Restless Bandits in the Field: Real-World Study for Improving Maternal and Child Health Outcomes

Anonymous Author(s)
Affiliation
Address
email

Abstract

The widespread availability of cell phones has enabled non-profits to deliver critical health information to their beneficiaries in a timely manner. This paper describes our work in assisting non-profits employing automated messaging programs to deliver timely preventive care information to new and expecting mothers during pregnancy and after delivery. Unfortunately, a key challenge in such information delivery programs is that a significant fraction of beneficiaries tend to drop out. Yet, non-profits often have limited health-worker resources (time) to place crucial service calls for live interaction with beneficiaries to prevent such engagement drops. To assist non-profits in optimizing this limited resource, we developed a Restless Multi-Armed Bandits (RMABs) system. One key technical contribution in this system is a novel clustering method of offline historical data to infer unknown RMAB parameters. Our second major contribution is evaluation of our RMAB system in collaboration with an NGO, via a real-world service quality improvement study. The study compared strategies for optimizing service calls to 23003 participants over a period of 7 weeks to reduce engagement drops. We show that the RMAB group provides statistically significant improvement over other comparison groups, reducing $\sim 30\%$ engagement drops. To the best of our knowledge, this is the first study demonstrating the utility of RMABs in real world public health settings. We are transitioning our system to the NGO for real-world use.

1 Introduction

The ubiquity of cell phones has allowed non-profits to deliver targeted health information via voice or text messages to beneficiaries in underserved communities, often with significant demonstrated benefits to those communities [15, 22]. We focus in particular on non-profits that target improving maternal and infant health in low-resource communities in the global south. These non-profits deliver ante- and post-natal care information via voice and text to prevent adverse health outcomes [2, 12, 13].

Unfortunately, such information delivery programs are often faced with a key shortcoming: a large fraction of beneficiaries who enroll may drop out or reduce engagement with the information program. Yet non-profits often have limited health-worker time available on a periodic (weekly) basis to help prevent engagement drops. More specifically, there is limited availability of health-worker time where they can place crucial service calls (phone calls) to a limited number of beneficiaries, to encourage beneficiaries' participation, address complaints and thus prevent engagement drops.

Optimizing limited health worker resources to prevent engagement drops requires that we prioritize beneficiaries who would benefit most from service calls on a periodic (e.g., weekly) basis. We model this resource optimization problem using Restless Multi-Armed Bandits (RMABs), with each beneficiary modeled as an RMAB arm. RMABs have been well studied for allocation of limited resources motivated by a myriad of application domains including preventive interventions for

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

healthcare [21], planning anti-poaching patrols [25], machine repair and sensor maintenance [11] and communication systems [27]. However, RMABs have rarely seen real world deployment, and to the best of our knowledge, never been deployed in the context of large-scale public health applications.

This paper presents first results of an RMAB system in real world public health settings. Based on available health worker time, RMABs choose m out of N total beneficiaries on a periodic (e.g., weekly) basis for service calls, where the m are chosen to optimize prevention of engagement drops. The paper presents two main contributions. First, previous work often assumes RMAB parameters as either known or easily learned over long periods of deployment. We show that both assumptions do not hold in our real-world contexts; instead, we present clustering of offline historical data as a novel approach to infer unknown RMAB parameters.

Our second contribution is a real world evaluation showing the benefit of our system, conducted in partnership with an Indian NGO, **ABC** (real name withheld for anonymity), focused on maternal and child care. **ABC** conducts a large-scale health information program, with concrete evidence of health benefits, which has so far served over a million mothers. In this program, an automated voice message is delivered to an expecting or new mother (beneficiary) over cell phone on a weekly basis throughout pregnancy and for a year post birth in a language and time slot of her preference.

Unfortunately, **ABC**'s information delivery program also suffers from engagement drops. Therefore, in collaboration with **ABC** we conducted a service quality improvement study to maximize the effectiveness of their service calls to ensure beneficiaries do not drop off from the program or stop listening to weekly voice messages. More specifically, the current standard of care in **ABC**'s program is that any beneficiary may initiate a service call by placing a so called "missed call". This beneficiary-initiated service call is intended to help address beneficiaries' complaints and requests, thus encouraging engagement. However, given the overall decreasing engagement numbers in the current setup, key questions for our study are to investigate an approach for effectively conducting additional **ABC**-initiated service calls (these are limited in number) to reduce engagement drops. To that end, our service quality improvement study comprised of 23,003 real-world beneficiaries spanning 7 weeks. Beneficiaries were divided into 3 groups, each adding to the current standard of care. The first group exercised **ABC**'s current standard of care (CSOC) without additional **ABC**-initiated calls. In the second, the RMAB group, **ABC** staff added to the CSOC by initiating service calls to 225 beneficiaries on average per week chosen by RMAB. The third was the Round-Robin group, where the exact same number of beneficiaries as the RMAB group were called every week based on a systematic sequential basis.

Results from our study demonstrate that RMAB provides statistically significant improvement over CSOC and round-robin groups. This improvement is also practically significant — the RMAB group achieves a $\sim 30\%$ reduction in engagement drops over the other groups. Moreover, the round-robin group does not achieve statistically significant improvement over the CSOC group, i.e., RMAB's optimization of service calls is crucial. To the best of our knowledge, this is the first large-scale empirical validation of use of RMABs in a public health context. Based on these results, the RMAB system is currently being transitioned to **ABC** to optimize service calls to their ever growing set of beneficiaries. Additionally, this methodology can be useful in assisting engagement in many other awareness or adherence programs, e.g., Chen et al. [5], Thirumurthy and Lester [29].

2 Problem Statement

We assume the planner has access to an offline historical data set of beneficiaries, $\mathcal{D}_{train}$. For each beneficiary i, $\mathcal{D}_{train}[i]$ consists of a tuple, $\langle f, \mathcal{E} \rangle$, where f corresponds to beneficiary i's static feature vector, and $\mathcal{E}$ is an episode storing the trajectory of (s, α, s') pairs for that beneficiary, where s denotes the start state, α the action taken (passive v/s active), and s' the next state that the beneficiary lands in after executing α in state s. We assume that these (s, α, s') samples are drawn according to fixed, latent transition matrices $P^a_{ss'}[i]$ and $P^p_{ss'}[i]$ (corresponding to the active and passive actions respectively), unknown to the planner, and potentially unique to each beneficiary.

Additionally, we have a different beneficiary cohort $\mathcal{D}_{test}$, with N beneficiaries, marked $\{1, \ldots, N\}$, that the planner must plan service calls for. The transition parameters corresponding to beneficiaries in $\mathcal{D}_{test}$ are unknown to the planner, but assumed to be drawn at random from a distribution similar to the joint distribution of features and transition parameters of beneficiaries in the historical data distribution. The planner has access to the feature vector f for beneficiaries in $\mathcal{D}_{test}$.

