Limited Resource Allocation in a Non-Markovian World: The Case of Maternal and Child Healthcare

Panayiotis Danassis[1] , **Shresth Verma**[2] , **Jackson A. Killian**[1] , **Aparna Taneja**[2] and **Milind Tambe**[1,2]
[1]Harvard University
[2]Google Research
pdanassis@seas.harvard.edu, vermashresth@google.com, jkillian@g.harvard.edu, aparnataneja@google.com, milind_tambe@harvard.edu

Abstract

The success of many healthcare programs depends on participants' adherence. We consider the problem of scheduling interventions in low resource settings (e.g., placing timely support calls from health workers) to increase adherence and/or engagement. Past works have successfully developed several classes of Restless Multi-armed Bandit (RMAB) based solutions for this problem. Nevertheless, all past RMAB approaches assume that the participants' behaviour follows the Markov property. We demonstrate significant *deviations from the Markov assumption* on real-world data on a maternal health awareness program from our partner NGO, ARMMAN. Moreover, we extend RMABs to *continuous state* spaces, a previously understudied area. To tackle the generalised non-Markovian RMAB setting we (i) model each participant's trajectory as a *time-series*, (ii) leverage the power of time-series forecasting models to learn complex patterns and dynamics to predict future states, and (iii) propose the *Time-series Arm Ranking Index* (TARI) policy, a novel algorithm that selects the RMAB arms that will benefit the most from an intervention, given our future state predictions. We evaluate our approach on both synthetic data, and a secondary analysis on *real data* from ARMMAN, and demonstrate significant increase in engagement compared to the SOTA, deployed Whittle index solution. This translates to 16.3 hours of additional content listened, 90.8% more engagement drops prevented, and reaching more than twice as many high dropout-risk beneficiaries.

1 Introduction

According to the latest estimates from 2020 [GatesFoundation, 2020], the global Maternal Mortality Ratio (MMR) is 152 deaths per 100,000 live births, more than double the UN Sustainable Development Goal (SDG) 3.1's target [UN, 2023]. For context, the MMR in the USA is estimated to be 35, while in Western Europe is 5. Lack of access to preventive care information, especially in the global south, is a major contributing factor for these deaths. For example, India's MMR is estimated to be 130 deaths per 100,000 live births – almost 90% of which are avoidable if women receive the right kind of intervention [ARMMAN, 2023a]. To reduce MMR in India, our partner NGO, ARMMAN (armman.org), employs an automated call-based information program to disseminate critical healthcare information to pregnant women and recent mothers in underserved communities. Such programs have repeatedly demonstrated significant benefits (e.g., see [HelpMum, 2023; ARMMAN, 2023b; Verma *et al.*, 2023; Mate *et al.*, 2022; Kaur *et al.*, 2020; Pfammatter *et al.*, 2016]), as they raise awareness regarding the need for regular care, potential risk factors, and complications. According to ARMMAN, one of the biggest challenges these programs face is that of dwindling adherence, as a large fraction of beneficiaries often drop out. It is thus crucial to provide timely interventions through support calls, or home visits from healthworkers so as to minimize disengagement.

In this paper, we study the problem of scheduling healthcare interventions under limited healthcare worker resources. We model this resource optimization problem as a Restless Multi-armed Bandit (RMAB) problem [Whittle, 1988], in

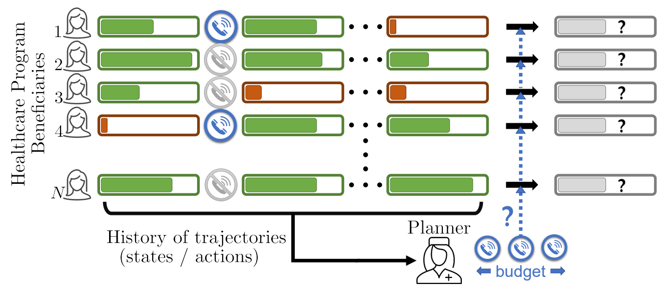


Figure 1: Scheduling healthcare interventions: at each timestep, a planner selects k out of N arms (healthcare program beneficiaries) to schedule an intervention (e.g., a healthcare worker will call or visit). Each bar represents the state (level of engagement) of each arm at each timestep, which can change even when the arm is not pulled. Green (red) bars represent beneficiaries with state above (below) the desired engagement threshold (e.g., for ARMMAN the threshold is set to listening to 25% of the automated message). Intervening on a beneficiary will increase their engagement in expectation (e.g., see beneficiary #4). The planner observes the states and adjusts its policy to maximize the number of engaging beneficiaries.

which a planner can act on k out of N arms (beneficiaries) at each timestep (Figure 1). Contrary to stochastic bandits [Auer *et al.*, 2002], in RMABs each arm has a state, the reward depends on said state, and the state changes even when the arm is not pulled. Past works have developed RMAB-based solutions for several classes of sequential scheduling problems with limited resources. Examples include anti-poaching patrols [Qian *et al.*, 2016], machine maintenance [Glazebrook *et al.*, 2006], online advertising [Meshram *et al.*, 2016], and, healthcare [Wang* *et al.*, 2023]. Notably, all past RMAB approaches assume that the arms' behaviour follows the Markov property, where state transitions are history independent [Puterman, 2014]. We challenge this assumption, as human behaviour is likely to contain *temporal dependencies*, i.e., depend on past states, observations, and actions [Chierichetti *et al.*, 2012; Early *et al.*, 2022; Meiss *et al.*, 2010]. Using real-world data on a maternal health awareness program from our partner NGO, ARMMAN, we demonstrate significant deviations from the Markov assumption. Specifically, the log-likelihood of observing the historical trajectories increases (up to 23%), as we increase the order h of the underlying model (see Section 3).

Even under the Markov assumption, computing an optimal policy for RMABs is PSPACE-hard [Papadimitriou and Tsitsiklis, 1994]. Instead, state-of-the-art (SOTA) approaches commonly adopt the Whittle index policy [Whittle, 1988], an approximate solution that estimates the expected future value (Whittle index) of acting on an arm, and then proceeds to act on the top-k arms with the largest value.

