Dynamic Learning of Patient Response Types: An Application to Treating Chronic Diseases

Diana M. Negoescu Kostas Bimpikis Margaret L. Brandeau Dan A. Iancu*

this version: September 15, 2014

Abstract

Currently available medication for treating many chronic diseases is often effective only for a subgroup of patients, and biomarkers accurately assessing whether an individual belongs to this subgroup do not exist. In such settings, physicians learn about the effectiveness of a drug primarily through experimentation, i.e., by initiating treatment and monitoring the patient's response. Precise guidelines for discontinuing treatment are often lacking or left entirely at the physician's discretion. We introduce a framework for developing adaptive, personalized treatments for such chronic diseases. Our model is based on a continuous-time, multi-armed bandit setting, and acknowledges that drug effectiveness can be assessed by aggregating information from several channels: by continuously monitoring the (self-reported) state of the patient, but also by (not) observing the occurrence of particular infrequent health events, such as relapses or disease flare-ups. Recognizing that the timing and severity of such events carries critical information for treatment design is a key point of departure in our framework compared with typical (bandit) models used in healthcare. We show that the model can be analyzed in closed form for several settings of interest, resulting in optimal policies that are intuitive and have practical appeal. We showcase the effectiveness of the methodology by developing a treatment policy for multiple sclerosis. When compared with standard guidelines, our scheme identifies non-responders earlier, leading to improvements in quality-adjusted life expectancy, as well as significant cost savings.

1 Introduction

Recent years have witnessed an alarming rise in the costs associated with the delivery of healthcare in the U.S., both in terms of total expenditure (e.g., as a percentage of gross domestic product), but also in spending recognized as wasteful, redundant or inefficient (Young and Olsen 2010). In conjunction with advances in the field of medicine and the use of information technology, this has

*Negoescu (negoescu@umn.edu) is with the Industrial and Systems Engineering Department at University of Minnesota. Bimpikis (kostasb@stanford.edu) and Iancu (daniancu@stanford.edu) are with the Stanford Graduate School of Business, and Brandeau (brandeau@stanford.edu) is with the Department of Management Science and Engineering at Stanford University. We are particularly thankful to Allie Dunworth Leeper and Stephen Chick for helpful comments. All remaining errors are ours.

placed increasing pressure for healthcare innovation, aimed, among other issues, at the implementation of healthcare solutions delivering better outcomes in a cost-effective manner.

Despite this renewed impetus, however, the design of adaptive treatment policies for chronic conditions has often been perceived as slow,[1] with some of the complicating factors intrinsically related to the specifics of disease progression and available medication.

To start, the currently available disease modifying therapies (DMTs) for several chronic illnesses are only effective in a subset of the population ("responders"), and biomarkers that accurately assess *a priori* whether a given patient belongs to this subgroup are not available. In such cases, the main way to evaluate DMT efficacy is by initiating treatment, and continuously monitoring the patient through self-reported surveys, periodic check-ups, or more in-depth scans and evaluations.

If the role of treatment were the reversal of an obvious short-term abnormality, such monitoring would provide sufficient evidence for how well the patient is responding. However, the primary goal of DMTs for *chronic* illnesses is to prevent disease progression in the long run, which often translates in limiting the occurrence of particular infrequent negative health events, which can have severe implications on a patient's quality of life. As such, the (non)occurrence or the exact *timing* and severity of such episodes often convey critical information concerning a DMT's effectiveness on the patient. Quantifying the impact of such information and translating it into actionable guidelines for medical decision-making is often not straightforward.

A primary example of a chronic disease with these features is multiple sclerosis (MS), an autoimmune inflammatory disease of the central nervous system that is a leading cause of disability in young adults. MS is an incurable disease, and DMTs attempt to slow its progression by decreasing the frequency and severity of clinical attacks, known as "relapses" (see, e.g., Cohen et al. 2004, NMSS 2014). However, while all the available drugs represent advances for MS management, none is fully effective (Molyneux et al. 2000, Rovaris et al. 2001), and the question of identifying patients that are not responsive to treatment is a centrally important one. In the words of the National Clinical Advisory Board of the National Multiple Sclerosis Society (NMSS 2004),

> *"[...] whatever the relative merits of these drugs, all can only be considered partially effective agents. This reality raises the difficult problem of the identification of a suboptimal response or treatment failure in an individual case and, once identified, leads to consideration of the appropriate avenues for alternative treatments."*

As the quote highlights, the problem of identifying patients who do not respond to DMTs is not only relevant, but also quite challenging. For a newly diagnosed case, current guidelines recommend immediately starting treatment, and assessing effectiveness by continually monitoring the disease progression, through MRI scans and self-reported assessments of disability, such as the Expanded Disability Status Scale (EDSS) (NMSS 2008). The guidelines emphasize the critical role played by learning, and explicitly recognize that the timing and frequency of relapses, as well as more contin-

[1]For instance, in an editorial paper, Murphy and Collins (2007) state that "despite the activity in evaluating adaptive treatment strategies, the development of data collection and analytic methods that directly inform the construction of adaptive treatment strategies lags behind".

uous measurements such as EDSS and/or MRI can all be informative.[2] However, they stop short of providing a systematic way to use this information, and suggest only simple rules for discontinuing treatment. To the best of our knowledge, these rules were not the outcome of a quantitative framework, and have not been tested for efficiency (see Cohen et al. 2004). Furthermore, while several studies have attempted to identify early predictors of non-response (Horakova et al. 2012, Prosperini et al. 2009, Romeo et al. 2013), the results have not been used to inform the design of optimal treatment plans in a quantitative fashion.

Further underscoring the need for fast and accurate identification of non-responders is the fact that DMTs can cause significant side effects, such as persistent flu-like symptoms, injection site necrosis, and liver damage, which result in poor compliance and large drop-out rates (Prosser et al. 2004). Additionally and quite importantly, treatment is expensive, with mean annual costs of $13 billion, and lifetime costs of $3.4 million per diagnosed case in the U.S. (Adelman et al. 2013, Kobelt et al. 2006). This has resulted in a significant amount of debate around policies for MS treatment, in the US and elsewhere.[3]

This example and the preceding discussion give rise to several interesting research questions. Given the available medications, what is the optimal treatment plan for chronic diseases such as multiple sclerosis? Does it involve a discontinuation rule, i.e., is it optimal to start a patient on treatment, and then stop at a particular point in time? If so, how can a medical decision maker optimally aggregate all the information acquired during therapy to design the discontinuation rule(s)? Would such optimal rules outperform current existing medical guidelines?

This paper can be viewed as one step towards answering such questions. We propose a framework that can be used to inform treatment decisions for chronic diseases that have the features described above, i.e., treatment is effective only for a subset of patients which is *a priori* unidentifiable; the frequency and severity of side effects and major health events depends on a patient's response type; and information regarding the effectiveness of treatment is obtained gradually over time. Our main contributions can be summarized as follows:

- We formulate the problem of determining an optimal adaptive treatment policy as a continuous-time stochastic control problem. A key point of departure from other work in medical decision-making is that we incorporate information from three channels: the *day-to-day monitoring* of disease progression, as well as the *timing* and *severity* of major health events. Furthermore, our framework explicitly models the trade-off between the immediate and the long-term

[2] "[...] the effects of current therapies on attack rates and MRI measures of newly accumulated lesion burdens [...] are the events that are most readily available to the clinician when considering treatment failure or suboptimal response in an individual patient" (NMSS 2004).

[3] The National Institute of Health in the UK launched an innovative risk sharing scheme in 2002, according to which patients would be closely monitored to evaluate the cost-effectiveness of the drugs used in standard treatment, with an agreement that prices would be reduced if overall patient outcomes were worse than predicted. The scheme became controversial when reports from observational cohorts suggested that the outcomes were far below expectations – implying that treatment was generally not cost-effective – yet the drug providers did not reduce their prices as per the agreement (Boggild et al. 2009, Raftery 2010, Sudlow and Counsell 2003). It is worth noting that personalized discontinuation rules for patients were not considered, though such rules might have reduced total costs and also improved patient outcomes.

impact of treatment,[4] and provides a systematic way for incorporating newly acquired information in the design of an optimal treatment plan.

- Our model can be analyzed in closed form for several settings of interest, resulting in intuitive optimal policies, with practical appeal. In a medical context, we show that the resulting treatment policies are often discontinuation rules, implying that DMTs should always be administered in full doses at inception, and then stopped at a well-chosen point in time. However, we also highlight settings where, depending on the severity of the health events, treatment may have to continue indefinitely, at adjusted doses.

- We apply our framework to multiple sclerosis, for which we develop and test an adaptive treatment policy. In a detailed simulation study, we find that our policy outperforms the standard guidelines, yielding improved health outcomes for non-responders, without impacting the responders. Furthermore, our treatment plan generates cost savings of about 15%, suggesting a potential cost reduction of more than $104 million annually in the United States.

While we apply our model to MS primarily because of the availability of data, we note that the treatment of many other chronic diseases could benefit from this analysis. Examples include rheumatoid arthritis, where increased disability is associated with higher mortality (Pincus et al. 1984); Crohn's disease, where treatment often involves the same classes of medications as for multiple sclerosis; and depression and other mental illnesses, where psychiatrists must choose between various treatments without knowing *a priori* which one might be effective.

1.1 Relevant Literature

Our model builds on the theory of continuous-time multi-armed bandits (Karatzas and Shreve (1998), Berry and Fristedt (1985), Mandelbaum et al. (1987), and more recently Cohen and Solan (2013), and Harrison and Sunar (2014)). The canonical setup involves a decision maker who is allocating scarce resources between two risky alternatives over time, and receiving rewards governed by a continuous-time stochastic process. Closest to our work are the recent papers on strategic experimentation (Bolton and Harris (1999), Keller et al. (2005), and Keller and Rady (2010)), which study free riding among a team of agents in an experimentation context. We adapt their framework in a medical decision-making setting, and extend their model and analysis by allowing the decision maker to learn from observing the rewards generated by two stochastic processes, whose parameters depend on the choice of treatment: a Wiener process (Brownian motion) that models the day-to-day side effects experienced by the patient, and a Poisson process that captures the arrival of major health events, i.e., disease flare-ups and progression.

From an application standpoint, our paper is mostly related to the clinical trials literature, and in particular to the growing number of studies that consider adaptive rules for assigning patients to treatments. Berry (1978) and Berry and Pearson (1985) consider Bayesian designs for maximizing the number of successes in clinical trials. More recently, Cheng and Berry (2007) consider

[4]For instance, when treatment induces immediate negative side-effects, but has long-term benefits, or vice-versa.

constrained Bayesian optimal adaptive design, where each arm has a minimum probability of being chosen, Press (2009) introduce two-arm Bernoulli designs with ethically motivated cost functions, and Wen and Haoda (2011) study clinical trials with delayed response. Ahuja and Birge (2012) consider learning from multiple patients simultaneously, and propose an adaptive clinical trials design that performs significantly better than the best existing static designs. Bertsimas et al. (2014) propose a data-driven approach for the analysis and design of clinical trials, with the goal of discovering promising drug combinations for cancer patients. These approaches typically assume that the outcome of a clinical trial is binary (success/failure), and that the decision maker can learn from multiple patients since outcomes are positively correlated. That makes it difficult to implement these approaches in the context of a chronic disease, where a given patient's response to treatment is independent from another's, and information about the quality of treatment is obtained gradually over time, with no single event providing sufficient indication for or against a given treatment plan. In contrast, our approach allows the decision maker to incorporate objective evidence from the timing and severity of health events, as well as expert opinion in the form of a Bayesian prior.

Our work is also related to the growing literature in the medical decision-making community on developing adaptive treatment policies. Markov decision processes with fully or partially observed states, and dynamic linear Gaussian systems have been used in optimizing treatment decisions (see, e.g., Denton et al. 2009, Helm et al. 2014, Mason et al. 2014, Zhang et al. 2012). However, such models are difficult to use in our setting, as they are not designed to handle the arrival of information through multiple distinct channels, such as frequent quality-of-life reports, as well as the occurrence and severity of infrequent life events. In the medical literature on adaptive treatments, Murphy and Collins (2007), Pineau et al. (2007), and Almirall et al. (2012) propose simple adaptive treatment schemes in the context of psychiatric conditions such as depression, anxiety disorders, and drug abuse. However, the benchmarks used for classifying patients as (non)responders in these studies are not derived by optimizing an objective, and therefore carry no guarantee of yielding treatments that improve patients' quality of life. In contrast, our approach explicitly seeks to optimize an objective related to quality-adjusted life expectancy, so that the optimal solution yields a treatment policy with concrete and measurable benefits.

2 Model Formulation

We first introduce our model in an abstract setting, and then discuss the connection and relevance to the medical applications motivating our work. In an effort to make the paper accessible to a broad audience, we deliberately keep the exposition style less formal, placing more emphasis on the intuition and connection with the applications. Readers interested in the mathematical details can refer to Bolton and Harris (1999), Keller and Rady (2010) and references therein, which form the basis of our model.

We consider a continuous-time frame, indexed by $t \in [0, \infty)$. A single decision maker (DM) is

faced with the problem of choosing how to allocate the current period $[t, t+dt)$ between two possible alternatives ("arms"). Each arm brings the DM immediate rewards, which accrue continuously over time, but also induces particular "life events", which occur more rarely and generate lump-sum rewards. More precisely, the first arm, considered "safe", generates instantaneous rewards governed by a Brownian motion, with drift rate μ_0 and volatility σ, and induces life events from a Poisson process with rate λ_0. The second arm, which is "risky", can be of either good (G) or bad (B) type, unbeknownst to the DM. Depending on the type $\theta \in \{G, B\}$, this arm yields corresponding instantaneous Brownian rewards with drift rate μ_θ and volatility σ, and induces life events according to a Poisson process with rate λ_θ.

When allocating the time interval $[t, t+dt)$ between the two arms, the DM can use any fractional split. More precisely, by allocating a fraction $\alpha_t \in [0, 1]$ of the period to the risky arm, and $1 - \alpha_t$ to the safe arm, the DM receives instantaneous rewards of $d\pi^1(t)$ and $d\pi^0(t)$, respectively, where

$$d\pi^1(t) \stackrel{\text{def}}{=} \alpha_t \mu_\theta \, dt + \alpha_t^{1/2} \sigma \, dZ^1(t), \tag{1a}$$

$$d\pi^0(t) \stackrel{\text{def}}{=} (1 - \alpha_t)\mu_0 \, dt + (1 - \alpha_t)^{1/2} \sigma \, dZ^0(t). \tag{1b}$$

Here, $dZ^0(t)$ and $dZ^1(t)$ are independent, normally distributed random variables, with mean 0 and variance dt, and $\theta \in \{G, B\}$, depending on the risky arm's type. To understand the scaling used here, note that the DM's instantaneous rewards from the risky and safe arm are normally distributed, with mean $\alpha_t \mu_\theta dt$ and variance $\alpha_t \sigma^2 dt$, and mean $(1-\alpha_t)\mu_0 dt$ and variance $(1-\alpha_t)\sigma^2 dt$, respectively. As such, the total instantaneous reward exactly corresponds to a fraction α_t of the risky reward, and $(1 - \alpha_t)$ of the safe reward.

In addition to the instantaneous rewards, the DM may also receive a "lump-sum" reward L_t, in case a life event occurs during period $[t, t + dt)$. When the allocation used is α_t, life events occur according to a Poisson process,[5] with rate $\lambda(t, \theta) \stackrel{\text{def}}{=} (1 - \alpha_t)\lambda_0 + \alpha_t \lambda_\theta$. The lump-sum rewards are stochastic, and can depend on the allocation α_t used during the current period, and the type θ of the risky arm.[6]

The DM knows all the underlying parameters governing the arms and the reward structure, i.e., μ_0, λ_0, σ, $\mu_\theta, \lambda_\theta$ and the distribution for $L_t(\alpha_t, \theta)$, for $\theta \in \{G, B\}$, but does *not* know the type θ of the risky arm. At time $t = 0$, he starts with some initial belief $p(0)$ that the risky arm is good, which he then updates during the rest of the planning horizon, depending on the observed instantaneous and lump-sum rewards. This generates an updated belief p_t at time t.

The DM's goal is to find an allocation policy $\{\alpha_t\}_{t \geq 0}$, such that α_t depends on all available information at time t and maximizes the total expected discounted rewards Π over a particular[7]

[5] This is consistent with an interpretation of α_t as a probability, so that $\lambda(t, \theta)$ denotes the probability of having any arrival of a life event, from either the safe or the risky arm.

[6] We consider two models for the lump-sum rewards: one in which they are constant (i.e., time-invariant and independent of the combination of arms used), and one in which they are independent draws from two-point distributions, with values and probabilities depending on α_t and θ.