We now define the service call planning problem as follows. The planner has upto m resources available per round, which the planner may spend towards delivering service calls to beneficiaries. Beneficiaries are represented by N arms of the RMAB, of which the planner may pull upto m arms (i.e., m service calls) at each time step. We consider a round or timestep of one week which allows planning based on the most recent engagement patterns of the beneficiaries.

3 Methodology

We use clustering techniques that exploit historical data D_{train} to estimate an offline RMAB problem instance relying solely on the beneficiaries' static features and online state transition data. While there is limited service call data (active transition samples) for any single beneficiary, clustering on the beneficiaries allows us to combine their data to infer transition probabilities for the entire group, thus overcoming the challenge of limited samples (time-steps) per beneficiary. Clustering offers the added advantage of reducing computational cost for resource limited NGOs; since all beneficiaries within a cluster share identical transition probability values we can compute their Whittle index all at once. We test four such clustering methods, but adopt the PPF method (described below) in our final study, and defer the description of other methods to Appendix D. Finally, we adopt the Whittle solution approach for RMABs (detailed in Appendix E) to plan actions and pre-compute all of the possible $2k$ index values possible (corresponding to combinations of k possible clusters and 2 states).

Passive Transition-Probability based Clustering (PPF): The key motivation here is to group together beneficiaries with similar transition behaviors, irrespective of their features. To this end, we use k-means clustering on passive transition probabilities (to avoid issues with missing active data) of beneficiaries in D_{train} and identify cluster centers. We then learn a map ϕ from the feature vector f to the cluster assignment of the beneficiaries that can be used to infer the cluster assignments of new beneficiaries at test-time solely from f. We use a random forest model as ϕ.

4 Service Quality Improvement Study

For our quality improvement study, we consider the cohort of beneficiaries registered in the program between Feb 16, 2021 and March 15, 2021 (as D_{test}). The 23003 beneficiaries are randomly distributed across 3 groups, each group adding to the CSOC as follows:

Current-Standard-of-Care (CSOC) Group: The beneficiaries in this group follow the original standard of care, where there are no **ABC** initiated service calls. The listenership behavior of beneficiaries in this group is used as a benchmark for the RR and RMAB groups.

RMAB group: In this group, beneficiaries are selected for **ABC**-initiated service call per week via the Whittle Index policy described in Section B. Even though all beneficiaries within a cluster are modeled by identical MDP parameters, their states may evolve independently, and so the Whittle indices are tracked for each beneficiary separately, leading to an RMAB with 7668 arms.

Round Robin (RR) group: By default, NGOs including **ABC** often conduct service calls using some systematic set order – the idea here is to have an easily executable policy, that services enough of a cross-section of beneficiaries and can be scaled up or down per week based on available resources. To recreate this setting, we generate service calls to beneficiaries based on the ascending order of their date of enrollment for this RR group, as recommended by **ABC**. If this method succeeds compared to CSOC, then a simple manual strategy is enough; RMAB style optimization may not be needed.

We ensure absence of selection bias in our randomized group assignment via Analysis of Variance (ANOVA) test (see Appendix H). Beneficiaries across all three groups receive the same automated voice messages regarding pregnancy and post-birth care throughout the program, and no health related information is withheld from any beneficiary. The study only aims to evaluate the effectiveness of **ABC**-initiated outbound service calls with respect to improving engagement with the program across the three groups. No interviews or research data or feedback was collected from the beneficiaries.

4.1 Results and Statistical Analysis

We present our key results from the study in Figure 1. The results are computed at the end of 7 weeks from the start of the quality improvement study on April 26, 2021. Figure 1 measures the impact of service calls by the RMAB and RR policies in comparison to the CSOC Group. Beneficiaries' engagement with the program typically starts to dwindle with time.

In Figure 1, we measure the impact of a service call policy as the cumulative drop in engagement prevented compared to the CSOC Group (see Appendix J for details). We consider drop in engagement instead of the raw engagement numbers themselves, because of the slight difference in the numbers of beneficiaries in engaging (E) state at the start of the study. Figure 1 shows that the RMAB policy prevents a total 622 instances of a drop in automated health message engagement, at the end of 7 weeks, as compared to CSOC. RR group, on the other hand, only prevents 101 engagement drops by the end of week 7. Given that there are a total of 1944 engagement drops in the CSOC group, we show in Table 1, that the RMAB group has 32.0% and 28.3% less cumulative engagement drops as compared to the CSOC and RR groups respectively by the end of the study.

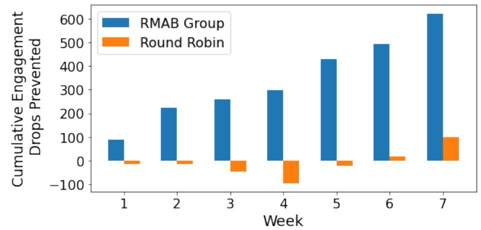


Figure 1: Cumulative weekly engagement drops prevented (compared to CSOC) by RMAB far exceed those prevented by RR.

To investigate the benefit from use of RMAB policy over policies in the RR and CSOC groups, we use regression analysis [1]. The results are summarized in Table 1. We find that RMAB has a statistically significant treatment effect in reducing cumulative engagement drop (negative $\beta, p < 0.05$) as compared to CSOC group. However, the treatment effect is not statistically significant when comparing RR with CSOC group ($p = 0.740$). Additionally, comparing RMAB group with RR, we find β, the RMAB treatment effect, to be significant ($p < 0.1$). This shows that RMAB policy has a statistically significant effect on reducing cumulative engagement drop as compared to both the RR policy and CSOC. RR fails to achieve statistical significance against CSOC. Together these results illustrate the importance of RMAB's optimization of service calls, and that without such optimization, service calls may not yield any benefits.

Table 1: Statistical significance for service call policy impact at week 7 is tested using a linear regression model. We use: $^*p < 0.05$; $^\dagger p < 0.1$

	RMAB vs CSOC	**RR vs CSOC**	**RMAB vs RR**
% reduction in cumulative engagement drops	32.0%	5.2%	28.3%
p-value	0.044*	0.740	$0.098^\dagger$
Coefficient β	-0.0819	-0.0137	-0.0068

5 Conclusions and Lessons Learned

We present an RMAB based system to assist these non-profits in optimizing their limited service resources to support their massive programs delivering key health messages to a broad population of beneficiaries. To the best of our knowledge, ours is the first study to demonstrate the effectiveness of such RMAB-based resource optimization in real-world public health contexts. These encouraging results have initiated the transition of our RMAB software to **ABC** for real-world deployment. We hope this work paves the way for use of RMABs in many other health service applications.