If we want to capture non-Markovian[1] behaviors using SOTA Whittle index based approaches, we would run into computation and data limitations. First, there will be a combinatorial explosion of the state space (an h-order Markov process can be viewed as a first order Markov process on the expanded state space, where each 'super' state consists of h original states, i.e., $s' = \times_{i=1}^{h} s_i$). Second, since the Whittle index policy requires to know the underlying Markov decision process (MDP) – which grows exponentially in both (i) the order of the process and (ii) the discretization of the state – it would need ever larger datasets to calculate the empirical transition probabilities.

We are the first to cast limited resource optimization problems into the generalised non-Markovian RMAB setting. In order to provide a *practical*, and *scalable* solution, we take inspiration from the core idea of the Whittle index – pull arms with the highest expected gains from pulling – but we drop the cumbersome MDPs. Instead, we opt to *independently* model each participant's trajectory as a *time-series*, leveraging the power of time-series models to learn complex patterns and dynamics to predict future states. Additionally, we develop the *Time-series Arm Ranking Index* (TARI) policy, a novel algorithm that selects the arms that will benefit the most from an intervention, given our model's future state predictions.

Finally, as we are no longer limited by the complexity of the Whittle index, in this work, *we extend RMABs to continuous state spaces* – a previously understudied area – without the need for the discretization of the state, thus bypassing approximation losses (e.g., see [Sinha and Mahajan, 2022]). Combining continuous states and non-Markovian transitions offers additional expressiveness that can more accurately capture behavior transitions and patterns, as we showcase in our results.

[1]By non-Markovian we refer to any Markov process of order 2 or higher. For details please see Section 2.

Figure 2: A beneficiary receiving preventive information (photo courtesy of ARMMAN).

Supplementary Material. Please see [Danassis *et al.*, 2023] for the full version, including the source code.

1.1 Our Contributions

(1) We demonstrate significant deviations from the Markov assumption in real-world data from a *deployed* maternal and child health awareness program by ARMMAN.

(2) We are the *first* to cast limited resource optimization problems into the generalised *non-Markovian, continuous state* restless multi-armed bandit setting, enabling us to capture *temporal dependencies* in human behaviour.

(3) We model each arm as a time-series, and develop a novel algorithm, the *Time-series Arm Ranking Index* (TARI) policy, that acts on arms which will benefit the most from an intervention, given our model's future state predictions, resulting in a *practical*, and *scalable* solution.

(4) We perform a secondary analysis on *real-data* (2252 participants, 23 weeks) from a maternal health awareness program (mHealth), in partnership with an Indian NGO, ARMMAN. Compared to the SOTA, deployed Whittle index policy, TARI results in 16.3 hours of additional content listened, 90.8% more engagement drops prevented, and reaching more than twice as many high dropout-risk beneficiaries.

1.2 Discussion & Related Work

Restless multi-armed bandits (RMABs). Prior work in RMAB assumes that arms follow the Markov property. Even in Markovian settings, and when transition dynamics are fully known, RMABs suffer from the curse of dimensionality. Planning an optimal policy is PSPACE-hard [Papadimitriou and Tsitsiklis, 1994]. As such, SOTA approaches usually deploy approximate planning solutions, most notably the ***Whittle index policy*** [Whittle, 1988], which solves the Lagrangian relaxation of the problem. The resulting Lagrange multipliers capture the 'resource-efficient value for acting' on an arm (more accurately, the opportunity cost [Buchanan, 1991] for not acting). Then the Whittle index policy proceeds to greedily act on the arms with the largest Lagrange multipliers (see the supplement for a formal definition). The Whittle index approach has been shown to be asymptotically optimal (i.e., when $N \to \infty$ with fixed $\frac{k}{N}$) [Weber and Weiss, 1990], and

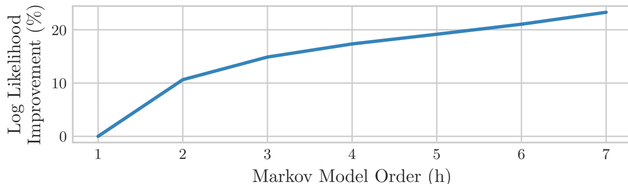


Figure 3: Relative (with respect to $h = 1$) improvement in log likelihood. As we increase the order h of the underlying Markov model, the probability of observing the trajectories in our dataset increases. This suggests non-Markovian behaviour.

has been shown to perform well empirically in many applications (e.g., [Qian *et al.*, 2016; Hsu, 2018; Mate *et al.*, 2022; Kadota *et al.*, 2016]). Nevertheless, the Whittle index remains a heuristic, and asymptotic optimality does not necessarily translate to practically relevant problem sizes and planning horizons, as was recently demonstrated in [Ghosh *et al.*, 2023]. Critically, the Whittle index is only optimal under several assumptions (see [Ghosh *et al.*, 2023]), which are often hard to validate, and part of active research. Finally, despite being a heuristic, the approach can be prohibitively slow, thus it often requires a problem-specific fast method for computing the index. As such, using the traditional Whittle index on the expanded state space (i.e., $s' = \times_{i=1}^{h} s_i$) in non-Markov settings is quite challenging and does not guarantee high-quality results, even if the problem can be approximated sufficiently well with low order Markov processes. In this work, we are the first to generalise RMABs to non-Markovian settings. We maintain the key idea of the Whittle index policy (acting on arms that will benefit the most from an intervention), but we use the power of time-series prediction models to capture complex patterns and dynamics to predict the effect of an intervention.