[7] We consider two models, one with T corresponding to the first occurrence of a life event, and one with $T = +\infty$.

planning horizon T,

$$\Pi \stackrel{\text{def}}{=} \mathbb{E}\left[\int_0^T e^{-rt}\left[d\pi^1(t) + d\pi^0(t) + \lambda(t,\theta)L_t dt\right]\right]. \tag{2}$$

Some observations regarding the problem formulation are in order. First, note that the integrand in the expression for Π contains three terms. The first two, $d\pi^1(t)$ and $d\pi^0(t)$, correspond to the instantaneous rewards received from the risky and safe arms, given in (1a) and (1b), respectively. The third term corresponds to the lump-sum reward, received upon the occurrence of a life event during period $[t, t+dt)$ (recall that the process governing such events is Poisson, with rate $\lambda(t,\theta)$). The integral is taken over the total (instantaneous and lump-sum) rewards, discounted at a fixed rate $r > 0$. The expectation in (2) is with respect to the stochastic processes $dZ^0(t), dZ^1(t)$, α_t, L_t, and also p_t. The latter reflects the DM's use of the belief p_t, at time t, with regard to the type θ of the risky arm.[8]

Additionally, note that, in choosing a policy α_t to maximize the expected rewards, the DM is faced with the classical trade-off between "exploration" and "exploitation" (Barto (1998), Powell (2007), Powell and Ryzhov (2012)), i.e., between acquiring information about an unknown alternative, which *may* entail higher rewards, versus using a safe option. In this sense, α_t critically trades off the rate at which new information is gained, with the risks entailed by the experimentation generating such information.[9] It is important to emphasize that new information in our model is acquired through *three* potential channels: (1) by observing the instantaneous rewards from the risky arm, $d\pi^1(t)$, (2) by (*not*) observing life events, and (3) by evaluating the magnitude of the lump-sum rewards associated with a life event (when L_t depends on θ). Whenever $\alpha_t > 0$, these channels all convey meaningful information to the DM, potentially tilting his belief p_t towards (or away from) deeming the risky arm as good.

2.1 Discussion in the Context of Chronic Diseases

We conclude the section by discussing how this mathematical framework can be applied to the design of an adaptive treatment policy for chronic diseases such as multiple sclerosis (MS).

The arms. In a medical context, the arms of our model correspond to available treatments, and the DM is a physician choosing the optimal treatment policy for a new patient. The "rewards" correspond to a patient's health utility, with an arm's instantaneous reward denoting the impact of that treatment on the patient's immediate quality of life. In contrast, a "life event" is a major change, such as a relapse in MS or a heart attack, severe infection or panic attack in anxiety disorders.[10] In our model, this implies that the relevant lump-sum "rewards" L_t are actually negative, i.e., they are *disutilities*.

[8] This effectively means that the (conditional) expectation, at time t, of quantities depending on θ should be taken with respect to a corresponding two-point distribution given by p_t. For instance, $\mathbb{E}[\mu_\theta] = p_t\mu_G + (1-p_t)\mu_B$.

[9] With $\alpha_t = 0$, the DM would only gain instantaneous and lump-sum rewards from the safe arm, hence completely eliminating his exposure to the risky arm, but also his ability to update the belief p_t.

[10] Other examples applicable to chronic illnesses include kidney failure, liver failure, malignancy or death.

Safe arm. A "safe" arm represents a treatment with homogenous response in the population. In MS, this typically consists of medication aimed at reducing or controlling MS-specific symptoms (such as bowel and bladder function, spasticity and pain), without modifying the disease-progression. Note that, in our model, such a treatment may still yield stochastic outcomes in terms of both instantaneous health utility and life events, as one would expect in practice. The critical assumption is that the parameters governing these outcomes $(\mu_0, \sigma, \lambda_0)$ are known to the physician. This is reasonable, since physicians often have more information about the natural disease progression when patients are not subjected to treatment, e.g., from studies of large historical cohorts of patients (Cohen et al. 2004, Prosser et al. 2003, Scalfari et al. 2010).

Risky arm. In contrast, the "risky" arm is only effective in a subset of the population, i.e., when the type is good ($\theta = G$). We assume that the physician is unable to determine *a priori* whether a new patient belongs to this subset. This is in keeping with the fact that precise biomarkers do not exist for many chronic diseases. For instance, in MS, treatments such as interferon-β are effective only in a subgroup of patients (Cohen et al. 2004, Horakova et al. 2012, Prosser et al. 2003). In such cases, the only way to assess the impact of a drug or therapy is by subjecting the patient to treatment, and relying on periodic examinations or self-reported assessments, such as the EDSS in MS. When patients respond to treatment, their general condition may improve (i.e., $\mu_G > \mu_0$), the likelihood/frequency of severe life events may be diminished (i.e., $\lambda_G < \lambda_0$), and the severity of such events may also decrease. When patients do *not* respond, their condition may remain the same or even deteriorate, e.g., due to side effects from treatment.[11] We note that a central assumption underlying our model is that physicians are able to (separately) assess the parameters governing how responders and non-responders are impacted by treatment, i.e., $\mu_\theta, \lambda_\theta$ and L_t, when it depends on θ. This is reasonable since medical studies often track groups of patients for a longer period of time, and then retrospectively assign them to responder and non-responder groups (e.g., Horakova et al. 2012).

Fractional allocations. Our model allows the possibility of a fractional allocation of treatment, i.e., $\alpha_t \in (0, 1)$. This may be essential for some medical settings, e.g., when cocktails are drugs are considered (Rudick et al. 2006). However, as we will show, the optimal policies are "bang-bang" in many cases (i.e., $\alpha_t \in \{0, 1\}$), so this modeling choice does not limit the use of our framework in settings where a single treatment with fixed dosage can be applied at any point in time.

Objective. We consider two versions of the DM's objective function. We first take T as the first occurrence of a major health event. This is appropriate in settings where the risky treatment improves the immediate quality of life of a patient, but carries a significantly higher risk of rare severe side effects, or even death. For instance, studies have shown that certain rheumatoid arthritis treatments improve pain and disability, but may cause malignancies or severe infections (Galloway

[11] We note that several relationships between μ_0, λ_0, $\mu_{G,B}$, and $\lambda_{G,B}$ may be encountered in a medical application, depending on the particular disease and treatments. For instance, in MS, a standard "risky" treatment is interferon-β, which has severe side-effects, irrespective of whether the patient is a responder or not (i.e., $\mu_B \approx \mu_G < \mu_0$). However, the likelihood of experiencing relapses is considerably reduced for responders, while it remains relatively unchanged for non-responders, i.e., $\lambda_G < \lambda_0 \approx \lambda_B$ (Cohen et al. 2004, Horakova et al. 2012).

et al. 2011, Mariette et al. 2011). This modeling choice is also appropriate when major health events do not have such negative outcomes, as long as the set of feasible treatment options is considerably changed in the aftermath. In addition, such an objective would also introduce a certain degree of risk aversion in the model, by effectively allowing less experimentation with the risky arm. This is often desirable in medical practice, as physicians prefer learning about treatment effectiveness without causing too much harm from initial exploration.

We also consider a model with $T = \infty$. This may be more appropriate when major events are a natural course of the disease progression, with or without treatment, and the main role of successful therapy is to postpone or reduce the frequency and/or severity of such events, possibly at the cost of inducing unpleasant side effects in the short run. This is the case with relapses in the relapsing-remitting stage of MS, which occur regardless of treatment or response status, but whose frequency is reduced when treatment is successful (Cohen et al. 2004, Horakova et al. 2012, Prosser et al. 2004, Romeo et al. 2013). While most MS models assume that treatment only affects the likelihood of a relapse (Lee et al. 2012), successful treatments in many other chronic diseases may also reduce the severity (i.e., magnitude) of negative health events. This is the case, for example, in patients suffering from depression (Driessen et al. 2010, Fournier et al. 2010) or Crohn's disease (Lichtenstein et al. 2009), or patients infected with the herpes simplex virus (Reichman et al. 1984, Wald et al. 2002), where the severity of episodes or symptoms relapses is often an indication of treatment ineffectiveness. To capture this possibility, we allow the "rewards" L_t received upon a life event to depend on the risky treatment type, θ. Note that, apart from capturing the therapy-disease interaction in a more realistic way, this feature also makes the magnitude of a health event informative for physicians, as it can tilt the belief of a risky treatment being effective.

Both models include a fixed discount rate $r > 0$, in keeping with the recommendations of the US Panel on Cost-Effectiveness in Health and Medicine that costs and quality-adjusted life years should be discounted when estimating the effectiveness of health care interventions (see Gold 1996).

Simplifying assumptions. Our model makes a number of simplifying assumptions. First, we only consider two arms (treatments). This may be appropriate in settings where at most two treatments are usually administered simultaneously because adding a third treatment could cause severe adverse events (see, e.g., Abalos et al. 2007, Garcia et al. 2012, Hirsh and Lee 2002, Neutel et al. 2008), and constitutes a reasonable first step when dealing with multiple alternatives.

Second, we assume Brownian rewards, Poisson arrivals, and time-invariant drifts and rates. Although this is primarily for the sake of analytical tractability, it approximates well the reality of some diseases. For instance, relapses in MS are relatively rare events, occurring on average once or twice per year, and the disease progresses slowly (Lee et al. 2012, Prosser et al. 2004, Scalfari et al. 2010). As such, the assumption that the rates of rewards and relapses are constant is fairly reasonable, particularly when the treatment plan is re-evaluated after the event of a relapse (see our comments about the objective).

Third, we assume that the reward and life event rates are known for both responders and non-responders. This is reasonable in the context of treatments that have already been approved

for use in the general population, as extensive clinical trials must be carried out before obtaining approval from government agencies, such as the US Food and Drug Administration; furthermore, medical studies often report rates for responders versus non-responders in treatments that have been on the market for a longer time (Horakova et al. 2012, Romeo et al. 2013). In principle, our framework could be extended for use in settings where no such prior knowledge of the rates exists, e.g, by defining an undesirable set of rates and considering patients with rates in this set as "non-responders"; we leave the details of such an extension for future work.

Fourth, we allow "rewards" to be observed and quantified continuously, which may be difficult and even impractical for a physician. This is also primarily for analytical tractability, and the model could be extended to allow belief updates only at particular points of time, such as when patients undergo periodic evaluations or when major health events occur (see our analysis in Section 3.2).

Fifth, we assume that the volatility in the observed rewards of a patient is the same under both arms. This is not needed for the analysis, as the results extend to a setup with known volatilities that depend on patient response type. We adopt it primarily for simplicity and added realism – volatilities must be estimated from the variance in self-reported surveys of quality of life, and there is insufficient data in the medical literature to suggest that this variance depends on patient response (Prosser et al. 2003). This assumption also likely leads to an underestimate of the performance of our adaptive policies, since more precise measurements (e.g., distinguishing volatilities) would improve the accuracy of identifying responders from non-responders. This is thus a conservative assumption, placing a lower bound on the performance of our policies.

Finally, we tacitly assume that the sole DM in question is the physician, and that patients adhere to the recommended treatments. This may be unrealistic, particularly when dealing with treatments that have severe side effects and documented large proportions of drop-outs, such as MS (Prosser et al. 2004). However, this further highlights the importance of adopting a dynamic treatment policy that can identify non-responders earlier, and take them off treatment. We leave for further research a model that takes into account incentives for encouraging patients' compliance.

Although our model simplifies the reality of chronic illnesses, it has the advantage of allowing exact analytical results, with simple and intuitive interpretations, as we discuss next.

3 Analysis

We briefly formalize the DM's problem, and then provide exact analytical results characterizing the optimal policy for several cases of interest. Since our emphasis is on interpreting the results in the context of chronic diseases (and particularly MS), we defer all the proofs to the Appendix.

Let $\mathcal{F}_t$ denote the σ-algebra generated by the allocations, samples, events and lump-sum rewards observed by time t, i.e., $\mathcal{F}_t \stackrel{\text{def}}{=} \sigma\big(\{\alpha_\tau, d\pi^0(\tau), d\pi^1(\tau), N_\tau, L_\tau\}_{0\leq\tau<t}\big)$, where N_τ denotes the number of life events in time interval $[0,\tau)$. This induces a filtration $\{\mathcal{F}_t\}_{t\geq 0}$.

In the context of our model, it can be readily seen that the belief that the risky arm is good, i.e., $p_t \stackrel{\text{def}}{=} \mathbb{P}\{\theta = G|\mathcal{F}_t\}$, is a sufficient statistic of the history up to time t, since the distributions

of rewards and relapses for each arm type are known, and there is uncertainty only with respect to the risky arm's type (Bolton and Harris 1999). Thus, it is natural to use p_t as the state of the system at time t. Then, α_t is the action (control) of the decision maker at time t, i.e, the fraction of treatment allocated to the risky arm. Finally, with $\mathcal{A} \stackrel{\text{def}}{=} \{(\alpha_t)_{t\geq 0} \,|\, \alpha_t : \mathcal{F}_t \to [0,1]\}$ denoting the set of all sequential, non-anticipative policies that are adapted to the available information, the DM's problem can be compactly formulated as

$$\max_{\alpha\in\mathcal{A}} \mathbb{E}^{\alpha}\left[\int_0^T e^{-rt}\big[d\pi^1(t) + d\pi^0(t) + \lambda(t,\theta)L_t dt\big]\right],$$

where the expectation is with respect to the stochastic processes $dZ^0(t), dZ^1(t), L_t, \alpha_t$ and p_t.

Throughout the analysis, we make no explicit assumption about the relationships between the values of λ_0 and $\lambda_{G,B}$ (or μ_0 and $\mu_{G,B}$), but we do require that $\frac{\mu_B}{r+\lambda_B} \leq \frac{\mu_0}{r+\lambda_0} \leq \frac{\mu_G}{r+\lambda_G}$, and that $\mu_B - D\lambda_B \leq \mu_0 - D\lambda_0 \leq \mu_G - D\lambda_G$. This corresponds to the natural condition that using the risky arm indefinitely should give expected discounted rewards below those of the safe arm if the risky arm is bad, and above if it is good.[12]

3.1 Optimizing Up to the First Life Event

We first find the DM's optimal policy when optimizing over a planning horizon up to the (random) time of the first life event, i.e., for $T = \inf\{t \geq 0 \,|\, N_t > 0\}$. As a first step, we characterize the evolution of the DM's belief in the absence of a life event during period $[t, t+dt)$, as a function of the current belief p_t and the DM's action α_t.

Lemma 1. *If a negative event does not occur in the time interval $[t, t+dt)$, the change in the DM's belief, $p_{t+dt} - p_t$, is distributed normally, with mean $\alpha_t p_t(1-p_t)(\lambda_B - \lambda_G)dt$ and variance $\alpha_t\phi(p_t)$, where $\phi(p) \stackrel{\text{def}}{=} \left(\frac{p(1-p)(\mu_G-\mu_B)}{\sigma}\right)^2 dt$.*

We make several observations about the result. First, note that the mean drift in the belief distribution is positive, and depends on the difference between the rates of life events in the good and bad scenarios, respectively. This is intuitive, because the absence of a negative event in the present time interval can be viewed as "good news" for the DM. Furthermore, the non-occurrence of a life event becomes more informative (i.e., increases the belief faster) as the rate difference, $\lambda_B - \lambda_G$, grows. The term $p(1-p)$ implies that the belief change is largest when the prior is least informative: as p gets closer to the extremes (0 or 1), it takes a much stronger signal to move the belief, as compared to when $p \approx 0.5$. Finally, as expected, the change in belief is affected by the intensity with which the risky treatment is applied, as measured by α.

With this result, we can now provide a characterization of the DM's optimal adaptive policy.

[12] The requirement is typically true in the case of MS, since $\mu_G \approx \mu_B < \mu_0$ due to side effects, and $\lambda_G < \lambda_0 \approx \lambda_B$, meaning the treatment induces less frequent relapses and disease flare-ups if it is effective, and has no effect on their frequency if it is ineffective.

Theorem 1. *When T corresponds to the time of the first life event, the optimal policy is given by*

$$\alpha_t^*(p_t) = \begin{cases} 0 & \text{if } p_t < p^* \\ 1 & \text{otherwise,} \end{cases} \tag{3}$$

where

$$p^* \stackrel{\text{def}}{=} \frac{(\xi^* - 1)\left(\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B}\right)}{(\xi^* - 1)\left(\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B}\right) + (\xi^* + 1)\left(\frac{\mu_G}{r+\lambda_G} - \frac{\mu_0}{r+\lambda_0}\right)},$$

$$\xi^* \stackrel{\text{def}}{=} \frac{2(\lambda_B - \lambda_G)\sigma^2}{(\mu_G - \mu_B)^2} + \sqrt{1 + \frac{4(\lambda_B + \lambda_G + 2r)\sigma^2}{(\mu_G - \mu_B)^2} + \frac{4(\lambda_B - \lambda_G)^2\sigma^4}{(\mu_G - \mu_B)^4}}.$$

Theorem 1 confirms that the optimal policy is a threshold one; in particular, fractional allocations are not needed, and the DM can always select a single arm at each point of time. Note that the optimal threshold p^* depends on renormalized reward rates involving ratios of drift rates to the sum of discount and life event rates (e.g., $\frac{\mu_0}{r+\lambda_0}$). It can be readily verified that p^* is increasing in μ_0, reflecting the intuitive fact that, ceteris paribus, a safe treatment with higher instantaneous rewards makes the risky treatment less appealing. Furthermore, when $\lambda_G \leq \min(\lambda_B, \lambda_0)$ and $\mu_0 \geq \mu_G$, as in the case of MS, it can also be verified that p^* is decreasing in λ_0 (see Figure 1). This shows that a DM behaving optimally should be more prone to experimenting with a risky alternative when life events under the safe alternative become more frequent/likely. This effect is exacerbated here since the DM is optimizing only up to the first life event, so as λ_0 increases, the problem horizon shrinks. Finally, the threshold p^* is strictly increasing in σ and r, confirming that increased volatility and/or an increasing degree of myopic behavior lead to strictly less experimentation with the risky alternative.