Some key lessons learned from this research, complementing those outlined in [8, 30, 34] include the following (see Appendix K for an elaborate discussion). First, social-impact driven engagement and design iterations with the NGOs on the ground is crucial to understanding the right AI model for use and appropriate research challenges. In short, domain partnerships with NGOs to achieve real social impact automatically revealed requirements for use of novel application of an AI model (RMAB) and new research problems in this model. Second, data and compute limitations of non-profits are a real world constraint, and must be seen as genuine research challenges in AI for social impact, rather than limitations. Third, in deploying AI systems for social impact, there are many technical challenges that may not need innovative solutions, but they are critical to deploying solutions at scale. Finally we hope this work serves as a useful example of deploying an AI system for social impact in partnership with non-profits in the real world and will pave the way for more such impactful solutions.

References

[1] Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics*. Princeton university press.

[2] ARMMAN (2020). mmitra. `https://armman.org/mmitra/`.

[3] Avrachenkov, K. and Borkar, V. S. (2020). Whittle index based q-learning for restless bandits with average reward. *arXiv preprint arXiv:2004.14427*.

[4] Biswas, A., Aggarwal, G., Varakantham, P., and Tambe, M. (2021). Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*.

[5] Chen, R., Santo, K., Wong, G., Sohn, W., Spallek, H., Chow, C., and Irving, M. (2021). Mobile apps for dental caries prevention: Systematic search and quality evaluation. *JMIR mHealth and uHealth*, 9.

[6] Corotto, P. S., McCarey, M. M., Adams, S., Khazanie, P., and Whellan, D. J. (2013). Heart failure patient adherence: epidemiology, cause, and treatment. *Heart failure clinics*, 9(1):49–58.

[7] Dahiya, K. and Bhatia, S. (2015). Customer churn analysis in telecom industry. In *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions)*, pages 1–6. IEEE.

[8] Floridi, L., Cowls, J., King, T., and Taddeo, M. (2020). How to design ai for social good: Seven essential factors. *Science and Engineering Ethics*, 26.

[9] Gentile, C., Li, S., and Zappella, G. (2014). Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR.

[10] Gilbert, E. N. (1960). Capacity of a burst-noise channel. *Bell system technical journal*, 39(5):1253–1265.

[11] Glazebrook, K. D., Ruiz-Hernandez, D., and Kirkbride, C. (2006). Some indexable families of restless bandit problems. *Advances in Applied Probability*, 38(3):643–672.

[12] HelpMum (2021). Preventing maternal and infant mortality in nigeria. `https://helpmum.org/`.

[13] Johnson, J . (2017). Momconnect: Connecting women to care, one text at a time. `https://www.jnj.com/our-giving/momconnect-connecting-women-to-care-one-text-at-a-time`.

[14] Jung, Y. H. and Tewari, A. (2019). Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems*.

[15] Kaur, J., Kaur, M., Chakrapani, V., Webster, J., Santos, J., and Kumar, R. (2020). Effectiveness of information technology–enabled 'smart eating' health promotion intervention: A cluster randomized controlled trial. *PLOS ONE*, 15:e0225892.

[16] Killian, J. A., Wilder, B., Sharma, A., Choudhary, V., Dilkina, B., and Tambe, M. (2019). Learning to prescribe interventions for tuberculosis patients using digital adherence data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[17] Li, C., Wu, Q., and Wang, H. (2021). Unifying clustered and non-stationary bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1063–1071. PMLR.

[18] Li, S., Chen, W., and Leung, K.-S. (2019). Improved algorithm on online clustering of bandits. *arXiv preprint arXiv:1902.09162*.

[19] Martin, L. R., Williams, S. L., Haskard, K. B., and DiMatteo, M. R. (2005). The challenge of patient adherence. *Therapeutics and clinical risk management*, 1(3):189.

[20] Mate, A., Killian, J., Xu, H., Perrault, A., and Tambe, M. (2020). Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 34.

[21] Mate, A., Perrault, A., and Tambe, M. (2021). Risk-aware interventions in public health: Planning with restless multi-armed bandits. *Autonomous Agents and Multii-Agent Systems (AAMAS)*.

[22] Pfammatter, A., Spring, B., Saligram, N., Davé, R., Gowda, A., Blais, L., Arora, M., Ranjani, H., Ganda, O., Hedeker, D., Reddy, S., and Ramalingam, S. (2016). mhealth intervention to improve diabetes risk behaviors in india: A prospective, parallel group cohort study. *Journal of Medical Internet Research*, 18:e207.

[23] Pilote, L., Tulsky, J. P., Zolopa, A. R., Hahn, J. A., Schecter, G. F., and Moss, A. R. (1996). Tuberculosis Prophylaxis in the Homeless: A Trial to Improve Adherence to Referral. *Archives of Internal Medicine*, 156(2):161–165.

[24] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley.

[25] Qian, Y., Zhang, C., Krishnamachari, B., and Tambe, M. (2016). Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In Jonker, C. M., Marsella, S., Thangarajah, J., and Tuyls, K., editors, *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, pages 123–131. ACM.

[26] Shaaban, E., Helmy, Y., Khedr, A., and Nasr, M. (2012). A proposed churn prediction model. *International Journal of Engineering Research and Applications*, 2(4):693–697.

[27] Sombabu, B., Mate, A., Manjunath, D., and Moharir, S. (2020). Whittle index for AoI-aware scheduling. In *2020 12th International Conference on Communication Systems & Networks*. IEEE.

[28] Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S., and Lee, S.-K. (2010). Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthcare informatics research*, 16(4):253–259.

[29] Thirumurthy, H. and Lester, R. T. (2012). M-health for health behaviour change in resource-limited settings: applications to hiv care and beyond. *Bulletin of the World Health Organization*, 90:390–392.

[30] Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D., Ezer, D., Haert, F., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake, M., Othman, M., Glasmachers, T., Wever, W., and Clopath, C. (2020). Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11:2468.

[31] Tuldrà, A., Ferrer, M. J., Fumaz, C. R., Bayés, R., Paredes, R., Burger, D. M., and Clotet, B. (1999). Monitoring Adherence to HIV Therapy. *Archives of Internal Medicine*, 159(12):1376–1377.

[32] Weber, R. R. and Weiss, G. (1990). On an index policy for restless bandits. *Journal of applied probability*, pages 637–648.

[33] Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298.

[34] Wilder, B., Onasch-Vera, L., Diguiseppi, G., Petering, R., Hill, C., Yadav, A., Rice, E., and Tambe, M. (2021). Clinical trial of an ai-augmented intervention for hiv prevention in youth experiencing homelessness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14948–14956.

[35] Xie, Y., Li, X., Ngai, E., and Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3, Part 1):5445 – 5449.

[36] Yang, L., Liu, B., Lin, L., Xia, F., Chen, K., and Yang, Q. (2020). Exploring clustering of bandits for online recommendation system. In *Fourteenth ACM Conference on Recommender Systems*, pages 120–129.