Time-series Forecasting. Time series forecasting has been an active area of research over the past few years, with applications in diverse areas such as energy consumption, sensor network monitoring, traffic planning, variations in air pollution, weather forecast, disease propagation, and so on (e.g., [Matsubara *et al.*, 2014; Wu *et al.*, 2020; Zhou *et al.*, 2021]). Solutions range from traditional statistical methods (e.g., [Ariyo *et al.*, 2014]), to deep learning-based models (e.g., [Lai *et al.*, 2018; Bai *et al.*, 2018; Liu *et al.*, 2021]), and, more recently, Transformer-based solutions [Wu *et al.*, 2020; Zhou *et al.*, 2021; Wu *et al.*, 2021; Wen *et al.*, 2022] which mostly focus on the more challenging long-term forecasting problem (although their effectiveness has recently come into question [Zeng *et al.*, 2023]). Our contributions are not in developing a SOTA time series prediction model. Instead, we utilise off-the-shelf models in a novel way, to solve non-Markovian RMAB problems. Importantly, our approach is *model agnostic*. In fact, we have evaluated a variety of architectures, including LSTM [Hochreiter and Schmidhuber, 1997], BiLSTM [Graves and Schmidhuber, 2005], Transformer models [Vaswani *et al.*, 2017], adding time-based vector representations [Kazemi *et al.*, 2019], attention layers, and more (see the supplement for the details). Future advancements in the area of time series forecasting can easily translate to better performance for the proposed approach.

From the application perspective, [Nishtala *et al.*, 2020] and [Nisthala *et al.*, 2021] aim to predict the dropout risk in a similar maternal health awareness setting. Both works are about classification to high and low risk of dropout, and not time-series regression. The former does not optimize or schedule interventions, while the latter aims to identify a smaller subset of beneficiaries and the use the traditional Whittle index in a classic, binary-state Markov setting.

2 Problem Formulation: Non-Markovian Restless Multi-armed Bandits (NMRMAB)

We consider scheduling problems in which a planner must act on k out of N independent, *continuous* state ($[0, 1]$) arms each round. The planner fully observes the state of each arm, then all arms undergo a history-dependent (i.e., non-Markovian) state transition. The planner's goal is to maximize the number of processes in 'engaging' state over the time horizon H.

Let $s_{i,t} \in \mathcal{S}$, and $a_{i,t} \in \mathcal{A}$ denote the state and action taken on arm i, respectively, in timestep t. We assume that states are continuous in $[0, 1]$, and represent the 'level of engagement' of a beneficiary, with higher numbers representing a higher level of engagement. The action set consists of two actions: active ($a_{i,t} = 1$), and passive ($a_{i,t} = 0$). A non-Markovian Restless Multi-armed Bandit (NMRMAB) problem instance is a 4-tuple $\{\mathcal{N}, k, (X_t^{i \in N})_{t=1}^{H}, R\}$, where $\mathcal{N}$ is the set of *independent* arms, k is the budget constraint such that $\sum_i a_{i,t} = k, \forall t$, denoting how many arms can be pulled at a given time-step, $(X_t^{i \in \mathcal{N}})_{t=1}^{H}$ is an associated transition process for arm i for time horizon H, and $R : (X_t)_{t=1}^{H} \to \Re$ is the reward function. In our setting, the reward at timestep t is given by $R(\times_{i<t}(s_i, a_i), s_t) = \mathbb{1}_{s_t \geq s^*}$, where s^* is a domain-specific engagement threshold (e.g., ARMMAN considers $s^* = 0.25$). The planner's goal is to maximize the total reward, i.e., $\sum_{t \in [1 \dots H]} \sum_{i \in \mathcal{N}} R(\cdot)$. Finally, we assume $(X_t^{i \in \mathcal{N}})_{t=1}^{H}$ to be a higher order Markov process. For an h-order Markov process, the next state depends on the proceeding h states. More formally:

Definition 2.1 (Order h Markov Process)**.** Let $x_1, \dots, x_t$ be the elements of the process. A Markov process of order h is a process $(X_t)_{t=1}^{\infty}$, such that $\forall t$:

$$\Pr[X_t = x_t \mid x_{t-1}, x_{t-2}, \dots, x_1] =$$
$$\Pr[X_t = x_t \mid x_{t-1}, x_{t-2}, \dots, x_{t-h}]$$

3 Non-Markovian Behaviour in Maternal mHealth Data

The proposed modeling raises the question as to whether human activity is indeed non-Markovian in our domain. To answer this question, following related literature [Chierichetti *et al.*, 2012], we compute the log-likelihood of the participants' trajectories for a process of order h, based on the data on a maternal health awareness program from our partner NGO, ARMMAN. Specifically, we start by computing the empirical transition probabilities, assuming the underlying process is of order h. Let $h = 1$. This is easily achieved by maintaining counters $C_{(s_t, a_t) \to s_{t+1}}$. The transition probabilities are simply $C_{(s_t,a_t)\to s_{t+1}} / \sum_x C_{(s_t,a_t)\to x}$. For $h > 1$, an h-order Markov process can be viewed as a first order Markov process on the expanded state space $s' = \times_{i=1}^{h} s_i$. Thus, we can employ the same approach to calculate empirical probabilities. Given that we are dealing with continuous states, to maintain said counters, we first need to discretize them. We opted for a binary discretization, as in related literature [Verma *et al.*, 2023], into 'engaging' and 'non-engaging' states, using $s*$ as a threshold (see also Section 5.3). Then, for each trajectory x in our dataset, we compute $l(h) \triangleq -\log \mathcal{L}(h \mid x) \triangleq -\log \Pr(X = x \mid \text{model of order } h)$. Finally, we average over all trajectories. Due to the increase in the number of counters and data required, we model up to seventh order processes ($h = 7$). Figure 3 shows the relative improvement $-\left(\frac{l(h)-l(h=1)}{l(h=1)}\right)$ in negative log-likelihood for order h Markov processes (x-axis). There is a clear improvement for higher order models, specifically about 10% for $h = 2$, and up to 23% for $h = 7$. This suggests that, under the often-used binary-discretization model, participants' behavior across the duration of the program is indeed not Markovian.

4 Methodology

4.1 RMABs as a Time-series Forecasting Problem

In this work, we are the *first* to propose a time-series forecasting (TSF) based framework for supervised representation learning of arms' trajectories in non-Markovian RMABs.

