

# Beyond Listenership: AI-Predicted Interventions Drive Improvements in Maternal Health Behaviours

Arpan Dasgupta <sup>a,\*<sup>1</sup></sup>, Sarvesh Gharat <sup>b,1</sup>, Neha Madhiwalla <sup>c</sup>, Aparna Hegde <sup>c</sup>, Milind Tambe <sup>a, d</sup> and Aparna Taneja <sup>a</sup>

<sup>a</sup>Google DeepMind

<sup>b</sup>IIT Bombay

<sup>c</sup>ARMMAN

<sup>d</sup>Harvard University

**Abstract.** Automated voice calls with health information are a proven method for disseminating maternal and child health information among beneficiaries and are deployed in several programs around the world. However, these programs often suffer from beneficiary dropoffs and poor engagement. In previous work, through real-world trials, we showed that an AI model, specifically a restless bandit model, could identify beneficiaries who would benefit most from live service call interventions, preventing dropoffs and boosting engagement. However, one key question has remained open so far: does such improved listenership via AI-targeted interventions translate into beneficiaries' improved knowledge and health behaviors? We present a first study that shows not only listenership improvements due to AI interventions, but also simultaneously links these improvements to health behavior changes. Specifically, we demonstrate that AI-scheduled interventions, which enhance listenership, lead to statistically significant improvements in beneficiaries' health behaviors such as taking iron or calcium supplements in the postnatal period, as well as understanding of critical health topics during pregnancy and infancy. This underscores the potential of AI to drive meaningful improvements in maternal and child health.

## 1 Introduction

Maternal mortality remains a major global health concern, particularly in developing countries, and its reduction is a critical target highlighted by the World Health Organization (WHO) under the Sustainable Development Goals (SDGs) [1]. Timely access to reliable health information plays a crucial role in improving maternal and infant health outcomes, particularly in underserved communities where traditional healthcare services, and access to resources may be limited. Mobile health (mHealth) programs [23, 10] have emerged as an effective way to bridge this gap in several public health domains, leveraging the widespread use of basic mobile phones to deliver essential health education at scale.

mMitra [3] is one such program conducted by the NGO ARMMAN [2]. mMitra is the second largest maternal mHealth program in the world. The program enrolls expectant and new mothers from underserved communities in Mumbai, Palghar, Pune and Nashik in India. Once enrolled, the beneficiary receives pre-recorded automated

voice calls every week throughout pregnancy and until one year post child birth. These messages are timed according to their gestational age and contain critical health information around pregnancy, antenatal and post natal care, child care practices, family planning and many other relevant critical topics. In fact, ARMMAN validated that regular listenership of mMitra automated voice call messages has a significant positive impact on not only improved awareness and self care and infant care practices among mothers, but also led to improved health outcomes for both the mothers and babies [11, 17].

However, despite the known benefits and outreach of the program, ARMMAN noticed a decline in beneficiary engagement both in terms of listenership and drop offs from the program. Adherence to public health programs is a well known challenge [12, 13, 18]. A typical solution to improve adherence to such programs is the use of targeted interventions. In particular, interventions such as calls or visits from community health workers can help keep beneficiaries engaged by providing personalized support and encouragement. ARMMAN, deployed a similar intervention whereby health workers would reach out to beneficiaries via live service calls to encourage them to listen to messages regularly, as well as explain to them the importance of listening to the messages. However, due to limited support staff, it is not feasible for the health workers to reach out to every beneficiary, making it essential to prioritize and identify beneficiaries who would benefit the most from receiving an intervention at a given time.