A Related Work

Patient adherence monitoring in healthcare has been shown to be an important problem [19], and is closely related to the churn prediction problem, studied extensively in the context of industries like telecom [7], finance [26, 35], etc. The healthcare domain has seen several studies on patient adherence for diseases like HIV [31], cardiac problems [6, 28], Tuberculosis [16, 23], etc. These studies use a combination of patient background information and past adherence data, and build machine learning models to predict future adherence to prescribed medication [1]. However, such models treat adherence monitoring as a single-shot problem and are unable to appropriately handle the sequential resource allocation problem at hand. Additionally, the pool of beneficiaries flagged as high risk can itself be large, and the model can't be used to prioritize calls on a periodic basis, as required in our settings.

The Restless Multi-Armed Bandit (RMAB) framework has been popularly adopted to tackle such sequential resource allocation problems [14, 33]. Computing the optimal solution for RMAB problems is shown to be PSPACE-hard. Whittle proposed an index-based heuristic [33], that can be solved in polynomial time and is now the dominant technique used for solving RMABs. It has been shown to be asymptotically optimal for the time average reward problem [32], and other families of RMABs arising from stochastic scheduling problems [11]. Several works as listed in Section 1, show applicability of RMABs in different domains but these unrealistically assume perfect knowledge of the RMAB parameters, and have not been tested in real-world contexts. Avrachenkov and Borkar [3], Biswas et al. [4], present a Whittle Index-based Q-learning approach for unknown RMAB parameters. However, their techniques either assume identical arms or rely on receiving thousands of samples from each arm, which is unrealistic in our setting, given limited overall stay of a beneficiary in an information program — a beneficiary may drop out or stop engaging with the program few weeks post enrolment unless a service call convinces them to do otherwise. Instead, we present a novel approach that applies clustering to the available historical data to infer model parameters.

Clustering in the context of Multi-Armed Bandit and Contextual Bandits has received significant attention in the past [9, 17, 18, 36], but these settings do not consider restless bandit problems.

B Preliminaries

B.1 Background: Restless Multi-Armed Bandits

An RMAB instance consists of N independent 2-action Markov Decision Processes (MDP) [24], where each MDP is defined by the tuple $\{\mathcal{S}, \mathcal{A}, R, \mathcal{P}\}$. $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the set of possible actions, R is the reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ and $\mathcal{P}$ represents the transition function. We use $P^{\alpha}_{s,s'}$ to denote the probability of transitioning from state s to state s' under the action α. The policy π, is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ that selects the action to be taken at a given state. The total reward accrued can be measured using either the discounted or average reward criteria to sum up the immediate rewards accrued by the MDP at each time step. Our formulation is amenable to both, although we use the discounted reward criterion in our study.

The expected *discounted reward* starting from state s_0 is defined as $V^{\pi}_{\beta}(s_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R(s_t, \pi(s_t), s_{t+1} | \pi, s_0)\right]$ where the next state is drawn according to $s_{t+1} \sim P^{\pi(s_t)}_{s_t, s_{t+1}}$, $\beta \in [0, 1)$ is the discount factor and actions are selected according to the policy mapping π. The planner's goal is to maximize the total reward.

We model the engagement behavior of each beneficiary by an MDP corresponding to an arm of the RMAB. Pulling an arm corresponds to an active action, i.e., making a service call (denoted by $\alpha = a$), while $\alpha = p$ denotes the passive action of abstaining from a call. The state space $\mathcal{S}$ consists of binary valued states, s, that account for the recent engagement behavior of the beneficiary; $s \in [NE, E]$ (or equivalently, $s \in [0, 1]$) where E and NE denote the 'Engaging' and 'Not Engaging' states respectively. For example, in our domain, **ABC** considers that if a beneficiary stays on the automated voice message for more than 30 seconds (average message length is 1 minute), then the beneficiary

[1] Similarly, in our previous preliminary study (anonymous 2020) published in a non-archival setting, we used demographic and message features to build models for predicting beneficiaries likely to drop-off from **ABC**'s information program.

has engaged. If a beneficiary engages at least once with the automated voice messages sent during a week, they are assigned the engaging (E) state for that time step and non-engaging (NE) state otherwise. For each action $\alpha \in \mathcal{A}$, the beneficiary states follow a Markov chain represented by the 2-state Gilbert-Elliot model [10] with transition parameters given by $P^{\alpha}_{ss'}$, as shown in Figure 2. With slight abuse of notation, the reward function $R(.)$ of n^{th} MDP is simply given by $R_n(s) = s$ for $s \in \{0, 1\}$.

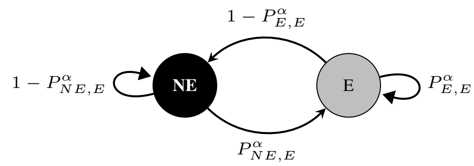


Figure 2: The beneficiary transitions from a current state s to a next state s' under action α, with probability $P^{\alpha}_{ss'}$.

We adopt the Whittle solution approach described previously for solving the RMAB. It hinges around the key idea of a "passive subsidy", which is a hypothetical reward offered to the planner, in addition to the original reward function for choosing the passive action. The Whittle Index is then defined as the infimum subsidy that makes the planner indifferent between the 'active' and the 'passive' actions, i.e.,:

$$W(s) = inf_{\lambda}\{\lambda : Q_{\lambda}(s, p) = Q_{\lambda}(s, a)\} \tag{1}$$

C Data Description

Beneficiaries enroll into **ABC**'s information program with the help of health workers, who collect the beneficiary's demographic data such as age, education level, income bracket, phone owner in the family, gestation age, number of children, preferred language and preferred slots for the automated voice messages during enrolment. These features are referred to as Beneficiary Registration Features in rest of the paper. Beneficiaries provided both written and digital consent for receiving automated voice messages and service calls. **ABC** also stores listenership information regarding the automated voice messages together with the registration data in an anonymized fashion.

