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: February 27, 2017

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 typically 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 to 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 where drug effectiveness is assessed by aggregating information from several channels: by continuously monitoring the 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 provides critical information for treatment decisions 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 may have practical appeal. We illustrate the effectiveness of the methodology by developing a set of efficient treatment policies for multiple sclerosis, which we then use to benchmark several existing treatment guidelines.

1 Introduction

Costs associated with the delivery of healthcare in the U.S. have risen sharply in recent years, 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 & Olsen 2010). In conjunction with advances in the field of medicine and the use of information technology, this has created increasing pressure for healthcare solutions that deliver better outcomes in a cost-effective manner.

*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.

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

Currently available disease modifying therapies (DMTs) for several chronic diseases 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.[2] In such cases, the main way to evaluate DMT efficacy is by initiating treatment and then 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* diseases is to prevent disease progression in the long run, which often translates to limiting the occurrence of infrequent negative health events (e.g., disease flare-ups) that can severely diminish 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 for 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 incurable; 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). While newly available drugs represent advances for MS management, none is fully effective (Rovaris et al. 2001), and the question of identifying patients who are not responsive to treatment is centrally important. 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 quite challenging. For a newly diagnosed patient, 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 of learning, and explicitly

[1] For instance, in an editorial paper, Murphy & 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."

[2] Biomarkers exist for some chronic diseases, e.g., breast cancer. Our focus is on diseases for which the existing biomarkers do not perfectly classify patients as responders and non-responders and, thus, there is scope for experimentation with the available treatments. As we mention below one such disease is multiple sclerosis. It is worthwhile to note that the discovery of biomarkers for MS is a very active field of research, e.g., Derfuss (2012).

recognize that the timing and frequency of relapses, as well as more continuous measurements such as EDSS and/or MRI, can all be informative.[3] 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 are 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, 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 $60,000 per diagnosed case in the U.S. (Hartung et al. 2015). This has resulted in a significant amount of debate around policies for MS treatment, in the U.S. and elsewhere.[4]

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

This paper can be viewed as one step toward answering such questions. We propose a framework that can be used to inform treatment decisions for chronic diseases that have the features described above: treatment is effective only for a subset of patients that is *a priori* unidentifiable; the frequency and/or 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 both the *day-to-day monitoring* of disease progression, as well as the *timing* of major health events. Our framework thus implicitly trades off the immediate and the long-term impact of treatment, providing a systematic way to incorporate new information in the design of optimal treatment plans.

[3] "[...] 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).

[4] 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 & 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.

- When choosing between two treatments with linear dose-response, our model can be analyzed in closed form, resulting in intuitive optimal policies that take the form of discontinuation rules. In an extension discussed in Appendix A of the Online Companion, we show how these analytical results can be used to derive optimal policies for choosing among several treatments by solving very simple (one-dimensional, convex) optimization problems. We also discuss conditions under which the optimal policy is no longer a simple discontinuation rule, such as when dose-response curves are nonlinear or when the *severity* of major health events is indicative of treatment effectiveness.

- We apply our results to multiple sclerosis, for which we develop and test adaptive treatment policies for administering interferon-β. Our framework allows an explicit trade-off between the benefits of treatment and its associated costs, and can be used to generate an entire frontier of cost-effective treatment policies, depending on the amount a decision-maker is willing to pay for one additional quality-adjusted life year (QALY), expressed as a willingness-to-pay (WTP) value. We use these policies to benchmark and test the performance of three treatment guidelines: a "no-treatment" policy, which does not prescribe any interferon, a "standard" policy, which administers interferon to all patients and discontinues treatment upon progression to an EDSS score of 6-7.5 (Río et al. 2011), and a "consensus" policy, which discontinues treatment when patients experience two or more relapses in a year or progress to an EDSS score of 6-7.5 (Cohen et al. 2004).

  Our first finding suggests that a no-treatment policy is optimal if the WTP does not exceed \$150,000/QALY. Furthermore, the gains from interferon treatment are generally not large in absolute terms, and come at steep costs: even the best adaptive policy, requiring a WTP exceeding \$800,000/QALY, can only increase the QALYs by 3.45% relative to a no-treatment alternative, while increasing costs by 16.2%. Since interventions are generally considered cost-effective when the costs per QALY gained do not exceed three times the country's per-capita GDP (Drummond 2005, Hunink et al. 2014), this suggests that interferon treatment is not necessarily cost-effective, and that no-treatment may be optimal under lower WTP. However, this finding should be interpreted with caution: even though the increases in QALY may not be large in absolute terms, they may nonetheless be significant, particularly for a chronic disease as debilitating as MS.

  Our results provide validation of the consensus criteria proposed by Cohen et al. (2004): the resulting policy is close to being efficient at intermediate values of WTP, and achieves net monetary benefits close to a fully adaptive policy. As such, these simple discontinuation rules may represent a viable alternative to implementing a complex optimal adaptive policy, particularly at intermediate WTP values.

  Finally, we find that none of the treatment guidelines are satisfactory at very large WTP: the consensus and no-treatment policies generate low QALYs, and the standard policy is inefficient. An adaptive policy derived from our framework under a WTP of \$800,000/QALY

generates the most QALYs without incurring significant costs, and attains a good balance between administering sufficient treatment to responders and identifying non-responders early.

While we apply our model to MS and interferon-β primarily because of data availability, we note that the treatment of many other chronic diseases could benefit from our framework. Such 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 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 (Bank & Küchler 2007, Berry & Fristedt 1985, Cohen & Solan 2013, Harrison & Sunar 2015, Mandelbaum 1987). Closest to our work are papers on strategic experimentation (Bolton & Harris 1999, Keller & Rady 2010, Keller et al. 2005), 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.

Our paper is related to the clinical trials literature, and in particular to the growing number of studies that consider adaptive rules for assigning patients to treatments (e.g., Ahuja & Birge 2016, Berry 1978, Berry & Pearson 1985, Bertsimas et al. 2014). 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. It is difficult to implement such an approach in the context of a chronic disease, however, as one 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 evidence for or against a given treatment plan.

Also closely related is a growing literature (e.g., Denton et al. 2009, Helm et al. 2015, Mason et al. 2014, Zhang et al. 2012) that uses Markov decision processes with fully or partially observed states and dynamic linear Gaussian systems to derive adaptive treatment policies. We formulate the problem using a continuous-time bandit model and derive closed-form expressions for optimal treatment decisions that as we discuss in the paper may have several advantages.

In the medical literature, Murphy (2003), Murphy (2005), Murphy & Collins (2007), Pineau et al. (2007), and Almirall et al. (2012) among others propose adaptive treatment schemes in the context of psychiatric conditions such as depression, anxiety disorders, and drug/alcohol abuse. In particular, Murphy (2003) motivates and lays the foundation for developing *dynamic treatment regimes*, i.e., a set of rules for choosing effective treatments that are tailored to the individual characteristics of patients. The methodology in these studies varies from non-quantitative approaches, e.g., pre-defining a protocol to switch therapies after a certain time if a criterion is not met (Almirall

et al. 2012), to reinforcement learning (Pineau et al. 2007) and developing statistical frameworks for optimizing a general outcome while achieving a desired level of power or bias (Murphy 2005). Although we share the same motivation and use similar methodological tools as some of these studies, i.e., dynamic programming (albeit, in continuous time), the emphasis in these papers is not on deriving explicit optimal treatment policies nor on preserving the computational tractability of the resulting framework. In contrast, our explicit characterization of the optimal adaptive policies allows us to compute useful comparative statics with respect to features of the underlying environment and to derive optimal policies for the case of multiple treatments by solving very simple (one-dimensional) convex optimization problems in an offline fashion (see Appendix A).

Finally, our work is also related to empirical cost-effectiveness studies for MS, which have found DMTs to be very expensive for the benefits they provide, with costs of up to $1.6 million per additional QALY gained (Noyes et al. 2011, Phillips 2004, Tappenden et al. 2009). Although we use a similar disease evolution model, the key point of departure is that we assess cost-effectiveness based on optimal adaptive treatments that utilize all available information, instead of heuristic treatment guidelines. We find that the optimal interferon treatment is not necessarily cost-effective, and we quantify the WTP values under which no-treatment is optimal.

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 El Karoui & Karatzas (1994), Bolton & Harris (1999), Keller & Rady (2010), and references therein, which form the basis of our model.

2.1 Model Framework

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"): a "safe" alternative, with known characteristics, and a "risky" alternative, which can be of either good (G) or bad (B) type, unbeknownst to the DM.

Each arm brings the DM immediate rewards that accrue continuously over time. More precisely, the "safe" arm generates instantaneous rewards governed by a Brownian motion with drift rate μ_0 and volatility σ, and a risky arm of type $\theta \in \{G, B\}$ generates instantaneous Brownian rewards with drift rate μ_θ and volatility σ. When the DM allocates a fraction $\alpha_t \in [0, 1]$ of the time interval $[t, t+dt)$ to the risky arm and the remaining fraction $1-\alpha_t$ to the safe arm, the total instantaneous

rewards received are $d\pi^1(t) + d\pi^0(t)$, where

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

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

Here, $dZ^0(t)$ and $dZ^1(t)$ are independent, normally distributed random variables, with mean 0 and variance dt. To understand the scaling used, 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 equals a fraction α_t of the risky reward and $1 - \alpha_t$ of the safe reward.

In addition to the instantaneous rewards, each arm also induces relatively rare "life events," as well as a special "stopping event" that terminates the decision process. The occurrence of any life event generates a deterministic "reward" of $-D$. The frequency of such events depends on the allocation used by the DM. More precisely, when $\alpha_t \in [0, 1]$ of the period is allocated to the risky arm, life events occur according to a Poisson process with rate $(1 - \alpha_t)\lambda_0 + \alpha_t \lambda_\theta$, where λ_0 (λ_θ) denotes the rate of life events under a safe arm (a risky arm of type θ, respectively). Similarly, the stopping event occurs at a time T that is exponentially distributed with rate $(1-\alpha_t)\eta_0 + \alpha_t \eta_\theta$, and generates a reward of magnitude V. We assume that, conditional on the risky arm's type θ and on the allocation α_t, the stopping event is independent from the Poisson process for life events.[5]

The DM knows all the underlying parameters governing the arms and the reward structure, i.e., μ_0, λ_0, σ, D, V, μ_θ, and λ_θ, 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 a non-anticipative allocation policy $\{\alpha_t\}_{t \geq 0}$ that maximizes the total expected discounted rewards Π up to the stopping event, i.e.,

$$\Pi \stackrel{\text{def}}{=} \mathbb{E}\left[\int_0^T e^{-rt}\left[d\pi^1(t) + d\pi^0(t) - (N_{t+dt} - N_t)\, D\right] + e^{-rT} V\right], \tag{2}$$

where N_t denotes the total number of life events occurring in $[0, t)$.

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 expected lump-sum reward received upon the occurrence of a life event during period $[t, t + dt)$. The integral is taken over the total (instantaneous and lump-sum) rewards discounted at a fixed rate $r > 0$, and the term outside the integral corresponds to the reward received upon the occurrence of the stopping event at time T. The expectation in (2) is

[5] An alternative formulation could have considered the stopping event as a "special instance" of a life event. Since splitting a Poisson process would yield an exponentially distributed time for the stopping event, this would be equivalent to our current model.

with respect to the stochastic processes $dZ^0(t), dZ^1(t)$, α_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.[6]

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" (Powell & 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. With a choice $\alpha_t = 0$, the DM would only gain instantaneous and lump-sum rewards from the safe arm, hence completely eliminating the exposure to the risky arm as well as the ability to update the belief p_t. It is important to emphasize that new information in our model is acquired through two channels: (1) by observing the instantaneous rewards $d\pi^1(t)$ from the risky arm, and (2) by (*not*) observing life events and the stopping event. Whenever $\alpha_t > 0$, these channels all convey meaningful information to the DM, potentially tilting his belief p_t toward (or away from) deeming the risky arm as good.

2.2 Application in the Context of Chronic Diseases

We now discuss how our 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 the patient. Depending on the focus, the "rewards" could either correspond to a patient's health utility, or to a cost-adjusted health utility that also accounts for the cost of treatment (see our more detailed discussion in Section 4). As such, an arm's instantaneous reward denotes the impact of treatment on the patient's immediate (cost-adjusted) quality of life. "Life events" correspond to sudden health episodes associated with normal disease progression, which bring about immediate disutility (and costs) to the patient, without altering the fundamental underlying disease evolution or the efficacy of treatments. Examples of life events include *relapses* in MS or panic attacks in anxiety disorders. Depending on the circumstance and the exact disease modeled, the "stopping event" could be a special instance of a life event or an entirely separate event, which changes the disease evolution or the treatment options (e.g., a heart attack, kidney failure, malignancy or death). We elaborate on these distinctions further when discussing the *objective*. We note that our base-case model only allows choosing between two arms/treatments. We discuss the important extension to multiple arms in Appendix A.

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 disease progression. In our model, such a treatment may still yield stochastic outcomes in terms of both instantaneous (cost-adjusted) health utility and life/stopping events, as one would expect in practice. The crit-

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

ical assumption is that the parameters governing these outcomes $(\mu_0, \sigma, \lambda_0, \eta_0)$ 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 (Scalfari et al. 2010).

Risky arm. The "risky" arm is only effective for 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, treatments for MS 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 for MS. When patients respond to treatment, their condition may improve (i.e., $\mu_G > \mu_0$), the frequency of life events may be diminished (i.e., $\lambda_G < \lambda_0$), and the likelihood of a major health event or disease progression may also decrease (i.e., $\eta_G < \eta_0$). When patients do *not* respond, their condition may remain the same or even deteriorate slightly, e.g., due to side effects from treatment. 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 η_θ. This is reasonable since medical studies often track patients for a relatively long period of time, and retrospectively assign them to responder and non-responder groups (e.g., Horakova et al. 2012).

Lump-sum rewards. Our assumption that the lump-sum "reward" $-D$ received upon life events is independent of the type θ is particularly pertinent for diseases such as MS and anxiety disorder. For instance, relapses in MS correspond to periods of acute disease activity when patients experience neurological symptoms such as sudden paralysis or loss of vision. Such episodes generate immediate disutility and have similar severity/consequences in all patients, but occur less often among patients responding to treatment (Horakova et al. 2012, Kremenchutzky et al. 2006). Our framework can be extended to *stochastic* rewards that are independent of θ, as only their expected value would matter. We discuss the extension to rewards dependent on θ in Appendix B.

Fractional allocations. Our model assumes that fractional allocations of treatment are possible, and that the response (i.e., the reward) is directly proportional to the allocation. Fractional allocations allow modeling cocktails of drugs (Rudick et al. 2006) or administering a lower dosage of a drug, e.g., by adjusting the frequency and/or the magnitude of doses. The assumption that the response is linear renders our model analytically tractable and is a reasonable first-order approximation, as dose-response functions are often S-shaped and thus linear in a central band of values (see, e.g., the MS study of OWIMS (1999)). We discuss this limitation further in Section 5, and we examine its impact numerically in Appendix D.

Objective. Considering a planning horizon T that corresponds to an exponentially distributed "stopping" event allows modeling flexibility, without sacrificing analytical tractability. The horizon T could capture the first occurrence of a life event, which is appropriate when the risky treatment improves the immediate quality of life of a patient but incurs a higher risk of severe side effects.

For instance, studies have shown that certain rheumatoid arthritis treatments improve pain and disability, but may cause malignancies or severe infections (Mariette et al. 2011). More broadly, T could correspond to any major event that permanently alters the state of the patient, the disease evolution, or the response/rewards from treatment. Examples could include progression to severe disease (e.g., in MS, transitioning from the relapsing-remitting phase to the secondary-progressive phase (Lee et al. 2012)) or the release of a new drug that alters the set of feasible treatment options or drastically reduces the cost of treatment, impacting the reward rates in a cost-adjusted objective. The rewards V received upon the stopping event can be interpreted as continuation values, which allows using our model as a building block for studying diseases with more complex dynamics, involving potentially non-stationary reward rates or phase transitions. For more details, we refer to our case study in Section 4, which implements this idea.

Our model includes a fixed discount rate $r > 0$, in keeping with the recommendations of the U.S. Panel on Cost-Effectiveness in Health and Medicine that costs and quality-adjusted life years should be discounted when estimating the cost-effectiveness of healthcare interventions (Gold 1996).

Simplifying assumptions. To preserve analytical tractability, our model makes a number of simplifying assumptions: arms/treatments are characterized by Brownian rewards and Poisson arrivals with known and stationary parameters; dose-response curves are linear; information collection and treatment updating can be conducted very frequently; and patients fully adhere to treatment recommendations. In Section 5, we discuss these limitations more extensively, providing several extensions and robustness checks, and outlining interesting directions for future research.