In the context of chronic illnesses, these results confirm that the optimal treatment policy is a discontinuation rule: the patient is given the "risky" treatment (e.g., interferon-β in the case of MS) as long as the belief that she is responding is above a threshold. Once the belief falls below this threshold, the patient is taken off treatment, and since no "learning" occurs while on the safe treatment exclusively, the process of experimentation essentially stops. For MS, the results are in accordance[13] with the current best practice recommendations of the National Medical Advisory Board of the National Multiple Sclerosis Society (NMSS 2004), which state that "higher-dosed, more frequently administered formulations of interferon beta may provide better short-term clinical efficacy than lower, less frequently dosed formulations of interferon beta in relapsing MS." Our results exactly quantify when treatment should be stopped, and furthermore suggest that physicians should more readily recommend DMT when the safe alternative (e.g., symptom management through non-DMT medications) becomes less effective, either in terms of reducing

[13]Note that we are not necessarily suggesting using this model to construct a treatment policy for MS. In fact, the infinite-horizon model discussed in Section 3.2 will be more suitable for such purposes.

relapse frequency or in ameliorating current symptoms. Interestingly, the results also suggest that, upon the re-evaluation of treatment for patients who have not been receiving disease modifying therapy, a physician should more readily recommend therapy for patients whose condition (e.g., as reported through self-evaluations) has been relatively stable.

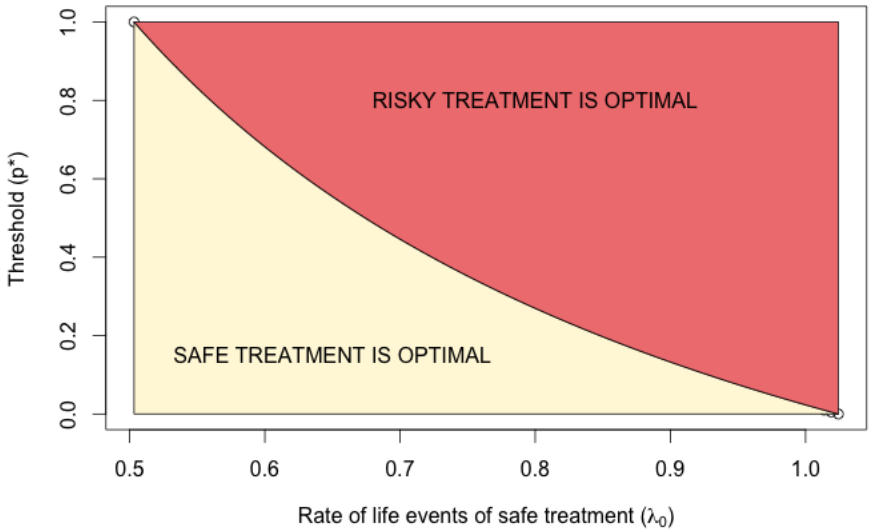

Figure 1: Optimal threshold p^* as a function of λ_0, the rate of (negative) life events under the safe treatment. Parameter values consistent with MS values were used for all other parameters (see Table 2 for details).

3.2 Optimizing With An Infinite Planning Horizon

When $T = \infty$, the distribution of the lump-sum "rewards" received upon a life event becomes relevant. We distinguish two cases, depending on whether the rewards' magnitude is influenced by the arms used.

Case 1: Independent Lump-Sum Rewards. We assume that, if a life event occurs in time interval $[t, t + dt)$, the DM incurs a reward of magnitude $-D$, regardless of the allocation α_t to the risky arm (α_t still affects the rate at which life events occur). We adopt the negative sign for convenience, to bring our model closer to the medical applications, where life events are typically associated with *disutilities* for patients. Here, we also assume that the model parameters satisfy $\mu_B - D\lambda_B \leq \mu_0 - D\lambda_0 \leq \mu_G - D\lambda_G$, consistent with the interpretation that a good arm dominates the safe alternative (which, in turn, dominates a bad arm), in terms of the total (instantaneous plus lump-sum) rewards per unit of time.

As discussed in Section 2.1, such a model is particularly pertinent for MS, where life events (i.e., relapses) occur among all patients, but less often among patients responding to treatment. Relapses are periods of acute disease activity, when patients experience neurological symptoms such as sudden paralysis or loss of vision, and thus bring patients a sudden decrement in their quality of life (Cutter et al. 1999, Horakova et al. 2012, Lublin et al. 1996).

In the absence of a life event, the DM's belief update process is still characterized by Lemma 1, as information is still acquired only from the instantaneous rewards. However, the belief update is different upon the occurrence of a life event, as characterized in the next result.

Lemma 2. *Upon the occurrence of a life event at time t, the belief p_t (just prior to the occurrence) jumps to the value $j(\alpha_t, p_t) \stackrel{\text{def}}{=} \frac{p_t(\alpha_t\lambda_G+(1-\alpha_t)\lambda_0)}{(1-\alpha_t)\lambda_0+\alpha_t\lambda(p_t)}$, where $\lambda(p_t) \stackrel{\text{def}}{=} p_t\lambda_G + (1-p_t)\lambda_B$.*

As expected, the occurrence of a life event results in a jump in the DM's belief (note that this update is in addition to the typical update performed after observing the instantaneous rewards). It can be readily verified that this updated belief $j(\alpha_t, p_t)$ is increasing in p and λ_G, and decreasing in λ_B. This confirms the intuition that, ceteris paribus, the occurrence of a life event makes it more believable that a risky arm is good when (a) the prior belief that it was good was larger, (b) life events become more likely under a good arm, or (c) life events become less likely under a bad arm.

Furthermore, when $\lambda_G < \lambda_B$, it can be verified that $j(\alpha_t, p_t)$ is increasing in λ_0 and decreasing in α_t. This has interesting implications in the context of MS, where the inequality holds. It suggests that, as long as there is reason to believe that treatment decreases the relapse rate in responders, then the occurrence of a relapse should tilt physicians towards considering the patient to be a responder when the relapse rate without treatment increases (e.g., due to natural disease progression). Furthermore, it agrees with the intuition that physicians who administer more aggressive treatments are more likely to become skeptical about the efficiency of treatment upon the occurrence of a relapse.

We next characterize the DM's optimal policy under the new modeling assumptions with an infinite planning horizon, and constant, deterministic lump-sum rewards. The next result confirms that the policy remains a threshold one.

Theorem 2. *The optimal treatment policy is given by*

$$\alpha_t^*(p_t) = \begin{cases} 0 & \text{if } p_t < p^* \\ 1 & \text{otherwise,} \end{cases} \tag{4}$$

where

$$p^* \stackrel{\text{def}}{=} \frac{\nu^* \left[(\mu_0 - \mu_B) - D(\lambda_0 - \lambda_B)\right]}{(1+\nu^*)\left[(\mu_G - \mu_0) - D(\lambda_G - \lambda_0)\right] + \nu^* \left[(\mu_0 - \mu_B) - D(\lambda_0 - \lambda_B)\right]} \tag{5a}$$

$$\nu^* \stackrel{\text{def}}{=} \left\{ \nu > 0 \,|\, \lambda_B + r + \left(\lambda_B - \lambda_G - \frac{(\mu_G - \mu_B)^2}{2\sigma^2}\right)\nu - \frac{(\mu_G - \mu_B)^2}{2\sigma^2}\nu^2 = \lambda_B \left(\frac{\lambda_B}{\lambda_G}\right)^{\nu} \right\}. \tag{5b}$$

The expression of the optimal threshold policy cannot be written explicitly in this case, since equation (5b) does not have a closed-form solution. However, a positive solution is always guaranteed to exist (as shown in the Appendix), and the optimal policy can be easily found numerically.

Similar to our previous model, the fact that the optimal policy is "bang-bang", interpreted in a medical context, has the appeal of simplifying treatment by only requiring a maximal dose or a zero dose of a drug whose effect on the patient is still unknown. Furthermore, it suggests that

administering higher or more frequent doses early in the treatment process can be more effective when dealing with chronic diseases, consistent with current recommendations for MS (NMSS 2004).

When $\lambda_G < \lambda_B$ and $\mu_G \geq \mu_B$ (as is the case in MS), the threshold p^* is strictly decreasing (increasing) in λ_0 (μ_0). This confirms that the risky arm becomes more preferable as the safe alternative becomes less attractive, due to either instantaneous rewards or increased rates of (negative) life events. In conjunction with the results of Lemma 2, this further emphasizes that an increased frequency of relapses under the safe treatment makes a physician more prone to increase dosages of risky treatments, and this effect is even stronger upon the occurrence of a relapse. As in the case of optimizing rewards up to the first life event, the threshold for this case is increasing in r and σ as well, confirming that more myopic behavior or increased volatility leads to less experimentation with the risky arm. Furthermore, when $\lambda_G < \min(\lambda_0, \lambda_B)$ and $\mu_0 > \max(\mu_G, \mu_B)$, which is consistent with MS (see our earlier discussion in Section 2.1), it can be verified that p^* is strictly decreasing in D, confirming the intuition that a physician should more readily recommend the risky treatment if the severity of relapses increases.

It is also instructive to compare the threshold p^* derived above with the threshold corresponding to the model in Section 1. This is done numerically, in Figure 2. As expected, when the disutility from life events is sufficiently small, the safe treatment becomes more appealing than the risky treatment under an infinite-horizon model, and hence a DM becomes more conservative, removing the patient from the risky drug earlier. As the disutility D increases, however, experimenting with the risky drug in the hope of reducing the frequency of relapses becomes more appealing under an infinite-horizon model.

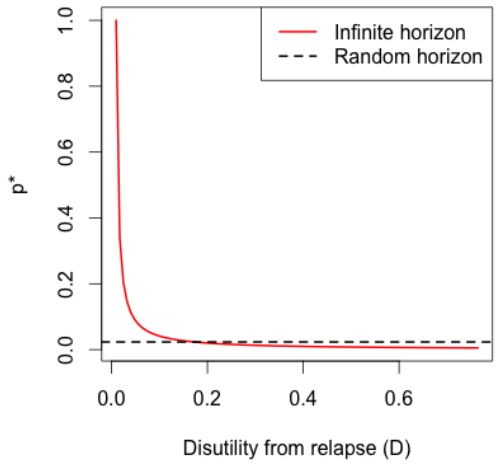

Figure 2: Comparison of optimal thresholds under random and infinite horizons. Here, $\mu_0 > \max(\mu_G, \mu_B)$, and all parameters except D were same as those in Table 2, consistent with the case of MS.

Case 2: Lump-Sum Rewards Dependent on the Type θ. We now extend our model to a case where the severity of life events can depend on the unknown type θ. More precisely, we assume that life events can be either "mild" or "severe". A mild event brings a "reward" (i.e., disutility, in the context of medical applications) of size $-D_M$, while a "severe" event brings a reward of

magnitude $-D_S$, where $D_M < D_S$. Furthermore, the likelihood of a mild/severe event depends on the allocation α_t used by the DM: if α is allocated to the risky arm, then the probability of a mild event is $\bar{p}_\theta = (1-\alpha)p_0 + \alpha p_\theta$ with $\theta \in \{B, G\}$, where p_θ is the probability that a given relapse is mild given response type θ (and, typically, $p_G > p_B$).

In the context of chronic diseases, allowing the severity of health events to depend on the response type essentially amounts to an assumption that a successful treatment has the potential to reduce the magnitude/impact of major negative health events, apart from just decreasing their likelihood. This may be a relevant modeling feature for diseases such as depression (Driessen et al. 2010, Fournier et al. 2010) or Crohn's disease (Lichtenstein et al. 2009).

We now discuss the belief updating and optimal policies in the current model. As with our two prior models, when no event occurs during the time interval $[t, t+dt)$, the belief is updated according to Lemma 1. However, when a life event occurs, the belief update now also depends on the magnitude of the "rewards" received by the DM. This is characterized in the following lemma.

Lemma 3. *If a mild (severe) event occurs at time t, when a fraction α_t of treatment is allocated to the risky arm, then the belief p_t jumps to a value of $j_M(\alpha_t, p_t)$ ($j_S(\alpha_t, p_t)$, respectively), where*

$$
\begin{aligned}
j_M(\alpha_t, p_t) &\stackrel{\text{def}}{=} \frac{p_t \bar{p}_G \bar{\lambda}_G}{(1-p_t)\bar{p}_B \bar{\lambda}_B + p_t \bar{p}_G \bar{\lambda}_G}, \\
j_S(\alpha_t, p_t) &\stackrel{\text{def}}{=} \frac{p_t (1-\bar{p}_G) \bar{\lambda}_G}{(1-p_t)(1-\bar{p}_B)\bar{\lambda}_B + p_t (1-\bar{p}_G)\bar{\lambda}_G}, \\
\bar{\lambda}_G &\stackrel{\text{def}}{=} (1-\alpha_t)\lambda_0 + \alpha_t \lambda_G, \\
\bar{\lambda}_B &\stackrel{\text{def}}{=} (1-\alpha_t)\lambda_0 + \alpha_t \lambda_B.
\end{aligned}
$$

As in Case 3.2, the occurrence of a life event, whether mild or severe, results in a jump in the DM's belief. It can also be verified that the updated beliefs $j_M(\alpha_t, p_t)$ and $j_S(\alpha_t, p_t)$ are increasing in p and λ_G, and decreasing in λ_B, again confirming the intuition that, ceteris paribus, the occurrence of a life event makes it more believable that a risky arm is good when (a) the prior belief that it was good was larger, (b) life events become more likely under a good arm, or (c) life events become less likely under a bad arm. In addition, j_M (j_S) is increasing (decreasing) in p_G, confirming that a mild (severe) event is more (less) indicative of the risky arm being good when the probability of a mild event under a good arm is larger. As before, the exact reverse comparative statics hold with respect to p_B. Furthermore, when $\lambda_G < \lambda_B$, both $j_M(\alpha_t, p_t)$ and $j_S(\alpha_t, p_t)$ are increasing in λ_0 and decreasing in p_0; this suggests that the DM is increasingly likely to deem the risky arm good as the safe arm becomes less attractive (either due to higher frequency or increased likelihood of severe events).

Unfortunately, finding a closed-form expression for the optimal policy in this case is no longer analytically tractable, due to the nonlinear dependency on α_t induced by the belief jumps. However, by examining the extreme case $p_t = 1$, we can derive the following insights.

Theorem 3. *1. If $D_M < D_S$, $\lambda_0 > \lambda_G$, and $p_0 < p_G$, then the optimal allocation for $p_t = 1$,*

i.e., $\alpha_t^*(1)$, *is given by the expression*

$$\alpha_t^*(1) = \frac{\frac{\mu_G - \mu_0}{D_S - D_M} + (\lambda_0 - \lambda_G)(\frac{D_S}{D_S - D_M} - p_0) + \lambda_0(p_G - p_0)}{2(p_G - p_0)(\lambda_0 - \lambda_G)}$$

2. *Furthermore, if* $-(\lambda_0 - \lambda_G)(D_M p_0 + D_S(1 - p_0)) - \lambda_0(p_G - p_0)(D_S - D_M) < \mu_G - \mu_0$ *and* $\mu_G - \mu_0 < (D_S - D_M)[p_G(\lambda_0 - \lambda_G) - \lambda_G(p_G - p_0)] - D_S(\lambda_0 - \lambda_G)$ *both hold, then* $\alpha_t^*(1) \in (0, 1)$*, and the optimal policy is not bang-bang even when* $p_t = 1$.

While Theorem 3 does not yield a general form for the optimal allocation, it does provide the insight that even when the risky arm is guaranteed to be "good", a fractional allocation may still be strictly better than a complete allocation to the risky arm. This occurs in cases when a good risky arm has smaller instantaneous rewards compared to the safe arm, but also lower relapse rates, so that mixing the two arms might achieve "the best of both worlds".

To put this insight into a medical context, recent studies on MS have estimated that $D_S = 0.0252$, $D_M = 0.0076$, $\lambda_0 = 0.083$, $\lambda_G = 0.0416$ (Horakova et al. 2012, Lee et al. 2012). Such studies usually assume that the probability of a relapse being severe is independent of treatment status. However, it is easy to verify that if $\mu_G - \mu_0 = -0.001$, $p_0 = 0.2$ and $p_G = 0.8$, then the condition of part 2 of Theorem 3 is satisfied, and hence $\alpha^*(1) \in (0, 1)$. In other words, a physician should strictly favor lower DMT dosages combined with the safe treatment, as this would always allow some benefits (in terms of immediate alleviation of symptoms) compared to using the DMT exclusively. Although DMT dosages in MS are typically fixed, and not adjusted for individual patients (NMSS 2004, 2008), Theorem 3 can nonetheless be useful in developing guidelines on how to combine treatments or how to adaptively increase or decrease dosage over time for other chronic diseases, such as depression, Crohn's disease or herpes infection.