This is essentially a limited resource allocation problem which involves complex sequential decision making, to identify beneficiaries every week who are at risk of drop off the program and may benefit from receiving an intervention. Restless multi-armed bandits (RMABs) have been shown to be an effective approach to address such sequential resource allocation problems [20, 15]. In fact, the research team together with ARMMAN developed a RMAB model [26] to identify the beneficiaries that would most benefit from receiving an intervention at a given time. This model was thoroughly evaluated in the lab, followed by a real world field study [24] and then underwent further iterations and finally evolved into an algorithm that was deployed [25] by the NGO and used on a weekly basis. The model had in fact served **350,000** mothers until 2024. Despite, the RMAB model having demonstrated clear gains in listenership and reduced drop-off rates in [16, 26], a crucial gap remained in understanding whether this improved engagement translated into better knowledge comprehension and, consequently, changes in health im-

\* Corresponding Author. Email: arpandg@google.com

<sup>1</sup> Equal contribution.

proving behaviors.

Therefore, the primary objective of this paper is to demonstrate the impact of enhanced engagement due to AI-predicted interventions into measurable knowledge gains and facilitating improvements in behavioral and health outcomes. To achieve this, in collaboration with the NGO, we designed and conducted a randomized controlled trial (RCT), with two arms, and a total of over 34,000 beneficiaries. The intervention arm received additional support through interventions predicted by the RMAB model, while the control arm did not. Both groups, however, continued to receive standard weekly automated voice calls. The impact of the AI interventions was determined by comparing the knowledge and behavioral outcomes between these two groups of beneficiaries.

Although numerous commendable public health programs exist that enhance adherence and awareness across a variety of health topics [13, 18, 12], few [28] have been able to establish the impact of AI interventions on health outcomes. In this paper, we establish statistically significant improvements in both the awareness and health behaviors of beneficiaries due to the AI-targeted interventions (Table 1), e.g., our AI (RMAB) interventions led to significantly enhanced iron and calcium intake post-delivery.

An additional major contribution of this paper is the experimental methodology for end-to-end evaluation of health outcomes of AI interventions. Careful evaluation of AI interventions in the field is extremely time consuming, and in our case, given sequential decision making in the interventions, how to conduct such an evaluation is itself not clearly known. We present a novel approach, in Section 3, which includes careful comparison of the Whittle index based interventions in restless bandits with corresponding “dummy Whittle index” based control group for reliable and robust counterfactual reasoning. This requires a thorough and detailed experimental methodology which we present in detail in our paper.

Moreover, while this study was conducted within the specific context of the mMitra program, we believe the demonstrated success in converting increased engagement into measurable knowledge and behavioral gains offers a valuable model. Similar public health programs, whether focused on preventative care, managing chronic diseases, or promoting lifestyle modifications, could harness AI-targeted interventions to improve participant engagement and knowledge retention, and ultimately, the adoption of healthier behaviors and achievement of better health outcomes thereby amplifying their impact on public health, particularly across marginalized communities with limited access to resources.

## 2 Related Work and Background

**RMABS:** The allocation of finite resources is a recurring challenge across various domains necessitating strategic planning. Restless multi-armed bandits (RMABs) serve as a prevalent instrument for addressing such sequential resource allocation problems within uncertain environments and have demonstrated their utility in applications such as anti-poaching surveillance [20], multi-channel communication optimization [15], task scheduling [4, 30], unmanned aerial vehicle routing [31], and others. In collaboration with ARMMAN, Mate et al. [16] presented a RMAB based approach to determine the allocation of service calls, where each beneficiary is modeled as a Markov decision process. This initial model was evaluated thoroughly in a field study Mate et al. [16] and deployed [25].

**Decision Focused Learning (DFL):** While the above approach optimized for predictive accuracy to learn model parameters, followed

by solving a separate optimization problem using the learnt parameters. Decision Focused Learning on the other hand, optimizes for decision quality [21], by integrating these two stages into an end-to-end system. In particular, focusing on how well predictions lead to effective actions rather than solely on the accuracy of the predictions themselves. The previous model with ARMMAN was then evolved into a DFL based model [26] to improve the listenership in the mMitra program via DFL predicted interventions which was again evaluated in a field study and eventually **deployed** [24].

**Whittle Index** is a heuristic policy tool [27] used primarily for solving RMABs, which are notoriously complex sequential resource allocation tasks. The Whittle Index assigns a numerical value (the index) to each arm, representing the benefit of acting on that arm, naturally providing a ranking or prioritization of which arms should be acted on for maximal benefit, following the budget constraints.