The offline data of beneficiaries consists of beneficiary features, automated call data and service call data. The beneficiary features are collected at registration time and are unique determined by a `Beneficiary ID`. The features available in this data are:

- Age
- Education level - Ordinal value from 1-7 specifying increasing education level
- Income Binned - Ordinal value from 1-7 specifying increasing income groups (e.g. 1 is 0-5000 monthly income group and 7 is 30000 and above income group)
- Phone Owner in the family - family, neighbor, husband, woman herself
- Gestation Age - At time of registration
- Number of children
- Preferred Language - choice among two languages for automated voice messages made offered by **ABC**
- Preferred slots - Preference for time at which calls are to be made

The automated call data has complete log of automated calls made by **ABC**. Every call is uniquely determined by a `Call ID` and is characterised by:

- `Beneficiary ID` of the beneficiary who is called
- Date and time of call
- Duration of call listened
- Gestation age of beneficiary at time of call

k	Average RMSE	Standard Deviation	Missing Data	Non-meaningful Clusters
10	0.061	232.57	0.0	0.20
20	0.041	145.59	0.0	0.30
30	0.032	99.32	0.2	0.30
40	0.027	77.50	0.2	0.20
50	0.021	55.48	0.3	0.30
75	0.016	50.34	0.4	0.27
100	0.013	45.25	0.6	0.27
200	0.006	33.40	0.9	0.14

Table 2: Average RMSE, cluster size variance, missing data ratio, and non-meaningful clusters over all beneficiaries for different k. Total Beneficiaries = 4238

The duration column is used to determine whether at any timestep, the beneficiary is in engaging or non-engaging state. The service call data consists of beneficiary ID and the week in which the call is made. Transitions (s, α, s') can be obtained by combining the automated call data (obtain states s, s') and service call data (obtain action α). For beneficiaries in D_{test}, we only have beneficiary features available at test time. The automated call data is obtained every week and the service calls are generated by the RMAB or Round Robin algorithm.

D Clustering Methods

We use a historical dataset, $\mathcal{D}_{train}$ from **ABC** consisting of 4238 beneficiaries in total. We experiment to find an optimal number of clusters, k. We tried k = {10, 20, 30, 40, 50, 75, 100, 200} on Passive Transition-Probability based Clustering (PPF) and evaluated results on the following parameters:

1. **Representation:** Cluster centers that are representative of the underlying data distribution better resemble the ground truth transition probabilities. This is of prime importance to the planner, who must rely on these values to plan actions, and is measured by RMSE error in Table 2.
2. **Balanced cluster sizes:** A low imbalance across cluster sizes is desirable to preclude the possibility of arriving at few, gigantic clusters which will assign identical whittle indices to a large group of beneficiaries. Table 2 shows the variance among cluster sizes for each k.
3. **Missing data:** The historical dataset $\mathcal{D}_{train}$ contains very few active transition samples which results in missing active transition probabilities for many beneficiaries. Clustering beneficiaries together alleviates this issue but with increasing number of clusters, the missing data problem aggravates. The ratio of clusters with missing data for a given k is presented in table 2.
4. **Meaningful transition probabilities:** Interventions naturally tend to have a positive impact on the engagement behaviour of beneficiaries, hence $P^a_{NE,E} > P^p_{NE,E}$ and $P^a_{E,E} > P^p_{E,E}$. Clusters with transition probabilities conforming to these constraints are thus plausible and more desirable. Such clusters are termed as *meaningful* in our evaluation. Table 2 contains ratio of non-meaningful clusters for a given k.

We observe that as k increases, the RMSE and standard deviation decrease because of better fitting and smaller clusters respectively; the missing data problem worsens though. Based on the performance of different values of k on these factors, with the check for non-meaningful clusters, we found k = 40 to be the most optimal value.

Figure 3 shows our overall solution methodology.

1. Features-only Clustering (FO): This method relies on the correlation between the beneficiary feature vector f and their corresponding engagement behavior. We employ k-means clustering on the feature vector f of all beneficiaries in the historic dataset D_{train}, and then derive the representative transition probabilities for each cluster by pooling all the (s, α, s') tuples of beneficiaries assigned to that cluster. At test time, the features f of a new, previously unseen beneficiary in D_{test} map the beneficiary to their corresponding cluster and estimated transition probabilities.

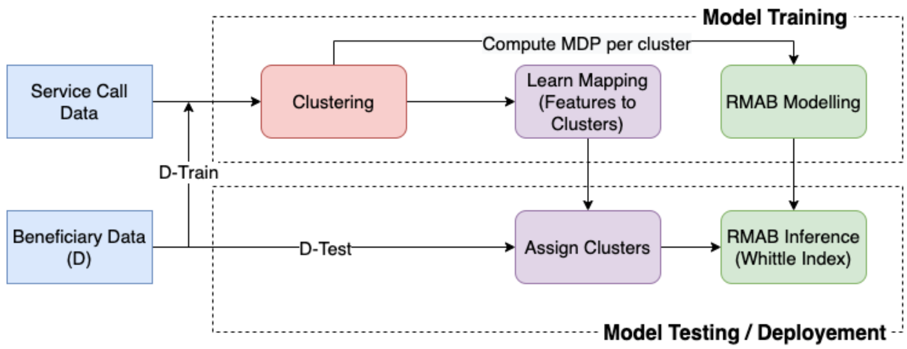


Figure 3: RMAB Training and Testing pipelines proposed

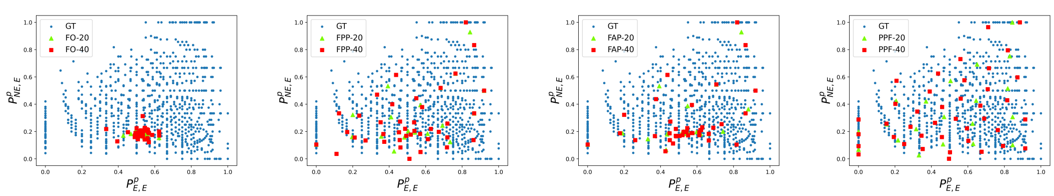
2. Feature + All Probabilities (FAP) In this 2-level hierarchical clustering technique, the first level uses a rule-based method, using features to divide beneficiaries into a large number of pre-defined buckets, B. Transition probabilities are then computed by pooling the (s, α, s') samples from all the beneficiaries in each bucket. Finally, we perform a k-means clustering on the transition probabilities of these B buckets to reduce them to k clusters ($k \ll B$). However, this method suffers from several smaller buckets missing or having very few active transition samples.

3. Feature + Passive Probabilities (FPP): This method builds on the FAP method, but only considers the passive action probabilities to preclude the issue of missing active transition samples.

The rule-based clustering on features involved in *FPP* and *FAP* methods can be thought of as using one specific, hand-tuned mapping function ϕ. In contrast, the *PPF* method *learns* such a map ϕ from data, eliminating the need to manually define accurate and reliable feature buckets.

D.1 Evaluation of Clustering Methods

We use a historical dataset, $\mathcal{D}_{train}$ from **ABC** consisting of 4238 beneficiaries in total, who enrolled into the program between May-July 2020. We compare the clustering methods empirically, based on the criteria described below.



(a) FO clustering (b) FPP clustering (c) FAP clustering (d) PPF clustering

Figure 4: Comparison of passive transition probabilities obtained from different clustering methods with cluster sizes $k = \{20, 40\}$ with the ground truth transition probabilities. Blue dots represent the true passive transition probabilities for every beneficiary while red or green dots represent estimated cluster centres.