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

3 Analysis

Letting $\mathcal{F}_t$ denote the information set available to the DM at time t,[7] it can be seen that the belief that the risky arm is good, $p_t \overset{\text{def}}{=} \mathbb{P}\{\theta = G|\mathcal{F}_t\}$, is a sufficient statistic of the history up to time t. Thus, we take p_t as the state of the system, and we take the fraction of treatment allocated to the risky arm α_t as the DM's action (control). Finally, with $\mathcal{A} \overset{\text{def}}{=} \{(\alpha_t)_{t\geq 0} \,|\, \alpha_t : \mathcal{F}_t \to [0,1]\}$ 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

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

where the expectation is with respect to the stochastic processes $dZ^0(t), dZ^1(t), N_t, \alpha_t$ and p_t. As a first step in our analysis, we characterize the evolution of the DM's belief during time interval $[t, t+dt)$, as a function of the current belief p_t and action α_t. We start with the case when no life

[7] Formally, $\mathcal{F}_t$ is the sigma-algebra generated by the allocations, rewards, events, and lump-sum rewards up to time t, i.e., $\mathcal{F}_t \overset{\text{def}}{=} \sigma\big(\{\alpha_\tau, d\pi^0(\tau), d\pi^1(\tau), N_\tau, L_\tau\}_{0\leq\tau<t}\big)$.

event or stopping event occurs during the interval.

Lemma 1. *When no life event or stopping event occurs during time interval $[t, t+dt)$:*
(i) the posterior belief p_{t+dt} conditional on an observed instantaneous reward from the risky arm $d\pi^1(t) = y$ is given by Bayes' rule, and takes a value of

$$p_{t+dt} = \frac{p_t F(\mu_G/\sigma) e^{-(\bar{\lambda}_G + \bar{\eta}_G)dt}}{p_t F(\mu_G/\sigma) e^{-(\bar{\lambda}_G + \bar{\eta}_G)dt} + (1-p_t) F(\mu_B/\sigma) e^{-(\bar{\lambda}_B + \bar{\eta}_B)dt}}, \tag{3}$$

where

$$\bar{\xi}_\theta \overset{\text{def}}{=} (1-\alpha_t)\xi_0 + \alpha_t \xi_\theta, \quad \forall\, \xi \in \{\lambda, \eta\}, \quad \forall\, \theta \in \{G, B\} \tag{4a}$$

$$F(\mu) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi dt}} \exp\Big(-\frac{1}{2dt} \big(y/\sigma - \sqrt{\alpha}\mu dt \big)^2 \Big); \tag{4b}$$

(ii) the change in the DM's belief $p_{t+dt} - p_t$ is normally distributed, with mean $\alpha_t p_t (1-p_t)\big((\lambda_B + \eta_B) - (\lambda_G + \eta_G)\big)dt$ and variance $\alpha_t \big(\frac{p_t(1-p_t)(\mu_G - \mu_B)}{\sigma}\big)^2 dt$.

We make several observations about this result. First, note that when no life event or stopping event occurs during $[t, t+dt)$, the belief evolution only depends on the characteristics of the *merged* process of life and stopping events, which is Poisson under our assumptions. Thus, all the results depend on the sum of the rates of life events and stopping events, i.e., $\lambda_i + \eta_i$, $\forall\, i \in \{0, G, B\}$.

Part (i) of the result provides the update rule for the DM's belief. Note that changes only occur when the risky arm is used (i.e., when $\alpha_t > 0$), and the posterior only depends on the observed instantaneous reward from the *risky* arm ($d\pi^1$), but not from the *safe* arm ($d\pi^0$). This is intuitive, since the safe arm conveys no information about the risky arm's type. Note that the result implicitly requires the ability to separately observe the risky rewards, which may be problematic when the DM only observes the total instantaneous rewards $d\pi^0 + d\pi^1$ and $\alpha_t \in (0,1)$. As our later results will show, this issue becomes moot in our setting, since the optimal policy will always entail $\alpha_t \in \{0,1\}$, so the DM will never observe a mix of safe and risky instantaneous rewards.

Part (ii) establishes that the belief change is normally distributed, with parameters that depend on the arms' characteristics. The belief drifts upward—i.e., the risky arm is deemed more likely to be good—if and only if events under a good arm are less likely than under a bad arm, i.e., $\lambda_G + \eta_G < \lambda_B + \eta_B$. This is intuitive, since the absence of an event under such conditions can be viewed as "good news" for the DM. Consistent with this observation, note that "more learning" occurs—i.e., the mean belief update grows—as the difference between the rates of events under a good and a bad arm increases. More learning also occurs as $p_t(1-p_t)$ grows, i.e., as the DM has more uncertainty *a priori* about the arm's type, as measured through the variance of the prior: as p_t gets closer to the extremes (0 or 1), it takes a much stronger signal to alter the belief as compared to when p_t is close to 0.5. Lastly, as expected, more learning occurs as the DM experiments more aggressively with the risky arm, i.e., as α_t grows. However, such aggressive experimentation also leads to a larger variance in the updates, i.e., "more noise." Updates also get noisier as the DM

has more uncertainty *a priori* concerning the arm's type (i.e., as $p_t(1-p_t)$ grows), as the difference in mean rewards under a good and bad arm is larger (i.e., $(\mu_G - \mu_B)^2$ grows) or as the rewards get less noisy (i.e., σ decreases).

Our next result completes the characterization of the DM's belief update, by focusing on the case when a life event occurs during the interval $[t, t+dt)$.

Lemma 2. *When a life event occurs during $[t, t+dt)$:*
(i) the posterior belief p_{t+dt} conditional on an observed instantaneous reward from the risky arm $d\pi^1 = y$ is given by Bayes' rule, and takes a value of

$$p_{t+dt} = \frac{p_t F(\mu_G/\sigma)(1-e^{-\bar{\lambda}_G dt})e^{-\bar{\eta}_G dt}}{p_t F(\mu_G/\sigma)(1-e^{-\bar{\lambda}_G dt})e^{-\bar{\eta}_G dt} + (1-p_t)F(\mu_B/\sigma)(1-e^{-\bar{\lambda}_B dt})e^{-\bar{\eta}_B dt}}; \tag{5}$$

(ii) the change in the DM's belief $p_{t+dt} - p_t$ is normally distributed, with

$$\text{mean} = \frac{\alpha_t p_t(1-p_t)(\lambda_G - \lambda_B)}{\bar{\lambda}(p_t)} + \alpha_t p_t(1-p_t)\bar{\lambda}_G\bar{\lambda}_B \frac{\eta_B - \eta_G + (\mu_G - \mu_B)\mu(p_t)/\sigma^2}{(\bar{\lambda}(p_t))^2} dt$$
$$\text{variance} = \alpha_t \left(\frac{p_t(1-p_t)\bar{\lambda}_G\bar{\lambda}_B(\mu_G - \mu_B)}{\sigma}\right)^2 dt,$$

where $\bar{\lambda}_G, \bar{\lambda}_B, \bar{\eta}_G, \bar{\eta}_B, F$ are defined in (4a)-(4b), and $\xi(p) \overset{\text{def}}{=} p\xi_G + (1-p)\xi_B$, $\forall\, \xi \in \{\bar{\lambda}, \mu\}$.

Part (i) of the result provides an expression for the posterior of the DM's belief, which now depends separately on the rates of life events and stopping events under a good/bad arm. Part (ii) shows that the belief update remains normally distributed, but with modified mean and variance.

The mean update in (ii) now involves two terms. The first term is independent of dt and constitutes a jump in the belief caused by the occurrence of a life event. It can be readily verified that the posterior belief accounting for this jump, i.e., $j(\alpha_t, p_t) \overset{\text{def}}{=} p_t + \alpha_t p_t(1-p_t)(\lambda_G - \lambda_B)/\bar{\lambda}(p_t)$, is increasing in p_t and λ_G, and decreasing in λ_B. This confirms the intuition that, *ceteris paribus*, the occurrence of a life event makes it *relatively* more likely that a risky arm is good when the prior belief that it was good was larger, or when life events become more (less) likely under a good (bad) arm. Note that $j(\alpha_t, p_t) < p_t$ if and only if $\lambda_G < \lambda_B$, so that a life event makes it more likely that the arm is good if only if life events are more likely under a good arm than under a bad one. When $\lambda_G < \lambda_B$, it can also be verified that $j(\alpha_t, p_t)$ is decreasing in α_t, so that a DM who experiments more aggressively becomes more skeptical about the risky arm upon the occurrence of a life event.

The second term, which is directly proportional to dt, is a further drift in the belief caused by the instantaneous rewards. These rewards also give rise to variability (i.e., variance) in the belief update, and it can be checked that this grows as the DM has more uncertainty *a priori* concerning the arm's type (i.e., as $p_t(1-p_t)$ grows), as the good and bad arm differ more in their instantaneous rewards (i.e., as $(\mu_G - \mu_B)^2$ grows), as the processes describing life events get more noisy (i.e., $\lambda_0, \lambda_G, \lambda_B$ grow), or as the rewards get less noisy (i.e., σ decreases).

With these results, we can now provide a characterization of the DM's optimal policy. We

restrict our subsequent analysis to the "interesting" case: we allow belief updates to be noisy (i.e., $\mu_G \neq \mu_B$), and we assume that no arm can be eliminated *a priori* (i.e., a good arm dominates the safe arm, which in turn dominates a bad arm). This is summarized in Assumption 1 below.

Assumption 1. *The primitives for the framework satisfy*

$$\mu_G \neq \mu_B \quad \textit{and} \quad A_B < A_0 < A_G,$$

where $A_\theta \stackrel{\text{def}}{=} (\mu_\theta - \lambda_\theta D + \eta_\theta V)/(r+\eta_\theta)$, $\forall\, \theta \in \{0, G, B\}$ *denote the total rewards per unit time for a safe, good and bad arm, respectively.*

Theorem 1. *Let Assumption 1 hold. Then, the DM's optimal policy is given by*

$$\alpha_t^*(p_t) = \begin{cases} 0 & \text{if } p_t < p^* \\ 1 & \text{otherwise,} \end{cases} \tag{6}$$

where

$$p^* \stackrel{\text{def}}{=} \frac{w_B\big(A_0 - A_B\big)}{w_B(A_0 - A_B) + w_G(A_G - A_0)} \tag{7a}$$

$$w_B = \frac{r+\eta_B}{r+\eta_B+\lambda_B}\nu^* \tag{7b}$$

$$w_G = \frac{r+\eta_G}{r+\eta_G+\lambda_G}(1+\nu^*) \tag{7c}$$

and $\nu^* \stackrel{\text{def}}{=} -\frac{1}{2} + \frac{\sigma^2(\lambda_B+\eta_B-\lambda_G-\eta_G)}{(\mu_G-\mu_B)^4} + \frac{\sqrt{((\mu_G-\mu_B)^4-2\sigma^2(\lambda_B+\eta_B-\lambda_G-\eta_G))^2+8\sigma^2(r+\eta_B+\lambda_B)(\mu_G-\mu_B)^4}}{2(\mu_G-\mu_B)^4}$.

Theorem 1 confirms that the optimal policy is a threshold policy; 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^* only depends on suitably weighted relative differences of the (per unit time) rewards for each arm type, A_0, A_G, A_B. In particular, p^* depends on the safe arm only through A_0. It can be readily verified that p^* is increasing in μ_0, reflecting the intuitive fact that, *ceteris paribus*, a safe arm with higher instantaneous rewards makes the risky arm 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 and in D. This shows that a DM behaving optimally should be more prone to experimenting with a risky arm when life events under the safe arm become more frequent/likely or when they have more severe consequences. The threshold p^* is also 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. Finally, note that p^* does not depend on the prior belief p_0 that the arm is good. This is a useful feature in an optimal policy, since it suggests a certain separation between the (objective) effectiveness of an arm and the (potentially subjective) prior.

We conclude this section with a brief discussion of extensions and implications of the results. We start by noting that, although our approach focuses on two arms/treatments, some of the

results generalize. In Appendix A, we discuss the important case where several arms with binary (good/bad) types exist. Although we are unable to explicitly characterize the optimal policy, we argue that it remains indexable—involving a single arm used at any point of time—and we use our analytical results above to provide an algorithm that calculates the optimal policy to within an arbitrary precision. Our proposed algorithm only requires an offline solution for a small number of one-dimensional convex optimization problems, and an online updating of the beliefs using Lemmas 1 and 2, making it appealing in settings with many arms or frequent updating.

Second, we note that the "bang-bang" structure of the optimal policy relies on several of our modeling assumptions. The "bang-bang" structure no longer holds, for instance, when the response to the DM's allocation is nonlinear (see our Appendix D) or when the lump-sum rewards received from life events depend on the risky arm's type (see Appendix B). In such cases, a strictly fractional allocation that trades off the benefits of the safe arm with those of the risky arm turns out to be optimal, and this is true even in cases when the risky arm is exactly known to be good or bad.

In the context of chronic diseases that are consistent with our framework, our results suggest that, given our modeling assumptions, the optimal treatment policy is a discontinuation rule: the patient is given the "risky" treatment 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. We next illustrate how the findings of our simple analytical model can be potentially used for a disease with more complex and realistic dynamics.

4 Case study: Multiple Sclerosis

We illustrate our framework with a case study of 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 (Kremenchutzky et al. 2006). No cure currently exists, and treatments are only effective for some patients, with no accurate biomarkers for assessing responsiveness *a priori* (Romeo et al. 2013). While the search for biomarkers in MS remains an active area of research (Derfuss 2012), physicians currently rely mostly on MRI scans and surveys in which patient-reported symptoms are used to compute the Expanded Disability Status Scale (EDSS) score, which can be translated into quality-of-life utilities for assessing disease evolution and treatment effectiveness (Prosser et al. 2003).

We focus on the most common form of MS, relapsing-remitting multiple sclerosis (RRMS), which comprises about 80% of cases. The initial stage of the disease, which typically lasts for 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. 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).

Although MS is incurable, disease-modifying therapies (DMTs) attempt to slow progression and reduce relapses (NMSS 2014). The most common treatments are injectable DMTs such as interferon-β preparations and glatiramer acetate, and more recently oral DMTs such as dimethyl fumarate (approved for use in the U.S. in 2013), teriflunomide (approved in 2012), fingolimod (approved in 2010) and natalizumab (approved in 2004) (see 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 (Cohen et al. 2010). The response profile to interferon has been well documented (Horakova et al. 2012, Romeo et al. 2013) but 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 the benefits of administering interferon-β treatment in addition to conducting symptom management without DMT. This decision problem is especially important because patients receiving interferon-β experience a significant decrease in quality of life due to side effects, such as pain at the local injection site, flu-like symptoms, depression, and allergic reactions. Furthermore, interferon-β generates significant and rapidly escalating healthcare costs, currently amounting to \$60,000 per year for each diagnosed case in the U.S., and estimated to increase at rates 5 to 7 times larger than prescription drug inflation (Hartung et al. 2015).

Despite the potential benefits, the problem of determining an optimal policy for administering interferon has not received much attention in the medical literature. For a newly diagnosed patient, current guidelines recommend immediately starting treatment and suggest simple discontinuation rules (Cohen et al. 2004, Río et al. 2011). These rules were not the outcome of a quantitative framework and have not been tested for efficiency.

With this motivation, we first describe a detailed disease model from the medical literature, which we use as a basis to design adaptive treatments using our results in Section 3. The goals of our numerical study are: (1) to quantify how close existing treatment guidelines are to being optimal and cost-effective; and (2) to understand the potential benefits of using a sophisticated treatment policy that relies on very frequent belief and treatment updating.

4.1 Disease Model

We implement a disease model similar to those in the medical literature (Lee et al. 2012, Prosser et al. 2004). Disease progression is modeled as a Markov chain, with states given by the patient's EDSS score[8] and whether she is currently experiencing a relapse (see Figure 1 for details).

[8]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

As in Lee et al. (2012), our simulation follows a hypothetical cohort of 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 to be around 52% (Horakova et al. 2012). We utilize a one-month time step, and simulate patients over a 50-year time horizon.