**Behavioral and Health Outcomes:** To date, the primary observable objective optimized by the deployed model and similar intervention scheduling programs for ARMMAN has been the mother’s listenership of automated voice calls; consequently, program performance has consistently been evaluated through improvements in listenership metrics. However, a correlation between AI-scheduled interventions and behavioral outcomes had not been demonstrated, until [9] which showed a preliminary investigation and positive trends in some questions in a health study. However, it failed to establish concrete statistical significance due to several limitations. This paper establishes more conclusive results by improving upon the methodology of conducting the survey, asking more targeted questions, as well improvising the evaluation methodology via a more robust and reliable counterfactual comparison.

## 3 Study Setup

Our initial methodological step involved segmenting registered beneficiaries into distinct cohorts, categorized by their respective program enrollment dates. Within each cohort, we then implemented a randomized allocation procedure, creating intervention and control groups. It is important to note that all participants, regardless of group assignment, continued to receive the automated voice messages every week containing essential health information as per schedule. However, only those assigned to the intervention group were considered for supplementary interventions. The DFL algorithm played a crucial role in determining, on a weekly basis, which intervention group members would receive personalized live service calls from health workers. Finally, to evaluate the impact of these interventions on behavioral and health knowledge, the NGO conducted a comprehensive survey on representative subsets of both the intervention and control groups.

### 3.1 Cohorts

The study was conducted in three cohorts with a combined number of 34453 beneficiaries, with details on the number of enrolments and respective timelines in table 2. As explained later in Section 3.3.1, the program treats these cohorts as entirely distinct groups. And they also serve primarily as a mechanism for establishing intervention eligibility timelines.

### 3.2 Experiment Arms

Within each cohort, beneficiaries were randomly assigned to either the intervention or control arm, ensuring a similar distribution of key

Question	Percentage improvement	p-Value
“Did you take iron pills after delivery?”	<b>21.74 %</b>	<b>0.0981</b>
“Did you take calcium pills after delivery?”	<b>28.00 %</b>	<b>0.0413</b>
“What was the baby’s weight at birth?”	<b>8.97%</b>	<b>0.0080</b>

Table 1: Key improvement on health knowledge and behaviour questions due to AI (RMAB) interventions with p-Values

Cohort	Beneficiaries Registered	Date of Registration	Timeline of Intervention
1	12749	1st Oct to 31st Oct 2023	5th Feb to 3rd Mar 2024
2	9122	1st Nov to 30th Nov 2023	4th Mar to 24th Mar 2024
3	12582	1st Dec to 31st Dec 2023	25th Mar to 21st Apr 2024

Table 2: Registration and Intervention Timelines and details of the three cohorts

attributes between the two groups as described below. This procedure mirrors covariate adaptive randomization, a technique designed to balance the distribution of relevant covariates across experimental groups [14]. We specifically balanced the distribution across arms based on gestational age to ensure similar distribution of expectant women and women post birth, as well as listenership behaviours as described below.

### Engagement States

- For each beneficiary and a given automated voice call, we denote the engagement state  $E@T$  at a threshold  $T$  as  $E@T = 1$  if the beneficiary listened to the call for at least  $T$  seconds, and  $E@T = 0$  otherwise.
- We compute  $E@T_w$  for each beneficiary, representing the engagement state over  $w$  weeks leading up to the cohort’s anticipated intervention start date.
- To ensure comparable listenership patterns between the two arms, we aim to achieve approximately equivalent values of  $E@T_w$  for thresholds  $T \in \{1, 5, 10, 30, 100\}$  and time windows  $w \in \{1, 2, 3\}$ .