4 Case study: Multiple Sclerosis

We illustrate our analytical results by developing an adaptive treatment policy for MS. In MS, affected individuals experience increasing disability to the point of becoming bedridden, as well as blurred vision, muscle weakness, dizziness, fatigue and various sensory abnormalities (Cutter et al. 1999). No biomarkers exist to accurately assess treatment responsiveness. Instead, practitioners rely on MRI scans or surveys in which patient-reported symptoms are used to compute the Expanded Disability Status Scale (EDSS), which can be translated into quality-of-life utilities (Prosser et al. 2003).

The most common form of MS, found in about 80% of cases, is relapsing-remitting multiple sclerosis (RRMS) (Sorensen 2005), on which we focus in our case study. The initial stage of the disease, which typically lasts for an extended period of time (10 years on average) is characterized by clearly defined relapses that occur on average once per year (Prosser et al. 2004), from which patients may or may not fully recover (Lublin et al. 1996). After this stage, patients typically enter the progressive stage of the disease, characterized by gradual worsening of disability (Kremenchutzky

et al. 2006). Typically, relapse rates decrease over time for all patients regardless of treatment, with rates for responders generally lower than for non-responders (Horakova et al. 2012). Mortality for MS patients depends on both age and current level of disability (Prosser et al. 2004).

MS is an incurable disease, and disease-modifying therapies (DMTs) attempt to slow progression and reduce relapses (NMSS 2014). The most common treatments used involve injectable DMTs, such as interferon-β preparations and glatiramer acetate, and more recently, oral DMTs such as dimethyl fumarate (approved for use in the US in 2013), teriflunomide (approved in 2012), fingolimod (approved in 2010) and natalizumab (approved in 2004) (see Molyneux et al. 2000, NMSS 2014, Rovaris et al. 2001, for more details). Interferon-β is often the first treatment prescribed, as the newer therapies, especially fingolimod and natalizumab, have been associated with an increased risk of severe side effects, such as potentially fatal infections, tumor development, lowering of cardiac rate, and encephalitis (inflammation of the brain) (see, e.g., Cohen et al. 2010, Goodin et al. 2008, Kleinschmidt-DeMasters and Tyler 2005). Furthermore, whereas the response profile to interferon has been well documented (Horakova et al. 2012, Romeo et al. 2013), the long-term effectiveness of oral medications has not been established (Carroll 2010).

Our goal in this section is to build a support tool that can inform medical decision makers about choosing between interferon-β treatment (risky option) and symptom management without DMT (safe option). This decision problem is especially important because patients receiving interferon-β treatment experience a significant decrease in quality of life due to side effects, such as pain at local injection site, flu-like symptoms, depression or allergic reactions. Furthermore, treatment also generates significant health care costs. We use symptom management as the safe treatment due to the lack of data on the long-term effectiveness of oral medications.

We first describe a detailed disease model inspired by the medical literature, which we use to simulate a hypothetical cohort of 10,000 responders and 10,000 non-responders to interferon-β. We consider an equal proportion of responders and non-responders because about 52% of patients are estimated to be non-responders (Horakova et al. 2012). We then compare the outcomes for these patients under two treatments: one corresponding to the current guidelines for interferon therapy, and the second corresponding to an adaptive policy derived using our results in Section 3.2.

4.1 Disease Model

We implement a disease model similar to that used in Prosser et al. (2004) and Lee et al. (2012). The disease progression is modeled as a Markov chain, with states given by the patient's Kurtzke Expanded Disability Status Scale (EDSS),[14] and whether or not she is currently experiencing a relapse (see Figure 3 for details).

As in Lee et al. (2012) and previous MS models, our simulation follows a hypothetical cohort of

[14]We choose to focus on EDSS instead of MRI in our study for several pragmatic reasons. First, EDSS is considerably more widespread, and there is no consensus in the medical community concerning the use of MRI for monitoring therapeutic response in MS (see Cohen et al. 2004). Second, MRI scans may not be available for a large subset of the population, or may be difficult or costly to administer frequently. Third, there is insufficient data in medical studies concerning the difference in MRI scans between responders and non-responders.

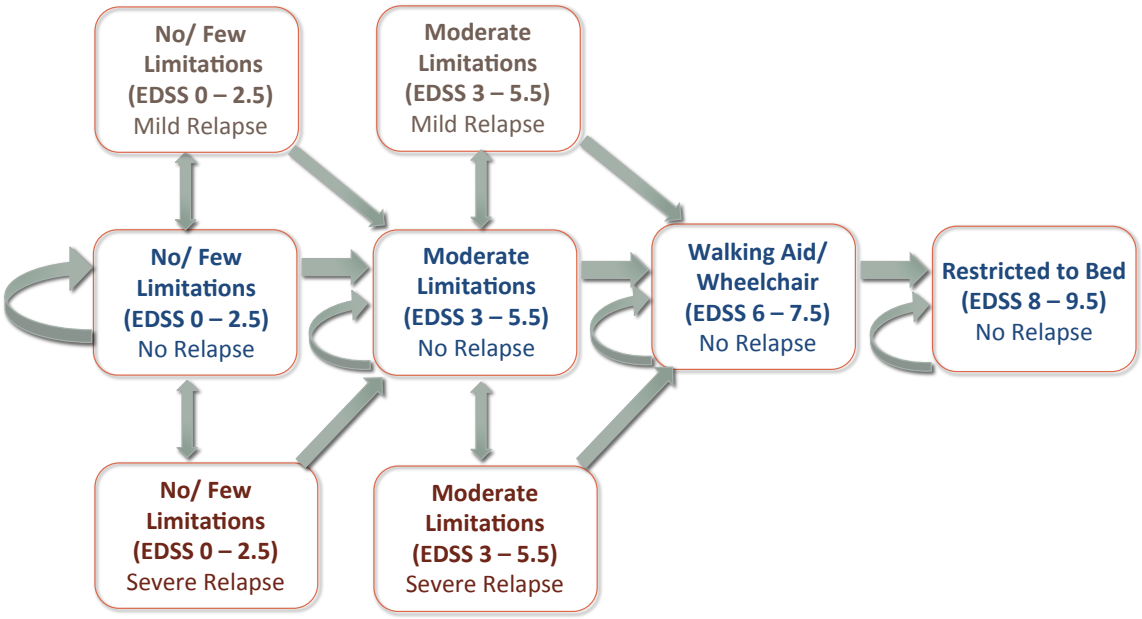

Figure 3: Multiple sclerosis disease model. All patients start in the lowest disability state (EDSS 0-2.5). In each month, a given patient can (1) remain in the same state without a relapse, (2) progress to the next level of disability, or (3) experience a relapse (while in the two lowest disability states), which can be either mild or severe, and which lasts exactly one month.

37-year-old RRMS patients with an initial EDSS score of 0-2.5. The cohort includes 10,000 responders and 10,000 non-responders, consistent with studies documenting the proportion of responders to interferon-β in the population (Horakova et al. 2012). We utilize a one-month time step, and simulate patients over a 10-year time horizon. Each patient in our simulation can transition from the score of 0-2.5 (no or few limitations) to a score of 3-5.5 (mild to moderate mobility limitations), and from there to a score of 6-7.5 (requiring a walking aid), and finally to a score of 8-9.5 (restricted to bed). While in EDSS score states 0-2.5 or 3-5.5, patients can experience a relapse, which can be either mild/moderate or severe, and which lasts for exactly one month, after which they can either remain in their pre-relapse disability level or progress to the next disability level. Once in EDSS state 6-7.5, patients are assumed to have entered the secondary-progressive stage of the disease, characterized by no relapses and gradual destruction of neurons. We also make several other modeling assumptions, consistent with medical studies evaluating the effectiveness of MS treatments (Lee et al. 2012, Prosser et al. 2004).[15]

Each state in the chain is associated with a mean quality-adjusted life year (QALY) value capturing a patient's (quality-of-life) utility, with a year in perfect health having a mean QALY of

[15]More precisely, we allow the probability of progressing to the next EDSS state to depend on the current EDSS state and on response type (when the patient is on treatment), but not on whether the patient is experiencing a relapse. The probability of a relapse is the same for EDSS states 0-2.5 and 3-5.5, and equal to zero for higher disability states. When the patient is on treatment, we also allow this probability to depend on response type. We model relapses as either mild/moderate or severe, independently of EDSS state, treatment or response type; deaths can occur from all EDSS levels depending on patient's age, with MS-related deaths only occurring while in EDSS state 8-9.5.

1, and death having a mean QALY of 0. The realized QALY in a given EDSS state is normally distributed around the mean value, with a variance that is consistent with quality-of-life surveys (Prosser et al. 2003). The mean QALYs of low disability (EDSS) states are higher than those of high disability states; also, at any given disability level, the mean QALYs of non-relapse states are higher than those of relapse states. A responding patient on treatment spends less time in relapse states, and advances slower through the progression model than a non-responding patient or a patient not on treatment. However, being on treatment reduces the QALY associated with each state, for both responders and non-responders.

In addition to QALYs, each state of the chain also has an associated cost value, summarizing the direct and indirect monthly costs of treatment. When measuring treatment outcomes, we conduct the analysis from a societal perspective, by aggregating the QALYs or costs across all patients, and discounting at an annual rate of 3% (Gold 1996).

The parameter values used in our study are summarized in Tables 1-3, and are taken from published epidemiological studies, consistent with Lee et al. (2012) and Horakova et al. (2012). To the best of our knowledge, no study reports difference in QALY on treatment between responders and non-responders. We assume that a responder has a small (0.0012 per month on average) increase in quality of life compared to being a non-responder, and we vary this value in sensitivity analysis.

Monthly costs (in 2011 USD)	Value	Range
Interferon-β treatment	3,062	2,297-3,828
Direct costs / disability level		
EDSS 0 - 2.5	536	402-607
EDSS 3 - 5.5	1,037	778-1,296
EDSS 6 - 7.5	2,460	1,845-3,075
EDSS 9 - 9.5	4,327	3,245-5,408
Direct costs per relapse		
Mild/ Moderate	104	0-200
Severe	5,215	3,911-6,519
Indirect costs/ disability level		
EDSS 0 - 2.5	1,421	1,066-1,776
EDSS 3 - 5.5	2,964	2,223-3,705
EDSS 6 - 7.5	3,124	2,343-3,905
EDSS 9 - 9.5	3,182	2,387-3,978

Table 1: Direct and indirect monthly costs for base case analysis. The source for all values is Lee et al. (2012).

4.2 Treatment Policies

We compare the standard interferon-β treatment described in Lee et al. (2012) to an adaptive treatment policy based on the model described in Section 3.2.

Under the standard treatment policy, we assume that all patients are started on interferon-β, and remain on treatment until reaching EDSS disability state 6-7.5, at which point they discontinue treatment. This is consistent with current recommendations of maintaining the patient on treatment indefinitely (Río et al. 2011), and has been modeled similarly in previous studies (Lee

Parameter	Value	Range	Source
Utility Means		(monthly)	
Baseline utilities/ disability level			
EDSS 0 - 2.5	0.0687	0.0515-0.0833	(1)
EDSS 3 - 5.5	0.0566	0.0424-0.0708	(1)
EDSS 6 - 7.5	0.0444	0.0333-0.0555	(1)
EDSS 9 - 9.5	0.0409	0.0307-0.0512	(1)
Reduction in utility from treatment in first 6 months	0.0096	0.0038-0.0154	(1)
Reduction in utility from treatment after first 6 months	0.001	0 - 0.0096.	(1,2)
Change in utility on treatment, due to response type			
Responder	+0.00058	0-0.002	–
Non-responder	-0.00058	-0.002-0	–
Reduction in utility from relapse			
Mild/ Moderate	0.0076	0.0053-0.0099	(1)
Severe	0.0252	0.0198-0.0305	(1)
Utility Standard Deviations			
If no relapse	0.02136	0.0029 - 0.038	(1)
If relapse	0.001	–	–

Table 2: Utility parameters for base case analysis. The sources are (1) Lee et al. (2012), and (2) Prosser et al. (2004).

Parameter	Value	Range	Source
Probability of progressing to next disability level		(monthly)	
If not in treatment, or non-responder			
EDSS 0 - 2.5	0.004438	0.0033-0.0055	(1)
EDSS 3 - 5.5	0.009189	0.0070-0.0115	(1)
EDSS 6 - 7.5	0.003583	0.0027-0.0045	(1)
EDSS 9 - 9.5	0.000952	0.0007-0.0012	(1)
Treatment effect for responders			
Relative rate of progression	0.5	0.38-1.00	(2)
Relative rate of relapse	0.5	0.33-0.90	(2)
Probability of treatment discontinuation	0.0087	0-0.0174	(1)
Probability of relapse	0.0799	0.0566-0.0944	(1,2,3)
Probability of severe relapse (given a relapse)	0.23	0.14-0.56	(1)

Table 3: Probabilities and rates for base case analysis. The sources are (1) Lee et al. (2012), (2) Horakova et al. (2012), and (3) Prosser et al. (2004).

et al. 2012, Prosser et al. 2004). Patients in earlier disability states may also abandon treatment in any month within the first three years with a fixed probability, consistent with the abandonment rate observed in randomized clinical trials (Cohen et al. 2010, Prosser et al. 2004). We assume that once a patient discontinues treatment, she will remain off treatment for the rest of her life (Lee et al. 2012, Prosser et al. 2004).

In the adaptive treatment policy, we consider a safe treatment consisting of symptom management, e.g, through non-DMT agents and standard care in case of relapses. The risky treatment represents interferon-β, with a "good" ("bad") type corresponding to a patient being a responder (non-responder), and a "bad" scenario corresponding to a non-responder. We start with a prior belief of 0.9 that the patient is a responder, and update this belief after each month spent in treatment, depending on the observed quality-of-life utility for that month (if not a relapse month), whether a relapse occurred, and whether the patient progressed to a higher disability state. Our base-case

prior is chosen so high in order to ensure that all patients are initially exposed to treatment, which is consistent with medical practice (NMSS 2008).

To determine the adaptive treatment policy, we use an objective corresponding to total discounted QALYs over an infinite horizon, as described in Section 3.2, where the decrement in health utility from a relapse is the expected disutility from a relapse, averaged over its severity. One notable difference between the optimization model in Section 3.2 and the disease model is that the former assumes constant rates of rewards and relapses, while, in reality these rates change over time. In fact, treatment can improve not only the relapse rates but also the rates of disease progression, so that using only the relapse rate in computing the optimal threshold would therefore render the adaptive policy somewhat myopic. To attenuate this, we add the rates of the progression to the rates of relapse; more precisely, when computing the optimal threshold p^* from Theorem 2 of Section 3.2, we take λ_0 as the sum of the monthly relapse rate under disease management and the natural disease progression rate for the current disability state of the patient, while λ_B (λ_G) is the sum of the monthly relapse rate and progression rate for non-responders (responders). As in the standard treatment, we assume that the patient can choose to abandon treatment with the same fixed probability in each month during the first three years when it is prescribed, and that once treatment is abandoned by patient choice, it is never restarted. All patients who reach EDSS state 6-7.5 discontinue treatment, as in the standard policy.

4.3 Results

4.3.1 Base case analysis

In our base case analysis, the optimal threshold for the belief above which treatment is initiated under our adaptive policy is about 64.5% for a patient in EDSS state 0-2.5, and 57.5% for a patient in EDSS state 3-5.5. These thresholds increase slightly over time (by about 0.06% per year) due to the increase in mortality rate as patients age. Compared to the standard interferon treatment, our policy gained slightly more QALYs, and incurred substantially lower costs. The benefits were present for both treatment responders and non-responders, with the latter group benefiting substantially more from the adaptive policy. For a responder, the average number of QALYs gained over the 10 simulated years was 6.459 for the standard policy and 6.453 for the adaptive policy, with an average cost of $482,358 under the standard policy and $472,178 under the adaptive policy, respectively (2.1% reduction in costs and 0.09% reduction in QALYs). For a non-responder, the average number of QALYs gained was 6.246 for the standard treatment and 6.294 for the adaptive treatment policy, with corresponding costs of $496,963 and $420,767, respectively (15.3% reduction in costs and 0.8% increase in QALYs). We note that, while the increase in QALYs may seem small on first sight, it becomes very significant when compared to the maximum possible improvement for non-responders. If the decision had perfect hindsight in our model, so that non-responders were *never* subjected to treatment, their gain in QALYs compared to the standard treatment would only be 2%. As such, the performance of our policy, which learns the information from observations, becomes quite relevant.