Each patient 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 (bedridden). While in EDSS states 0-2.5 or 3-5.5, patients can experience relapses, which can be either mild/moderate or severe, and which last 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. Consistent with medical studies, we also assume that: (a) relapses do not occur in states with EDSS score above 6, (b) the probability that relapses are severe is independent of EDSS state, treatment or response type, (c) disease progression probabilities are independent of whether the patient is currently experiencing a relapse, and (d) deaths can occur from all states depending on the patient's age, with MS-related deaths only occurring in EDSS state 8-9.5. Furthermore, when simulating patients on treatment, we allow patients in earlier disability states (with EDSS lower than 6) to abandon treatment in any month within the first three years with a fixed probability,

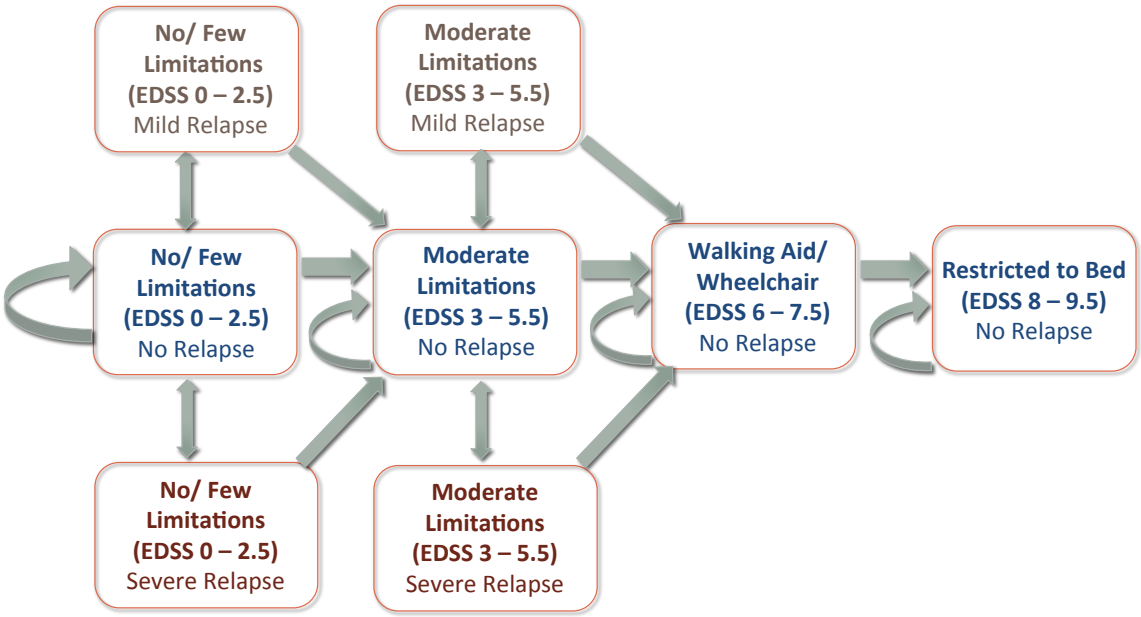


Figure 1: Multiple sclerosis disease model. All patients start in the lowest disability state (EDSS 0-2.5). In each month, a given patient can: (1) remain in the same state, without a relapse, (2) experience a relapse (in the two lowest disability states), which can be either mild or severe, and which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

medical studies concerning the difference in MRI scans between responders and non-responders.

consistent with the abandonment rate observed in related medical studies (Cohen et al. 2010, O'Rourke & Hutchinson 2005, Prosser et al. 2004). We assume that once a patient discontinues treatment, she will remain off treatment for the rest of her life.

Values for the transition probabilities are shown in Table 1, and are consistent with medical studies (Lee et al. 2012, Prosser et al. 2004). These probabilities depend on whether the patient is on treatment and on the patient's response type, with successful treatment reducing the disease progression and relapse rates by roughly 50% in responders compared to non-responders or patients not on treatment. We use age-specific mortality rates published by the U.S. Centers for Disease Control and Prevention (Arias 2014).

Each state in the chain is associated with a mean quality-adjusted life year (QALY) value capturing a patient's (quality-of-life) utility (Table 2), with a year in perfect health having a utility of 1, and death having a utility of 0. The realized QALYs in a given state are normally distributed, with a variance that is consistent with quality-of-life surveys (Prosser et al. 2003). Mean utilities of low disability states are higher than those of high disability states, and mean utilities of non-relapse states are higher than those of relapse states. Furthermore, being on treatment *reduces* the QALYs associated with each state for both responders and non-responders, due to side effects, and this effect is more pronounced during the first six months of treatment. To the best of our knowledge, no study reports a difference in quality of life between responders and non-responders on treatment. We assume that a responder has a small increase in quality of life compared to a non-responder (0.0012 per month on average), and we vary this value in sensitivity analysis.

In addition to QALYs, each state also has an associated cost (Table 3), representing the direct and indirect monthly costs—which occur regardless of whether a patient is on treatment—as well as the cost of interferon-β treatment and the cost of managing (severe) relapses.

4.2 Treatment Policies

Consistent with medical practice, we assume that all treatments apply standard care and symptom management throughout the entire lifetime of the patient. Thus, treatment policies choose between

Table 1: Transition Probabilities for the Disease Model in Figure 1. Sources: (1) Lee et al. (2012), (2) Horakova et al. (2012), (3) Prosser et al. (2004).

Parameter	Value	Range	Source
If not in treatment or non-responder			
Monthly probability of disease progression			
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)
Monthly probability of relapse (EDSS states 0-2.5, 3-5.5)	0.0799	0.0566-0.0944	(1,2,3)
Conditional probability of severe relapse	0.23	0.14-0.56	(1)
Responder on treatment			
Probability of progression (relative to non-responder)	0.5	0.38-1.00	(2)
Probability of relapse (relative to non-responder)	0.5	0.33-0.90	(2)
Monthly probability of treatment discontinuation	0.0087	0-0.0174	(1)

two arms, with the "safe" arm corresponding to standard care, and the "risky" arm corresponding to prescribing interferon-β in addition to standard care. We consider the following policies:

- *No treatment:* A policy that does not prescribe interferon.
- *Standard:* A policy that immediately starts all patients on interferon-β, and only discontinues treatment when patients reach EDSS disability state 6-7.5. This policy is consistent with current recommendations to maintain patients on treatment indefinitely (Río et al. 2011), and has been modeled similarly in previous studies (Lee et al. 2012, Prosser et al. 2004);
- *Consensus Criteria:* A policy proposed by Cohen et al. (2004), where all patients are started on interferon-β, but treatment is discontinued if patients experience two or more relapses in a year, or progress to an EDSS state of 6-7.5;
- *Adaptive:* A set of adaptive treatment policies based on our model, which we describe next.

Table 2: Utility Values for each Disease State. Sources: (1) Lee et al. (2012), (2) Prosser et al. (2003). (*) Values were converted from yearly to monthly values.

Parameter	**Value**	**Range**	**Source**
Utility Means (in QALYs per month)			
Baseline utilities by 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	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 (in QALYs)			
Mild or Moderate	0.0076	0.0053-0.0099	(1)
Severe	0.0252	0.0198-0.0305	(1)
Utility Standard Deviation (in QALYs per month*)	0.0087	0.0034 - 0.0262	(2)

Table 3: Direct, Indirect, and Treatment Costs for MS. Sources: Lee et al. (2012), Noyes et al. (2011).

Monthly costs (in USD)	**Value**	**Range**
Interferon-β treatment	2,061	1,000-3,828
Direct costs by 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 or Moderate	104	0-200
Severe	5,215	3,911-6,519
Indirect costs by 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

We henceforth refer to the first three policies above (no treatment, standard, and consensus criteria) as *treatment guidelines*.

4.2.1 Implementation of Adaptive Treatments

Our adaptive policies are derived by applying the analytical results in Section 3 to the Markov model in Section 4.1. Since this procedure is required whenever implementing our framework for a complex disease model, we elaborate on the critical steps below.

We identify "life events" with relapses, and the "stopping event" with disease progression. We apply our model on a *disease state* basis, and in *age-dependent* fashion. Consistent with practice, we assume that interferon-β is not prescribed for patients with EDSS scores exceeding 6, and thus we design treatment only for patients with EDSS scores of 0-2.5 or 3-5.5, which we henceforth refer to as s_1 and s_2, respectively. Since mortality is age-dependent, we allow treatment decisions to depend on the patient's age when transitioning into a particular EDSS state. Applying our model thus generates policies in the form of belief thresholds $p^*(s, x)$ for both EDSS scores $s \in \{s_1, s_2\}$ and for various initial age values x. This requires fitting our model parameters—the instantaneous reward rates $\mu_{0,G,B}$, the standard deviation σ, the relapse rates $\lambda_{0,G,B}$, the disutility associated with a relapse D, the stopping event rates $\eta_{0,G,B}$, the terminal lump-sum reward V, and the discount rate r—for each relevant state, as we discuss next.

Objective. We follow the typical approach in healthcare economics to combine the two objectives of QALYs and costs into a single objective, using the net monetary benefit (NMB) conversion (Drummond 2005, Hunink et al. 2014). This requires pre-defining a willingness-to-pay threshold (WTP), which reflects how much policymakers are willing to spend in order to gain one additional QALY. Then, the NMB of an intervention that achieves additional QALYs of Q at cost C is calculated as $\text{NMB} = Q \times \text{WTP} - C$. Policies can then be compared in terms of their associated NMB, with a higher WTP value shifting the weight from costs to QALYs. We adopt this objective, and solve the model for various values of the WTP parameter; this allows us to recover an entire frontier of cost-effective treatments.

Instantaneous rewards (μ, σ). A month in each disease state $s \in \{s_1, s_2\}$ is associated with an average QALY value and an average cost. We therefore have

$$\mu(s) = (\text{mean monthly QALY in } s) \times \text{WTP - (monthly costs in } s).$$

For instance, for a patient in EDSS state 3-5.5 who is not on treatment, the reward $\mu_0(s_2)$ is determined by a baseline utility of 0.0566 QALYs/month (per Table 2) and a baseline cost of \$4,001/month (\$1,037+\$2,964, per Table 3). For all patients on treatment, we use the instantaneous reward rates based on the QALYs after the first six months. For instance, a responder on treatment in EDSS state 3-5.5 has a reward rate $\mu_B(s_2)$ determined by a mean QALY of $0.0566-0.001+0.00058$ (per Table 2), and a monthly cost of $\$4,001 + \$2,061$ (per Table 3). We set σ equal to the standard deviation of utility in Table 2.

Disutility from relapse (D). Each relapse causes a decrement in the utility and an increment in the cost for the month in which it occurs, by the amounts listed in Table 2. Therefore,

$$D = (\text{decrement in monthly QALY}) \times \text{WTP} + (\text{increment in monthly cost}).$$

For instance, if a severe relapse occurs while in EDSS 0-2.5, the mean QALY per month decreases by 0.0252, and the cost increases by \$5,215.

Relapse and progression rates (λ, η). Table 1 provides the values for the monthly probabilities of relapse and progression in the disease Markov model. Using our assumption that relapse and progression events occur according to Poisson processes, these monthly probabilities can be converted into monthly rates using the relationship $p_{\text{monthly}} = 1 - e^{-\text{rate}_{\text{monthly}}}$. We note that the rates for responders on treatment are lower, according to Table 1.

Discount rate (r). We take a societal perspective, aggregating costs and QALYs across all patients, and discount at an annual rate of 3% (Gold 1996). We incorporate patient death by viewing it as an exponentially distributed event with a terminal reward equal to 0. Since the mortality rate is the same for both responders and non-responders, we can easily account for the death event by directly adding the mortality rate to the discount rate (this can be verified formally in the context of Theorem 1). For simplicity, since mortality rates increase with age, we set the mortality for a patient of age x equal to the average mortality rate over years $x+1, x+2, \ldots, x+\tau$, where τ is the average time spent in a state by responders. The discount rates used in each state $s \in \{s_1, s_2\}$ thus depend on the patient's age x upon initially transitioning into state s.

Terminal reward (V). The lump-sum terminal reward in any state corresponds to the expected NMB upon transitioning from that state to the next disease state. Our implementation requires such rewards for any state $s \in \{s_1, s_2\}$ and for every initial patient age x upon entering state s, i.e., we need to specify $V(s, x)$. Rewards may also depend on the patient's response type θ when the patient is subjected to treatment. To account for this, we calculate the terminal rewards separately for responders and non-responders (i.e., assuming perfect identification), and then weight these by the probability of the patient being a responder.

To determine $V(s_2, x)$, we first simulate the Markov model separately for each type θ, calculating the expected remaining QALYs $Q_{6-7.5}(x + \tau_\theta)$ and costs $C_{6-7.5}(x + \tau_\theta)$ from the random time τ_θ when the patient transitions into the next disease stage (with EDSS 6-7.5) until her death. We then set $V(s_2, x) = \mathbb{E}_{p_\theta}[Q_{6-7.5}(x + \tau_\theta) \times \text{WTP} - C_{6-7.5}(x + \tau_\theta)]$, where the weights are taken with respect to the prior probability that the patient is a responder. For $V(s_1, x)$ we proceed similarly, calculating rewards from the random transition into state s_2 onwards.

Initial prior probability of the patient being a responder. We start every patient in state s_1 with a prior of 0.52 of being a responder, in accordance with the distribution of responders and non-responders in the population. When simulating our adaptive policies, we update this belief as long as the patient is on treatment. The updates are done monthly, depending on the observed quality-of-life utility and whether a relapse occurred during the month, using the results in Lemma 2(i) and Lemma 1(i), respectively.

We note that, although our implementation captures some of the features of the complex MS model in Figure 1, such as disease progression and age-dependent mortality, it nonetheless remains an approximation. For instance, it ignores the different magnitudes of side effects in the first six months of treatment, the different variances in QALYs in a relapse month, and the patient abandonment rates, and does not rigorously account for mortality rates. However, our simulation accounts for all of these features, making the performance assessment for all policies under consideration considerably more accurate.

4.3 Results

We start by determining optimal adaptive treatments based on our implementation, for different values of the WTP parameter. This generates an efficient frontier of policies that systematically trade off QALYs and costs, and is useful for benchmarking potential alternatives. We consider WTP values from \$50,000/QALY to \$800,000/QALY, consistent with MS studies that report costs in excess of \$500,000/QALY gained (Noyes et al. 2011). For each WTP, we find the optimal belief thresholds at which treatment should be discontinued for each relevant EDSS state (0-2.5 and 3-5.5), and for every patient age.

Table 4 shows these thresholds for a typical 37-year old patient. As expected, we find that the propensity to recommend treatment increases with WTP and with age. Note that for a typical patient starting with an EDSS of 0-2.5 and a 52% prior probability of being a responder, the optimal adaptive treatment would prescribe interferon only for a WTP above \$200,000/QALY. Since interventions are generally considered cost-effective when the cost per QALY is less than three times the country's per-capita GDP (Drummond 2005, Hunink et al. 2014), this suggests that interferon treatment might not be considered cost-effective, and that a no-treatment policy is optimal if the WTP is no more than \$150,000/QALY.

We next simulate all the policies under consideration—no-treatment, standard, consensus, and all our adaptive policies—using the detailed Markov disease model. It is important to emphasize that under this simulation, all the policies and information sets/beliefs are updated on a *monthly* basis, so that all the results correspond to the realized performance under this frequency. In

Table 4: Optimal Discontinuation Thresholds for a Patient Aged 37 in State s_1 (EDSS score 0-2.5) and State s_2 (EDSS score 3-5.5), for Various WTP Values.

WTP (\$/QALY)	$p^*(s_1, 37)$	$p^*(s_2, 37)$
50K	1	1
100K	1	1
150K	0.77	1
200K	0.48	0.67
250K	0.36	0.45
300K	0.28	0.34
500K	0.16	0.18
800K	0.10	0.11

particular, although all our adaptive policies were calculated under the assumption of a continuous-time model, they are implemented and assessed under discrete-time updates.

A visual summary of the results is shown in Figure 2, which displays the costs and QALYs per patient averaged over responders and non-responders, assuming a 52% fraction of responders in the population. To put the results into perspective, the figure also displays the performance of a *perfect hindsight* policy, which correctly classifies all patients *a priori*, and only prescribes interferon to responders. In our simulation, this policy would result in 16.389 QALYs and costs of $1,143,031.

The no-treatment policy yields the smallest number of QALYs on average (15.794), but is also the least expensive ($1,036,656). The standard policy yields an average of 16.333 QALYs, an increase of 3.4% compared to no-treatment, but is also the most expensive of all policies, with an average cost of $1,281,692 per patient. The consensus criteria policy falls in between, achieving 15.984 QALYs at a cost of $1,096,950.

As expected, the adaptive policies form an efficient frontier that dominates all policies except the perfect hindsight one. At WTP values below $150,000/QALY, a no-treatment policy is equivalent to adaptive policies. The consensus policy is strictly dominated by adaptive policies for a WTP value of $205,000-$220,000/QALY, although the differences are not very substantial (with QALYs increased by 0.4% or costs reduced by 1.2%). The standard policy is dominated by an adaptive policy with a WTP of $800,000/QALY, which increases QALYs slightly (by 0.006), but significantly reduces costs (by $77,000, or 6%).

Figure 3 compares the policies in terms of their achieved NMB, expressed as a percentage of the NMB of the best adaptive policy. Consistent with our prior observations, note that the highest NMB is achieved by the no-treatment policy at low WTP, by the consensus policy at intermediate

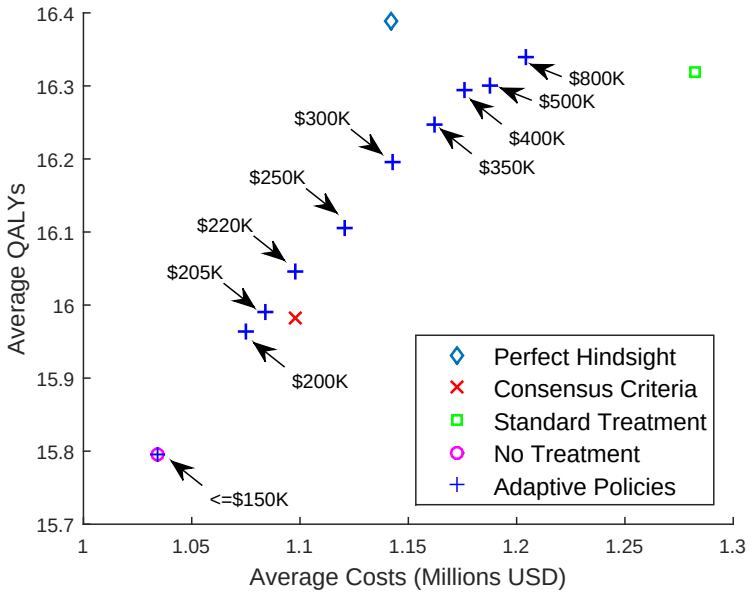


Figure 2: Population-averaged costs and QALYs for all policies. The arrows indicate the values of the WTP parameter under which the adaptive policies were obtained.