Based on previous studies and discussions with the NGO [25], we observed it was critical to ensure a balanced distribution of above attributes (based on gestational age and  $E@T_w$ ) to avoid any biases in the cohort distribution and ensure a fair evaluations of both arms. This was achieved, by constructing a feature vector  $Y$  for each beneficiary by concatenating the above attributes. Subsequently, we partition the beneficiaries into two equally sized groups, employing  $Y$  as the stratification criterion [26]. This is accomplished by treating  $Y$  as a categorical label and utilizing a stratified splitting mechanism to produce two balanced subsets. Specifically, we leverage the stratified option within the `train_test_split` function from the sklearn library [19]. Given the sufficient number of beneficiaries within each cohort, we successfully achieve an exact split, resulting in perfectly balanced groups.

### 3.3 Conducting AI predicted Interventions

Interventions in mMitra are conducted via live service calls made by healthcare workers to encourage regular listenership of the automated voice messages.

#### 3.3.1 Number of Interventions per week.

Interventions began on February 2024. Interventions were only given to beneficiaries that had been present for at least 3 weeks in the program. We perform interventions on beneficiaries of cohort 1, 2 and 3

one after the other, intervening on a total of 35% of beneficiaries in each cohort to simulate the realistic scenario of intervening on only a subset of the beneficiaries, with exact respective timelines of each cohort detailed in Table 2.

The NGO discouraged repeated calls if a woman’s listening behavior remains unchanged. This consideration makes the model more aligned with real-world human behavior, ultimately leading to a more effective and respectful intervention strategy. Hence, both in the deployed model and this study each beneficiary can be intervened on only once.

The NGO conducted approximately 330 interventions per week for cohorts 1 and 3 and about 285 for Cohort 2. This was done to keep the approximate proportion of interventions similar for each week, and since Cohort 2 had lesser registrations, the number of weeks is lesser. In total about 12000 interventions were conducted.

#### 3.3.2 Eligibility for Interventions.

Beneficiaries are deemed eligible for interventions when the following criteria are met:

1. Active Program Status: They are currently receiving automated voice messages.
2. Recent Engagement: They have answered at least one automated voice message within the four weeks preceding the intervention period (to ensure phone number is still active/valid).
3. No Prior Intervention: They have not previously received an intervention.

Based on our discussion with the NGO, these eligibility requirements serve to ensure the efficient and equitable allocation of the program’s constrained intervention resources.

#### 3.3.3 Conducting Interventions.

Each week, the DFL [26] algorithm, determines the set of beneficiaries from the intervention arm who will receive an intervention, which we refer to as the list  $I_D$ . We also simulate the DFL algorithm on the control (dummy) arm to determine the set of beneficiaries that would have been selected for an intervention (assuming we conducted the same number of interventions as in the intervention arm). As in the intervention arm, beneficiaries in the control arm cannot be selected multiple times. These beneficiaries from the control arm are stored into a list  $I_C$ . The intervention list  $I$  refers to the combined lists  $I_D$  and  $I_C$ . The idea behind this is that we later compare the behavior of beneficiaries from  $I_D$  and  $I_C$ , as we can think of beneficiaries from  $I_C$  as the counterfactual counterparts of those from  $I_D$ , inspired by Boehmer et al. [6] which established this method of evaluating index-based policies in a statistically meaningful way. The interventions concluded on 21st April 2024.

### 3.4 Health Knowledge and Behavioral Assessment Process

#### 3.4.1 Survey Questions Design and Objective

The health knowledge and behaviour assessing survey contained 23 questions (Table 4 and Table 5), designed by the NGO to evaluate their knowledge across various health-related topics. This assessment is aligned with the content of the automated voice messages, aiming to gauge the beneficiaries' comprehension of the delivered information. Specifically, the survey encompassed categories such as program engagement, breastfeeding, micro nutrient supplementation, perinatal care, early infancy care, family involvement in maternal health and perceived benefits of the program. As generic maternal and child health information is widely available, the survey focused on those messages which are unique to the mMitra program. Table 5 and Table 4 provide a comprehensive list of these questions. Some questions have an Ante-natal Care (ANC) equivalent which implies asking about care pre-delivery. For each question, beneficiaries received a score based on their responses.

