for publication.

[38] Alagoz, O., A.J. Schaefer, L.M. Maillart and M.S. Roberts (2002). Determining the optimal timing of living-donor liver transplantation using a Markov decision process (MDP) model. *Medical Decision Making*, 22, 558 (abstract).

[39] Roberts, M.S. and D.C. Angus (2002). The optimal timing of liver transplantation: Final report R01 HS09694. University of Pittsburgh, Pittsburgh, PA.