Figure 4 shows the histogram of QALYs gained under the two treatment policies for responders and non-responders over the 10-year time horizon, and Figure 5 shows the corresponding histogram of costs for the two groups. The large cost savings for non-responders occur because these patients are identified earlier and spend fewer months (on average) on treatment under the adaptive policy (56.6) than under the standard policy (87.1). In contrast, responders spend a similar amount of time on treatment under the standard and the adaptive policies (91.3 and 86.7 months, respectively).

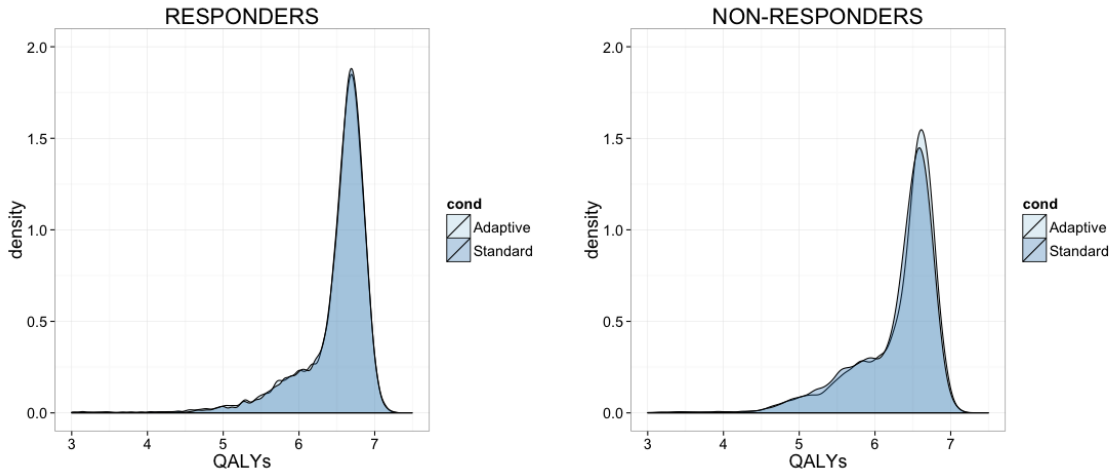

Figure 4: Histogram of QALYs gained over 10 years under the adaptive and standard treatment policies for responders and non-responders

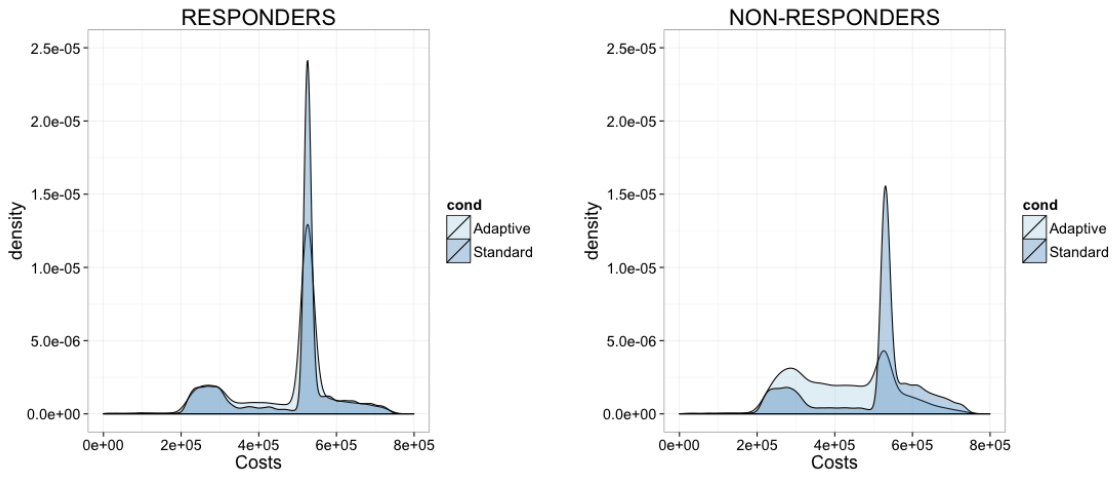

Figure 5: Histogram of costs incurred over 10 years under the adaptive and standard treatment policies for responders and non-responders

These results can be observed consistently for each year in our 10-year simulation, as displayed in Figure 6 and Figure 7, which also show the costs and QALYs for the "no treatment" policy. Not giving treatment to any patient provides a lower bound on the QALYs of responders and an upper bound on the QALYs of non-responders. Note that both costs and QALYs decrease over

time, under both standard and adaptive policies, and under both response types; this is due to disease progression, and the facts that all patients are taken off treatment once they reach EDSS state 6-7.5. However, non-responders experience more QALYs and incur significantly lower costs under the adaptive policy than under the standard policy every year after their first, and responders experience similar QALYs but incur slightly reduced costs. The difference between policies becomes larger especially in the first 4 years, as the adaptive policy learns the type of the patient.

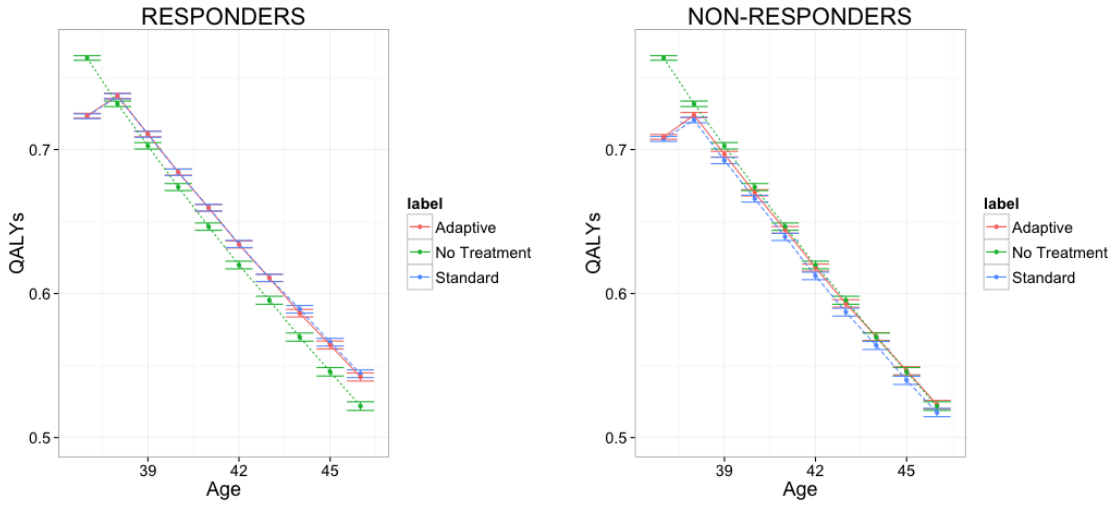

Figure 6: QALYs experienced under the adaptive and standard treatment policies for responders and non-responders: Yearly means and 99% CI of means

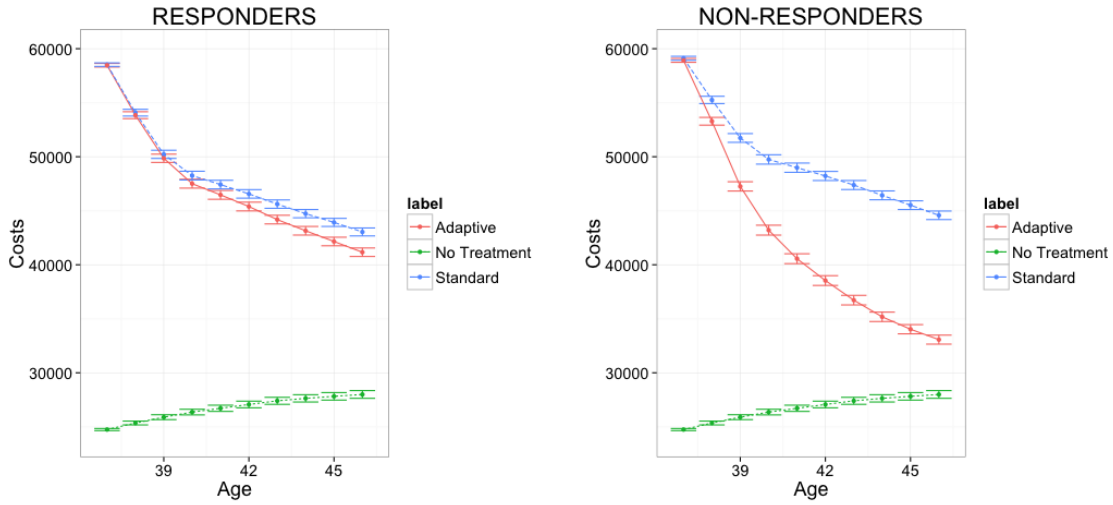

Figure 7: Costs incurred under the adaptive and standard treatment policies for responders and non-responders: Yearly means and 99% CI of means

4.3.2 Sensitivity Analysis

To test the robustness of our results, we performed a univariate sensitivity analysis in which we varied each parameter one at a time, using the lower and upper values listed in Tables 1-3, taken from published literature (Lee et al. 2012), while keeping the same adaptive policy derived from our base-case values. The results of this analysis are shown in Figure 8. In this scatter plot, each point represents the result of a simulation with 10,000 responders and 10,000 non-responders for a given set of parameters. In each plot, there are two points for each parameter shown in Tables 1-3, corresponding to the lowest and the highest range value, respectively. Baseline costs and QALYs for each EDSS state were varied together, in order to preserve the monotonicity of values for progressive disease states.

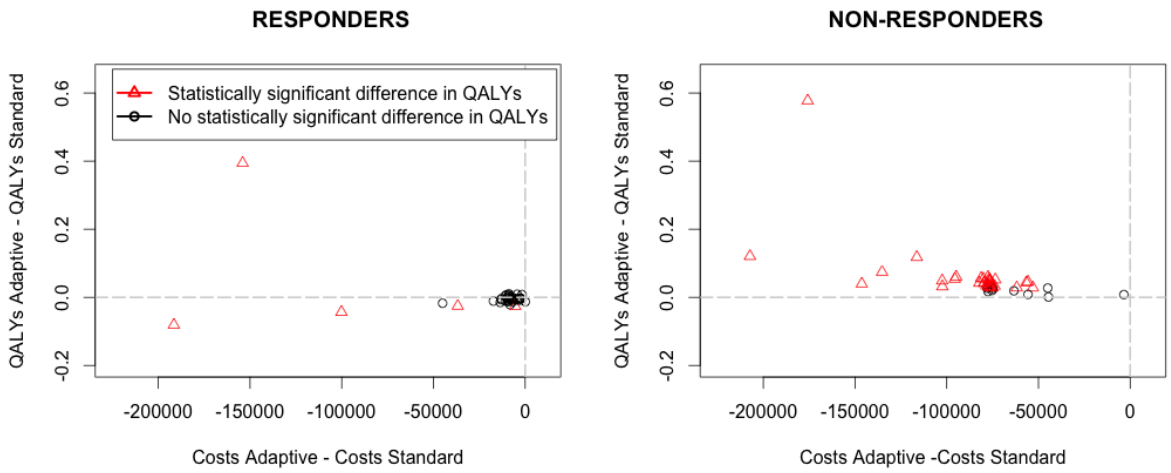

Figure 8: Results of univariate sensitivity analysis: Incremental costs and QALYs for responders and non-responders (adaptive treatment policy compared to standard treatment policy).

Note that, for non-responders, most of the points lie in the upper left quadrant, suggesting that the adaptive policy outperforms the standard policy, by incurring lower costs and generating more QALYs. Furthermore, the points displayed as triangles, which correspond to simulations where the difference in QALYs is statistically significant (i.e., absolute value of the difference in means is higher that the sum of the 95% confidence intervals of the means), all lie in the upper left quadrant, indicating that the adaptive policy is likely cost-saving. For responders, the points are closer to the origin than for non-responders, indicating that the expected costs and QALYs are likely similar under the two treatment policies, except in a few instances.

In particular, the QALYs of responders were most sensitive to the size of the disutility from treatment side effects, the duration of side effects, the standard deviation in the measurement noise, and the choice of prior. A high disutility from treatment after the initial 6 months, or maintaining the high initial 6-month disutility value throughout the entire modeling horizon leads to treatment being detrimental for both responders and non-responders because the benefits from treatment (slower progression and fewer relapses) are outweighed by side effects. In this case, our

adaptive policy significantly improves costs and outcomes for both responders and non-responders, by taking them off treatment earlier. A high value of measurement noise makes learning the response type more difficult and can cause responders to be mislabeled as non-responders and be taken off treatment prematurely. In our simulation, a high value of 0.038 standard deviation in the monthly health utility value (compared to 0.021 in our base case) led to an average loss of 0.043 QALYs for responders (from 6.43 to 6.39 over the 10 years), and no loss for non-responders. The prior, when chosen too low, can lead to the adaptive policy not recommending treatment for anyone: for example, a prior of 0.5 would make the adaptive policy a "no treatment" policy, since the optimal thresholds as computed by our optimization would be above this prior for both initial disability states when treatment can be prescribed. If the DM is very risk averse about harming responders, a high initial prior (0.9) is recommended.

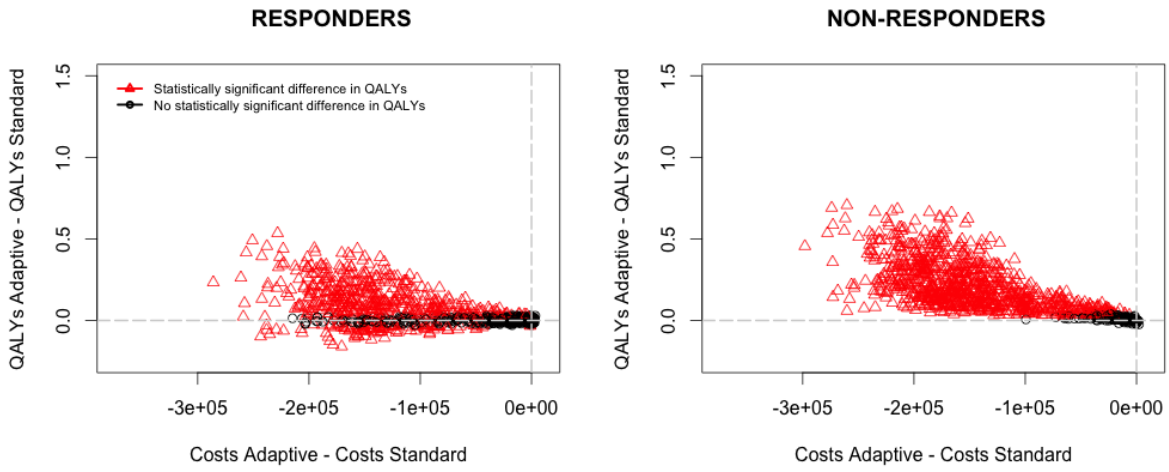

Figure 9: Results of probabilistic sensitivity analysis: Incremental costs and QALYs for responders and non-responders (adaptive treatment policy compared to standard treatment policy)

We also performed a probabilistic sensitivity analysis in which we randomly generated 1,000 sets of parameters. For each set, each parameter was randomly sampled from triangle distributions where the mode was given by the base case value, and the lowest and highest values corresponded to the ranges in Tables 1 - 3. The results of this analysis are shown in Figure 9. As can be seen, all statistically significant differences in mean QALYs between the adaptive and standard policies were positive for non-responders, indicating that our adaptive policy is especially cost-saving for non-responders. For responders, 79% of the differences in QALYs between the adaptive and standard policy were positive, suggesting that the adaptive policy is beneficial even for responders in almost 80% of the samples. At a 99% confidence level, the total QALYs (sum of responder and non-responder QALYs) of the standard policy exceeded those of the adaptive policy in only 1 sample out of 1000. Thus, while our policy does sometimes run the risk of harming the responders, it seems to generate substantial improvements in the overall population QALYs.

Given that approximately 52% of MS patients are non-responders (Horakova et al. 2012), we

estimate that, by using our adaptive treatment policy, more than $40,000 and 7 quality-adjusted days could be saved on average for each newly diagnosed MS patient over a 10-year horizon. Moreover, our treatment policy is cost-saving: because approximately 2,500 patients are diagnosed with RRMS in the US every year, we estimate that, on average over responders and non-responders, more than $104 million (8.5%) and 49 quality-adjusted years (0.3%) could be saved annually in the US by using an adaptive policy, without decreasing the average quality of life of RRMS patients.

5 Conclusions

This paper introduced a quantitative framework that can be used to inform treatment policies for chronic diseases sharing the following features: (1) there is *a priori* uncertainty about the extent to which a patient will respond to an available treatment; (2) observations of the effectiveness of treatment are noisy, and (3) there exists a risk of disease flare-ups, which depends on how well the patient is responding. We derived closed-form expressions for the optimal treatment policy in this context, and characterized cases when this corresponds (qualitatively) to the discontinuation rules used in practice. We then used our framework in a case study on multiple sclerosis (MS). When comparing the performance of an adaptive policy derived from our analytical results with the standard treatment policy, we found that the former clearly outperforms the latter, in terms of both patient outcomes as well as cost, by identifying non-responders early in the process.