Among the 23 questions, 13 were structured as single-choice (Yes/No) questions. The remaining 8 were questions where scores were determined based on multiple possible correct answers, with varying weights assigned to different answers. Beneficiaries were offered the choices, and could select one or more of these answers, and their final score was the cumulative score of all correctly identified answers within that question.

Again, a key difference compared to the previous study ([9]) was the way the questions were structured by the NGO, to reduce ambiguity in questions and responses, thereby ensuring a more reliable way to quantitatively evaluate performance across the two groups.

#### 3.4.2 Conducting the Survey

The surveys were then conducted by the NGO on the beneficiaries from the intervention list  $I$  between 20th June to 30th October in 2024. We also excluded from  $I_C$  and  $I_D$ , the beneficiaries who had given birth before the interventions (to be able to conduct post-natal surveys). Additionally, among of all the people who were scheduled to receive interventions in  $I_D$ , only a fraction actually answered the intervention call (say this set is  $I_{D'}$ ). This combined set of  $I_C$  and  $I_{D'}$  is then considered for the survey.

This is a key difference from the previous study [9] where the authors surveyed and included the entire set  $I_D$  irrespective of whether the intervention actually happened or not, thereby impacting the reliability of the comparison, since  $I_D$  then contained both people who had and had not answered the intervention call.

In this study, the subset of beneficiaries  $I_{D'}$  and  $I_C$  are then called by a health worker to answer the questions from the survey. Further, to ensure fairness, the interviewers were blinded and had no idea whether the interviewee belonged to Control or intervention group. However, again, the survey calls may not be answered by all beneficiaries. This makes it difficult to evaluate the outcome of the study as we know the survey results for only a subset of the beneficiaries that are willing and available to answer to the survey questions (in particular, this group of beneficiaries is not chosen uniformly at random). Hence, we have to re-balance the control and intervention group for the final comparison. The exact numbers at each stage of the study can be seen in Table 3.

## 4 Results

### 4.1 Comparison Methodology

As explained above in Section 3.4.2, a direct comparison of the people finally surveyed (control vs intervention) does not suffice for comparison, since not everyone chosen for intervention ( $I_D$ ) and/or survey ( $I$ ), ends up answering the call, and this set of people is non-random, as described in Section 3.4.2. This motivates the need for a more robust and fair comparison between the two groups.

Prior work [9] selected survey respondents from the intervention group ( $I_D$ ) and formed a matched control group by minimizing Mahalanobis distance using manually selected features like pre-intervention listenership, gestational age, and number of previous children. Firstly, this approach doesn't account for imbalances from those who never answered the intervention call potentially diluting the impact of the interventions (as highlighted in 3.4.2 and secondly, they employed handcrafted features for counterfactual matching, which could again potentially impact the comparison's reliability. Consequently, they were unable to establish conclusive impact of AI interventions on behavioral outcomes.

In this study, however, we instead perform counterfactual matching of beneficiaries using the Whittle Index [27], which is the index or ranking used to determine who gets an intervention by the DFL-RMAB problem, guiding resource allocation by prioritizing arms that offer the most expected marginal reward. The matching is done by greedily matching the Whittle index across the control and intervention arms, making sure that the difference is below a threshold of 0.01.

Whittle indices are computed both for intervention and control group, but the actual interventions only go out to the intervention group. For the control group, the whittle indices only help us identify beneficiaries who would have been intervened on by the model had we conducted interventions on that group. The advantage of using this mechanism is that, firstly, Boehmer et al. [6] demonstrated, this approach to be a statistically valid mechanism of evaluating a RCT involving an index based policy. Furthermore, the Whittle index generated by the DFL algorithm [26] inherently learns and weighs relevant demographic features and listenership behaviors in the prediction process. Given that interventions are ultimately assigned by ranking these Whittle indices, the index value itself signifies the algorithm's preference for intervening on a particular beneficiary. This implies a consistent preference for beneficiaries with similar index values, regardless of whether they are in the intervention or control group. Therefore, to ensure a robust and reliable counterfactual comparison, beneficiaries in the intervention set ( $I'_D$ ) who answered the survey were matched with a beneficiary with a similar Whittle index in the control set ( $I_C$ ) in the same cohort and who answered the survey. This is a **key methodological improvement that led to a more stable and reliable counterfactual comparison**.