WTP, and by the standard policy at high WTP. Furthermore, all policies except standard achieve a high NMB uniformly, i.e., for any WTP value.

These results can be used by policy makers to quantify the benefits of interferon treatment, and weigh them against the corresponding costs. Our findings suggest that gains from interferon treatment are not large in absolute terms, and come at steep costs: even the best adaptive policy can increase QALYs by only 3.45% relative to the no-treatment alternative, while increasing costs by 16.2%.[9] We find that the WTP required for such improvements exceeds \$800,000/QALY; this confirms earlier studies reporting costs larger than \$500,000/QALY for interferon (Noyes et al. 2011), showing that this persists even when considering optimal adaptive treatments instead of heuristic treatment policies. This reinforces our earlier observation that interferon treatment is not necessarily cost-effective, and suggests that even in environments with larger WTP (e.g., above \$150,000/QALY), not prescribing interferon may be the optimal action. However, this recommendation should be interpreted with caution—even though improvements in QALYs may not be large in absolute terms, they may nonetheless be significant in relative terms, and particularly for chronic diseases as debilitating as MS. Furthermore, patients (and policy makers alike) may not easily accept the cost-benefit analysis inherent in such a no-treatment recommendation.

Second, our results provide empirical validation of the consensus criteria proposed by Cohen et al. (2004). We find that the resulting policy is close to being efficient at intermediate values of WTP, and achieves net monetary benefits close to a fully adaptive policy. Thus, these simple discontinuation rules may represent a viable alternative to implementing a complex optimal adaptive policy, particularly at intermediate values of WTP.

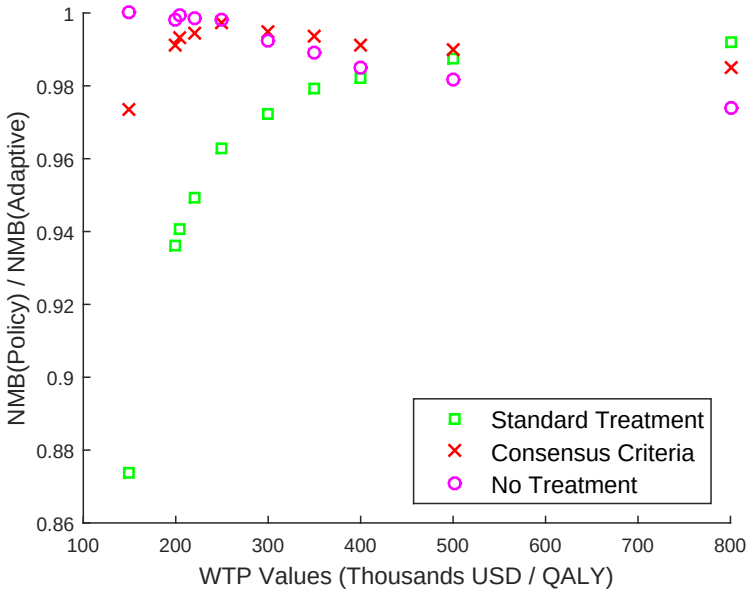


Figure 3: Net Monetary Benefits (NMB) achieved by each policy relative to the optimal adaptive policy.

[9]Similar observations can be made with respect to the perfect hindsight policy, which increases QALYs by 3.7% and costs by 10.2% relative to no-treatment.

Finally, the results suggest that none of the treatment guidelines is satisfactory at very large WTP: the consensus and no-treatment policies generate low QALYs, while the standard policy is inefficient, dominated in both QALYs and costs by an adaptive policy. To better illustrate the differences in performance, and to understand how the benefits are distributed among responders and non-responders, we examine this case in more detail.

4.3.1 An Optimal Policy at High WTP

We compare the three treatment guidelines with the adaptive policy calculated for a WTP of \$800,000/QALY (Section E of the Appendix provides a detailed description of this policy). The simulation results by response type are summarized in Table 5. As can be seen, the optimal adaptive policy achieves QALYs for responders that are close to those of the standard policy (which is optimal for this type), and considerably exceeds the no-treatment and consensus policies, by 6.6% and 4.1%, respectively. At the same time, the adaptive policy also achieves higher QALYs than the standard policy for non-responders, by identifying them and removing them from treatment earlier; no-treatment and consensus, which are both more aggressive in removing patients from treatment, are only marginally superior to the adaptive policy for non-responders, with QALY improvements of less than 0.6%.

The outcomes of the policies are primarily driven by the number of months each patient spends in treatment. The standard policy incurs the most costs, as it keeps patients on treatment for the longest time on average (289 months for responders, and 208 months for non-responders). In contrast, the adaptive policy keeps responders on treatment for 279 months on average and non-responders for 78 months, achieving the highest overall gain in QALYs. Perhaps the best illustration of the effectiveness of the adaptive policy is Figure 4, which plots the proportion of patients on treatment over time, by response type. The proportion of non-responders on treatment converges to zero for the adaptive policy, whereas the proportion of responders on treatment remains relatively high. The optimal policy thus attains a good balance between the optimal treatment for responders (standard) and for non-responders (no-treatment).

These results can be observed consistently for each year in our simulation, as displayed in Figures 5 and 6. Note that both costs and QALYs decrease over time under both the standard

Table 5: QALYs and Costs for Each Policy, by Patient Type. "(Non-)Resp" denotes a (non-)responder to interferon-β. The adaptive policy and all NMBs are calculated for a WTP of \$800,000.

	Standard		**No-treatment**		**Consensus**		**Adaptive (800,000)**	
	Resp	Non-resp	Resp	Non-resp	Resp	Non-resp	Resp	Non-resp
QALY means	16.939	15.677	15.794	15.794	16.225	15.724	16.917	15.712
QALY stderr	0.037	0.035	0.035	0.035	0.035	0.035	0.036	0.036
Cost means (\$)	1,241,224	1,325,533	1,036,656	1,036,656	1,099,763	1,150,516	1,236,178	1,170,188
Cost stderr (\$)	3,778	4,095	4,214	4,214	3,963	4,227	3,912	4,208
NMB means (\$)	12,309,976	11,216,067	11,598,544	11,598,544	11,880,237	11,428,684	12,297,348	11,399,254
NMB stderr (\$)	15,673	15,311	17,026	17,026	16,346	17,162	16,159	17,210
NMB means (\$)	11,784,899		11,598,544		11,663,491		11,866,262	
NMB stderr (\$)	13,377		17,026		14,615		14,586	

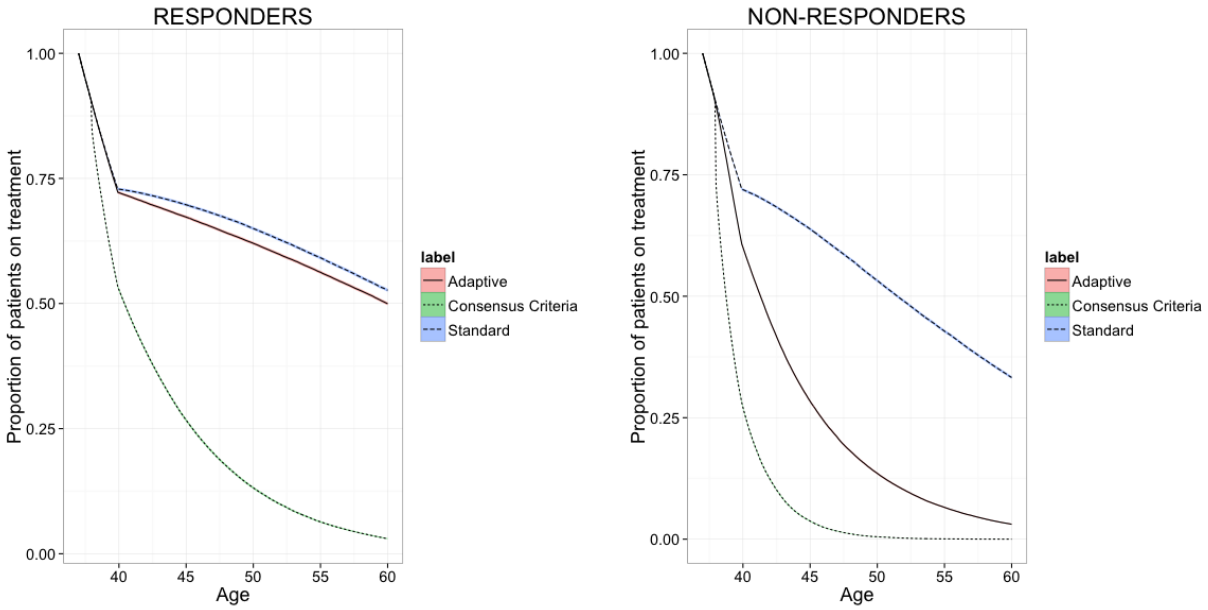


Figure 4: Proportion of patients on treatment until age 60 under the consensus, standard and adaptive policy (for WTP = \$800,000/QALY), for responders and non-responders.

and the adaptive policy and under both response types; this is due to disease progression, and the fact that all patients are taken off treatment once they reach EDSS state 6-7.5. Also, consistent with reality, the disutility incurred by patients due to side effects is higher for the first six months on treatment, which is why both groups display a non-monotonic pattern in the first year.

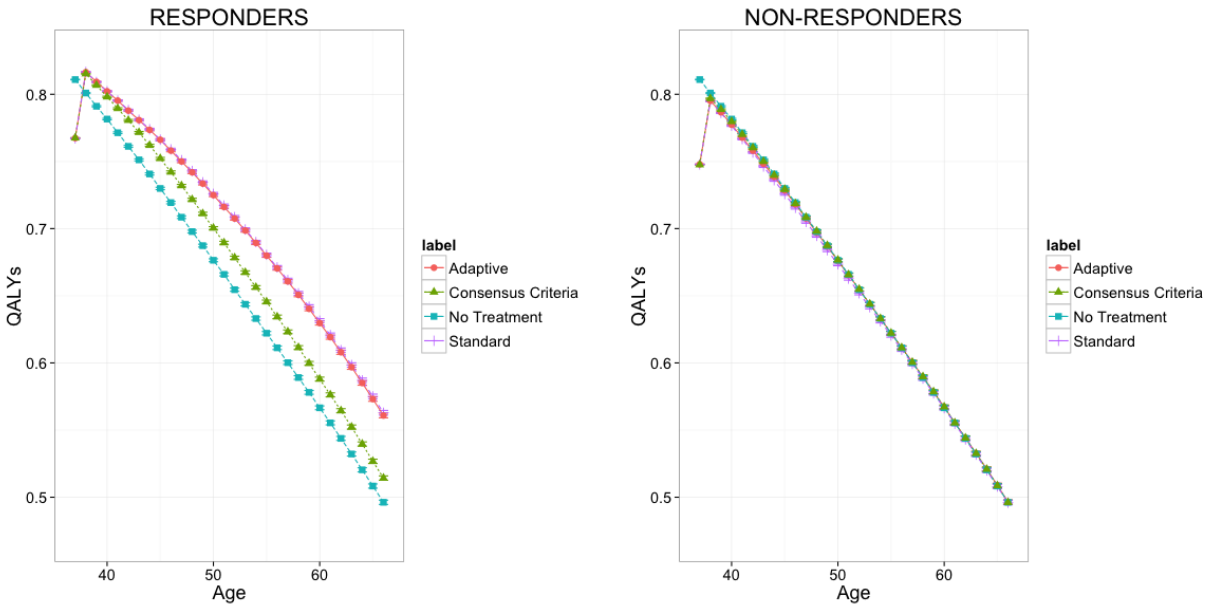


Figure 5: QALYs experienced under all treatment policies for responders and non-responders: Yearly means and 95% confidence intervals for the means.

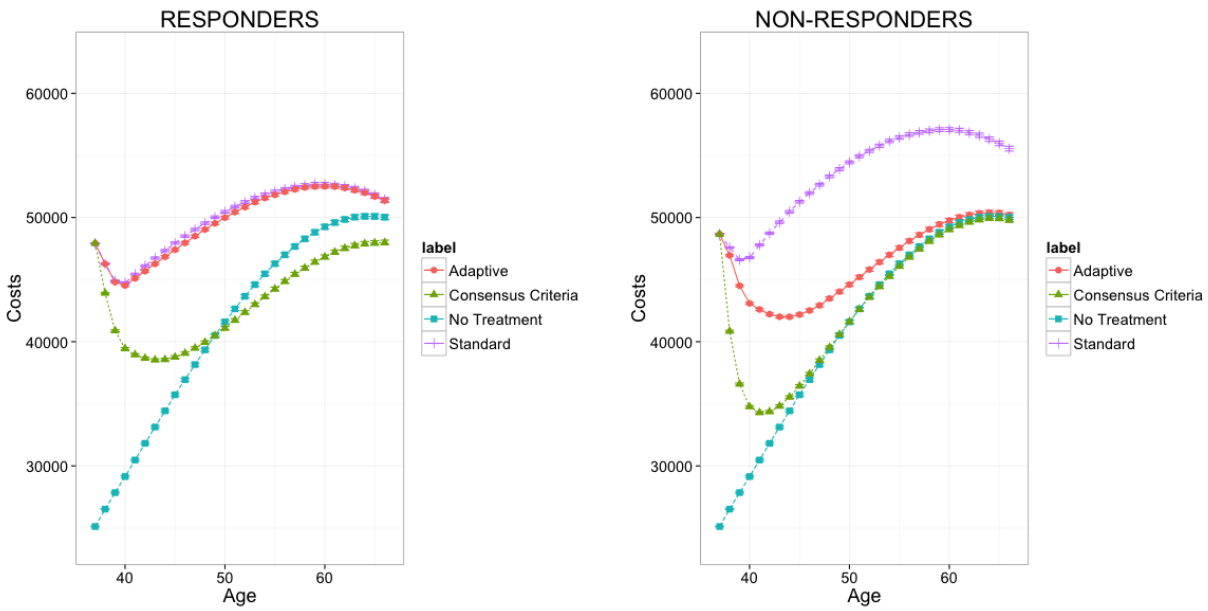


Figure 6: Costs incurred under all treatment policies for responders and non-responders: Yearly means and 95% confidence intervals for the means.

To test the robustness of our findings, we performed a probabilistic sensitivity analysis in which we randomly generated 1,000 problem instances. For each instance, each parameter was randomly sampled from a triangle distribution with mode given by the base case value, and the lowest and highest values corresponding to the ranges in Tables 1-3. The results of the analysis are shown in Figure 7. 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. As can be seen, all statistically significant differences in mean QALYs (and costs) between the adaptive and standard policies were positive (respectively, negative) for non-responders, indicating that our adaptive policy is especially cost-effective for non-responders compared to the standard policy.

5 Conclusions, Limitations, and Future Directions

Our paper introduced a quantitative framework that can inform treatment policies for chronic diseases sharing the following features: (1) there is *a priori* uncertainty about whether a patient will respond to an available treatment; (2) observations of the effectiveness of treatment are noisy, and (3) learning about treatment effectiveness occurs both from monitoring day-to-day disease progression, but also from observing the timing and severity of less frequent, major health events.

We showed that the problem of choosing between two treatments with linear dose-response can be analyzed in closed form, resulting in intuitive optimal policies that take the form of discontin-

uation rules. We also discussed how our analytical results can be used to optimally select among several treatments by solving a small number of one-dimensional, convex optimization problems, and provided conditions when the optimal treatment is no longer a simple discontinuation rule.

Finally, we used our framework to develop a set of treatment policies for administering interferon to patients suffering from multiple sclerosis. Our policies explicitly traded off treatment benefits and costs through a parameter capturing the policy makers' willingness to pay (WTP) for every quality-adjusted-life-year (QALY) gained. Using these policies as benchmarks, we then assessed several treatment guidelines used in practice for administering interferon, which lead to three conclusions that can inform policy makers and medical practitioners:

1. At WTP values below \$150,000/QALY, we found that a no-treatment policy is optimal.

2. At WTP values between \$150,000/QALY and \$500,000/QALY, we found that a policy based on the consensus criteria discussed in Cohen et al. (2004) delivers a good balance between QALYs and costs, and is almost efficient. Considering its simplicity relative to our adaptive policies, it thus emerges as the preferred treatment guideline at intermediate WTP values.

3. At WTP values above \$500,000/QALY, none of the treatment guidelines considered deliver adequate performance; an adaptive policy derived from our framework under a WTP of \$800,000/QALY attained a better balance between administering sufficient treatment to responders and identifying non-responders early.