While we illustrated our framework on MS, we believe that the ideas developed here could be useful for physicians or policy makers treating other chronic diseases that involve non trivial trade-offs between the short-term and long-term risks and benefits of treatment. Such examples might include celiac disease, rheumathoid arthritis, Crohn's disease, depression and other mental illnesses. We envision this model as part of a decision support tool that physicians could consult when evaluating the effectiveness of treatment for patients: the input would consist of information pertaining to observed quality-of-life measurements and occurrence of relapses or progression events since the patient's last evaluation, and the output would be an estimate of the likelihood that the patient is responding to treatment, along with a recommendation on how to continue treatment. Such a decision support tool could assist physicians in making decisions that explicitly balance trade-offs in a quantitative manner, leading to better outcomes as well as economic efficiency compared to current practice.

References

Abalos, E, L Duley, DW Steyn, DJ Henderson-Smart. 2007. Antihypertensive drug therapy for mild to moderate hypertension during pregnancy. *Cochrane Database of Systematic Reviews* **1** 1–69.

Adelman, G, SG Rane, KF Villa. 2013. The cost burden of multiple sclerosis in the United States: a systematic review of the literature. *Journal of Medical Economics* **16**(5) 639–647.

Ahuja, V, JR Birge. 2012. Fully adaptive designs for clinical trials: Simultaneous learning from multiple patients. *The 34th Annual Meeting of the Society for Medical Decision Making*. SMDM.

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

Barto, AG. 1998. *Reinforcement learning: An introduction*. MIT press.

Berry, DA. 1978. Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical Association* **73**(362) 339–345.

Berry, DA, B Fristedt. 1985. *Bandit problems: Sequential allocation of experiments*. London: Chapman and Hall.

Berry, DA., LM. Pearson. 1985. Optimal designs for clinical trials with dichotomous responses. *Statistics in Medicine* **4**(4) 497–508.

Bertsimas, D, A O'Hair, S Relyea, J Silberholz. 2014. An analytics approach to designing clinical trials for cancer. *Working paper* .

Boggild, M, J Palace, P Barton, Y Ben-Shlomo, T Bregenzer, C Dobson, R Gray. 2009. Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator. *BMJ: British Medical Journal* 1359–1363.

Bolton, P, C Harris. 1999. Strategic experimentation. *Econometrica* **67**(2) 349–374.

Carroll, WM. 2010. Oral therapy for multiple sclerosis–sea change or incremental step. *N Engl J Med* **362**(5) 456–8.

Cheng, Y, DA Berry. 2007. Optimal adaptive randomized designs for clinical trials. *Biometrika* **94**(3) 673–689.

Cohen, A, E Solan. 2013. Bandit problems with Lévy processes. *Mathematics of Operations Research* **38**(1) 92–107.

Cohen, B. A., O. Khan, D. R. Jeffery, K. Bashir, S. A. Rizvi, E. J. Fox, M. Agius, R. Bashir, T. E. Collins, R. Herndon, P. Kinkel, D. D. Mikol, M. A. Picone, V. Rivera, C. Tornatore, H. Zwibel. 2004. Identifying and treating patients with suboptimal responses. *Neurology* **63**(12 suppl 6) S33–S40.

Cohen, JA, F Barkhof, G Comi, HP Hartung, BO Khatri, X Montalban, J Pelletier, R Capra, P Gallo, G Izquierdo, et al. 2010. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *New England Journal of Medicine* **362**(5) 402–415.

Cutter, GR, ML Baier, RA Rudick, DL Cookfair, JS Fischer, J Petkau, K Syndulko, BG Weinshenker, JP Antel, C Confavreux, et al. 1999. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain* **122**(5) 871–882.

Denton, BT, M Kurt, ND Shah, SC Bryant, SA Smith. 2009. Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making* **29**(3) 351–367.

Driessen, E, P Cuijpers, SD Hollon, JM Dekker. 2010. Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. *Journal of Consulting and Clinical Psychology* **78**(5) 668.

Fournier, JC, RJ DeRubeis, SD Hollon, S Dimidjian, JD Amsterdam, RC Shelton, J Fawcett. 2010. Antidepressant drug effects and depression severity: A patient-level meta-analysis. *JAMA* **303**(1) 47–53.

Galloway, JB, KL Hyrich, LK Mercer, WG Dixon, B Fu, AP Ustianowski, KD Watson, M Lunt, DPM Symmons, et al. 2011. Anti-TNF therapy is associated with an increased risk of serious infections in patients with rheumatoid arthritis especially in the first 6 months of treatment: Updated results from

the British Society for Rheumatology Biologics Register with special emphasis on risks in the elderly. *Rheumatology* **50**(1) 124–131.

Garcia, DA, TP Baglin, JI Weitz, MM Samama, et al. 2012. Parenteral anticoagulants: Antithrombotic therapy and prevention of thrombosis: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest* **141**(2 Suppl) e24S–43S.

Gold, MR. 1996. *Cost-effectiveness in health and medicine*. Oxford University Press.

Goodin, DS, BA Cohen, P O'Connor, L Kappos, JC Stevens. 2008. Assessment: The use of natalizumab (Tysabri) for the treatment of multiple sclerosis (an evidence-based review) Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. *Neurology* **71**(10) 766–773.

Harrison, J, N Sunar. 2014. Investment timing with incomplete information and multiple means of learning. *Working paper* .

Helm, JE, MS Lavieri, MP Van Oyen, JD Stein, DC Musch. 2014. Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Working paper* .

Hirsh, J, AYY Lee. 2002. How we diagnose and treat deep vein thrombosis. *Blood* **99**(9) 3102–3110.

Horakova, D, T Kalincik, O Dolezal, J Krasensky, M Vaneckova, Z Seidl, E Havrdova. 2012. Early predictors of non-response to interferon in multiple sclerosis. *Acta Neurologica Scandinavica* **126**(6) 390–397.

Karatzas, I, ES Shreve. 1998. *Brownian motion and stochastic calculus*. Springer.

Keller, G, S Rady. 2010. Strategic experimentation with Poisson bandits. *Theoretical Economics* **5**(2) 275–311.

Keller, G, S Rady, M Cripps. 2005. Strategic experimentation with exponential bandits. *Econometrica* **73**(1) 39–68.

Kleinschmidt-DeMasters, BK, KL Tyler. 2005. Progressive multifocal leukoencephalopathy complicating treatment with natalizumab and interferon beta-1a for multiple sclerosis. *New England Journal of Medicine* **353**(4) 369–374.

Kobelt, G, J Berg, D Atherly, O Hadjimichael. 2006. Costs and quality of life in multiple sclerosis: A cross-sectional study in the United States. *Neurology* **66**(11) 1696–1702.

Kremenchutzky, M, GPA Rice, J Baskerville, DM Wingerchuk, GC Ebers. 2006. The natural history of multiple sclerosis: a geographically based study: Observations on the progressive phase of the disease. *Brain* **129**(3) 584–594.

Lee, S, DC Baxter, B Limone, MS Roberts, CI Coleman. 2012. Cost-effectiveness of fingolimod versus interferon beta-1a for relapsing remitting multiple sclerosis in the United States. *Journal of Medical Economics* **15**(6) 1088–1096.

Lichtenstein, Gary R, Stephen B Hanauer, William J Sandborn. 2009. Management of Crohn's disease in adults. *The American Journal of Gastroenterology* **104**(2) 465–483.

Lublin, FD, SC Reingold, et al. 1996. Defining the clinical course of multiple sclerosis: Results of an international survey. *Neurology* **46**(4) 907–911.

Mandelbaum, A, et al. 1987. Continuous multi-armed bandits and multiparameter processes. *The Annals of Probability* **15**(4) 1527–1556.

Mariette, X, M Matucci-Cerinic, K Pavelka, P Taylor, R van Vollenhoven, R Heatley, C Walsh, R Lawson, A Reynolds, P Emery. 2011. Malignancies associated with tumour necrosis factor inhibitors in registries

and prospective observational studies: A systematic review and meta-analysis. *Annals of the Rheumatic Diseases* **70**(11) 1895–1904.

Mason, JE, BT Denton, ND Shah, SA Smith. 2014. Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients. *European Journal of Operational Research* **233**(3) 727–738.

Molyneux, PD, L Kappos, C Polman, C Pozzilli, F Barkhof, M Filippi, T Yousry, D Hahn, K Wagner, M Ghazi, et al. 2000. The effect of interferon beta-1b treatment on MRI measures of cerebral atrophy in secondary progressive multiple sclerosis. *Brain* **123**(11) 2256–2263.

Murphy, SA, LM Collins. 2007. Customizing treatment to the patient: Adaptive treatment strategies. *Drug and Alcohol Dependence* **88**(Suppl 2) S1.

Neutel, JM, SS Franklin, P Lapuerta, A Bhaumik, A Ptaszynska. 2008. A comparison of the efficacy and safety of irbesartan/HCTZ combination therapy with irbesartan and HCTZ monotherapy in the treatment of moderate hypertension. *Journal of human hypertension* **22**(4) 266–274.

NMSS. 2004. Changing therapy in relapsing multiple sclerosis: Considerations and recommendations of a task force of the National Multiple Sclerosis Society. *National Multiple Sclerosis Society* .

NMSS. 2008. Disease management consensus statement. *National Multiple Sclerosis Society* .

NMSS. 2014. Brochure – The MS disease modifying medications. *National Multiple Sclerosis Society* .

Pincus, T, LF Callahan, WG Sale, AL Brooks, LE Payne, WK Vaughn. 1984. Severe functional declines, work disability, and increased mortality in seventy-five rheumatoid arthritis patients studied over nine years. *Arthritis & Rheumatism* **27**(8) 864–872.

Pineau, J, MG Bellemare, AJ Rush, A Ghizaru, SA Murphy. 2007. Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* **88** S52–S60.

Powell, WB. 2007. *Exploration vs. Exploitation*. John Wiley & Sons, Inc., 323–350.

Powell, WB, IO Ryzhov. 2012. *Optimal learning*, vol. 841. John Wiley & Sons.

Press, WH. 2009. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences* **106**(52) 22387–22392.

Prosperini, L, V Gallo, N Petsas, G Borriello, C Pozzilli. 2009. One-year MRI scan predicts clinical response to interferon beta in multiple sclerosis. *European Journal of Neurology* **16**(11) 1202–1209.

Prosser, LA, KM Kuntz, A Bar-Or, MC Weinstein. 2003. Patient and community preferences for treatments and health states in multiple sclerosis. *Multiple Sclerosis* **9**(3) 311–319.

Prosser, LA, KM Kuntz, A Bar-Or, MC Weinstein. 2004. Cost-effectiveness of interferon beta-1a, interferon beta-1b, and glatiramer acetate in newly diagnosed non-primary progressive multiple sclerosis. *Value in Health* **7**(5) 554–568.

Raftery, J. 2010. Multiple sclerosis risk sharing scheme: A costly failure. *BMJ* **340**.

Reichman, RC, GJ Badger, GJ Mertz, L Corey, DD Richman, JD Connor, D Redfield, MC Savoia, MN Oxman, Y Bryson, et al. 1984. Treatment of recurrent genital herpes simplex infections with oral acyclovir: a controlled trial. *JAMA* **251**(16) 2103–2107.

Río, J, M Comabella, X Montalban. 2011. Multiple sclerosis: current treatment algorithms. *Current Opinion in Neurology* **24**(3) 230.

Romeo, M, F Martinelli-Boneschi, M Rodegher, F Esposito, V Martinelli, G Comi, San Raffaele Multiple Sclerosis Clinical Group. 2013. Clinical and MRI predictors of response to interferon-beta and glati-

ramer acetate in relapsing-remitting multiple sclerosis patients. *European Journal of Neurology* **20**(7) 1060–1067.

Rovaris, M, G Comi, MA Rocca, JS Wolinsky, M Filippi, et al. 2001. Short-term brain volume change in relapsing–remitting multiple sclerosis: Effect of glatiramer acetate and implications. *Brain* **124**(9) 1803–1812.

Rudick, RA, WH Stuart, PA Calabresi, C Confavreux, SL Galetta, EW Radue, FD Lublin, B Weinstock-Guttman, DR Wynn, F Lynn, et al. 2006. Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *New England Journal of Medicine* **354**(9) 911–923.

Scalfari, A, A Neuhaus, A Degenhardt, GP Rice, PA Muraro, M Daumer, GC Ebers. 2010. The natural history of multiple sclerosis, a geographically based study 10: Relapses and long-term disability. *Brain* **133**(7) 1914–1929.

Sorensen, PS. 2005. Multiple sclerosis: Pathophysiology revisited. *The Lancet Neurology* **4**(1) 9–10.

Sudlow, CLM, CE Counsell. 2003. Problems with UK government's risk sharing scheme for assessing drugs for multiple sclerosis. *British Medical Journal* **326**(7385) 388.

Wald, A, D Carrell, M Remington, E Kexel, J Zeh, L Corey. 2002. Two-day regimen of acyclovir for treatment of recurrent genital herpes simplex virus type 2 infection. *Clinical Infectious Diseases* **34**(7) 944–948.

Wen, L, F Haoda. 2011. Bayesian optimal adaptive designs for delayed-response dose-finding studies. *Journal of Biopharmaceutical Statistics* **21**(5) 888 – 901.

Young, PL, LA Olsen. 2010. *The Healthcare Imperative:Lowering Costs and Improving Outcomes: Workshop Series Summary*. The National Academies Press.

Zhang, J, BT Denton, H Balasubramanian, ND Shah, BA Inman. 2012. Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management* **14**(4) 529–547.

Appendix: Proofs

Note: In the following proofs, we will often suppress the subscript t for ease of notation.

Lemma 1. *If a negative event does not occur in the time interval $[t, t+dt)$, the change in the DM's belief, $p_{t+dt} - p_t$, is distributed normally, with mean $\alpha_t p_t(1-p_t)(\lambda_B - \lambda_G)dt$ and variance $\alpha_t \phi(p_t)$, where $\phi(p) \stackrel{\text{def}}{=} \left(\frac{p(1-p)(\mu_G - \mu_B)}{\sigma}\right)^2 dt$.*

Proof of Lemma 1. The proof is similar to Bolton and Harris (1999), except we need to incorporate the fact that the lack of a life event during the interval $[t, t+dt)$, in addition to the rewards $d\pi^1(t)$, provides information for the update. The rewards $d\pi^1(t)$ are observationally equivalent to $d\tilde{\pi}^1(t) = \sqrt{\alpha}\tilde{\mu}dt + dZ^1(t)$, where $\tilde{\mu} = \mu/\sigma$. Using Bayes' rule, we have:

$$\begin{aligned} p(t+dt) &= \frac{\mathbb{P}(\text{reward, no event } |\theta = G)\,\mathbb{P}(\theta = G)}{\mathbb{P}(\text{reward, no event})} \\ &= \frac{pF(\tilde{\mu}_G)e^{-\bar{\lambda}_G dt}}{pF(\tilde{\mu}_G)e^{-\bar{\lambda}_G dt} + (1-p)F(\tilde{\mu}_B)e^{-\bar{\lambda}_B dt}} \end{aligned}$$

where $\bar{\lambda}_G \stackrel{\text{def}}{=} (1-\alpha)\lambda_0 + \alpha\lambda_G$, $\bar{\lambda}_B \stackrel{\text{def}}{=} (1-\alpha)\lambda_0 + \alpha\lambda_B$, $F(\tilde{\mu}) = \frac{1}{\sqrt{2\pi dt}} \exp\left(-\frac{1}{2dt}\left(d\tilde{\pi}^1(t) - \sqrt{\alpha}\tilde{\mu}dt\right)^2\right)$.

After Taylor-expanding the $e^{-\bar{\lambda}dt}$ terms and dropping terms of order dt^2 or higher, dp becomes

$$dp = \frac{p(1-p)\left[\tilde{F}(\tilde{\mu}_G) - \tilde{F}(\tilde{\mu}_B) - dt\left(\tilde{F}(\tilde{\mu}_G)\bar{\lambda}_G - \tilde{F}(\tilde{\mu}_B)\bar{\lambda}_B\right)\right]}{p\tilde{F}(\tilde{\mu}_G) + (1-p)\tilde{F}(\tilde{\mu}_B) - dt\left[p\tilde{F}(\tilde{\mu}_G)\bar{\lambda}_G + (1-p)\tilde{F}(\tilde{\mu}_B)\bar{\lambda}_B\right]} \tag{6}$$

where $\tilde{F}(\tilde{\mu}) = \exp\left(\sqrt{\alpha}\tilde{\mu}d\pi^1 - 1/2\alpha\tilde{\mu}^2 dt\right)$ and we suppressed the dependence on t.