Preliminaries: TSF Problem Formulation

Let $\mathcal{X} = \{X_t^1, \ldots, X_t^M\}_{t=1}^{h}$ denote the historical data of an M-variate series, where h is the look-back window length and X_t^i is the value of the ith variate at timestep t. Formally, the TSF task is, given $\mathcal{X}$, to predict the future T values, i.e., $\hat{\mathcal{Y}} = \{\hat{Y}_t^1, \ldots, \hat{Y}_t^M\}_{t=h+1}^{h+T}$. There are two methods for predicting $\hat{\mathcal{Y}}$ when $T > 1$: Iterated multi-step (IMS) forecasting [Taieb *et al.*, 2012] where the model learns to predict a single-step forward, and then is recursively called to obtain multi-step forecasts. Alternatively, with direct multi-step (DMS) forecasting [Chevillon, 2007], one trains a model that directly optimizes the multi-step forecasting objective, and is only called once. Usually IMS forecasts result in smaller variance than DMS, but the error could accumulate over long prediction horizons [Zeng *et al.*, 2023]. For simplicity, and given the relatively short forecasting horizon in our domain, we opted for the IMS approach (see also Section 4.2).

Labeled Dataset Using a Sliding Window

To produce a labeled dataset for both training (D_{train}), and evaluation (D_{test}), we used a fixed-length sliding window approach. First, we assume that the planner has access to an offline historical dataset of beneficiaries' trajectories, as is the case for example with our partner NGO, ARMMAN. We then run a sliding window of length h on each trajectory in the dataset to get training samples of history ($\mathcal{X}$), and next state ($\hat{\mathcal{Y}}$), as depicted in Figure 4. For the application at hand, $\mathcal{X} = \{s_t, a_t\}_{t=1}^{h}$, i.e., pairs of state/actions. This intuitively corresponds to an h-order Markov model approximation. h is a hyper-parameter that depends on the application and needs

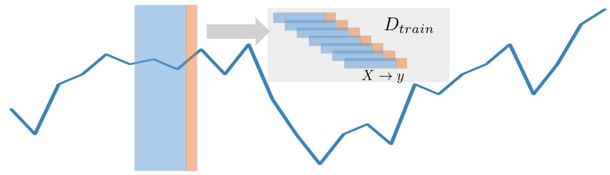


Figure 4: We use a fixed length sliding window, sliding right one time-step at a time, to generate a supervised learning training dataset from participants' time series trajectories.

to be tuned depending on (i) the level of non-Markovian behaviour of beneficiaries, and (ii) the achieved error of the model (larger h does not necessarily translate to lower error). Finally, $\hat{\mathcal{Y}} = \{s_t\}_{t=h+1}$, i.e., we only predict the next state. Optionally, depending on the domain, we can enhance the input with auxiliary tokens relevant to the task and the arms' behavior (e.g., socio-demographic features).

4.2 Proposed Approach: Time-series Arm Ranking Index (TARI) Policy

The proposed Time-series Arm Ranking Index (TARI) maintains the core idea of the traditional Whittle index: estimate the expected future value of acting on an arm, and then greedily act on arms that will benefit the most from an intervention. However, we adjust the methodology to account for (i) the additional complexity due to the non-Markovian setting, and (ii) the scarcity of resources. Specifically for the latter, a key challenge in real-world applications (and especially ones related to healthcare), is that they are resource constrained. For example, despite the scale of ARMMAN's operation, with *millions* of active users, due to limited availability of healthworkers beneficiaries typically receive *at most one intervention* (phone call by a health worker) within a period of 3 months. Taking this into account, *TARI estimates the marginal long-term improvement in engagement if you act once on a arm (and never act again), compared to never acting*.

Specifically, we model each arm independently as a time series, and train a model to predict the next state (s_{t+1}), given (i) a history of state/actions ($\times_{i<t}^{h}(s_i, a_i)$) of length h, (ii) the current state (s_t), and (iii) the potential action (a_t, where $a_t = 1$ corresponds to acting, and $a_t = 0$ not acting). This is the offline training part. We use this model online, in an iterated multi-step manner (see Section 4.1), to generate a long term forecast ($s_{t+1}, s_{t+2}, \ldots, s_{t+H}$), which then we use to compute the TARI index for planning as follows.

For each arm n *independently*, we estimate two quantities by recursively using our TSF model: (i) The time u_n until arm n switches to non-engaging,[2] if we act once at timestep t and never act again (line 4 in Algorithm 1, and green box in Figure 5), and (ii) the time v_n until arm n switches to non-engaging, if we never act (line 11 in Algorithm 1, and orange box in Figure 5). Then the TARI index for arm n is simply given by the ratio of the two numbers (Equation 1). This in-

[2]Non-engaging means that the continuous state s drops below an application-specific threshold (s^*). ARMMAN considers $s^* = 0.25$.

Algorithm 1: Time-series Arm Ranking Index

Data: Historical dataset of beneficiaries' trajectories
Offline: Train TSF model $\mathcal{M}$ to predict the next state
Online: Decision timestep t:
for $arm \in \mathcal{N}$ **do**
  a **= 1**, $u = 1$, $s = s_t$, history=$\times_{i<t}^{h}(s_i, a_i)$
  $s' = \mathcal{M}(\text{history}, s, a)$
  while $s' \geq s^*$ *and* $u \leq H$ **do** // While the state is above the engagement threshold. Forecast ahead at most until the time horizon.
    history.append((s, a))
    $s = s'$, a **= 0**, $u = u + 1$
    $s' = \mathcal{M}(\text{history}, s, a)$

  a **= 0**, $v = 1$, $s = s_t$, history=$\times_{i<t}^{h}(s_i, a_i)$
  $s' = \mathcal{M}(\text{history}, s, a)$
  while $s' \geq s^*$ *and* $v \leq H$ **do**
    history.append((s, a))
    $s = s'$, a **= 0**, $v = v + 1$
    $s' = \mathcal{M}(\text{history}, s, a)$

  TARI(arm) = $\frac{u}{v}$

tuitively gives the 'value' of acting. Finally, at each timestep, just like with the traditional Whittle index, we act on the top-k arms with the highest TARI value. The proposed approach is depicted in Figure 5 and Algorithm 1.

$$\text{TARI}(n) = \frac{u_n}{v_n} \tag{1}$$

The TARI policy offers significant advantages. Arms are modeled independently, which allows for scalability. Furthermore, it is computationally efficient to train and compute in non-Markovian continuous state settings, contrary to the traditional Whittle index which requires supporting and computing over MDPs that grow exponentially in both (i) the order of the underlying Markov process, and (ii) the discretization of the state.

