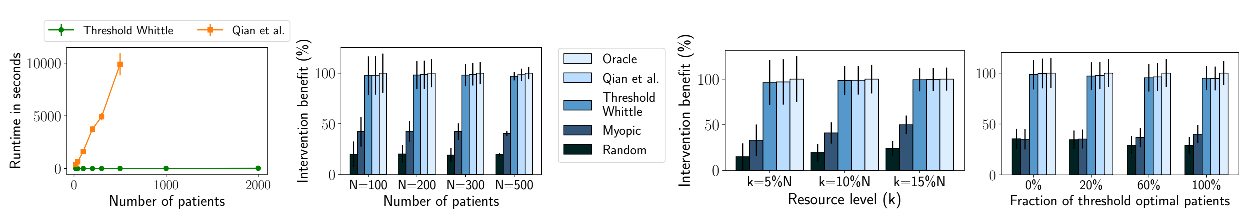
We sincerely thank the reviewers for their thoughtful and constructive feedback. We address specific questions below.

[R1] (1) Assumption that workers learn states only when acting. This is the scenario today for health workers in Mumbai, India managing tuberculosis (TB) patients (this paper's direct motivation). A single worker monitors adherence and delivers basic care to large cohorts of geographically distributed patients via person-to-person phone calls; offline monitoring is unavailable. **(2) Determining arms to pull:** The Whittle index policy, defined by Whittle [35], pulls the k arms with the largest Whittle indices. We will make this explicit. **(3) Forward vs. reverse threshold policies in simulations:** The majority of patients have forward threshold optimality, which we rely on in the simulations. We will add details to the appendix. **(4) Comparison with Qian et al.:** For the optimality guarantees of the Whittle index to hold for our algorithm, the process must satisfy the conditions of Thms. 1 and 2. However, the real world data has a small fraction of patients who violate the condition of Thm. 2, resulting in the small gap in performance.

[R2] (1) Extending from 2 to M states: The 2-state model is well-established in literature (Gilbert-Elliot model, 1960) and is popularly studied (e.g. seminal work of Liu and Zhao [19] that we extend) because of its wide range of applications such as, to healthcare, anti-poaching, sensor maintenance, etc. Despite the wide applicability of this model, generalizing to an M-state model will make for interesting future work. **(2) Future work:** We will add avenues of future work to the camera-ready version. **(3) Link to combinatorial bandits:** Since RMABs also admit $\binom{N}{k}$ feasible actions per round, this connection seems natural. However, in an RMAB, rewards on each sub-arm are state-dependent. This would render existing combinatorial bandit algorithms – which maximize mean reward – sub-optimal in general.

[R3] (1) Complexity: Our work improves on the computational complexity of Qian et al., which has complexity per round of $\mathcal{O}(Nlog(\frac{1}{\epsilon})(|\mathcal{S}|T)^{2+\frac{1}{18}})$. Our algorithm has a one-time cost of $\mathcal{O}(|\mathcal{S}|^2T)$ to precompute the Whittle indices for all rounds, then has a per round cost of only $\mathcal{O}(Nmin\{k, log(N)\})$ to retrieve the top k indices. We will make this more explicit. **(2) Comparison with Qian et al.** Please see R1.(4). **(3) Indexability:** The guarantee that holds under indexability is the asymptotic optimality of the Whittle index policy as proven by Weber and Weiss (1990) [33] referenced on lines 39–40 of our paper. We will make this more explicit. **(4) Theorem conditions:** Thms. 2 and 3 give conditions under which the structure required for Thm. 1 is theoretically guaranteed. Following are two examples of processes for which conditions of Thm. 2 and Thm. 3 hold respectively (fwd: $P^a_{11} = 0.95, P^a_{01} = 0.9, P^p_{11} = 0.9, P^p_{01} = 0.4, \beta = 0.9$; rev: $P^a_{11} = 0.95, P^a_{01} = 0.4, P^p_{11} = 0.4, P^p_{01} = 0.35, \beta = 0.9$). Since these are sufficient but not necessary conditions, nothing can be concluded when neither is satisfied. However, we find from brute force checks that most processes, even those that violate condition of Thm.2. are either forward or reverse threshold optimal. **(5) Assuming P is known:** This is realistic in many settings, as P can be estimated from historical data collected either before or in early stages of planning. E.g., in the TB domain mentioned in R1.(1), this data is gathered from health workers' early round robin calling of patients. Further, since the offline planning portion of restless bandits is already PSPACE hard in general, it is often studied separately from the online version (Liu and Zhao [19]; Meshram et al. [21]). Additionally, since the optimal policy cannot be computed in general, regret bounds for general online restless bandit algorithms are typically defined with respect to an arbitrary reference policy with full information, rather than with respect to the optimal policy (e.g., Jung and Tewari [13]). This provides at least three reasons why developing strong algorithms for the version of the problem with known P is of significant interest. **(6) Empirical methodology:** We have updated our figures with confidence bounds (see Fig. 1 below). We have updated Fig. 5(d) of the main text to include 0% threshold optimal patients (Fig. 1(c) below); our algorithm shows strong performance. **(7) Preprocessing:** For the experiments derived from real-world data, preprocessing only involved imputing missing action information to align with natural constraint structure common in analogous domains (see lines 111–116). Further, sensitivity analysis in Appendix G confirms our conclusions for a wide range of imputations. We will clarify this in the final paper. **(8) Intervention benefit** Please see R4.(1). **(9) Link to combinatorial bandits** Please see R2.(3).

[R4] (1) Intervention benefit (described in text on Line 259) is calculated as: $I.B.(ALG) = \frac{\overline{R}^{ALG} - \overline{R}^{\text{No intervention}}}{\overline{R}^{\text{Oracle}} - \overline{R}^{\text{No intervention}}}$ where $\overline{R}$ is the average reward of the algorithm as defined on Line 70. **(2) Error bars** Updated Figs. with error bars are below.



(a) (b) (c)

Figure 1: Error bars show difference in performance between our algorithm and Qian et al. is not statistically significant.