Similar to Bolton and Harris (1999), one can show, by using Taylor expansions, that $\tilde{F}(\tilde{\mu}) = 1 + \sqrt{\alpha}\,\tilde{\mu}\,d\tilde{\pi} + o(dt)$, where by $o(x)$ we denote any function $f(x)$ such that $\lim_{x\to 0}\frac{f(x)}{x} = 0$. Substituting this into (6), we obtain, after some manipulations,

$$dp = \frac{p(1-p)\left(\sqrt{\alpha}(\tilde{\mu}_G - \tilde{\mu}_B)d\tilde{\pi} - (\bar{\lambda}_G - \bar{\lambda}_B)dt\right)}{1 + \sqrt{\alpha}\tilde{m}(p)d\tilde{\pi} - \bar{\lambda}(p)dt}, \tag{7}$$

where $\tilde{m}(p) \stackrel{\text{def}}{=} p\tilde{\mu}_G + (1-p)\tilde{\mu}_B$ and $\bar{\lambda}(p) \stackrel{\text{def}}{=} p\bar{\lambda}_G + (1-p)\bar{\lambda}_B$, and we drop all terms of order $dt^{\frac{3}{2}}$ or higher, as they go to zero in the limit. Also, it can be checked that

$$\frac{1}{1 + \sqrt{\alpha}\tilde{m}(p)d\tilde{\pi} - \bar{\lambda}(p)dt} = 1 - \sqrt{\alpha}\tilde{m}(p)d\tilde{\pi} + \bar{\lambda}(p)dt + o(dt).$$

Substituting this back into (7), we have

$$\begin{aligned} dp &= p(1-p)(\tilde{\mu}_G - \tilde{\mu}_B)(\sqrt{\alpha}d\tilde{\pi} - \alpha\tilde{m}(p)dt) - p(1-p)(\bar{\lambda}_G - \bar{\lambda}_B)dt + o(dt) \\ &= p(1-p)\frac{\mu_G - \mu_B}{\sigma}\sqrt{\alpha}dZ - \alpha p(1-p)(\lambda_G - \lambda_B)dt + o(dt). \end{aligned}$$

□

Theorem 1. *When T corresponds to the time of the first life event, the optimal policy is given by*

$$\alpha_t^*(p_t) = \begin{cases} 0 & \text{if } p_t < p^* \\ 1 & \text{otherwise,} \end{cases}$$

where

$$p^* \stackrel{\text{def}}{=} \frac{(\xi^* - 1)\left(\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B}\right)}{(\xi^* - 1)\left(\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B}\right) + (\xi^* + 1)\left(\frac{\mu_G}{r+\lambda_G} - \frac{\mu_0}{r+\lambda_0}\right)},$$

$$\xi^* \stackrel{\text{def}}{=} \frac{2(\lambda_B - \lambda_G)\sigma^2}{(\mu_G - \mu_B)^2} + \sqrt{1 + \frac{4(\lambda_B + \lambda_G + 2r)\sigma^2}{(\mu_G - \mu_B)^2} + \frac{4(\lambda_B - \lambda_G)^2\sigma^4}{(\mu_G - \mu_B)^4}}.$$

Proof of Theorem 1. With $u(p)$ denoting the optimal value function given state $p_t = p$, the Bellman recursion for our problem can be written as follows.

$$u(p) = \max_{\alpha \in [0,1]} \Big[(1-\alpha)\,\mu_0\,dt + \alpha\,\mathbb{E}[\mu]\,dt + e^{-rdt}\,\mathbb{E}\big[u(p+dp)\big]\,\mathbb{P}\big(\text{no event in } [t, t+dt)\big) + o(dt) \Big]. \quad (8)$$

The first two terms represent the immediate reward from allocating α to the risky treatment, while the third term denotes the discounted future reward. The final $o(dt)$ term comes from the expectation of total rewards conditional on a negative health event occurring in the next dt time period.

By Taylor-expanding the expression above, it can be shown that the optimal value function $u(p)$ satisfies the following differential equation:

$$u(p) = \frac{1}{r} \max_{\alpha} \Big[(1-\alpha)\mu_0 + \alpha m(p) + \alpha p(1-p)(\lambda_B - \lambda_G)u'(p) + \frac{1}{2}\alpha\phi(p)u''(p) - (1-\alpha)\lambda_0 u(p) - \alpha\lambda(p)u(p) \Big]. \quad (9)$$

A proof of this result is included in Lemma 4. Critically important, note that the expression inside the maximization in (9) is linear in α. Therefore, we immediately see that the optimal policy is bang-bang, i.e. $\alpha^* \in \{0, 1\}$, and we have:

$$\alpha^* = \begin{cases} 1, & \text{if } \frac{1}{2}\phi(p)u''(p) + p(1-p)(\lambda_B - \lambda_G)u'(p) - u(p)[\lambda(p) - \lambda_0] > \mu_0 - m(p), \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

In the region where $\alpha^* = 0$, the Bellman equation (9) implies that $u(p) = \frac{\mu_0}{r+\lambda_0}$ (as the u'' term disappears). When $\alpha^* = 1$, we obtain from (9) that u must satisfy the following second-order differential equation:

$$u(p)\big[r + \lambda(p)\big] = m(p) + p(1-p)(\lambda_B - \lambda_G)u'(p) + \frac{1}{2}\,\phi(p)\,u''(p), \quad (11)$$

where $m(p) \stackrel{\text{def}}{=} p\mu_G + (1-p)\mu_B$ and $\phi(p) \stackrel{\text{def}}{=} p^2(1-p)^2(\mu_G - \mu_B)^2/\sigma^2$.

We note that when $p = 1$, we have $\phi(1) = 0$, so that (11) immediately yields $u(1) = \frac{m(1)}{r+\lambda(1)} = \frac{\mu_G}{r+\lambda_G} > \frac{\mu_0}{r+\lambda_0}$, so it is optimal to play the risky arm. When $p = 0$, $\phi(0) = 0$ so that (11) implies $u(0) = \frac{m(0)}{r+\lambda(0)} = \frac{\mu_B}{r+\lambda_B} < \frac{\mu_0}{r+\lambda_0}$, so it is optimal to play the safe arm.

In fact, equation (11) can be solved explicitly for u^*. We first note that $p\frac{\mu_G}{r+\lambda_G} + (1-p)\frac{\mu_B}{r+\lambda_B}$ is a particular solution of (11). We then look for homogeneous solutions of the form

$$z(p) = (1-p)^{\frac{1+\xi}{2}} p^{\frac{1-\xi}{2}}.$$

Substituting $z(p)$ into (11), we obtain a quadratic equation in ξ, with one solution less than -1, and a positive solution $\xi^* \stackrel{\text{def}}{=} \frac{2(\lambda_B-\lambda_G)\sigma^2}{(\mu_G-\mu_B)^2} + \sqrt{1 + \frac{4(\lambda_B+\lambda_G+2r)\sigma^2}{(\mu_G-\mu_B)^2} + \frac{4(\lambda_B-\lambda_G)^2\sigma^4}{(\mu_G-\mu_B)^4}}$. The former solution is infeasible for us, since $z(p)$ would go to infinity as $p \to 1$. However, the solution ξ^* allows $z(p)$ to vanish in the limit $p \to 1$. We therefore now look for a solution of the form

$$u(p) = p\frac{\mu_G}{r+\lambda_G} + (1-p)\frac{\mu_B}{r+\lambda_B} + A(1-p)^{\frac{1+\xi^*}{2}} p^{\frac{1-\xi^*}{2}}.$$

By imposing the value matching and smooth pasting conditions, i.e., $u(p^*) = \frac{\mu_0}{r+\lambda_0}$ and $u'(p^*) = 0$, respectively, we obtain a system of two equations, which can be solved for p^* and A, yielding:

$$p^* = \frac{(\xi^*-1)\left(\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B}\right)}{(\xi^*-1)\left(\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B}\right) + (\xi^*+1)\left(\frac{\mu_G}{r+\lambda_G} - \frac{\mu_0}{r+\lambda_0}\right)}$$

$$A = \frac{\left(\frac{\mu_G}{r+\lambda_G} - \frac{\mu_B}{r+\lambda_B}\right)}{z^*(p^*)} \frac{2p^* - 1 + \xi^*}{2p^*(1-p^*)}.$$

Our optimal value function is therefore

$$u_t^*(p_t) = \begin{cases} \frac{\mu_0}{r+\lambda_0} & \text{if } p_t \in [0, p^*] \\ p_t\frac{\mu_G}{r+\lambda_G} + (1-p_t)\frac{\mu_B}{r+\lambda_B} + Az^*(p) & \text{otherwise.} \end{cases}$$

To finalize the proof, we must verify that the function above satisfies the optimality conditions stated in (10). For $p > p^*$, $p(1-p)(\lambda_B - \lambda_G)u'(p) + \frac{1}{2}\phi(p)u''(p) = u(p)(r+\lambda(p)) - m(p)$, so by (10) we need to check that $u(p)\big[r+\lambda(p)\big] - m(p) - u(p)(\lambda(p)-\lambda_0) > \mu_0 - m(p)$, which becomes $u(p) > \frac{\mu_0}{r+\lambda_0}, \forall\, p > p^*$, which holds since u is a convex increasing function on $[p^*, 1]$, increasing from $\frac{\mu_0}{r+\lambda_0}$ at p^*, to $\frac{\mu_G}{r+\lambda_G}$ at 1.

For $p < p^*$, checking (10) is equivalent to checking if $-\frac{\mu_0}{r+\lambda_0}(\lambda(p) - \lambda_0) < \mu_0 - m(p)$. This inequality is satisfied if and only if $p < \tilde{p}$, where

$$\tilde{p} \stackrel{\text{def}}{=} \frac{\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B}}{\frac{\mu_0}{r+\lambda_0} - \frac{\mu_B}{r+\lambda_B} + \left(\frac{r+\lambda_G}{r+\lambda_B}\right)\left(\frac{\mu_G}{r+\lambda_G} - \frac{\mu_0}{r+\lambda_0}\right)}.$$

Since p^* can be rewritten as $\frac{\left(\frac{\mu_0}{r+\lambda_0}-\frac{\mu_B}{r+\lambda_B}\right)}{\left(\frac{\mu_0}{r+\lambda_0}-\frac{\mu_B}{r+\lambda_B}\right)+\frac{\xi^*+1}{\xi^*-1}\left(\frac{\mu_G}{r+\lambda_G}-\frac{\mu_0}{r+\lambda_0}\right)}$, and $\frac{r+\lambda_G}{r+\lambda_B} < 1 < \frac{\xi^*+1}{\xi^*-1}$, it follows that $p^* < \tilde{p}$ and thus condition (10) is satisfied. Thus, our proposed value function satisfies the Bellman optimality conditions. □

Lemma 4. *In the context and notation of the proof of Theorem 1, the optimal value function $u(p)$ satisfies the following Bellman recursion:*

$$u(p) = \frac{1}{r}\max_{\alpha\in[0,1]}\Big[(1-\alpha)\mu_0 + \alpha m(p) + \alpha p(1-p)(\lambda_B-\lambda_G)u'(p) + \frac{1}{2}\alpha\phi(p)u''(p) - (1-\alpha)\lambda_0 u(p) - \alpha\lambda(p)u(p)\Big].$$

Proof. For conciseness, let A denote the event "no life event during $[t, t+dt)$". By Bayes' rule,

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A|\text{good})\mathbb{P}(\text{good}) + \mathbb{P}(A|\text{bad})\mathbb{P}(\text{bad}) \\ &= p\big[1-(1-\alpha)\lambda_0 dt - \alpha\lambda_G dt\big] + (1-p)\big[1-(1-\alpha)\lambda_0 dt - \alpha\lambda_B dt\big] \\ &= 1-(1-\alpha)\lambda_0 dt - \alpha\lambda(p)dt\end{aligned}$$

where $\lambda(p) \stackrel{\text{def}}{=} p\lambda_G + (1-p)\lambda_B$. Using Taylor series expansion, we also have:

$$u(p+dp) = u(p) + \alpha p(1-p)(\lambda_B-\lambda_G)u'(p)dt + \frac{1}{2}\alpha\phi(p)u''(p)dt + o(dt).$$

Replacing these into equation (8), and noting that $\mathbb{E}[\mu] = p\mu_G + (1-p)\mu_B \stackrel{\text{def}}{=} m(p)$, we obtain

$$u(p) = \max_{\alpha\in[0,1]}\Big[(1-\alpha)\mu_0 dt + \alpha m(p)dt + u(p) - ru(p)dt - (1-\alpha)\lambda_0 u(p)dt - \alpha\lambda(p)u(p)dt + \alpha p(1-p)(\lambda_B-\lambda_G)u'(p)dt + \frac{1}{2}\alpha\phi(p)u''(p)dt\Big],$$

where we have dropped all dt terms of power $3/2$ and higher. By simplifying the $u(p)$ terms in the equation above, dividing by dt and transferring terms, we obtain the desired result in (9). □

Lemma 2. *Upon the occurrence of a life event, the belief p_t (just prior to the occurrence) jumps to the value $j(\alpha_t, p_t) \stackrel{\text{def}}{=} \frac{p_t(\alpha_t\lambda_G + (1-\alpha_t)\lambda_0)}{\alpha_t\lambda(p_t) + (1-\alpha_t)\lambda_0}$, where $\lambda(p_t) \stackrel{\text{def}}{=} p_t\lambda_G + (1-p_t)\lambda_B$.*

Proof of Lemma 2. Let $\bar{\lambda}_G = (1-\alpha)\lambda_0 + \alpha\lambda_G$ and $\bar{\lambda}_B = (1-\alpha)\lambda_0 + \alpha\lambda_B$. By Bayes' rule, we have:

$$\begin{aligned} p_t &= \frac{\mathbb{P}\{\text{life event } |\theta = G\}\ \mathbb{P}\{\theta = G\}}{\mathbb{P}\{\text{life event}\}} \\ &= \lim_{dt\to 0} \frac{p_{t_-}(1 - e^{-\bar{\lambda}_G dt})}{(1-p_{t_-})(1-e^{-\bar{\lambda}_B dt}) + p_{t_-}(1-e^{-\bar{\lambda}_G dt})} \\ &= \frac{p_{t_-}\bar{\lambda}_G}{(1-p_{t_-})\bar{\lambda}_B + p_{t_-}\bar{\lambda}_G} \\ &= j(\alpha_t, p_t). \end{aligned}$$

□

Theorem 2. *The optimal treatment policy is given by*

$$\alpha_t^*(p_t) = \begin{cases} 0 & \text{if } p_t < p^* \\ 1 & \text{otherwise,} \end{cases} \tag{12}$$

where

$$p^* \overset{\text{def}}{=} \frac{\nu^*\left[(\mu_0 - \mu_B) - D(\lambda_0 - \lambda_B)\right]}{(1+\nu^*)\left[(\mu_G - \mu_0) - D(\lambda_G - \lambda_0)\right] + \nu^*\left[(\mu_0 - \mu_B) - D(\lambda_0 - \lambda_B)\right]} \tag{13a}$$

$$\nu^* \overset{\text{def}}{=} \left\{\nu > 0 \,|\, \lambda_B + r + \left(\lambda_B - \lambda_G - \frac{(\mu_G - \mu_B)^2}{2\sigma^2}\right)\nu - \frac{(\mu_G - \mu_B)^2}{2\sigma^2}\nu^2 = \lambda_B\left(\frac{\lambda_B}{\lambda_G}\right)^{\nu}\right\}. \tag{13b}$$

Proof of Theorem 2. Let $u(p)$ be the optimal value function given a current belief p. In this case, $u(p)$ satisfies the following Bellman recursion (see Lemma 5 for a proof):

$$\begin{aligned} ru(p) = \max_{\alpha}\Big[&(1-\alpha)\mu_0 + \alpha\, m(p) + \alpha p(1-p)(\lambda_B - \lambda_G)u'(p) + \frac{1}{2}\alpha\phi(p)u''(p) \\ &- \big[(1-\alpha)\,\lambda_0 + \alpha\,\lambda(p)\big]\,u(p) - D\big[(1-\alpha)\lambda_0 + \alpha\lambda(p)\big] + u\big(j(\alpha,p)\big)\big[(1-\alpha)\lambda_0 + \alpha\lambda(p)\big]\Big]. \end{aligned}$$

Let $f(\alpha)$ denote the function inside the maximization, and note that its second derivative is given by:

$$f''(\alpha) = \frac{\lambda_0^2(\lambda_B - \lambda_0)^2(1-p)^2p^2u''(j(\alpha,p))}{((1-\alpha)\lambda_0 + \alpha\lambda(p))^3} \tag{14}$$

In (14), all the terms besides u'' term are positive. Therefore, if u were convex, then f would be convex as well, and the maximum of f would occur at an extreme point of the feasible set, i.e., for $\alpha^* \in \{0, 1\}$. In particular, the optimal policy would again be "bang-bang".