### 4.2 Key Results

#### 4.2.1 Establishing Improved Listenership and Engagement

Towards the goal of assessing improvement in knowledge and behavioral outcomes across the two groups, we first evaluate whether the intervention successfully enhanced listenership across the two groups. The gain in listenership is quantified as the difference between post-intervention listenership and pre-intervention listenership, which is calculated by taking the average listenership per user

	Registered	After removing inactive	Intervened On [About 35%] (for Intervention)	Picked up Intervention (for Intervention)	After imposing delivery-date constraint (delivery after intervention : check section 3.4.2)	Picked up Survey
DFL	12749 + 9122 +		4495	3469	1496	701
Control	12582 = 34453	19970	4495	4495	1901	850

**Table 3:** Beneficiary Counts at Different Stages

Question	Correct Answer	Control Score	Intervention Score	p-Value
Did you know your baby's birth weight ?	Yes	0.991 ± 0.004	0.993 ± 0.003	1
How many iron tablets did you consume during pregnancy?	Recommended or more	0.836 ± 0.017	0.827 ± 0.018	0.776818
How many calcium tablets did you consume during pregnancy?	Recommended or more	0.836 ± 0.017	0.834 ± 0.018	0.99313
What is the name of the pill to be taken for preventing anemia?	Iron Pill	0.457 ± 0.023	0.431 ± 0.023	0.473308
What is the name of the pill to be taken for strengthening bones?	Calcium Pill	0.494 ± 0.023	0.497 ± 0.024	0.984697
<b>Are you currently taking iron tablets after delivery?</b>	<b>Yes</b>	<b>0.234 ± 0.019</b>	<b>0.283 ± 0.021</b>	<b>0.098111</b>
<b>Are you still taking calcium pills after delivery?</b>	<b>Yes</b>	<b>0.244 ± 0.019</b>	<b>0.306 ± 0.022</b>	<b>0.041262</b>
When you faced anxiety, did you share it with your husband?	Yes	0.1 ± 0.014	0.106 ± 0.015	0.877882
Should the baby be fed the first solid yellow milk that comes to the mother after childbirth?	Yes	0.881 ± 0.015	0.87 ± 0.016	0.665967
Should the baby be fed at night?	Yes	0.982 ± 0.006	0.962 ± 0.009	0.102588
How does the baby react when you talk to it (at 3 months)	Gives a Social Smile	0.697 ± 0.021	0.724 ± 0.021	0.406006
Have you heard calls from mMitra regularly?	Yes	0.713 ± 0.02	0.728 ± 0.021	0.662864
Have you ever discussed with your husband/family about the information you heard/told in the call?	Yes	0.258 ± 0.02	0.274 ± 0.021	0.633354

**Table 4:** Scores on single correct questions along with expected answers, standard error and p-values for cohorts 1 and 2.

over eight weeks post intervention and per user over eight weeks pre-intervention respectively. It was established by the NGO and prior work Dasgupta et al. [9], Verma et al. [25], that listenership typically gradually reduces over time, motivating the need for interventions.

We observe the same in Figure 1 which shows a general trend of decreasing listenership in the control group by 8.5%, in comparison to the intervention arm which gains by 3.6%, establishing that interventions led to reduced drop in listenership. The x-axis represents the week number, indicating the point in time at which the listenership gain is calculated. The y-axis represents the cumulative listenership gain in seconds, computed as the cumulative sum of the mean listenership gain for all beneficiaries who were intervened (or would have been intervened) in a given week. This approach ensures that for each week, we only consider the listenership gain of the relevant beneficiaries, providing a clear week-wise trend of the intervention’s impact.

