We thank the reviewers for their excellent feedback and for recognizing the value of our work to the NeurIPS audience.

R1, R2: Realism of the causal graph. In our model, the outcome Y_t is determined not only by A_t, but also by the unobserved moderating variable Z and possibly exogenous variables. Z may capture, for example, the disease state of a subject and introduce *correlations* between outcomes Y_s and Y_t for $s < t$. It is a common working assumption that previous actions and outcomes, A_s, Y_s for $s < t$, do not have direct *causal* effect on Y_t when Y_t represents the symptoms of chronic conditions where drugs do not affect the underlying disease state; given sufficient time between treatments, symptoms return to a baseline level until another treatment is started. In addition to rheumatoid arthritis [1], which is used as motivation in the manuscript, examples include depression [2] and Parkinson's disease [3].

R2: Usefulness of causal framework. The causal framework is necessary for distinguishing the observational outcome Y_t from the potential outcome $Y(a)$. Only under certain assumptions can $Y(a)$ be estimated from Y_t. Sufficient conditions for this are given in Section 4. The distinction is particularly important in the case where $Y(a)$ is not fully identifiable from observational data due to unobserved confounding, as discussed in Section 4.2.

R2: Observed and counterfactual outcomes. The relationship between counterfactual outcomes $Y(a)$ and observed outcomes Y_t is established in Theorem 1. The LHS of (4), $\rho(h_s)$, is a function of counterfactual outcomes, as defined in (2), and the RHS is a function of observed outcomes Y_t. As remarked after Theorem 1, under our assumptions, a model of $p(Y_t \mid H_t = h, A_t = a)$ is sufficient for estimating the distribution of $Y(a)$. To solve our policy optimization problem (1), it is not necessary to impute all counterfactuals. For example, in the binary case given by R2, if an action $a = 0$ has been tried, and $Y(0)$ observed, only the probability that $Y(1) > Y(0)$ is required to solve the problem.

R2: Relation to model-free RL. R2 is correct that the model-free method NDP compares similarly to the other methods in the antibiotics experiment. However, we show in Figure 1b that the qualitative behavior of NDP as a function of dataset size is very different from that of CDP and CG. Additionally, as shown in Appendix A.7 (Thm A7), NDP is suboptimal in the general case. We hope that these contributions are recognized. By a "transparent" tradeoff, we refer to the meaning of the parameters δ (CDP, CG) and λ (NDP). δ is directly interpretable as a probability threshold at which we are satisfied with the best-so-far treatment (as used in the antibiotics experiment). The value of λ does not have an immediate interpretation as a level of certainty of near-optimality—*the tradeoff for a fixed λ varies across datasets*.

R3, R4: Comparison with experts and the emulated expert. R3 is correct that it is feasible in practice to include more information in the policy so that it compares more favorably to experts in the first step. Similarly, the accuracy of the emulated doctor, remarked upon by R4, could be improved by using more information to emulate the doctor policy. In our experiment, we intentionally kept the patient representation small because the number of samples was fairly limited. The emulated expert in our study attempts to approximate the expert's policy with the same information given to the other algorithms. As such, it serves as an *imitation learning* baseline to complement the policy optimization approaches developed in this work. We will clarify this choice in the paper. We note, however, that our algorithm is trying to achieve a different goal than the expert. While experts may attempt to prescribe the best action on the first try, we try to minimize the expected number of tested treatments. In that sense, the expert can be thought of as a greedy agent, which our paper argues is not always optimal, if we're trying to minimize needless trial and error on patients.

R4: Impact of this work. We certainly agree that our method is not suitable for all applications. However, there is a large class of medical conditions and treatments which fall exactly under the specifications of our model. In fact, our motivation for this work is the result of working with active clinicians treating rheumatoid arthritis. The application of our method for this purpose is an ongoing project which will be aimed at the clinical research community, but we appreciate the reviewer's feedback which highlighted that the use cases of our method are not readily apparent from reading the paper. We will add a description of the RA problem to the paper to make it more concrete and demonstrate a motivating use case, along with a discussion of other uses such as in treating psychiatric disorders. Finally, we believe that the problem has applications also outside of medicine, such as for general recommendation systems.

R4: Short-term response. It is true that short-term response is critical for some applications and should not be discounted; this is a potential challenge also for reinforcement learning which optimizes long-term return, possibly sacrificing immediate rewards. Our goal is to find a near-optimal treatment in as few steps as possible which is an important consideration in other applications [2]. If an optimal treatment *can* be reliably identified in a single step, the algorithm is incentivized to do so. Short-term success is sacrificed only if there is great uncertainty about which treatment is likely to work and this can be reduced by a sub-optimal treatment. Our greedy approximation incentivizes short-term response by preferring actions higher that are likely to have a higher outcome (see 5.2). Such incentives could be incorporated into the dynamic programming solution as well, and is an interesting direction for future work.

[1] J. R. O'Dell. Therapeutic strategies for rheumatoid arthritis. *New England Journal of Medicine*, 350(25):2591–2602, 2004.

[2] M. F. Pradier, T. H. McCoy Jr, M. Hughes, R. H. Perlis, and F. Doshi-Velez. Predicting treatment dropout after antidepressant initiation. *Translational psychiatry*, 10(1):1–8, 2020.

[3] M. Stacy, A. Bowron, M. Guttman, R. Hauser, K. Hughes, J. P. Larsen, P. LeWitt, W. Oertel, N. Quinn, K. Sethi, et al. Identification of motor and nonmotor wearing-off in parkinson's disease: comparison of a patient questionnaire versus a clinician assessment. *Movement disorders: official journal of the Movement Disorder Society*, 20(6):726–733, 2005.