5 Simulation Setup

Training data are constructed in the manner described in Section 4.1. We use a 64%, 16%, 20% split for the training, validation, and testing datasets, respectively. All experiments are averaged over 10 independent runs.

5.1 Baselines

We compare the proposed TARI to four baselines: (i) the **Whittle index policy** [Whittle, 1988], (ii) **round-robin** selection, which often corresponds to the default policy for many NGOs [Mate *et al.*, 2022] including ARMMAN, (iii) **random**, where we act on arms selected uniformly at random, and (iv) **control**, where there are no support calls (no intervention, i.e., $a_{i,t} = 0, \forall i \in \mathcal{N}, \forall t \in H$).

We chose to compare to the Whittle index, as it is a popular, SOTA approach that has been deployed in the real-world

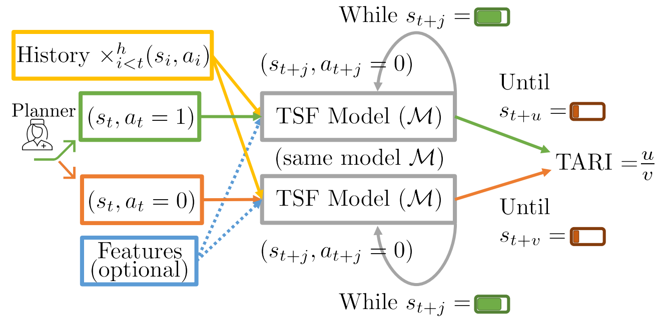


Figure 5: Graphical representation of the calculation of the proposed TARI. We consider two options for the planner: (i) act once (green box, $a_t = 1$) and then never act ($a_{t+j} = 0$), or (ii) never act (orange box). We recursively call the TSF model to predict the next state in the trajectory (feedback loop on the left), until the state switches from engaging (green bar) to non-engaging (red bar). Let this be at timestep u (i.e., $s_{t+u} < s^*$) and v (i.e., $s_{t+v} < s^*$) for options (i), and (ii), respectively. The ratio between the timesteps needed for the switch if we had acted, compared to not, constitutes the TARI. This is computed independently for each arm.

(e.g., see [Mate *et al.*, 2022; Verma *et al.*, 2023]). Informally, the Whittle index of an arm captures the added value from pulling said arm. Consider a 'passive subsidy' – a hypothetical exogenous compensation m rewarded for not pulling ($a = 0$) arm i. The Whittle index is defined as the smallest subsidy necessary to make the planner indifferent between pulling and not pulling (assuming indexability [Whittle, 1988]), i.e.,

$$W_i(s) \triangleq \inf_m \{Q_i^m(s \mid a = 0) = Q_i^m(s \mid a = 1)\} \tag{2}$$

The Whittle index policy computes the $W_i(s)$ of all arms and pulls the arms with the highest values of the index at each timestep. The augmented (with subsidy m) Bellman equations are solved via value iteration, and binary search is used to find the smallest m that satisfies Equation 2. To use the Whittle index in our setting, we must first discretize the continuous state. Following the convention in previous deployments (e.g., [Mate *et al.*, 2022]), we assume a binary state Markov model ('engaging' = 1, 'non-engaging' = 0, thresholded at s^*). For completeness, we also run simulations with a more fine-grained discretization of the state, and also incorporating history by using an expanded state space (i.e., $s' = \times_{i=1}^{h} s_i$). Of course this significantly increases computational and memory complexity, and data requirements (see Sections 1, 1.2, 3).

5.2 Synthetic Data

We generate trajectories containing an equal number of the following types of 'agents' (arms). (i) **Habit former**: The value of the continuous sate drops under passive action ($a = 0$), and increases with $a = 1$. If the state reaches 1 (formed a habit), it stays there for some duration, independent of the action. (ii) **Motivation based**: State drops over time. If we act, the state returns to baseline. (iii) **Random**: Random state independent of the action.

Drop rates, increase rates, habit duration, and state baselines include agent specific noise, drawn uniformly at random (UaR). Trajectories used for testing have *higher noise* than the ones used for training (see supplement). Historical trajectories are produced by simulating the participants under various simple intervention policies. Specifically, we act i times (drawn UaR in $[6, 24]$), every j timesteps (also drawn UaR in $[1, 14]$). Finally, every participant is associated with a *noisy* 'demographic' feature related to their type. These features are given as an additional input to our model (see Figure 5). Also, they are used by the Whittle index baseline, which learns (offline) empirical transition probabilities for each type of agent, and then uses the features online to map each arm to the corresponding probabilities.

5.3 Real Data on Maternal and Child Healthcare

We use data from a large-scale maternal and child healthcare program operated by our partner NGO, ARMMAN. The program serves pregnant women and early mothers in disadvantaged communities with median daily family income of $3.22 – below the global poverty line [TheWorldBank, 2023] – by disseminating timely health information (via *automated* voice calls) to reduce maternal, neonatal, and child mortality and morbidity. The main challenge the program faces is drop in engagement over time. Engagement is measured in terms of total number of automated voice messages listened. To mitigate this problem, a planner schedules support calls by *limited* healthcare workers.

We model this setting as a continuous state, fully observable RMAB problem. The state of each beneficiary represents the listening time. Each automated voice message has a maximum length of 120 seconds, which we normalise to $[0, 1]$. The planner's task is to recommend a subset of beneficiaries every week to receive support calls from healthcare workers, with the goal to maximize the number of beneficiaries above the engaging threshold s^*. ARMMAN considers a beneficiary to be engaging if they listen to more than 30 seconds of the automated message (i.e., if $s > 0.25 = s^*$). Transition dynamics are unknown, and we make no Markov assumptions. Finally, the dataset also includes socio-demographic features per beneficiary such as age, gestational age, family income, education, etc., that may be used as auxiliary information.