Several next steps can bring our research and findings closer to a treatment recommendation. First, our MS case study could be generalized to allow choosing among multiple drugs for symptom management and multiple disease modifying agents (NMSS 2014). Appendix A of the Online

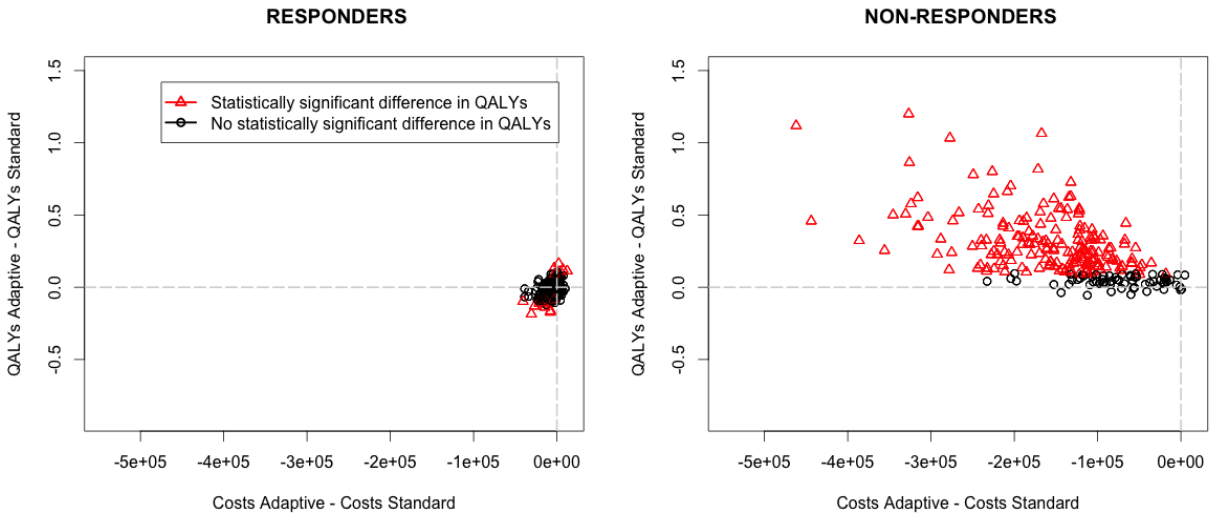


Figure 7: Results of probabilistic sensitivity analysis: Incremental costs and QALYs for responders and non-responders (adaptive treatment policy compared to standard treatment policy).

Companion provides a computationally tractable procedure for a model with multiple risky treatments/arms, which could be a building block in this direction. To enable this approach, one first requires a thorough understanding of the patient response to each drug, so as to calibrate the corresponding reward, relapse and progression rates. These could be obtained from clinical trials, such as those conducted as part of the drug approval process.

Second, one could extend our model by relaxing some of the assumptions made for analytical tractability. Of key importance here are the assumptions concerning the linear dose response and the ability to continuously measure rewards and update beliefs, which we discuss next.

Linear dose response. Although exact response curves are the subject of active research, response curves reported in the literature for many drugs tend to be S-shaped, exhibiting diminishing marginal returns at high dosage values. The assumption of linearity may nonetheless remain reasonable within a certain dosage range. For instance, clinical trials with interferon β-1a for MS suggest an approximately linear reduction in relapse rate when the dosage is below 66 micrograms per week, and a decreasing rate for higher dosage (OWIMS 1999).[10] In Appendix D, we discuss in detail the impact of the linearity assumption on optimal policies and performance when the underlying dose-response is S-shaped. We find that optimal policies are no longer "bang-bang," and that a strictly fractional treatment allocation may be optimal even when the patient is known to be a (non)responder. The optimality loss varies from 0% to 16%, depending on the degree of "nonlinearity," which suggests that embedding nonlinear response curves without sacrificing tractability may be a practically (and theoretically) meaningful future direction.

Continuous updates. Our framework allows for continuously measuring rewards and conducting belief and treatment updates. This is reasonable when the policies generated from our results are interpreted as upper bounds, which are then used either to suggest or otherwise benchmark simpler treatments with less frequent updates. Depending on the disease and treatment in question, these assumptions may also be(come) realistic. For instance, in MS, the use of wearable devices has shown to have great potential for the collection and relaying of real-time patient information (McIninch et al. 2015).[11] Combined with research aimed at understanding how disease progression and treatment response are related to observed mobility,[12] such developments could potentially make a near-continuous-time treatment policy feasible in the future (provided, of course, that the benefits outweigh the costs). Despite these examples, however, assuming continuous evaluations and treatment updates may not be reasonable when extensive medical exams are required (e.g., involving doctor visits, MRI scans, etc). Our model could be extended to allow belief updates only at particular points in time, provided that the information between these points can be suitably

[10] The Once Weekly Interferon for MS Study Group reports a reduction in relapse rates of 9.6%, 19%, 33% and 37% for respective weekly dosages of 30μg, 44μg, 66μg, and 132μg (OWIMS 1999).

[11] In a recent study conducted by the non-profit PatientsLikeMe and Biogen Idec, 248 FitBit One^{TM} devices were distributed to patients suffering from MS, and the personal mobility data of all the patients was collected and sent to centralized data servers. The results of the study were reported in the 67th American Academy of Neorology's Annual Meeting (April, 2015), revealing "a high degree of patient interest and perceived value in using activity tracking devices to help patients manage their MS" (McIninch et al. 2015).

[12] This research endeavor has recently been taken up in a collaboration by Biogen, Google X, and Cleveland Clinic (Bloomberg 2015).

aggregated. In Appendix C, we discuss the impact of monitoring frequency in more detail, and provide several theoretical and computational results that characterize the losses under less frequent updating. For our MS case study, we find that the loss from a monthly monitoring policy is less than 8%. However, we also find that as treatments for MS become more efficient at reducing the frequency of relapses in responders, these losses are likely to increase, prompting the need for more research that explicitly captures the costs of more frequent belief and treatment updating.

In addition to these, one other assumption worth relaxing would be the requirement that a risky (treatment) arm has exactly two types. In practice, more types may exist, e.g., corresponding to a patient fully, partially, or not responding to treatment. Our results would readily apply if the optimal treatment for each patient type still involved a binary choice between the same two alternatives, since then the various types could be aggregated into two "macro-types." When different patient types require different dosages or treatment options, our model would have to be extended to explicitly allow learning for all types simultaneously. This requires a multi-dimensional state that tracks the probability for each type, which considerably complicates the analysis.

Lastly, an important step in making the results implementable, is a clinical trial testing the performance of our adaptive policy against other guidelines. To that end, Appendix E of the paper's Online Companion provides an implementation-driven description of our proposed policy, which could guide such a design in conjunction with an appropriate selection of a cohort of patients.

To conclude, although we illustrated our framework with a case study on MS and interferon-β, we believe that the ideas could be used to inform the treatment of other chronic diseases, such as celiac disease, rheumatoid arthritis, Crohn's disease, or depression.

6 Acknowledgements

The authors are grateful to the department editor, the associate editor, and three anonymous referees, whose comments and suggestions have considerably improved the manuscript. Margaret Brandeau was supported by Grant Number R01-DA15612 from the National Institute on Drug Abuse.

References

Ahuja, V. & Birge, J. (2016), 'Response-adaptive designs for clinical trials: Simultaneous learning from multiple patients', *European Journal of Operational Research* **248**(2), 619 – 633.

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

Arias, E. (2014), 'National vital statistics reports'. Retrieved on December 4, 2015 from http://www.cdc.gov/nchs/data/nvsr/nvsr63/nvsr63_07.pdf.

Bank, P. & Küchler, C. (2007), 'On Gittins' index theorem in continuous time', *Stochastic Processes and Their Applications* **117**(9), 1357–1371.

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

Berry, D. & Fristedt, B. (1985), *Bandit problems: Sequential allocation of experiments*, London: Chapman and Hall.

Berry, D. & Pearson, L. (1985), 'Optimal designs for clinical trials with dichotomous responses', *Statistics in Medicine* **4**(4), 497–508.

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

Bloomberg (2015), 'Google, Biogen seek reasons for advance of multiple sclerosis'. Retrieved on September 19, 2015 from http://www.bloomberg.com/news/articles/2015-01-27/google-biogen-seek-reasons-for-advance-of-multiple-sclerosis.

Boggild, M., Palace, J., Barton, P., Ben-Shlomo, Y., Bregenzer, T., Dobson, C. & Gray, R. (2009), 'Multiple sclerosis risk sharing scheme: two year results of clinical cohort study with historical comparator', *BMJ: British Medical Journal* pp. 1359–1363.

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

Carroll, W. (2010), 'Oral therapy for multiple sclerosis–sea change or incremental step', *New England Journal of Medicine* **362**(5), 456–8.

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

Cohen, B. A., Khan, O., Jeffery, D. R., Bashir, K., Rizvi, S. A., Fox, E. J., Agius, M., Bashir, R., Collins, T. E., Herndon, R., Kinkel, P., Mikol, D. D., Picone, M. A., Rivera, V., Tornatore, C. & Zwibel, H. (2004), 'Identifying and treating patients with suboptimal responses', *Neurology* **63**(12 suppl 6), S33–S40.

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

Denton, B., Kurt, M., Shah, N., Bryant, S. & Smith, S. (2009), 'Optimizing the start time of statin therapy for patients with diabetes', *Medical Decision Making* **29**(3), 351–367.

Derfuss, T. (2012), 'Personalized medicine in multiple sclerosis: hope or reality?', *BMC medicine* **10**(1), 116.

Drummond, M. (2005), *Methods for the economic evaluation of health care programmes*, Oxford University Press.

El Karoui, N. & Karatzas, I. (1994), 'Dynamic allocation problems in continuous time', *The Annals of Applied Probability* **4**(2), 255–286.

Gold, M. (1996), *Cost-effectiveness in health and medicine*, Oxford University Press.

Harrison, J. M. & Sunar, N. (2015), 'Investment timing with incomplete information and multiple means of learning', *Operations Research* **62**(2), 442–457.

Hartung, D. M., Bourdette, D. N., Ahmed, S. M. & Whitham, R. H. (2015), 'The cost of multiple sclerosis drugs in the us and the pharmaceutical industry too big to fail?', *Neurology* **84**(21), 2185–2192.

Helm, J., Lavieri, M., Van Oyen, M., Stein, J. & Musch, D. (2015), 'Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support', *Forthcoming in Operations Research*.

Horakova, D., Kalincik, T., Dolezal, O., Krasensky, J., Vaneckova, M., Seidl, Z. & Havrdova, E. (2012), 'Early predictors of non-response to interferon in multiple sclerosis', *Acta Neurologica Scandinavica* **126**(6), 390–397.

Hunink, M., Weinstein, M., Wittenberg, E., Drummond, M., Pliskin, J., Wong, J. & Glasziou, P. (2014), *Decision making in health and medicine: integrating evidence and values*, Cambridge University Press.

Kaspi, H. & Mandelbaum, A. (1995), 'Lévy bandits: Multi-armed bandits driven by Lévy processes', *The Annals of Applied Probability* **5**(2), 541–565.

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

Keller, G. & Rady, S. (2015), 'Breakdowns', *Theoretical Economics* **10**, 175–202.

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

Kremenchutzky, M., Rice, G., Baskerville, J., Wingerchuk, D. & Ebers, G. (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., Baxter, D., Limone, B., Roberts, M. & Coleman, C. (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.

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

Mariette, X., Matucci-Cerinic, M., Pavelka, K., Taylor, P., van Vollenhoven, R., Heatley, R., Walsh, C., Lawson, R., Reynolds, A. & Emery, P. (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, J., Denton, B., Shah, N. & Smith, S. (2014), 'Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients', *European Journal of Operational Research* **233**(3), 727–738.

McIninch, J., Datta, S., DasMahapatra, P., Chiauzzi, E., Bhalerao, R., Spector, A., Goldstein, S., Morgan, L. & Relton, J. (2015), 'Remote tracking of walking activity in ms patients in a real-world setting (p3.209)', *Neurology*.

Murphy, S. (2005), 'An experimental design for the development of adaptive treatment strategies.', *Statistics in Medicine* **24**(10), 1455.

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

Murphy, S. A. (2003), 'Optimal dynamic treatment regimes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(2), 331–355.

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*.

Noyes, K., Bajorska, A., Chappel, A., Schwid, S., Mehta, L., Weinstock-Guttman, B., Holloway, R. & Dick, A. (2011), 'Cost-effectiveness of disease-modifying therapy for multiple sclerosis: A population-based study', *Neurology* **77**(4), 355–363.

O'Rourke, K. E. & Hutchinson, M. (2005), 'Stopping beta-interferon therapy in multiple sclerosis: an analysis of stopping patterns', *Multiple Sclerosis* **11**(1), 46–50.

OWIMS (1999), 'Evidence of interferon β-1a dose response in relapsing-remitting MS: The OWIMS study', *Neurology* **53**(4), 679.

Phillips, C. J. (2004), 'The cost of multiple sclerosis and the cost effectiveness of disease-modifying agents in its treatment', *CNS Drugs* **18**(9), 561–574.

Pincus, T., Callahan, L., Sale, W., Brooks, A., Payne, L. & Vaughn, W. (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., Bellemare, M., Rush, A., Ghizaru, A. & Murphy, S. (2007), 'Constructing evidence-based treatment strategies using methods from computer science', *Drug and Alcohol Dependence* **88**, S52–S60.

Powell, W. & Ryzhov, I. (2012), *Optimal learning*, Vol. 841, John Wiley & Sons.

Prosser, L., Kuntz, K., Bar-Or, A. & Weinstein, M. (2003), 'Patient and community preferences for treatments and health states in multiple sclerosis', *Multiple Sclerosis* **9**(3), 311–319.

Prosser, L., Kuntz, K., Bar-Or, A. & Weinstein, M. (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*.

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

Romeo, M., Martinelli-Boneschi, F., Rodegher, M., Esposito, F., Martinelli, V., Comi, G. & Group, S. R. M. S. C. (2013), 'Clinical and MRI predictors of response to interferon-beta and glatiramer acetate in relapsing-remitting multiple sclerosis patients', *European Journal of Neurology* **20**(7), 1060–1067.

Rovaris, M., Comi, G., Rocca, M., Wolinsky, J., Filippi, M. 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, R., Stuart, W., Calabresi, P., Confavreux, C., Galetta, S., Radue, E., Lublin, F., Weinstock-Guttman, B., Wynn, D., Lynn, F. et al. (2006), 'Natalizumab plus interferon beta-1a for relapsing multiple sclerosis', *New England Journal of Medicine* **354**(9), 911–923.

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

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

Tappenden, P., McCabe, C., Chilcott, J., Simpson, E., Nixon, R., Madan, J., Fisk, J. D. & Brown, M. (2009), 'Cost-effectiveness of disease-modifying therapies in the management of multiple sclerosis for the medicare population', *Value in Health* **12**(5), 657 – 665.

Young, P. & Olsen, L. (2010), *The Healthcare Imperative: Lowering Costs and Improving Outcomes: Workshop Series Summary*, The National Academies Press.

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

Technical Appendix for "Dynamic Learning of Patient Response Types: An Application to Treating Chronic Diseases"

by Diana M. Negoescu, Kostas Bimpikis, Margaret L. Brandeau, and Dan A. Iancu

Appendices

A Multiple Risky Arms

In this section, we explore the important extension where several risky arms with binary (i.e., good/bad) types exist. Our goal is to show how our previous results can be used to devise a very simple scheme that determines the optimal policy to within an arbitrary pre-specified precision, by solving a small number of one-dimensional convex optimization problems.

Consistent with our framework thus far, we assume there are $n+1$ arms numbered $0, \ldots, n$. Arm 0 corresponds to the "safe" arm, yielding instantaneous Brownian rewards with drift rate μ_0 and volatility σ. Every arm $i \geq 1$ is risky, and can be of either good or bad type $\theta_i \in \{G_i, B_i\}$. Depending on the type, the i-th arm thus yields instantaneous Brownian rewards with volatility σ and drift rate μ_{G_i} (if good) or μ_{B_i} (if bad), and induces life events according to a Poisson process with rate λ_{G_i} (if good) or λ_{B_i} (if bad). For notational convenience, let $\theta_0 \stackrel{\text{def}}{=} 0$. For simplicity, we ignore the stopping events, and consider an infinite planning horizon.

We assume that the DM's allocation during the interval $[t, t+dt)$ involves a single risky arm and the safe arm. The allocation entails a choice $i \in \{1, \ldots, n\}$ and a corresponding fraction $\alpha_t^i \in [0,1]$ allocated to the i-th risky arm, with the remaining fraction $\alpha_t^0 \stackrel{\text{def}}{=} 1 - \alpha_t^i$ allocated to the safe arm. This generates total instantaneous rewards of $d\pi^i(t)$ and $d\pi^0(t)$, respectively, where

$$d\pi^k(t) \stackrel{\text{def}}{=} \alpha_t^k \, \mu_{\theta_k} \, dt + \sqrt{\alpha_t^k} \, \sigma \, dZ^k(t), \quad k \in \{0, i\},$$

and $dZ^i(t)$ and $dZ^0(t)$ are independent, normally distributed random variables with mean 0 and variance dt. Additionally, under this allocation, life events occur according to a Poisson process with rate $\lambda(t,\theta) = \alpha_t^i \lambda_{\theta_i} + \alpha_t^0 \lambda_0$, with every occurrence generating a lump-sum reward $-D$.

We focus on an infinite planning horizon, so that the DM's objective is to maximize:

$$\Pi \stackrel{\text{def}}{=} \mathbb{E}\left[\int_0^\infty e^{-rt}\left(\sum_{i=0}^n d\pi^i(t) - D \cdot \lambda(t,\theta) dt\right)\right].$$

Furthermore, for simplicity, we discuss the case where learning occurs primarily through the instantaneous rewards, so that we assume $\lambda_{G_i} = \lambda_{B_i} = \lambda_i$ (similar ideas can be applied to the more general version of the problem). As before, we assume that belief updates can be noisy, and no arm can be *a priori* eliminated from consideration, summarized below.