1. Representation: Cluster centers that are representative of the underlying data distribution better resemble the ground truth transition probabilities. This is of prime importance to the planner, who must rely on these values to plan actions. Fig 4 plots the ground truth transition probabilities and the resulting cluster centers determined using the proposed methods. Visual inspection reveals that the *PPF* method represents the ground truth well, as is corroborated by the quantitative metrics of Table 3 that compares the RMSE error across different clustering methods.

2. Balanced cluster sizes: A low imbalance across cluster sizes is desirable to preclude the possibility of arriving at few, gigantic clusters which will assign identical whittle indices to a large groups of beneficiaries. Working with smaller clusters also aggravates the missing data problem in estimation of active transition probabilities. Considering the variance in cluster sizes and RMSE error for the different clustering methods with $k = \{20, 40\}$ as shown in Table 3, PPF outperforms the other clustering methods and was chosen for the pilot study.

Next we turn to choosing k, the number of clusters: as k grows, the clusters become sparse in number of active samples aggravating the missing data problem while a smaller k suffers from a higher RMSE. We found $k = 40$ to be optimal and chose it for the pilot study (details in Appendix D).

Table 3: Average RMSE and cluster size variance over all beneficiaries for different methods. Total Beneficiaries = 4238, $\mu_{20} = 211.9$, $\mu_{40} = 105.95$ (μ = average beneficiaries per cluster)

Clustering Method	Average RMSE		Standard Deviation	
	k = 20	**k = 40**	**k = 20**	**k = 40**
FO	0.229	0.228	143.30	74.22
FPP	0.223	0.222	596.19	295.01
FAP	0.224	0.223	318.46	218.37
PPF	0.041	0.027	145.59	77.50

As we got this RMAB system ready for real-world use, there was as an important observation for social impact settings: real-world use also required us to carefully handle several domain specific challenges, which were time consuming. For example, despite careful clustering, a few clusters may still be missing active probability values, which required employing a data imputation heuristic (details in Appendix F). Moreover, there were other constraints specific to **ABC**, such as a beneficiary should receive only one service call every η weeks, which was addressed by introducing "sleeping states" for beneficiaries who receive a service call (details in Appendix G).

E Whittle Index Implementation

For computing the whittle indices, we use the algorithm proposed by Qian et al. [25]. We perform a binary search over the passive subsidy λ that makes the Q-value at a given state indifferent to the active or passive action. Since transition probabilities are known (through estimation), we use value iteration to compute value function which can then be used to find Q-value. The value iteration is said to be converged when it doesn't change more than ϵ_{val_iter}. Similarly, we stop the binary search when the subsidy doesn't change more than ϵ_{bin_search}. We use $\epsilon_{val\_iter} = 1e-4$ and $\epsilon_{bin\_search} = 1e-5$ in our setup. The $2 * k$ indices thus computed for the k clusters, can then be looked up at all future time steps in constant time, making this an optimal solution for large scale deployment with limited compute resources.

F Estimating missing active transition probabilities

Since the number of beneficiaries in $\mathcal{D}_{train}$ who received a service call is much smaller compared to the total number of beneficiaries, there may be some clusters where the transition probabilities corresponding to the "active" action cannot be estimated from the data. We use the following heuristic to estimate these:

$$P^a_{s,s'}(i) = P^p_{s,s'}(i) + \delta, \qquad i \in M$$

where,

$$\delta = \frac{\sum\limits_{j \in K \setminus M} [P^a_{s,s'}(j) - P^p_{s,s'}(j)]}{\sum\limits_{j \in K \setminus M} 1}$$

$P^a_{s,s'}(i)$ corresponds to the transition probability for "active" action starting in state s and resulting in state s' for a cluster i, and $P^p_{s,s'}(j)$ corresponds to the "passive" action probability for cluster j. M is the set of clusters with missing active action transition probabilities $P^a_{s,s'}(i)$, and K is the set of all clusters.

G Sleeping states

To accommodate the frequency constraint presented by the NGO that prohibits beneficiaries from receiving a service call more than once every η timesteps, we augment the individual beneficiary MDPs to introduce additional sleeping states. To incorporate a frequency constraint of η, we introduce $2 * (\eta - 1)$ additional sleeping states, so that the augmented MDP consists of 2η states in total. The idea is to force the augmented MDP to transition through sleeping states for $(\eta - 1)$ time steps after an active action, before allowing it to reach the available states again. The set of 2η states can be listed as: $\{NE_1, E_1, NE_2, E_2 \ldots NE_\eta, E_\eta\}$, where the first $2 * (\eta - 1)$ states in the list are the sleeping states. We consider $\eta = 4$ henceforth in the paper. Formally, the passive and active transition matrices of the augmented MDP (indexed using the same order of states as the list above) can be given by:

$$\tilde{P}^p_{\mathsf{ss}'} = \begin{bmatrix} 0_{2\times 2}, & P^p_{ss'}, & 0_{2\times 2}, & 0_{2\times 2} \\ 0_{2\times 2}, & 0_{2\times 2}, & P^p_{ss'}, & 0_{2\times 2} \\ 0_{2\times 2}, & 0_{2\times 2}, & 0_{2\times 2}, & P^p_{ss'} \\ 0_{2\times 2}, & 0_{2\times 2}, & 0_{2\times 2}, & P^p_{ss'} \end{bmatrix}$$

$$\tilde{P}^a_{\mathsf{ss}'} = \begin{bmatrix} 0_{2\times 2}, & P^p_{ss'}, & 0_{2\times 2}, & 0_{2\times 2} \\ 0_{2\times 2}, & 0_{2\times 2}, & P^p_{ss'}, & 0_{2\times 2} \\ 0_{2\times 2}, & 0_{2\times 2}, & 0_{2\times 2}, & P^p_{ss'} \\ P^a_{ss'}, & 0_{2\times 2}, & 0_{2\times 2}, & 0_{2\times 2} \end{bmatrix}$$

The augmentation ensures that when passive, the state transitions to the adjoining bucket of sleeping states to the right, towards the available states each round as shown in Figure 5. When active, the state either resets to the starting sleeping state or transitions to the next sleeping state depending on whether currently the arm was available or sleeping. By design, the transition probabilities for the sleeping states of augmented MDP are identical for both actions and are distinct only in the available set of states.