Training Dataset

We use historical data from a large-scale quality improvement study performed by ARMMAN in 2022, obtained with beneficiary consent. The data follows 12000 participants (11256 with complete state information by the end) over a period of 31 weeks. In the study, a set of beneficiaries received interventions from a variety of policies (see supplement for details). Each beneficiary is represented by a single trajectory of states (engagement behavior) and actions (received, or not a call from a healthworker). Demographic features are used to infer the missing transition dynamics for the Whittle index baseline, as in [Verma *et al.*, 2023].

Notice on Data Usage

Our simulations are a secondary analysis on different evaluation metrics. All data are anonymized, and we have received approval from ARMMAN's ethics board. There is no actual deployment of TARI at ARMMAN.

5.4 Time Series Forecasting Models

We implemented a variety of time series forecasting models, ranging from simple LSTM, and BiLSTM architectures, to adding time-based vector representations and attention layers, to Transformer models. The majority of the models showed high performance. As such, we opted to use an LSTM-based architecture, as simpler models – being less computationally intensive and more sustainable in the long run – would be preferred by NGOs which are operating in a low resource environment. Our chosen model achieved MAE of 0.03 on average for one step prediction on synthetic data (excluding random agents), and 0.20 on real data. We used a history of $h = 7$ timesteps as input for the synthetic data, and $h = 8$ timesteps for the real data evaluation. For detailed results, including long horizon forecasts, please see the supplement. It is important to note that (i) as discussed in Section 1.2, our contribution is not in developing or improving SOTA TSF models, and (ii) the proposed TARI depends on relative trends (ratio of two predicted trajectories, see Equation 1), thus if the model consistently over- or under-predicts the two trajectories, the trend (ratio) will be consistent.

Counterfactuals in Non-Markovian Settings

Under no Markovian assumptions, we can not build a beneficiaries' model to compute counterfactual trajectories for the evaluation on real data. To compute such counterfactuals, we employed a *separate* TSF model trained on the *entire* dataset (train, validation, and test data). This model is *only used when trajectories deviate* from the historical data. Given that we only act on a small percentage of the beneficiaries (about 2%), the vast majority of the trajectories follow the historical real data, thus the model will not be used. Note that this second model is only needed for the purposes of the simulation. This is a fully observable RMAB, thus in real life we would actually observe the next state of each participant. We have evaluated an alternative approach to computing counterfactuals in the supplement.

6 Results on Synthetic Data

Engagement. We run a long-term simulation of one year (52 timesteps). Starting with the engagement, Figure 6 shows the percentage of engaging beneficiaries during the time horizon. TARI achieves 44.2% higher engagement on average (across timesteps and independent runs), and up to 107.3%, compared to the Whittle index policy (best baseline). Compared to Round-robin and Random it achieves 102.5% and 128.6% higher engagement on average, respectively. Additionally, TARI achieves significantly lower standard deviation (3.1%), compared to over 100% for all other baselines, as TARI never intervenes on Random agents.

Robustness. We evaluated the proposed approach under varying number of arms ($N = \{30, 90, 120, 600\}$) and budgets ($k = \{0.01, 0.1, 0.2\} \times N$, for $n = 90$). In both cases, TARI significantly outperformed all the baselines. Specifically, compared to the Whittle index (best baseline), TARI achieved on average $44.2 - 48.2\%$ higher engagement when varying the number of arms, and $37.8 - 88.7\%$ when varying

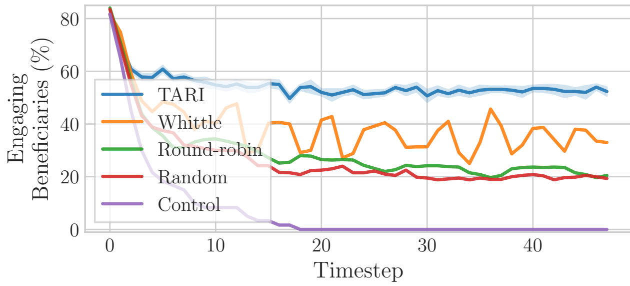


Figure 6: Synthetic data, $N = 90$ arms, budget $k = 9$. Percentage of engaging beneficiaries (excluding the random agents). The proposed approach achieves 44.2% higher engagement on average, and up to 107.3%, compared to Whittle index. Note that we opted not to show standard deviations for the baselines for better visualization, due to the high values (contrary, TARI achieves low s.d. of 3.1%).

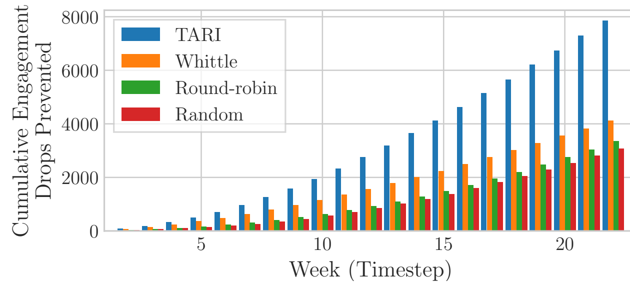
the budget (the smallest budget corresponds to just one intervention, hence the lower improvement). Furthermore, we enhanced the model for the Whittle index baseline to incorporate history ($s' = \times_{i=1}^{h} s_i$, for up to 4 past states), and a more fine-grained discretization of the state (up to 9 bins) – recall that the Whittle index baseline is not designed for continuous states, and thus requires discretization. In all of the cases, TARI achieved at least 44.2% higher engagement on average (and up to 144.2%). For detailed results, please see the supplement.

