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%.

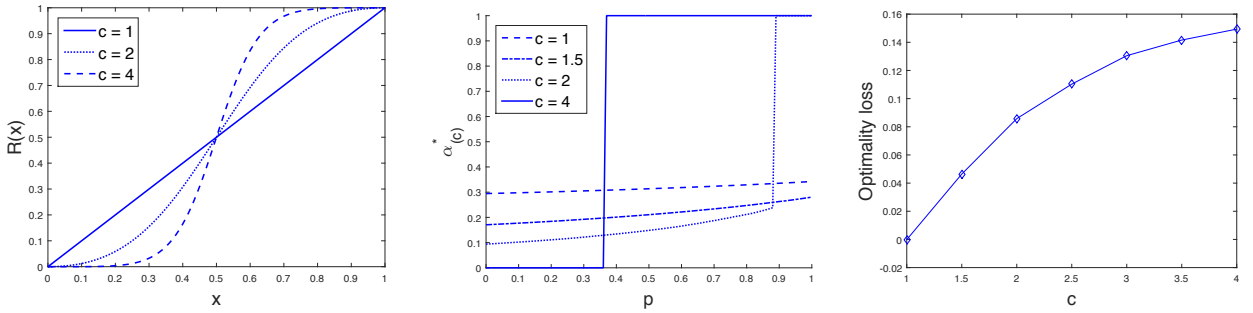
D Impact of Nonlinear Dose Response

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

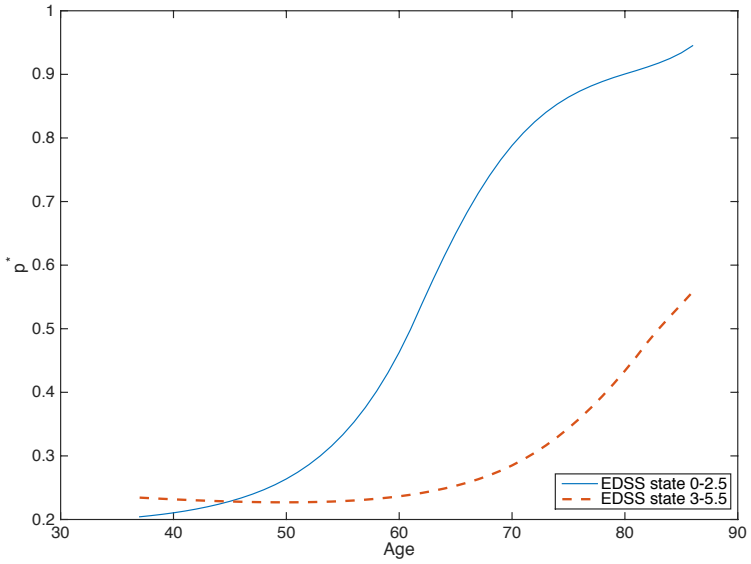


Figure 9: Optimal policy for a WTP = \$800,000/QALY. The plots correspond to the threshold values where treatment should be switched in each EDSS state, as a function of the patient's age.

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

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 $u(j(1,p)) = A_0 \overset{\text{def}}{=} \frac{\mu_0 - D\lambda_0 + \eta_0 V}{r + \eta_0}$, 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

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

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 $\frac{\partial^2 f_i}{\partial m^2} \geq 0$. □

References

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

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

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

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. (2015), 'Breakdowns', *Theoretical Economics* **10**, 175–202.

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