Assumption 2. *The model primitives satisfy the conditions* $\mu_{B_i} - D\lambda_i \leq \mu_0 - D\lambda_0 \leq \mu_{G_i} - D\lambda_i$ *and* $\mu_{G_i} \neq \mu_{B_i}$*, for any* $i \in \{1, \ldots, n\}$.

A sufficient statistic of the history up to time t is given by the vector $p_t \in [0,1]^n$, whose i-th component p_t^i denotes the probability that the i-th arm is good, conditional on all information up to time t. The update rule for p_t^i can be written exactly as in our benchmark model, depending on whether a life event occurs during $[t, t+dt)$, yielding results analogous to those in Lemmas 1 and 2. Note that while arm i is used, the beliefs for all arms $j \neq i$ are unaffected.

In this context, it can be readily verified that the evolution of p_t^i is driven by a Lévy process, and thus our model belongs to the class of Lévy bandits studied in Kaspi & Mandelbaum (1995). For such models, it is known that the optimal policy is indexable, i.e., one can define a Gittins index for every risky arm i, and the optimal policy is to use the arm with the largest index at every point in time (see, e.g., Theorem 3.1 in Kaspi & Mandelbaum 1995). Furthermore, if h_t^i denotes the stochastic process characterizing the rewards of arm i, then the Gittins index of arm i at time t is given by (see, e.g., Corollary 2.1 in Bank & Küchler 2007):

$$\inf\left\{ m \in \mathbb{R} \,:\, m \geq \mathbb{E}\left[\int_t^S e^{-r(u-t)}\, h_u^i \, du + e^{-r(S-t)} m \,\middle|\, \mathcal{F}_t^i \right] \right\}, \tag{8}$$

where S is any $\mathcal{F}^i$-stopping time satisfying $S \geq t$. In other words, the Gittins index is the smallest value of a deterministic "retirement reward" m that would make the DM indifferent between (i) immediately retiring at time t and earning a reward of m, or (ii) continuing to use the risky arm i and stopping optimally at some future time with a retirement reward of m.

Using this representation theorem in conjunction with our analytical framework enables us to characterize the Gittins index of a risky arm as the solution to a simple one-dimensional convex optimization problem. This is formalized in our next result.

Theorem 2. *Consider the* i*-th risky arm, whose prior probability of being good is* $p_t^i \equiv p$ *at time* t*. Its Gittins index is given by*

$$\mathcal{G}_t^i(p) = \begin{cases} -\infty, & \text{if } p < p_i^*\big(\tfrac{\mu_0 - D\lambda_0}{r}\big) \\ \min\big\{ m \in \mathbb{R} \;:\; m \geq f_i(p,m) \big\}, & \text{if } p \geq p_i^*\big(\tfrac{\mu_0 - D\lambda_0}{r}\big), \end{cases} \tag{9a}$$

$$\text{where} \quad f_i(p,m) \overset{\text{def}}{=} A_i(p) + B_i(p)\, \frac{\mu_{G_i} - \mu_{B_i}}{r}\, \frac{p_i^*(m)}{p_i^*(m) + \nu_i^*} \Big[\frac{p_i^*(m)}{1 - p_i^*(m)} \Big]^{\nu_i^*}, \tag{9b}$$

$$p_i^*(m) \overset{\text{def}}{=} \frac{\nu_i^* \big[r \cdot m - (\mu_{B_i} - D\lambda_i) \big]}{\mu_{G_i} - D\lambda_i - r \cdot m + \nu_i^* (\mu_{G_i} - \mu_{B_i})}, \tag{9c}$$

$$\nu_i^* \overset{\text{def}}{=} = \frac{-(\mu_{G_i} - \mu_{B_i}) + \sqrt{(\mu_{G_i} - \mu_{B_i})^2 + 8r\sigma^2}}{2(\mu_{G_i} - \mu_{B_i})}, \tag{9d}$$

$$A_i(p) \overset{\text{def}}{=} \frac{p\, \mu_{G_i} + (1-p)\, \mu_{B_i} - D\lambda_i}{r}, \tag{9e}$$

$$B_i(p) \overset{\text{def}}{=} (1-p) \Big(\frac{1-p}{p} \Big)^{\nu_i^*}. \tag{9f}$$

Furthermore, the function $f_i(p,m)$ *is convex in* m*, for any* $p \in [0,1]$.

We provide a proof of Theorem 2 in Appendix F. To gain some intuition behind the result, note that $p_i^*(\frac{\mu_0 - D\lambda_0}{r})$ exactly corresponds to the belief threshold in Theorem 1 for the special case of an infinite planning horizon, below which the DM would stop using the i-th risky arm and switch to a safe arm. Thus, the first part of expression (9a) confirms the intuitive fact that if the i-th risky arm is not worth experimenting with in isolation, i.e., $p < p_i^*\big(\frac{\mu_0 - D\lambda_0}{r}\big)$, it will not be worth experimenting with in the presence of other risky arms, so that $\mathcal{G}^i = -\infty$. The second part of (9a) states that, for an arm that is worth using in isolation, i.e., $p \geq p_i^*\big(\frac{\mu_0 - D\lambda_0}{r}\big)$, the Gittins index $\mathcal{G}^i(p)$ can be obtained by solving a single one-dimensional convex optimization problem. Since such problems can be solved very efficiently, for instance through a simple bisection method, this suggests the following algorithm for finding the optimal arm to play at any point of time.

Data: Number of risky arms (n); volatility (σ); mean rewards (μ_{θ_i}) and relapse rates (λ_i) for any type ($\theta_i \in \{G_i, B_i\}$) and for all arms ($i \in \{0, \ldots, n\}$); discretization error for prior values ($\epsilon > 0$).
Result: Values for the Gittins index of every arm i, at every discretized prior value p.
begin
  $\mathcal{P} \longleftarrow \{0, \epsilon, 2\epsilon, 3\epsilon, \ldots, 1\}$ `/* (discretized values for the prior) */`
  $\mathcal{G} \longleftarrow$ empty array of size $n \times |\mathcal{P}|$ `/* (table of Gittins index values) */`
  for $i = 1, \ldots, n$ **do**
    Calculate $p_i^*\big(\frac{\mu_0 - D\lambda_0}{r}\big)$ according to (9c)
    for $p \in \mathcal{P}$ **do**
      if $p < p_i^*\big(\frac{\mu_0 - D\lambda_0}{r}\big)$ **then**
        $\mathcal{G}(i,p) \longleftarrow -\infty$ `/* arm i not used at this prior value */`
      end if
      else
        $\mathcal{G}(i,p) \longleftarrow \min\{m : m \geq f_i(p,m)\}$.
      end if
    end for
  end for
end

Algorithm 1: Gittins Index Calculation

It is important to note that Algorithm 1 can be run entirely offline, before implementing the optimal policy in real time. In particular, for a given precision $\epsilon > 0$ governing the discretization, Algorithm 1 can generate the Gittins index for every risky arm i at every possible discretized belief value p, by solving $\mathcal{O}(\frac{n}{\epsilon})$ one-dimensional convex optimization problems. Once these indices are calculated, the DM can obtain an optimal discretized policy, as follows. The DM would first discretize time in increments of length δ, chosen small enough so that the probability of two or more life events during an interval of size δ is very small. At every time instant $k\delta$ ($k \in \{0, 1, \ldots\}$), the DM would start with a belief of value $\hat{p}^i$ that the arm is good (suitably initialized at time 0) and obtain the associated Gittins index via a simple look-up in the table provided by Algorithm 1, yielding $\mathcal{G}(i, \hat{p}^i)$. If all risky arms have index $-\infty$, the DM will switch to the safe arm and use it indefinitely. Otherwise, the DM would select the risky arm i^* with the largest Gittins index,

i.e., $i^* \in \arg\max_j \mathcal{G}(j, \hat{p}^j)$, and use an allocation $\alpha_t^{i^*} = 1$ in the time-period $[k\delta, (k+1)\delta)$. Once the instantaneous and lump-sum rewards are observed, the DM would update the belief for arm i^* according to Lemma 1(i) when no event occurs, or Lemma 2(i) upon a life event.

B Type-Dependent Lump-Sum Rewards

In this section, we extend our model to a case where the rewards received upon a life event can depend on the unknown type θ. This extension allows us to capture settings where a successful treatment also reduces the magnitude/impact of major negative health events, in addition to their likelihood/frequency—a feature that is relevant for diseases such as depression or Crohn's disease.

To that end, we assume that any life event can be either "mild" or "severe," with corresponding "rewards" (i.e., disutilities) of size $-D_M$ and $-D_S$, respectively, where $D_M < D_S$. Furthermore, when the DM's allocation is $\alpha \in [0,1]$, the probability that a given life event is *mild* is $\bar{q}_\theta \stackrel{\text{def}}{=} (1-\alpha)q_0 + \alpha q_\theta$ with $\theta \in \{B, G\}$. Here, q_0, q_G, q_B denote the probability of a mild life event under a safe, good and bad arm, respectively. For simplicity, we ignore stopping events and restrict attention to a model with an infinite planning horizon, i.e., we assume $\eta_0, \eta_G, \eta_B \to 0$.

We now discuss the belief updating and optimal policy. When no event occurs during $[t, t+dt)$, the belief is updated according to Lemma 1. When a life event occurs, the posterior now depends on whether the event was mild or severe. The following lemma provides the learning rule.

Lemma 3. *When a life event occurs during $[t, t+dt)$,*
(i) the posterior belief p_{t+dt} conditional on the observed event type (mild/severe) and on the instantaneous reward from the risky arm ($d\pi^1 = y$) is given by Bayes' rule, and takes a value of

$$p_{t+dt} = \begin{cases} \frac{p_t\, \bar{q}_G\, F(\mu_G/\sigma)\,(1-e^{-\bar{\lambda}_G dt})}{p_t\, \bar{q}_G\, F(\mu_G/\sigma)\,(1-e^{-\bar{\lambda}_G dt}) + (1-p_t)\, \bar{q}_B\, F(\mu_B/\sigma)\,(1-e^{-\bar{\lambda}_B dt})} & \text{if the event is mild} \\ \frac{p_t\,(1-\bar{q}_G)\, F(\mu_G/\sigma)\,(1-e^{-\bar{\lambda}_G dt})}{p_t\,(1-\bar{q}_G)\, F(\mu_G/\sigma)\,(1-e^{-\bar{\lambda}_G dt}) + (1-p_t)\,(1-\bar{q}_B)\, F(\mu_B/\sigma)\,(1-e^{-\bar{\lambda}_B dt})} & \text{if the event is severe;} \end{cases}$$

(ii) the change in the DM's belief $p_{t+dt} - p_t$ is normally distributed, with a mean of

$$\begin{cases} j_M(\alpha_t, p_t) - p_t + \alpha_t p_t (1-p_t) \bar{\lambda}_G \bar{\lambda}_B \frac{\eta_B - \eta_G + (\mu_G - \mu_B)\mu(p_t)/\sigma^2}{(\bar{\lambda}(p_t))^2} dt & \text{if the event is mild} \\ j_S(\alpha_t, p_t) - p_t + \alpha_t p_t (1-p_t) \bar{\lambda}_G \bar{\lambda}_B \frac{\eta_B - \eta_G + (\mu_G - \mu_B)\mu(p_t)/\sigma^2}{(\bar{\lambda}(p_t))^2} dt & \text{if the event is severe,} \end{cases}$$

and a variance of $\alpha_t \big(p_t(1-p_t)\bar{\lambda}_G \bar{\lambda}_B (\mu_G - \mu_B)/\sigma\big)^2 dt$, where $\bar{\lambda}_\theta, F$ are given in (4a)-(4b), and

$$j_M(\alpha_t, p_t) \stackrel{\text{def}}{=} \frac{p_t \bar{q}_G \bar{\lambda}_G}{(1-p_t)\bar{q}_B \bar{\lambda}_B + p_t \bar{q}_G \bar{\lambda}_G}, \qquad j_S(\alpha_t, p_t) \stackrel{\text{def}}{=} \frac{p_t (1-\bar{q}_G) \bar{\lambda}_G}{(1-p_t)(1-\bar{q}_B)\bar{\lambda}_B + p_t (1-\bar{q}_G)\bar{\lambda}_G}.$$

The proof follows similarly to Lemma 2, and is omitted. As in our main model, the occurrence of a life event results in a jump in the DM's belief, and the posteriors $j_M(\alpha_t, p_t)$ and $j_S(\alpha_t, p_t)$ obey similar comparative statics. A critical difference compared with our main model is that the posterior is no longer necessarily lower than the belief p_t. In particular, it can be readily checked

that $j_S(\alpha_t, p_t) < p_t$ always holds provided $\lambda_B > \lambda_G$, so that the occurrence of a *severe* event always lowers the DM's belief that the arm is good. When the event is *mild*, though, it is possible to have $j_M(\alpha_t, p_t) > p_t$, i.e., the likelihood of the arm being good may increase.

Interestingly, this seemingly innocuous change whereby beliefs can admit upward jumps bears important implications on the optimal policy. Although finding a closed-form expression is no longer feasible due to the nonlinear dependency on α_t induced by the jumps, by examining the extreme case $p_t = 1$ we can derive the following insights (we omit a proof for reasons of space).

Theorem 3 (Fractional allocation)**.** *Assume that belief updates can be stochastic (i.e., $\mu_G \neq \mu_B$), and a good arm dominates the safe arm, which dominates a bad arm in total rewards per unit time:*

$$\mu_B - \lambda_B(q_B D_M + (1-q_B)D_S) \leq \mu_0 - \lambda_0(q_0 D_M + (1-q_0)D_S) \leq \mu_G - \lambda_G(q_G D_M + (1-q_G)D_S).$$

(i) If $q_0 < q_G$, then the optimal allocation for $p_t = 1$, i.e., $\alpha_t^*(1)$, is given by the expression

$$\alpha_t^*(1) = \frac{\frac{\mu_G - \mu_0}{D_S - D_M} + (\lambda_0 - \lambda_G)(\frac{D_S}{D_S - D_M} - q_0) + \lambda_0(q_G - q_0)}{2(q_G - q_0)(\lambda_0 - \lambda_G)}.$$

(ii) Furthermore, if $-(\lambda_0 - \lambda_G)(D_M q_0 + D_S(1-q_0)) - \lambda_0(q_G - q_0)(D_S - D_M) < \mu_G - \mu_0$ and $\mu_G - \mu_0 < (D_S - D_M)[q_G(\lambda_0 - \lambda_G) - \lambda_G(q_G - q_0)] - D_S(\lambda_0 - \lambda_G)$ hold, then $\alpha_t^*(1) \in (0,1)$, and the optimal policy is not bang-bang even when $p_t = 1$.

Theorem 3 suggests that even when the risky arm is guaranteed to be "good," a fractional allocation may be strictly better than a complete allocation to the risky arm. To gain some intuition for the conditions in (ii), we note that they imply a lower bound on q_G coupled with lower and upper bounds on q_0, as well as an upper bound on λ_G, coupled with lower and upper bounds on λ_0. Thus, the conditions essentially require that the risky and safe arm deliver comparable performance in terms of instantaneous rewards, with the risky arm "sufficiently efficient" in reducing the magnitude of disutility from negative health events and the safe arm "not too efficient" for this purpose. Under these conditions, mixing the two arms can thus achieve "the best of both worlds."

C Monitoring Frequency

In this section, we explore the impact of the continuous monitoring assumption on our results. We consider a case where the allocation α_t and the belief concerning the arm type can only be updated at particular pre-determined points of time $t \in \{0, \Delta, 2\Delta, \dots\}$. The *monitoring interval* $\Delta > 0$ controls the frequency of monitoring. We restrict our attention to a model with an infinite planning horizon, binary allocation decisions ($\alpha_t \in \{0,1\}$), $\lambda_G < \lambda_B = \lambda_0 = 1$, and $\mu_B = \mu_G < \mu_0$.

We compare treatment decisions and performance for two adaptive policies, with monitoring intervals Δ and 2Δ. Let $J^{k\Delta}(p)$ and $\alpha^{k\Delta}(p)$ denote the optimal value function and the optimal policy under a monitoring interval $k\Delta, k \in \{1,2\}$. We then have the following result.

Lemma 4. *For any prior belief p that the arm is good, $J^{\Delta}(p) \geq J^{2\Delta}(p)$ and $\alpha^{\Delta}(p) \geq \alpha^{2\Delta}(p)$.*

Proof. Any policy that is feasible under 2Δ-monitoring is also feasible under Δ-monitoring, by ignoring the odd monitoring times $(2k+1)\Delta$, for $k \in \mathbb{N}$. Therefore, $J^{\Delta}(p) \geq J^{2\Delta}(p)$.