7 Results on Real Data on Maternal and Child Healthcare

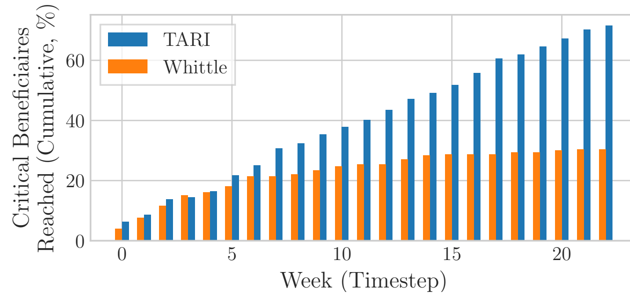
We run our analysis for more than 5 months (31 weeks of data, minus the 8 weeks we give as input to the TSF model, which corresponds to the observation period of ARMMAN). Due to limited resources, ARMMAN is able to provide support calls to about $2 - 4\%$ of beneficiaries at each week (timestep). In our dataset of 2252 beneficiaries (in the test set), the lower number corresponds to just 46 support calls per week. All policies were evaluated in the entire test set.

Figure 7a depicts the cumulative engagement drops ($s < s^* = 0.25$) prevented, compared to control. This figure clearly demonstrates the importance, and difficulty of scheduling effective support calls: the naive *round-robin approach (often used by NGOs [Mate* et al.*, 2022]) performs similarly to random*, i.e., following such a policy would incur all the cost associated with providing support calls, without any benefit. TARI achieves an 90.8% *improvement over Whittle*, and 133.9% over round-robin. Ensuring beneficiaries remain *consistently* engaged is crucial for the success of any healthcare program.

Real-world Significance. To put the real-world significance of TARI's engagement gains into concrete numbers, this corresponds to 3736 and 7851 additional messages listened compared to the Whittle index, and Control, respectively. Or, in other words, 16.3 *hours of additional content listened by the beneficiaries compared to the Whittle index, and* 62.6 *additional hours compared to Control*. Adhering to the program improves health literacy, which would ultimately lead to better health outcomes.



(a) Cumulative engagement drops prevented, compared to no intervention. TARI achieves an 90.8% *improvement over Whittle*.



(b) Cumulative percentage of high dropout-risk beneficiaries reached. TARI achieves a 135.2% *improvement over Whittle*.

Figure 7: Real data on maternal and child healthcare awareness from ARMMAN. $N = 2252$ beneficiaries (arms), budget $k = 0.02 \times N = 45$. Results are averaged over 10 independent runs.

Finally, an important open problem for ARMMAN is identifying and proactively reaching beneficiaries with high risk of dropping out from the program ('critical beneficiaries', see supplement). Figure 7b shows the cumulative percentage of critical beneficiaries that TARI, and Whittle chose to intervene. By the end of our observation period, TARI had reached 71.6% *of critical beneficiaries*, while Whittle only 30.4%, a 135.2% *improvement*. Please see the supplement for results on engagement, varying budgets, and more.

8 Conclusion

In this work, we study for the *first* time non-Markovian Restless Multi-armed Bandits (NMRMAB). Using real-world data on a maternal health awareness program from our partner NGO, ARMMAN, we demonstrate significant deviations from the Markov assumption. To solve the challenges that arise, we model arms as time-series, and propose the Time-series Arm Ranking Index (TARI) policy, a novel algorithm that selects the arms that will benefit the most from an intervention, given our future state predictions. Our evaluation shows a significant increase in engagement compared to the SOTA, deployed Whittle index solution, on real data from ARMMAN with 2252 participants. This translates to 16.3 hours of additional content listened, 90.8% more engagement drops prevented, and reaching more than twice ($\times 2.35$) as many high dropout-risk beneficiaries. While we focus on maternal and child healthcare as an indicative application, we expect the proposed approach to perform well on any other RMAB application that involves non-Markovian behaviours.

Acknowledgments

P.D. was funded in part by JPMorgan Chase & Co. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by JPMorgan Chase & Co. or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material should not be construed as an individual recommendation for any particular client and is not intended as a recommendation of particular securities, financial instruments or strategies for a particular client. This material does not constitute a solicitation or offer in any jurisdiction.

J.A.K. was supported by an NSF Graduate Research Fellowship under grant DGE1745303

References

[Ariyo *et al.*, 2014] Adebiyi A Ariyo, Adewumi O Adewumi, and Charles K Ayo. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, pages 106–112. IEEE, 2014.

[ARMMAN, 2023a] ARMMAN. Maternal and newborn health. https://armman.org/maternal-newborn-health/#1552759270935-9f057600-22c8, 2023. [Online; accessed 8-February-2023].

[ARMMAN, 2023b] ARMMAN. mmitra. https://armman.org/mmitra/, 2023. [Online; accessed 8-February-2023].

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[Bai *et al.*, 2018] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[Buchanan, 1991] James M. Buchanan. *Opportunity Cost*, pages 520–525. Palgrave Macmillan UK, London, 1991.

[Chevillon, 2007] Guillaume Chevillon. Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4):746–785, 2007.

[Chierichetti *et al.*, 2012] Flavio Chierichetti, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. Are web users really markovian? In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 609–618, New York, NY, USA, 2012. Association for Computing Machinery.

[Danassis *et al.*, 2023] Panayiotis Danassis, Shresth Verma, Jackson Killian, Aparna Taneja, and Milind Tambe. Limited resource allocation in a non-markovian world: The case of maternal and child healthcare. *arXiv preprint arXiv:2305.12640*, 2023. https://arxiv.org/abs/2305.12640.

[Early *et al.*, 2022] Joseph Early, Tom Bewley, Christine Evers, and Sarvapali Ramchurn. Non-markovian reward modelling from trajectory labels via interpretable multiple instance learning. *arXiv preprint arXiv:2205.15367*, 2022.

[GatesFoundation, 2020] GatesFoundation. Global progress and projections for maternal mortality: Maternal deaths per 100,000 live births. https://www.gatesfoundation.org/goalkeepers/report/2021-report/progress-indicators/maternal-mortality/, 2020. [Online; accessed 8-February-2023].

[Ghosh *et al.*, 2023] Abheek Ghosh, Dheeraj Nagaraj, Manish Jain, and Milind Tambe. Indexability is not enough for whittle: Improved, near-optimal algorithms for restless bandits. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23. International Foundation for Autonomous Agents and Multiagent Systems, 2023.