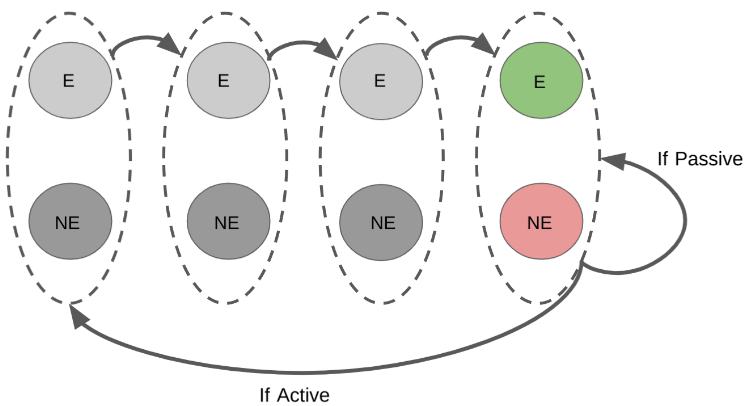


Figure 5: To capture the service call frequency constraint, the beneficiary transitions into the sleeping states for $\eta - 1 = 3$ timesteps (weeks), after receiving a service call, before becoming available again.

H ANOVA Test

To ensure that there is no selection bias in our random group assignment, we perform an Analysis of Variance (ANOVA) test to compare group means corresponding to all the beneficiary registration features. At 95% confidence interval, we find no significant difference in the mean values of features across all three groups. Joint p-value corresponding to every feature is reported in Table 4.

I Hyperparameters and Compute Specifications

We use $\epsilon_{val\_iter} = 1\text{e}{-4}$ and $\epsilon_{bin\_search} = 1\text{e}{-5}$ for whittle index computation. We used discount factor β defined in the MDP as 0.99. The random forest classifier used the following parameters:

Feature	μ_{RMAB}	μ_{RR}	μ_{CSOC}	Joint p-value
enroll_gest_age	20.323	20.454	20.325	0.472
age	24.799	24.733	24.878	2.609
number_of_children	1.675	1.686	1.705	1.373
education	3.531	3.490	3.477	2.961
income_binned	1.517	1.517	1.499	0.673
phone_owner_family	0.013	0.013	0.014	0.084
phone_owner_husband	0.103	0.108	0.107	0.636
phone_owner_neighbor	0.000	0.000	0.000	1.000
phone_owner_woman	0.884	0.879	0.879	0.616
language_1	0.519	0.517	0.507	1.261
language_2	0.475	0.477	0.487	1.262

Table 4: Mean feature values of beneficiaries in RMAB, RR and CSOC Groups along with joint p-value obtained from ANOVA test

number of trees in the forest = 200, quality of split = *entropy*, maximum depth of the trees = 30. All experiments are performed on an Intel(R) Xeon (R) CPU with 16 cores and 64 GB memory. The experiment for clustering and whittle index computation takes 10 mins. This is a one-time computation. All outputs are then stored offline and used during deployment with minimal overhead.

J Additional Experimental Details and Analysis

The quality improvement study started on April 26, 2021, with m beneficiaries selected from the RMAB and RR group each ($m \ll N$) per week for **ABC**-initiated service calls. Beneficiaries in $\mathcal{D}_{test}$ received automated voice messages few days post enrolment as per their gestational age. As per the current standard of care, any of these beneficiaries could also initiate a service call by placing a "missed call". **ABC** staff performing service calls were blind to the experimental groups that the beneficiaries belonged to. Recall, the goal of the service calls is to encourage the beneficiaries to engage with the health information message program in the future. For this study, number of service calls m was on average 225 per week for each of RMAB and RR groups to reflect real-world constraints on service calls. The study was scheduled for a total of 7 weeks, during which 20% of the RMAB (and RR) group had received a service call, which is closer to the percentage of population that may be reached in service calls by **ABC**. [2]

Table 5: Beneficiary distribution in the three groups and their start states during week 0 of the study.

Group	Engaging (E)	Non-Engaging (NE)	Total
RMAB	3571	4097	7668
RR	3647	4021	7668
CSOC	3661	4006	7667

Table 5 shows the absolute number of beneficiaries in states E or NE, where the state is computed using engagement data of one week (April 19 - April 26, 2021).

The drop in engagement under a policy π at time t can be measured as the change in engagement:

$$\Delta^{\pi}_{current}(t) := \sum_{n \in N} (R_n(s_0) - R_n(s_t)) \tag{2}$$

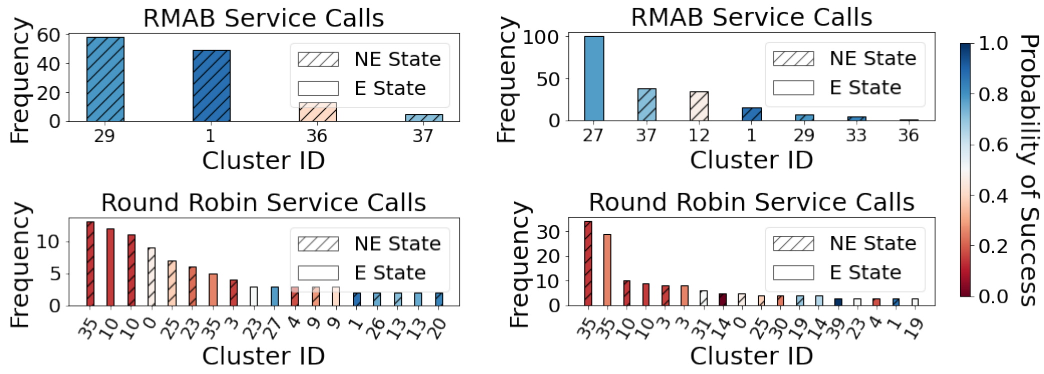
where $R_n(s_t)$ represents the reward for n^{th} beneficiary in state s_t at time step t and cumulative drop in engagement is:

$$\Delta^{\pi}_{cumulative}(t) := \sum_{n \in N} \sum_{\zeta=0}^{\zeta=t} (R_n(s_0) - R_n(s_\zeta)) \tag{3}$$

The cumulative drop in engagement prevented by a policy π, in comparison to the CSOC Group is thus simply:

$$\Delta^{\pi}_{cumulative}(t) - \Delta^{CSOC}_{cumulative}(t) \tag{4}$$

[2] Each beneficiary group also received very similar beneficiary-initiated calls, but these were less than 10% of the **ABC**-initiated calls in RMAB or RR groups over 7 weeks.



(a) Week 1 Service Calls

(b) Week 2 Service Calls

Figure 6: Distributions of clusters picked for service calls by RMAB and RR are significantly different. RMAB is very strategic in picking only a few clusters with a promising probability of success, RR displays no such selection.

and is plotted on the y-axis of Figure 1.