We claim that $\alpha^{\Delta}(p) = 0$ implies $\alpha^{2\Delta}(p) = 0$, which would complete our proof. Note that if $\alpha^{\Delta}(p) = 0$ for some p, then $J^{\Delta}(p) \leq J^{2\Delta}(p)$, since the former policy no longer updates the allocation (as no learning occurs once $\alpha^{\Delta}(p) = 0$), while the latter policy may update the allocation. Thus, we must have $J^{\Delta}(p) = J^{2\Delta}(p)$, and thus $\alpha^{2\Delta}(p) = \alpha^{\Delta}(p) = 0$ maximizes the value function. □

The lemma confirms the intuition that more frequent monitoring is beneficial: it yields higher value functions, and it allows the DM to experiment more aggressively with the risky arm, as any potential "mistakes" could be more readily corrected. We note that this finding also extends to a more general setting, such as when belief and allocation updating is also possible upon the occurrence of life events. This is summarized in the next corollary, whose proof follows a similar line of reasoning, and is omitted for space considerations.

Corollary 1. *Suppose monitoring occurs at every deterministic monitoring event as well as upon the occurrence of a life event. Then, $J^{\Delta}(p) \geq J^{2\Delta}(p)$, and $\alpha^{\Delta}(p) \geq \alpha^{2\Delta}(p)$, for any prior belief p.*

This setting may be particularly relevant in a medical context, since the infrequent life events may be inherently associated with a visit to the physician, which warrants additional testing and a potential treatment update.

To illustrate the effect of monitoring frequency on the optimal policy and value function, we generate and numerically solve several problem instances where we vary λ_G, λ_B, and Δ relative to λ_0. We set $\lambda_G = (1-\epsilon)\lambda_0$, $\lambda_B = (1+\epsilon)\lambda_0$, and let Δ be proportional to $1/\lambda_0$. Table 6 shows the optimality loss associated with a finite monitoring frequency as compared with continuous monitoring, i.e., $1 - J^{\Delta}/J^{0}$.

Table 6: Optimality Loss (%) Associated with Finite Monitoring Frequencies, Compared to Continuous Monitoring. Here, $\mu_0 = 0.958$ QALYs/year, $\mu_G = \mu_B = 0.938$ QALYs/year, $D = 0.559$ QALYs, $\lambda_0 = 1$ event/year, $\lambda_G = (1-\epsilon)\lambda_0$, $\lambda_B = (1+\epsilon)\lambda_0$.

Monitoring interval	$\epsilon = 0.1$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 0.75$
$\Delta = 1/(8\lambda_0)$	0.047	0.05	0.08	0.11
$\Delta = 1/(4\lambda_0)$	0.10	0.11	0.22	0.32
$\Delta = 1/(2\lambda_0)$	0.12	0.22	0.49	0.75
$\Delta = 1/\lambda_0$	0.15	0.47	1.1	1.7
$\Delta = 2/\lambda_0$	0.2	0.8	2	3.4
$\Delta = 4/\lambda_0$	0.3	1.3	3.6	6.7
$\Delta = 8/\lambda_0$	0.5	2.3	6.8	12.3
$\Delta = 16/\lambda_0$	0.7	4.0	11.7	19.4

As the table highlights, the efficiency losses resulting from infrequent monitoring are relatively small when the difference between the rates under a good and a bad arm is not too large (i.e., ϵ is small). However, the efficiency losses can become substantial as the relative benefits of successful treatment increase. For MS, λ_0 and λ_B are approximately 1, and λ_G is approximately 0.5. A monthly monitoring frequency thus is similar to the first line in Table 6, and therefore optimality losses are less than 8%.

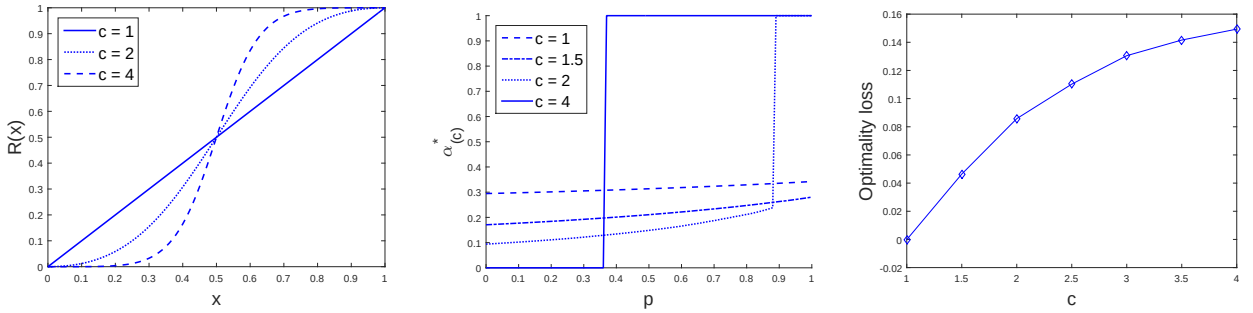
D Impact of Nonlinear Dose Response

In this section, we investigate the sensitivity of our results to the assumption of a linear dose response. We consider an S-shaped dose-response curve given by

$$\mathrm{R}(x) = \frac{x^c}{x^c + (1-x)^c}, \tag{10}$$

where c is a constant and x represents the dose. Figure 8i illustrates such curves for $c \in \{1, 2, 4\}$. Note that the dose-response curve is linear for $c = 1$, as in our base-case model in Section 2, and becomes increasingly nonlinear as c increases, approaching the threshold function $\mathbb{1}\{x > 0.5\}$.

Solving for an optimal policy with a nonlinear response curve under our general model is difficult analytically as well as computationally, and is outside the scope of our study. However, we can investigate the effects of a nonlinear response curve in a simplified version of our model, which can be solved numerically through value iteration. Namely, we consider an infinite planning horizon ($\eta_0, \eta_G, \eta_B \to 0$), where $\mu_G = \mu_B = \mu$ (learning can be achieved only by observing life events), and where the nonlinear response affects the frequency of negative health events. In other words, given an allocation of α to the risky arm and $1-\alpha$ to the safe arm, where $\alpha \in [0, 1]$, the instantaneous rewards received are $(1-\alpha)\mu_0 + \alpha\mu$ whereas the life events occur with rate $(1-\alpha)\lambda_0 + \mathrm{R}(\alpha)\lambda_\theta$, depending on the risky arm type $\theta \in \{G, B\}$.



(i) Dose-response curves $\mathrm{R}(x) = x^c/[x^c + (1-x)^c]$ for different values of c.

(ii) Optimal policies for different values of c.

(iii) Optimality losses incurred when applying policy for $c = 1$ in a model with true response c.

Figure 8: Impact of nonlinear dose-response curves on the optimal policy and value function. Here, $\lambda_0 = 1$ relapse/year, $\lambda_G = 0.85$ relapses/year, $\lambda_B = 1.75$ relapses/year, $D = 0.56$ QALYs, $\mu_0 = 0.64$ QALYs/year, $\mu_G = \mu_B = 0.62$ QALYs/year, and $r = 0.03$.

Let $\alpha^*_{(c)}$ denote the optimal policy corresponding to an S-shaped response curve with parameter c. Using our analytical results, we can derive the optimal policy for a linear response, i.e., $\alpha^*_{(1)}$. To numerically find $\alpha^*_{(c)}$ for a general c, we discretize the [0,1] spaces of the prior probability p and allocation α into intervals of size 0.01, and use a value iteration algorithm with daily time steps. The optimal policies $\alpha^*_{(c)}$ are depicted in Figure 8ii. Note that under a nonlinear response ($c > 1$), bang-bang policies are no longer optimal, and strictly splitting the allocation between the risky and the safe arm may be optimal even when the risky arm is known to be good or bad.

To measure the losses incurred by an incorrect linearity assumption, for each value of c, we simulate the policies $\alpha^*_{(1)}$ and $\alpha^*_{(c)}$ over a 10-year horizon, and record their respective performances $J(\alpha^*_{(1)})$ and $J(\alpha^*_{(c)})$, and the optimality loss $1-J(\alpha^*_{(1)})/J(\alpha^*_{(c)})$. The results, displayed in Figure 8iii, suggest that losses are relatively small under mild nonlinearities, e.g., below 8% for $c \leq 2$. As the response approaches a threshold function, losses approach 16% in a concave fashion. However, a threshold response is in some sense the "worst-case" nonlinearity, involving a jump in the profile that is unlikely for drug response curves. For instance, using the dose-response values for MS reported in OWIMS (1999) and fitting curves (10) for different values of c, it turns out that the linear curve provides the best fit under any Euclidean distance. These results suggest that a linear function can provide a reasonable first-order approximation when designing treatments in practice.

E Adaptive Policy at High Willingness-to-Pay

In this section, we provide a brief implementation-driven description of our proposed adaptive policy at high WTP (above \$800,000/QALY). For a new patient, our policy could be implemented as:

1. If the patient's EDSS score is higher than 6, no interferon-β treatment is administered.
2. Otherwise, initialize the belief $\hat{p}$ that the patient is a responder to a suitable value (such as the fraction of responders in the population at the patient's age, e.g., 52% at age 37).
3. On a monthly basis, and while the EDSS score is 0-2.5 or 3-5.5, repeat the following steps:
   (a) Using the patient's current age and EDSS score, obtain a threshold p^* from Figure 9.
   (b) If $\hat{p} < p^*$, discontinue treatment.
   (c) Otherwise,
      i. apply interferon-β for the next month;
      ii. at the end of the month, conduct a survey to assess the patient's quality-of-life (QALY) value during the preceding month;
      iii. using the QALY value and the parameters described in Section 4, update $\hat{p}$ according to formula (3) if there was no relapse during the preceding month, or according to formula (5) if there was a relapse;
      iv. update the patient's age and assess the patient's new EDSS score;
      v. go to step 3.

F Proofs

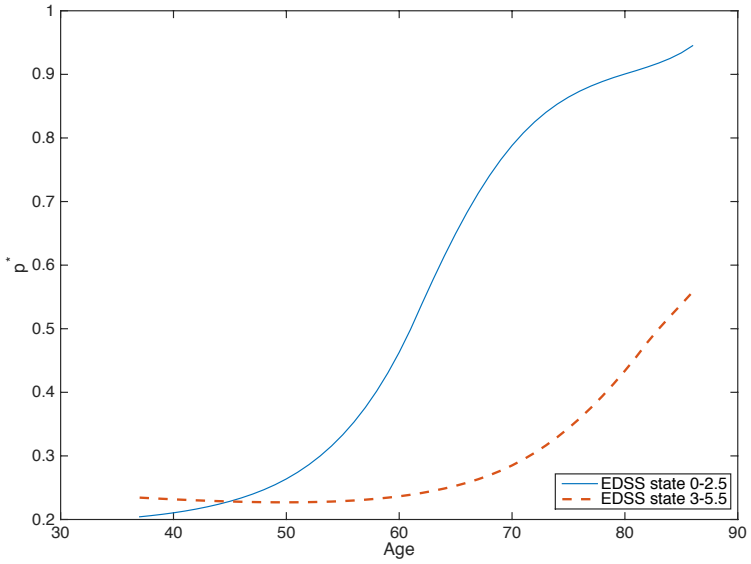


Figure 9: Optimal policy for a WTP = \$800,000/QALY. The plots correspond to the threshold values where treatment should be switched in each EDSS state, as a function of the patient's age.

Note: To simplify notation in our proofs, we suppress the subscript t whenever possible. Furthermore, since we frequently average quantities with respect to α_t or p_t, we define the following notation:

$$\bar{\xi}_\theta = \alpha_t \xi_\theta + (1-\alpha_t)\xi_0, \ \forall\, \theta \in \{G, B\} \qquad\qquad E_p[\xi] = p\xi_G + (1-p)\xi_B.$$

That is, an overbar will denote a convex combination of a quantity corresponding to the risky arm with the same quantity corresponding to the safe arm, with coefficients α_t and $1-\alpha_t$, e.g., $\bar{\lambda}_G = \alpha_t \lambda_G + (1-\alpha_t)\lambda_0$. Similarly, $\mathbb{E}_p[\cdot]$ will denote an expectation of a quantity pertaining to the risky arm taken with respect to p, e.g., $\mathbb{E}_p[\lambda_\theta] = p\lambda_G + (1-p)\lambda_B$.

Proof of Lemma 1. The proof is similar to Bolton & Harris (1999), except we need to incorporate the information provided by the lack of a life event during the interval $[t, t+dt)$. The rewards $d\pi^1(t)$ are observationally equivalent to $d\tilde{\pi}^1(t) = \sqrt{\alpha_t}\tilde{\mu}_\theta dt + dZ^1(t)$, with $\tilde{\mu}_\theta = \mu_\theta/\sigma$. Using Bayes' rule and omitting the subscript t, we have:

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

where $F(x) = \frac{1}{\sqrt{2\pi dt}} \exp\left\{-\frac{(d\tilde{\pi}^1(t) - \sqrt{\alpha}\, x\, dt)^2}{2dt}\right\}$. After Taylor-expanding the $e^{-(\bar{\lambda}_\theta + \bar{\eta}_\theta)dt}$ terms and

dropping terms of order dt^2 or higher, we have:

$$dp = p_{t+dt} - p = \frac{p(1-p)\left[\tilde{F}(\tilde{\mu}_G) - \tilde{F}(\tilde{\mu}_B) - dt\big(\tilde{F}(\tilde{\mu}_G)(\bar{\lambda}_G + \bar{\eta}_G) - \tilde{F}(\tilde{\mu}_B)(\bar{\lambda}_B + \bar{\eta}_B)\big)\right]}{p\tilde{F}(\tilde{\mu}_G) + (1-p)\tilde{F}(\tilde{\mu}_B) - dt\left[p\tilde{F}(\tilde{\mu}_G)(\bar{\lambda}_G + \bar{\eta}_G) + (1-p)\tilde{F}(\tilde{\mu}_B)(\bar{\lambda}_B + \bar{\eta}_B)\right]} \tag{11}$$

where $\tilde{F}(x) = \exp\left(\sqrt{\alpha}\, x\, d\pi^1 - 1/2\alpha\, x^2\, dt\right)$. Similar to Bolton & Harris (1999), one can show by using Taylor expansions that $\tilde{F}(x) = 1 + \sqrt{\alpha}\, x\, d\tilde{\pi} + o(dt)$, where by $o(x)$ we denote any function $f(x)$ such that $\lim_{x\to 0} \frac{f(x)}{x} = 0$. Substituting this into (11), we obtain, after some manipulation,

$$dp = \frac{p(1-p)\left(\sqrt{\alpha}(\tilde{\mu}_G - \tilde{\mu}_B)d\tilde{\pi} - (\bar{\lambda}_G + \bar{\eta}_G - \bar{\lambda}_B - \bar{\eta}_B)dt\right)}{1 + \sqrt{\alpha}\,\mathbb{E}_p[\tilde{\mu}_\theta]d\tilde{\pi} - \mathbb{E}_p[\bar{\lambda}_\theta + \bar{\eta}_\theta]dt}, \tag{12}$$

where we drop all terms of order $dt^{\frac{3}{2}}$ or higher. Also, it can be checked that

$$\frac{1}{1 + \sqrt{\alpha}\,\mathbb{E}_p[\tilde{\mu}_\theta]\, d\tilde{\pi} - \mathbb{E}_p[\bar{\lambda}_\theta + \bar{\eta}_\theta]dt} = 1 - \sqrt{\alpha}\,\mathbb{E}_p[\tilde{\mu}_\theta]\, d\tilde{\pi} + \mathbb{E}_p[\bar{\lambda}_\theta + \bar{\eta}_\theta]dt + o(dt).$$

Substituting this back into (12), we have

$$\begin{aligned} dp &= p(1-p)(\tilde{\mu}_G - \tilde{\mu}_B)(\sqrt{\alpha}d\tilde{\pi} - \alpha\mathbb{E}_p[\tilde{\mu}_\theta]dt) - p(1-p)(\bar{\lambda}_G + \bar{\eta}_G - \bar{\lambda}_B - \bar{\eta}_B)dt + o(dt) \\ &= p(1-p)\frac{\mu_G - \mu_B}{\sigma}\sqrt{\alpha}dZ - \alpha p(1-p)(\lambda_G + \eta_G - \lambda_B - \eta_B)dt + o(dt), \end{aligned}$$

and by identifying the mean and the variance, we reach the desired result. □

Proof of Lemma 2. Using notation similar to that in Lemma 1 and applying Bayes' rule, we have:

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

Using again the Taylor series expansion $\tilde{F}(\tilde{\mu}) = 1 + \sqrt{\alpha}\,\tilde{\mu}\, d\tilde{\pi} + o(dt)$, we have

$$\begin{aligned} dp &= \frac{p(1-p)\left(\bar{\lambda}_G - \bar{\lambda}_B + \sqrt{\alpha}\, d\tilde{\pi}\,(\tilde{\mu}_G\bar{\lambda}_G - \tilde{\mu}_B\bar{\lambda}_B) - dt(\bar{\lambda}_G\bar{\eta}_G - \bar{\lambda}_B\bar{\eta}_B)\right)}{\bar{\lambda}(p) + \sqrt{\alpha}d\tilde{\pi}(p\tilde{\mu}_G\bar{\lambda}_G + (1-p)\tilde{\mu}_B\bar{\lambda}_B) - dt(p\bar{\lambda}_G\bar{\eta}_G + (1-p)\bar{\lambda}_B\bar{\eta}_B)} \\ &= \frac{p(1-p)(\bar{\lambda}_G - \bar{\lambda}_B)}{\mathbb{E}_p[\bar{\lambda}_\theta]} + \frac{\alpha p(1-p)\bar{\lambda}_G\bar{\lambda}_B\big(\eta_B - \eta_G + (\tilde{\mu}_G - \tilde{\mu}_B)\tilde{\mu}(p)\big)}{(\mathbb{E}_p[\bar{\lambda}_\theta])^2}dt \\ &\quad + p(1-p)\bar{\lambda}_G\bar{\lambda}_B\sqrt{\alpha}(\tilde{\mu}_G - \tilde{\mu}_B)dZ. \end{aligned}$$

The final expression is normally distributed, with a mean and variance as in our result. □

Proof of Theorem 1. Since our bandit model is a special case of the Lévy bandits in Kaspi & Mandelbaum (1995), the optimal policy is a threshold policy. More precisely, there exists a threshold p^* such that the optimal allocation function $\alpha_t^*(p_t)$ is such that $\alpha_t^*(p_t) = 1$ for $p_t \geq p^*$ and equal to zero otherwise. We seek to determine this optimal threshold p^*.