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

[Graves and Schmidhuber, 2005] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[HelpMum, 2023] HelpMum. Preventing maternal and infant mortality in africa. https://helpmum.org/, 2023. [Online; accessed 8-February-2023].

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hsu, 2018] Yu-Pin Hsu. Age of information: Whittle index for scheduling stochastic arrivals. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2634–2638. IEEE, 2018.

[Kadota *et al.*, 2016] Igor Kadota, Elif Uysal-Biyikoglu, Rahul Singh, and Eytan Modiano. Minimizing the age of information in broadcast wireless networks. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 844–851. IEEE, 2016.

[Kaur *et al.*, 2020] Jasvir Kaur, Manmeet Kaur, Venkatesan Chakrapani, Jacqui Webster, Joseph Alvin Santos, and Rajesh Kumar. Effectiveness of information technology–enabled 'smart eating'health promotion intervention: A cluster randomized controlled trial. *PLoS One*, 15(1):e0225892, 2020.

[Kazemi *et al.*, 2019] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

[Lai *et al.*, 2018] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 95–104, 2018.

[Liu *et al.*, 2021] Minhao Liu, Ailing Zeng, Zhijian Xu, Qiuxia Lai, and Qiang Xu. Time series is a special sequence: Forecasting with sample convolution and interaction. *arXiv preprint arXiv:2106.09305*, 2021.

[Mate *et al.*, 2022] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12017–12025, 2022.

[Matsubara *et al.*, 2014] Yasuko Matsubara, Yasushi Sakurai, Willem G Van Panhuis, and Christos Faloutsos. Funnel: automatic mining of spatially coevolving epidemics. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 105–114, 2014.

[Meiss *et al.*, 2010] Mark R. Meiss, Bruno Gonçalves, José J. Ramasco, Alessandro Flammini, and Filippo Menczer. Agents, bookmarks and clicks: A topical model of web navigation. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, page 229–234, New York, NY, USA, 2010. Association for Computing Machinery.

[Meshram *et al.*, 2016] Rahul Meshram, Aditya Gopalan, and D Manjunath. Optimal recommendation to users that react: Online learning for a class of pomdps. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7210–7215. IEEE, 2016.

[Nishtala *et al.*, 2020] Siddharth Nishtala, Harshavardhan Kamarthi, Divy Thakkar, Dhyanesh Narayanan, Anirudh Grama, Aparna Hegde, Ramesh Padmanabhan, Neha Madhiwalla, Suresh Chaudhary, Balaraman Ravindran, and Milind Tambe. Missed calls, automated calls and health support: Using ai to improve maternal health outcomes by increasing program engagement. In *AI for Social Good Workshop*, 2020.

[Nisthala *et al.*, 2021] Siddhart Nisthala, Lovish Madaan, Aditya Mate, Harshavardhan Kamarthi, Anirudh Grama, Divy Thakkar, Dhyanesh Narayanan, Suresh Chaudhary, Neha Madhiwala, Ramesh Padhmanabhan, Aparna Hegde, Pradeep Varakantham, Balaram Ravindran, and Milind Tambe. Selective intervention planning using restless multi-armed bandits to improve maternal and child health outcomes. In *AAMAS workshop on Autonomous Agents for social good*, 2021.

[Papadimitriou and Tsitsiklis, 1994] Christos H Papadimitriou and John N Tsitsiklis. The complexity of optimal queueing network control. In *Proceedings of IEEE 9th Annual Conference on Structure in Complexity Theory*, pages 318–322. IEEE, 1994.

[Pfammatter *et al.*, 2016] Angela Pfammatter, Bonnie Spring, Nalini Saligram, Raj Davé, Arun Gowda, Linelle Blais, Monika Arora, Harish Ranjani, Om Ganda, Donald Hedeker, et al. mhealth intervention to improve diabetes risk behaviors in india: a prospective, parallel group cohort study. *Journal of medical Internet research*, 18(8):e207, 2016.

[Puterman, 2014] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Qian *et al.*, 2016] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, page 123–131, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.

[Sinha and Mahajan, 2022] Amit Sinha and Aditya Mahajan. Robustness of whittle index policy to model approximation. *Available at SSRN 4064507*, 2022.

[Taieb *et al.*, 2012] Souhaib Ben Taieb, Rob J Hyndman, et al. *Recursive and direct multi-step forecasting: the best of both worlds*, volume 19. Department of Econometrics and Business Statistics, Monash Univ., 2012.

[TheWorldBank, 2023] TheWorldBank. Fact sheet: An adjustment to global poverty lines. https://www.worldbank.org/en/news/factsheet/2022/05/02/fact-sheet-an-adjustment-to-global-poverty-lines, 2023. [Online; accessed 8-February-2023].

[UN, 2023] UN. Sustainable development goals. https://sdgs.un.org/goals/goal3, 2023. [Online; accessed 8-February-2023].

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Verma *et al.*, 2023] Shresth Verma, Gargi Singh, Aditya Mate, Paritosh Verma, Sruthi Gorantala, Neha Madhiwalla, Aparna Hegde, Divy Thakkar, Manish Jain, Milind Tambe, and Aparna Taneja. Increasing impact of mobile health programs: Saheli for maternal and childcare. In *Innovative Applications of Artificial Intelligence (IAAI)*, 2023.

[Wang* *et al.*, 2023] Kai Wang*, Shresth Verma*, Aditya Mate, Sanket Shah, Aparna Taneja, Neha Madhiwalla, Aparna Hegde, and Milind Tambe. Scalable decision-focused learning in restless multi-armed bandits with application to maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[Weber and Weiss, 1990] Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3):637–648, 1990.

[Wen *et al.*, 2022] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

[Whittle, 1988] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, 25(A):287–298, 1988.

[Wu *et al.*, 2020] Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion. Deep transformer models for time series forecasting: The influenza prevalence case. *arXiv preprint arXiv:2001.08317*, 2020.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

[Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.