We now find a convex u that satisfies our Bellman equation. Consider:

$$\begin{aligned} u(p) = \max\Big\{\frac{\mu_0 - D\lambda_0}{r},\quad &\frac{1}{r}\Big[m(p) - D\lambda(p) + \lambda(p)\big[u(j(1,p)) - u(p)\big] \\ &+ p(1-p)(\lambda_B - \lambda_G)u'(p) + \frac{1}{2}\phi(p)u''(p)\Big]\Big\}. \end{aligned} \tag{15}$$

When playing the risky arm is optimal, the optimal value function satisfies the equation:

$$m(p) - D\lambda(p) + \lambda(p)\big[u(j(1,p)) - u(p)\big] - ru(p) + p(1-p)(\lambda_B - \lambda_G)u'(p) + \frac{1}{2}\phi(p)u''(p) = 0.$$

A particular solution of this equation is given by $u(p) = \frac{m(p)-h\lambda(p)}{r}$. For the homogeneous solution to this equation, we try $u_h(p) = (1-p)\left(\frac{1-p}{p}\right)^{\nu}$. Replacing this in the homogeneous differential equation, we obtain the following equation in ν:

$$\lambda_B + r + \left(\lambda_B - \lambda_G - \frac{(\mu_G - \mu_B)^2}{2\sigma^2}\right)\nu - \frac{(\mu_G - \mu_B)^2}{2\sigma^2}\nu^2 = \lambda_B\left(\frac{\lambda_B}{\lambda_G}\right)^{\nu}. \tag{16}$$

Equation (16) does not have a closed-form analytic solution, but it always has a solution. The left-hand side is a quadratic function, which intersects the vertical axis at $\lambda_B + r$, and the horizontal axis on the positive side at the value

$$\zeta \overset{\text{def}}{=} \frac{\lambda_B - \lambda_G - \frac{(\mu_G-\mu_B)^2}{2\sigma^2} + \sqrt{(\lambda_B - \lambda_G - \frac{(\mu_G-\mu_B)^2}{2\sigma^2})^2 + 2\frac{(\mu_G-\mu_B)^2}{\sigma^2}(\lambda_B + r)}}{\frac{(\mu_G-\mu_B)^2}{\sigma^2}},$$

going to $-\infty$ as $\mu \to \infty$. The right-hand side is an increasing exponential function, which crosses the vertical axis at λ_B, strictly below where the left-hand side function crosses it. Therefore, there must exist a value $\nu^* \in [0, \zeta]$ that is a solution to (16).

Using this ν^*, we now look for the optimal $u(p)$ in the risky region, of the form

$$u(p) = \frac{m(p) - D\lambda(p)}{r} + C(1-p)\left(\frac{1-p}{p}\right)^{\nu^*}.$$

We look for C^* and p^* satisfying the value matching and smooth pasting conditions, i.e., $u(p^*) = \frac{\mu_0 - D\lambda_0}{r}$ and $u'(p^*) = 0$, respectively. These two equations provide a system of two equations, which can be solved for p^* and C^*. We find an optimal threshold

$$p^* = \frac{\nu^*\left[(\mu_0 - \mu_B) - D(\lambda_0 - \lambda_B)\right]}{(1+\nu^*)\left[(\mu_G - \mu_0) - D(\lambda_G - \lambda_0)\right] + \nu^*\left[(\mu_0 - \mu_B) - D(\lambda_0 - \lambda_B)\right]} \tag{17}$$

and $C^*(p^*) = \frac{[\mu_G - \mu_B - D(\lambda_G - \lambda_B)]p^*}{r(\nu^* + p^*)}\left(\frac{1-p^*}{p^*}\right)^{-\nu^*}$.

We can then propose the following optimal value function:

$$u(p) = \begin{cases} \frac{\mu_0 - D\lambda_0}{r} & \text{if } p < p^*, \\ \frac{m(p) - D\lambda(p)}{r} + C^*(p^*)(1-p)\left(\frac{1-p}{p}\right)^{\nu^*} & \text{if } p \geq p^*. \end{cases}$$

It is readily verifiable that this function satisfies the optimality condition (15), since on $[p^*, 1]$, this function is increasing from $\frac{\mu_0 - D\lambda_0}{r}$ to $\frac{\mu_G - D\lambda_G}{r}$ and is therefore greater than $\frac{\mu_0 - D\lambda_0}{r}$. On the interval $[0, p^*)$, our proposed function is a constant, which substituted into the lower branch of (15) gives

$\frac{m(p)-D\lambda(p)}{r}$. This is lower than $\frac{\mu_0-D\lambda_0}{r}$ for all $p<\bar{p}' \stackrel{\text{def}}{=} \frac{(\mu_0-\mu_B)-D(\lambda_0-\lambda_B)}{(\mu_G-\mu_0)-D(\lambda_G-\lambda_0)+(\mu_0-\mu_B)-D(\lambda_0-\lambda_B)}$, and it can be readily checked, by rewriting p^*, that $p^*<\bar{p}'$, so that (15) holds.

We note that the second derivative of the value function $u(p)$ we obtained is either 0 or $C^*\nu^*(1+\nu^*)(1-p)^{-1+\nu^*}\frac{1}{p^{2+\nu^*}}$, which is positive since C^* is positive, so u is indeed convex, and the optimal policy is indeed "bang-bang". □

Lemma 5. *The optimal value function satisfies*

$$ru(p)=\max_{\alpha}\Big[(1-\alpha)\mu_0+\alpha\, m(p)+\alpha p(1-p)(\lambda_B-\lambda_G)u'(p)+\frac{1}{2}\alpha\phi(p)u''(p)$$
$$-\big[(1-\alpha)\,\lambda_0+\alpha\,\lambda(p)\big]\,u(p)-D\big[(1-\alpha)\lambda_0+\alpha\lambda(p)\big]+u\big(j(\alpha,p)\big)\big[(1-\alpha)\lambda_0+\alpha\lambda(p)\big]\Big].$$

Proof of Lemma 5. Note that the value function satisfies the following Bellman equation:

$$\begin{aligned}
u(p)&=\max_{\alpha}\ \mathbb{E}_p\big[\text{Total rewards}\,|\,\text{no life event in dt}\big]\,\mathbb{P}[\text{no life event in dt}]\\
&\qquad+\mathbb{E}_p\big[\text{Total rewards}\,|\,\text{life event in dt}\big]\,\mathbb{P}[\text{life event in dt}]\\
&=\max_{\alpha}\Big[\ \Big[(1-\alpha)\mu_0 dt+\alpha m(p)dt+e^{-r\,dt}\mathbb{E}_p[u(p+dp)]\Big]e^{-((1-\alpha)\lambda_0+\alpha\lambda(p))dt}\\
&\qquad+\Big[-D+e^{-r\,dt}\mathbb{E}_p\big[u(j(\alpha,p))\big]\Big](1-e^{-((1-\alpha)\lambda_0+\alpha\lambda(p))dt})\Big]\\
&=\max_{\alpha}\Big[\ (1-\alpha)\mu_0 dt+\alpha m(p)dt+u(p)+\alpha p(1-p)(\lambda_B-\lambda_G)u'(p)dt\\
&\qquad+\frac{1}{2}\alpha\phi(p)u''(p)dt-(r+(1-\alpha)\lambda_0+\alpha\lambda(p))\,u(p)dt\\
&\qquad-D\,((1-\alpha)\lambda_0+\alpha\lambda(p))\,dt+u(j(\alpha,p))\,((1-\alpha)\lambda_0+\alpha\lambda(p))\,dt\Big],
\end{aligned}\tag{18}$$

where $m(p)\stackrel{\text{def}}{=}p\mu_G+(1-p)\mu_B$, $\lambda(p)\stackrel{\text{def}}{=}p\lambda_G+(1-p)\lambda_B$, and $\phi(p)\stackrel{\text{def}}{=}(p(1-p)(h-l)/\sigma)^2$, and all terms of order dt^2 and higher have been dropped. By canceling a $u(p)$ on both sides, and dividing by dt, we exactly recover the desired result. □

Lemma 3. *If a mild (severe) event occurs at time t, when a fraction α_t of treatment is allocated to the risky arm, then the belief p_t jumps to a value of $j_M(\alpha_t,p_t)$ ($j_S(\alpha_t,p_t)$, respectively), where*

$$\begin{aligned}
j_M(\alpha_t,p_t)&\stackrel{\text{def}}{=}\frac{p_t\,\bar{p}_G\,\bar{\lambda}_G}{(1-p_t)\bar{p}_B\bar{\lambda}_B+p_t\bar{p}_G\bar{\lambda}_G},\\
j_S(\alpha_t,p_t)&\stackrel{\text{def}}{=}\frac{p_t\,(1-\bar{p}_G)\bar{\lambda}_G}{(1-p_t)(1-\bar{p}_B)\bar{\lambda}_B+p_t(1-\bar{p}_G)\bar{\lambda}_G},\\
\bar{\lambda}_G&\stackrel{\text{def}}{=}(1-\alpha_t)\lambda_0+\alpha_t\lambda_G,\\
\bar{\lambda}_B&\stackrel{\text{def}}{=}(1-\alpha_t)\lambda_0+\alpha_t\lambda_B.
\end{aligned}$$

Proof of Lemma 3. For simplicity of notation, let "good" denote the event "$\theta=G$". Given a mild

event, the updated belief is given by Bayes' rule:

$$\begin{aligned} p_{t+} &= \frac{\mathbb{P}\{\text{mild life event} \,|\, \text{life event, good}\}\,\mathbb{P}\{\text{life event} \,|\, \text{good}\}\,\mathbb{P}\{\text{good}\}}{\mathbb{P}\{\text{mild life event}\}} \\ &= \lim_{dt\to 0} \frac{p_t\,\bar{p}_G\,(1-e^{-\bar{\lambda}_G dt})}{(1-p_t)\bar{p}_B(1-e^{-\bar{\lambda}_B dt}) + p_t\,\bar{p}_G\,(1-e^{-\bar{\lambda}_G dt})} \\ &= j_M(\alpha, p). \end{aligned}$$

Similarly, given a severe event, the update is:

$$\begin{aligned} p_{t+} &= \frac{\mathbb{P}\{\text{severe life event} \,|\, \text{life event, good}\}\,\mathbb{P}\{\text{life event} \,|\, \text{good}\}\,\mathbb{P}\{\text{good}\}}{\mathbb{P}\{\text{severe life event}\}} \\ &= \lim_{dt\to 0} \frac{p_t(1-\bar{p}_G)(1-e^{-\bar{\lambda}_G dt})}{(1-p_t)(1-\bar{p}_B)(1-e^{-\bar{\lambda}_B dt}) + p_t(1-\bar{p}_G)(1-e^{-\bar{\lambda}_G dt})} \\ &= j_S(\alpha, p). \end{aligned}$$

□

Theorem 3. *1. If $D_M < D_S$, $\lambda_0 > \lambda_G$, and $p_0 < p_G$, then the optimal allocation for $p_t = 1$, i.e., $\alpha_t^*(1)$, is given by the expression*

$$\alpha_t^*(1) = \frac{\frac{\mu_G - \mu_0}{D_S - D_M} + (\lambda_0 - \lambda_G)(\frac{D_S}{D_S - D_M} - p_0) + \lambda_0(p_G - p_0)}{2(p_G - p_0)(\lambda_0 - \lambda_G)}$$

2. Furthermore, if $-(\lambda_0 - \lambda_G)(D_M p_0 + D_S(1-p_0)) - \lambda_0(p_G - p_0)(D_S - D_M) < \mu_G - \mu_0$ and $\mu_G - \mu_0 < (D_S - D_M)[p_G(\lambda_0 - \lambda_G) - \lambda_G(p_G - p_0)] - D_S(\lambda_0 - \lambda_G)$ both hold, then $\alpha_t^*(1) \in (0,1)$, and the optimal policy is not bang-bang even when $p_t = 1$.

Proof of 3. Let $u(p)$ be the optimal value function given a current belief p. $u(p)$ satisfies the following Bellman recursion:

$$\begin{aligned} u(p) = \max_{\alpha}\ & \mathbb{E}_p[\text{total rewards} \,|\, \text{no life event in dt}]\,\mathbb{P}[\text{no life event in dt}] \\ & + \mathbb{E}_p[\text{total rewards} \,|\, \text{mild event in dt}]\,\mathbb{P}[\text{mild event in dt}] \\ & + \mathbb{E}_p[\text{total rewards} \,|\, \text{severe event in dt}]\,\mathbb{P}[\text{severe event in dt}] \\ = \max_{\alpha}\ & \Big[(1-\alpha)\mu_0 dt + \alpha m(p) dt + u(p) + \alpha p(1-p)(\lambda_B - \lambda_G) u'(p) dt \\ & + \frac{1}{2}\alpha\phi(p) u''(p) dt - (r + (1-\alpha)\lambda_0 + \alpha\lambda(p))\, u(p) dt \\ & + \Big(-D_M + u\big(j_M(\alpha,p)\big)\Big)\Big(\bar{p}_G p((1-\alpha)\lambda_0 + \alpha\lambda_G) + \bar{p}_B(1-p)((1-\alpha)\lambda_0 + \alpha\lambda_B)\Big) dt \\ & + (-D_S + u(j_S(\alpha,p)))\Big((1-\bar{p}_G)p((1-\alpha)\lambda_0 + \alpha\lambda_G) \\ & + (1-\bar{p}_B)(1-p)((1-\alpha)\lambda_0 + \alpha\lambda_B)\Big) dt \Big], \end{aligned}$$

where $m(p) \overset{\text{def}}{=} p\mu_G + (1-p)\mu_B$, $\lambda(p) \overset{\text{def}}{=} p\lambda_G + (1-p)\lambda_B$, $\phi(p) \overset{\text{def}}{=} \big(p(1-p)(\mu_G - \mu_B)/\sigma\big)^2$,

$\bar{p}_G \stackrel{\text{def}}{=} (1-\alpha)p_0 + \alpha p_G$, $\bar{p}_B \stackrel{\text{def}}{=} (1-\alpha)p_0 + \alpha p_B$, and all terms of order dt^2 and higher have been dropped. Canceling a $u(p)$ on both sides, and dividing by dt yields the maximization problem:

$$\begin{aligned} ru(p) = \max_{\alpha}\Big[& (1-\alpha)\mu_0 + \alpha m(p) + \alpha p(1-p)(\lambda_B - \lambda_G)u'(p) \\ & + \frac{1}{2}\alpha\phi(p)u''(p) - ((1-\alpha)\lambda_0 + \alpha\lambda(p))\,u(p) \\ & + (-D_M + u(j_M(\alpha,p)))\Big(\bar{p}_G(\alpha)p((1-\alpha)\lambda_0 + \alpha\lambda_G) + \bar{p}_B(\alpha)(1-p)((1-\alpha)\lambda_0 + \alpha\lambda_B)\Big) \\ & + (-D_S + u(j_S(\alpha,p)))\Big((1-\bar{p}_G(\alpha))p((1-\alpha)\lambda_0 + \alpha\lambda_G) \\ & + (1-\bar{p}_B(\alpha))(1-p)((1-\alpha)\lambda_0 + \alpha\lambda_B)\Big)\Big]. \end{aligned}$$

The second derivative with respect to α of the function inside the maximization has a nontrivial form, because of the nonlinear dependency of the jump term on α. This makes finding the optimal allocation α^* analytically intractable. However, by examining the extreme case when $p = 1$, we can see that the optimal policy in this case is not necessarily 'bang-bang'.

When $p = 1$, the function of α to be maximized is

$$\begin{aligned} f(\alpha) \stackrel{\text{def}}{=} & (1-\alpha)\mu_0 + \alpha\mu_G - D_M\left[(1-\alpha)p_0 + \alpha p_G\right]\left[(1-\alpha)\lambda_0 + \alpha\lambda_G\right] \\ & - D_S\left[1 - (1-\alpha)p_0 - \alpha p_G\right]\left[(1-\alpha)\lambda_0 + \alpha\lambda_G\right]. \end{aligned}$$

The first and second derivatives with respect to α of f are:

$$\begin{aligned} f'(\alpha) &= \mu_G - \mu_0 + (D_S - D_M)(p_G - p_0)((1-\alpha)\lambda_0 + \alpha\lambda_G) \\ &\quad -D_M(\lambda_G - \lambda_0)\left[(1-\alpha)p_0 + \alpha p_G\right] - D_S(\lambda_G - \lambda_0)(1 - (1-\alpha)p_0 - \alpha p_G), \\ f''(\alpha) &= -2(D_M - D_S)(\lambda_0 - \lambda_G)(p_0 - p_G). \end{aligned}$$

Under our assumptions on model parameters, the second derivative is negative, so f is a concave function, and therefore the (unconstrained) maximum occurs where the first derivative is 0. The corresponding critical α is given by:

$$\alpha^*(1) = \frac{\mu_G - \mu_0 + (\lambda_0 - \lambda_G)(D_M p_0 + D_S(1-p_0)) + \lambda_0(p_G - p_0)(D_S - D_M)}{2(p_G - p_0)(\lambda_0 - \lambda_G)(D_S - D_M)},$$

and for $\alpha^*(1)$ to be between 0 and 1, we require $-(\lambda_0 - \lambda_G)(D_M p_0 + D_S(1-p_0)) - \lambda_0(p_G - p_0)(D_S - D_M) < \mu_G - \mu_0 < (D_S - D_M)[p_G(\lambda_0 - \lambda_G) - \lambda_G(p_G - p_0)] - D_S(\lambda_0 - \lambda_G)$. □