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):

$$r\,u(p) = \max_{\alpha} \Big[\mathbb{E}_p[\bar{\mu}_\theta] - \mathbb{E}_p[\bar{\lambda}_\theta]\,D + \mathbb{E}_p[\bar{\eta}_\theta]\,V - \mathbb{E}_p[\bar{\lambda}_\theta + \bar{\eta}_\theta]\,u(p) + \mathbb{E}_p[\bar{\lambda}_\theta]\,u\big(j(\alpha,p)\big) + \alpha\,p(1-p)\,(\lambda_B + \eta_B - \lambda_G - \eta_G)\,u'(p) + \frac{1}{2}\,\alpha\,\phi(p)\,u''(p) \Big],$$

where $\phi(p) \stackrel{\text{def}}{=} \left[\frac{p(1-p)(\mu_G - \mu_B)}{\sigma}\right]^2$ and $j(\alpha, p) \stackrel{\text{def}}{=} p\bar{\lambda}_G / \mathbb{E}_p[\bar{\lambda}_\theta]$.

Consider a value of p such that $p \geq p^*$ and $j(1,p) < p^*$. The corresponding optimal actions are $\alpha_t^*(p) = 1$ and $\alpha_t^*(j(1,p)) = 0$. Since the value from using the safe arm until the stopping event is $u(j(1,p)) = A_0 \stackrel{\text{def}}{=} \frac{\mu_0 - D\lambda_0 + \eta_0 V}{r + \eta_0}$, the Bellman recursion for $u(p)$ becomes:

$$r\,u(p) = \mathbb{E}_p[\mu_\theta] - \mathbb{E}_p[\lambda_\theta]\,D + \mathbb{E}_p[\eta_\theta]\,V - \mathbb{E}_p[\lambda_\theta + \eta_\theta]\,u(p) + \mathbb{E}_p[\lambda_\theta]\,A_0 + p(1-p)\,(\lambda_B + \eta_B - \lambda_G - \eta_G)\,u'(p) + \frac{1}{2}\,\phi(p)\,u''(p) \Big],$$

A particular solution of this equation is given by

$$u_{\text{part}}(p) = pK_G + (1-p)K_B, \qquad \text{where } K_\theta \stackrel{\text{def}}{=} \frac{\mu_\theta - \lambda_\theta D + \eta_\theta V + \lambda_\theta A_0}{r + \lambda_\theta + \eta_\theta},\ \forall\,\theta \in \{G, B\}.$$

For the homogeneous solution to this equation, we use $u_{\text{hom}}(p) = (1-p)\left(\frac{1-p}{p}\right)^\nu$ for some fixed ν. Replacing this in the differential equation, we obtain the following quadratic equation for ν:

$$(\mu_G - \mu_B)^4 \nu(1+\nu) - 2\sigma^2(\lambda_B + \eta_B - \lambda_G - \eta_G)\nu - 2\sigma^2(r + \eta_B + \lambda_B) = 0.$$

which has the solutions:

$$\nu_{1,2} = -\frac{1}{2} + \frac{\sigma^2(\lambda_B + \eta_B - \lambda_G - \eta_G)}{(\mu_G - \mu_B)^4} \pm \frac{\sqrt{\big((\mu_G - \mu_B)^4 - 2\sigma^2(\lambda_B + \eta_B - \lambda_G - \eta_G)\big)^2 + 8\sigma^2(r + \eta_B + \lambda_B)(\mu_G - \mu_B)^4}}{2(\mu_G - \mu_B)^4}.$$

With ν^* denoting the positive root (corresponding to the plus sign), we look for $u(p)$ of the form

$$u(p) = u_{\text{part}}(p) + C(1-p)\left(\frac{1-p}{p}\right)^{\nu^*}.$$

Since $\mu_G \neq \mu_B$, the function $u(p)$ satisfies the value matching and smooth pasting conditions at the boundary $p = p^*$ (also see Cohen & Solan (2013) and Keller & Rady (2015)). Thus, we look for C and p^* so that $u(p^*) = A_0$ and $u'(p^*) = 0$, respectively. This provides a system of two equations, which can be solved for p^* and C. We thus find:

$$p^* = \frac{\nu^*(A_0 - K_B)}{\nu^*(A_0 - K_B) + (1+\nu^*)(K_G - A_0)}, \qquad C = \frac{(p^*)^{1+\nu^*}(K_G - K_B)}{(1-p^*)^{\nu^*}(p^* + \nu^*)}.$$

By rewriting the expression for p^* in terms of A_0, A_G, A_B (as defined in Assumption 1), we readily arrive at the desired result. □

Lemma 5. *Under the premises of Theorem 1, the optimal value function $u(p)$ satisfies:*

$$\begin{aligned} r\,u(p) = \max_{\alpha} \Big[&\mathbb{E}_p[\bar{\mu}_\theta] - D\,\mathbb{E}_p[\bar{\lambda}_\theta] + \mathbb{E}_p[V_\theta\,\bar{\eta}_\theta] - \mathbb{E}_p[\bar{\lambda}_\theta + \bar{\eta}_\theta]\,u(p) + \mathbb{E}_p\big[\bar{\lambda}_\theta\,u\big(j(\alpha,p)\big)\big] \\ &+ \alpha\,p(1-p)\,(\lambda_B + \eta_B - \lambda_G - \eta_G)\,u'(p) + \frac{1}{2}\,\alpha\,\phi(p)\,u''(p)\Big], \end{aligned}$$

where $\phi(p) \overset{\text{def}}{=} \left[\frac{p(1-p)(\mu_G - \mu_B)}{\sigma}\right]^2$ *and* $j(\alpha, p) \overset{\text{def}}{=} p\bar{\lambda}_G / \mathbb{E}[\bar{\lambda}_\theta]$.

Proof of Lemma 5. Let Π_t denote the DM's total rewards from t onwards, and $\mathcal{L}$ (respectively, $\mathcal{S}$) denote the occurrence of a life (respectively, stopping) event during period $[t, t+dt)$, with $\mathcal{L}^c$ ($\mathcal{S}^c$) denoting the complementary event. The value function satisfies the following Bellman equation:

$$u(p) = \max_{\alpha} \Big[\ \mathbb{E}\big[\Pi_t \,|\, \mathcal{L}^c, \mathcal{S}^c\big]\,\mathbb{P}[\mathcal{L}^c, \mathcal{S}^c] + \mathbb{E}\big[\Pi_t \,|\, \mathcal{L}, \mathcal{S}^c\big]\,\mathbb{P}[\mathcal{L}, \mathcal{S}^c] + \mathbb{E}\big[\Pi_t \,|\, \mathcal{S}\big]\,\mathbb{P}[\mathcal{S}]\ \Big], \tag{13}$$

where all the expectations are taken with respect to the filtration $\mathcal{F}_t$, and we omit subscript t for simplicity. In view of our standing assumptions, we have:

$$\mathbb{P}[\mathcal{L}^c, \mathcal{S}^c] = \mathbb{E}_p[e^{-(\bar{\lambda}_\theta + \bar{\eta}_\theta)dt}] = 1 - \mathbb{E}_p[\bar{\lambda}_\theta + \bar{\eta}_\theta]\,dt + o(dt), \tag{14a}$$

$$\mathbb{E}\big[\Pi_t \,|\, \mathcal{L}^c, \mathcal{S}^c\big] = \mathbb{E}_p[\bar{\mu}_\theta]\,dt + e^{-r\,dt}\,\mathbb{E}\big[u(p+dp) \big| \mathcal{L}^c, \mathcal{S}^c\big], \tag{14b}$$

$$\mathbb{P}[\mathcal{L}, \mathcal{S}^c] = \mathbb{E}_p[(1 - e^{-\bar{\lambda}_\theta dt})e^{-\bar{\eta}_\theta dt}] = \mathbb{E}_p[\bar{\lambda}_\theta]\,dt + o(dt), \tag{14c}$$

$$\mathbb{E}\big[\Pi_t \,|\, \mathcal{L}, \mathcal{S}^c\big] = -D + \mathbb{E}_p[\bar{\mu}_\theta]\,dt + e^{-r\,dt}\,\mathbb{E}\big[u(p+dp)\,|\,\mathcal{L}, \mathcal{S}^c\big] \tag{14d}$$

$$\mathbb{P}[\mathcal{S}] = \mathbb{E}_p[1 - e^{-\bar{\eta}_\theta dt}] = \mathbb{E}_p[\bar{\eta}_\theta]\,dt + o(dt), \tag{14e}$$

$$\mathbb{E}\big[\Pi_t \,|\, \mathcal{S}\big] = V + \mathbb{E}_p[\bar{\mu}_\theta]\,dt. \tag{14f}$$

By expanding the term $u(p+dp)$ in (14b) in a Taylor series around p, and using Lemma 1 to replace the mean and second moment of dp, we obtain:

$$\begin{aligned} \mathbb{E}\big[u(p+dp)\,|\,\mathcal{L}^c, \mathcal{S}^c\big] &= u(p) + u'(p)\,\mathbb{E}\big[dp\,|\,\mathcal{L}^c, \mathcal{S}^c\big] + \frac{1}{2}\,u''(p)\,\mathbb{E}\big[dp^2\,|\,\mathcal{L}^c, \mathcal{S}^c\big] + o(dt) \\ &= u(p) + u'(p)\,\alpha\,p(1-p)\,(\lambda_B + \eta_B - \lambda_G - \eta_G)dt + \frac{1}{2}\,u''(p)\,\alpha\,\phi(p)\,dt + o(dt). \end{aligned}$$

Similarly, by using Lemma 2 and expanding the term $u(p+dp)$ in (14d) in a Taylor series around $j(\alpha,p) \stackrel{\text{def}}{=} p + \alpha p(1-p)(\lambda_G - \lambda_B)/\mathbb{E}_p[\bar{\lambda}_\theta] = p\bar{\lambda}_G/\mathbb{E}_p[\bar{\lambda}_\theta]$, we have:

$$\begin{aligned}\mathbb{E}\big[u(p+dp)\,|\,\mathcal{L},\mathcal{S}^c\big] &= u\big(j(\alpha,p)\big) + u'\big(j(\alpha,p)\big)\,\mathbb{E}\big[dp\,|\,\mathcal{L}^c,\mathcal{S}^c\big] + \frac{1}{2}\,u''\big(j(\alpha,p)\big)\,\mathbb{E}\big[dp^2\,|\,\mathcal{L}^c,\mathcal{S}^c\big] + o(dp^2)\\ &= u\big(j(\alpha,p)\big) + u'\big(j(\alpha,p)\big)\,\alpha\,p(1-p)\,\bar{\lambda}_G\bar{\lambda}_B\frac{\eta_B - \eta_G + (\mu_G-\mu_B)\mathbb{E}_p[\mu_\theta]/\sigma^2}{(\mathbb{E}_p[\bar{\lambda}_\theta])^2}dt\\ &\quad + \frac{1}{2}\,u''\big(j(\alpha,p)\big)\,\alpha\left(\frac{p(1-p)\bar{\lambda}_G\bar{\lambda}_B(\mu_G-\mu_B)}{\sigma}\right)^2 dt + o(dt).\end{aligned}$$

Substituting these expressions together with (14a)-(14f) into (13), we finally obtain:

$$\begin{aligned}u(p) = \max_{\alpha}\Big[&\mathbb{E}_p[\bar{\mu}_\theta]\,dt + u(p) + u'(p)\,\alpha\,p(1-p)\,(\lambda_B + \eta_B - \lambda_G - \eta_G)dt + \frac{1}{2}\,u''(p)\,\alpha\,\phi(p)\,dt\\ &- ru(p)\,dt - \mathbb{E}_p[\bar{\lambda}_\theta + \bar{\eta}_\theta]\,u(p)\,dt - \mathbb{E}_p[\bar{\lambda}_\theta]\,D\,dt + u\big(j(\alpha,p)\big)\mathbb{E}_p[\bar{\lambda}_\theta]\,dt + \mathbb{E}_p[\bar{\eta}_\theta]\,V\,dt + o(dt)\Big].\end{aligned}$$

By canceling $u(p)$ on both sides, dividing by dt and taking the limit $dt \to 0$, we obtain the result. □

Proof of Theorem 2. Note that the representation result given in expression (8) implies that the Gittins index of an arm is independent of the other arms, and is only determined by the intrinsic value of continuing to play that arm compared against retiring to earn a deterministic reward.

Thus, we focus on the Gittins index for a given arm i in our model, having a prior with value $p_t^i \equiv p$ at time t. The problem of optimally choosing when to stop using this arm and switch to a retirement reward (received indefinitely thereafter) exactly corresponds to a special instance of our base-case model, namely when $\eta_0, \eta_B, \eta_G = 0$. In particular, by Theorem 1, the optimal policy is "bang-bang," and exactly corresponds to (optimally) stopping the use of the risky arm and switching to the safe arm, to earn a "retirement reward" given by the latter's expected discounted rewards, i.e., $\frac{\mu_0 - D\lambda_0}{r}$.

With this equivalence, the arguments in the proof of Theorem 1 become directly applicable. More precisely, assuming the deterministic "retirement" reward from infinitely using the safe arm is given by a generic value m (instead of $\frac{\mu_0 - D\lambda_0}{r}$), the optimal policy for playing the i-th risky arm is bang-bang, characterized by a threshold $p_i^*(m)$. Furthermore, the differential equation for the value function $u_i(p,m)$ in the region of beliefs $\big(p_i^*(m), p_i^*(m)+\varepsilon\big]$ for small enough $\varepsilon > 0$ becomes:

$$p\,\mu_{G_i} + (1-p)\,\mu_{B_i} - D\lambda - r\,u_i(p,m) + \frac{1}{2}\Big(\frac{p(1-p)(\mu_{G_i}-\mu_{B_i})}{\sigma}\Big)^2 u_i''(p,m) = 0.$$

It can be verified that a particular solution to this ODE is given by $\frac{p\,\mu_{G_i} + (1-p)\,\mu_{B_i} - D\lambda}{r}$, while the homogenous solution is given by $(1-p)\Big(\frac{1-p}{p}\Big)^{\nu_i^*}$, where

$$\nu_i^* \stackrel{\text{def}}{=} \frac{-(\mu_{G_i}-\mu_{B_i}) + \sqrt{(\mu_{G_i}-\mu_{B_i})^2 + 8r\sigma^2}}{2(\mu_{G_i}-\mu_{B_i})}.$$

Imposing the value-matching and smooth-pasting conditions at $p_i^*(m)$, we obtain:

$$p_i^*(m) = \frac{\nu_i^*\big[r \cdot m - (\mu_{B_i} - D\lambda_i)\big]}{\mu_{G_i} - D\lambda_i - r \cdot m + \nu_i^*(\mu_{G_i} - \mu_{B_i})}$$

$$u_i^*(p, m) = \begin{cases} m, & \text{if } p < p_i^*(m) \\ f_i(p, m), & \text{if } p \geq p_i^*(m), \end{cases}$$

where $f_i(p, m)$ is given by (9b). Thus, from the representation result in (8), we immediately have $\mathcal{G}_t^i = \inf\big\{m \in \mathbb{R} \,:\, m \geq u_i(p, m)\big\}$, which yields (9a). The convexity of $f_i(p, m)$ in m follows since

$$\frac{\partial^2 f_i}{\partial m^2} = B_i(p)\,\frac{\mu_{G_i} - \mu_{B_i}}{r}\,\frac{\nu_i^*\left(\frac{\nu_i^*(\phi_0 - \phi_{B_i})}{(1+\nu_i^*)(\phi_{G_i} - \phi_0)}\right)^{\nu_i^*}(\phi_{G_i} - \phi_{B_i})}{(\phi_0 - \phi_{B_i})(\phi_{G_i} - \phi_0)^2},$$

where $\phi_\xi \overset{\text{def}}{=} \mu_\xi - D\lambda_\xi$, $\forall\, \xi \in \{0, G_i, B_i\}$. Specifically, by Assumption 2, we have $\phi_{B_i} \leq \phi_0 \leq \phi_{G_i}$, so that all the terms above are positive, establishing that $\frac{\partial^2 f_i}{\partial m^2} \geq 0$. □