J.1 Statistical Analysis

Specifically, we fit a linear regression model to predict number of cumulative engagement drops at week 7 while controlling for treatment assignment and covariates specified by beneficiary registration features. The model is given by:

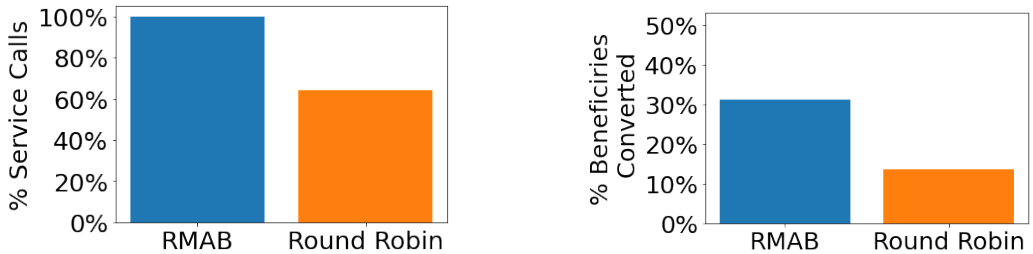
$$Y_i = k + \beta T_i + \sum_{j=1}^{J} \gamma_j x_{ij} + \epsilon_i$$

where for the i_{th} beneficiary, Y_i is the outcome variable defined as number of cumulative engagement drops at week 7, k is the constant term, β is the treatment effect, T_i is the treatment indicator variable, x_i is a vector of length J representing the i_{th} beneficiary's registration features, γ_j represents the impact of the j_{th} feature on the outcome variable and ϵ_i is the error term. For evaluating the effect of RMAB service calls as compared to CSOC group, we fit the regression model only for the subset of beneficiaries assigned to either of these two groups. T_i is set to 1 for beneficiaries belonging to the RMAB group and 0 for those in CSOC group. We repeat the same experiment to compare RR vs CSOC group and RMAB vs RR group.

J.2 RMAB Strategies

We analyse RMAB's strategic selection of beneficiaries in comparison to RR using Figure 6, where we group beneficiaries according to their whittle indices, equivalently their $\langle$`cluster, state`$\rangle$. Figure 6 plots the frequency distribution of beneficiaries (shown via corresponding clusters) who were selected by RMAB and RR in the first two weeks. For example, the top plot in Figure 6a shows that RMAB selected 60 beneficiaries from cluster 29 (NE state). First, we observe that RMAB was clearly more selective, choosing beneficiaries from just four (Figure 6a) or seven (Figure 6b) clusters, rather than RR that chose from 20. Further, we assign each cluster a hue based on their probability of transitioning to engaging state from their current state given a service call. Figure 6 reveals that RMAB consistently prioritizes clusters with high probability of success (blue hues) while RR deploys no such selection; its distribution emulates the overall distribution of beneficiaries across clusters (mixed blue and red hues).

Furthermore, Figure 7a further highlights the situation in week 1, where RMAB spent 100% of its service calls on beneficiaries in the non-engaging state while RR spent the same on only 64%. Figure 7b shows that RMAB converts 31.2% of the beneficiaries shown in Figure 7a from non-engaging to engaging state by week 7, while RR does so for only 13.7%. This further illustrates the need for optimizing service calls for them to be effective, as done by RMAB.



(a) (b)

Figure 7: (a) % of week 1 service calls on non-engaging beneficiaries (b) % of non-engaging beneficiaries of week 1 receiving service calls that converted to engaging by week 7

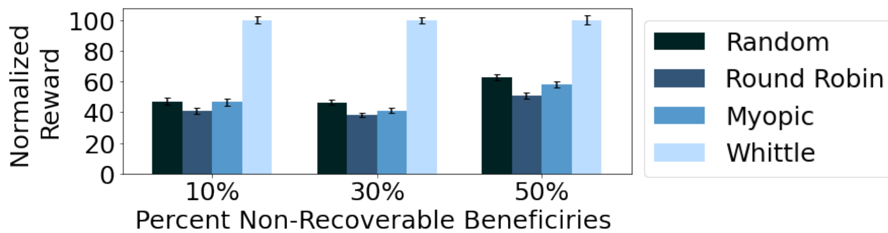


Figure 8: Performance of MYOPIC can be arbitrarily bad and even worse than RANDOM, unlike the Whittle policy.

J.3 Synthetic Results

We run additional simulations to test other service call policies beyond those included in the quality improvement study and confirm the superior performance of RMAB. Specifically, we compare to the following baselines: (1) RANDOM is a naive baseline that selects m arms at random. (2) MYOPIC is a greedy algorithm that pulls arms optimizing for the reward in the immediate next time step. WHITTLE is our algorithm. We compute a normalized reward of an algorithm `ALG` as: $\frac{100\times(\overline{R}^{\texttt{ALG}}-\overline{R}^{\text{CSOC}})}{\overline{R}^{\text{WHITTLE}}-\overline{R}^{\text{CSOC}}}$ where $\overline{R}$ is the total discounted reward. Simulation results are averaged over 30 independent trials and run over 40 weeks.

Figure 8 presents simulation of an adversarial example [20] consisting of $x\%$ of non-recoverable and $100-x\%$ of self-correcting beneficiaries for different values of x. Self-correcting beneficiaries tend to miss automated voice messages sporadically, but revert to engaging ways without needing a service call. Non-recoverable beneficiaries are those who may drop out for good, if they stop engaging. We find that in such situations, MYOPIC proves brittle, as it performs even worse than RANDOM while WHITTLE performs well consistently. The actual quality improvement study cohort consists of 48.12% non-recoverable beneficiaries (defined by $P^p_{01} < 0.2$) and the remaining comprised of self-correcting and other types of beneficiaries.

K Discussion of Lessons Learnt

Some key lessons learned from this research, which complement some of the lessons outlined in [8, 30, 34] include the following. First, social-impact driven engagement and design iterations with the NGOs on the ground is crucial to understanding the right AI model for use and appropriate research challenges. As discussed in footnote 1, our initial effort used a one-shot prediction model, and only after some design iterations we arrived at the current RMAB model. Next, given the missing parameters in RMAB, we found that the assumptions made in literature for learning such paramters did not apply in our domain, exposing new research challenges in RMABs. In short, *domain partnerships with NGOs to achieve real social impact automatically revealed requirements for use of novel application of an AI model (RMAB) and new research problems in this model.*

Second, *data and compute limitations of non-profits are a real world constraint, and must be seen as genuine research challenges in AI for social impact, rather than limitations.* In our domain, one key technical contribution in our RMAB system is deploying clustering methods on offline historical data to infer unknown RMAB parameters. Data is limited as not enough samples are available for any given beneficiary, who may stay in the program for a limited time. Non-profit partners also cannot bear the burden of massive compute requirements.

Our clustering approach allows efficient offline mapping to Whittle indices, addressing both data and compute limits, enabling scale-up to service 10s if not 100s of thousands of beneficiaries. Third, *in deploying AI systems for social impact, there are many technical challenges that may not need innovative solutions, but they are critical to deploying solutions at scale.* Indeed, deploying any system in the real world is challenging, but even more so in domains where NGOs may be interacting with low-resource communities. Finally we hope this work serves as a useful example of deploying an AI based system for social impact in partnership with non-profits in the real world and will pave the way for more such solutions with real world impact.