While listenership gradually reduces over time across both groups (as also shown in [25, 16]), however, as seen in Figure 1, across all weeks, we observe a general trend where the intervened group’s listenership drop is significantly lesser than that of the control group.

Interestingly, Cohort 3 shows an overall improvement in listenership in the control group compared to Control group of Cohorts 1 and 2. However, the gain still remains lower than the corresponding Cohort 3 in Intervention group, showing that the interventions have still had a significant positive impact on listenership among all cohorts.

It must be noted that the final goal of this study is to establish the benefit of the AI predicted interventions on health and behavioral outcomes. Hence, the comparison here is done only against the baseline with no-intervention. The efficacy of different AI methodologies in predicting interventions for improving listenership was already demonstrated in [16] and [26] and DFL was shown to significantly outperform several baselines and was hence deployed and used in this study.

#### 4.2.2 Establishing Impact on Health Knowledge and Behavioral Outcomes

The final objective of the study is to assess the improvements in health related knowledge and behaviours due to the AI predicted in-

terventions. Hence, the responses to the survey questions were compared across the single and multiple choice questions across both the intervention and control group. Key findings from the study are listed below, focusing in particular, on the improvement in health-related knowledge and behaviours among beneficiaries in Cohorts 1 and 2.

The complete list of comparison of response scores from all single choice and multiple choice questions across the intervention and control group for Cohort 1 and 2 can be found in Table 4 and Table 5 respectively, and for all cohorts combined in the Appendix. While we observe positive trends in several questions, many of these outcomes are not conclusive due to statistical insignificance. However, below we highlight the results for three questions, for which we observed statistically significant improvements in the intervention group.

#### Key Findings

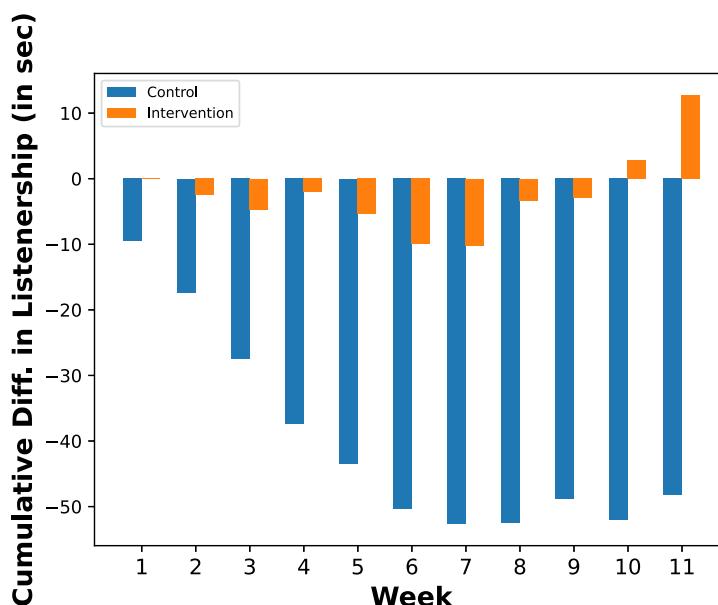
1. **"Are you still taking iron pills after delivery?"**
2. **"Are you still taking calcium pills after delivery?"**
3. **"What was the baby's weight at birth?"**

**Quantitative analysis** The improvements on these questions are shown in the Appendix, where the x-axis represents the week, while the y-axis denotes scores on that question. Here, cumulative weeks refer to the inclusion of all beneficiaries who received interventions up to a particular week. For example, in Cumulative Week 3, we consider all beneficiaries who were intervened in Weeks 1, 2, and 3. Similarly, for the control (dummy) arm, although no actual intervention takes place, we calculate scores on the beneficiaries identified in  $I_C$  until that week, plotted against the equivalent intervened arm beneficiaries from  $I_{D'}$ . The Y-axis represents the average score for a particular question corresponding to the respective groups. This score is derived from beneficiary responses to the specific knowledge-based questions asked.

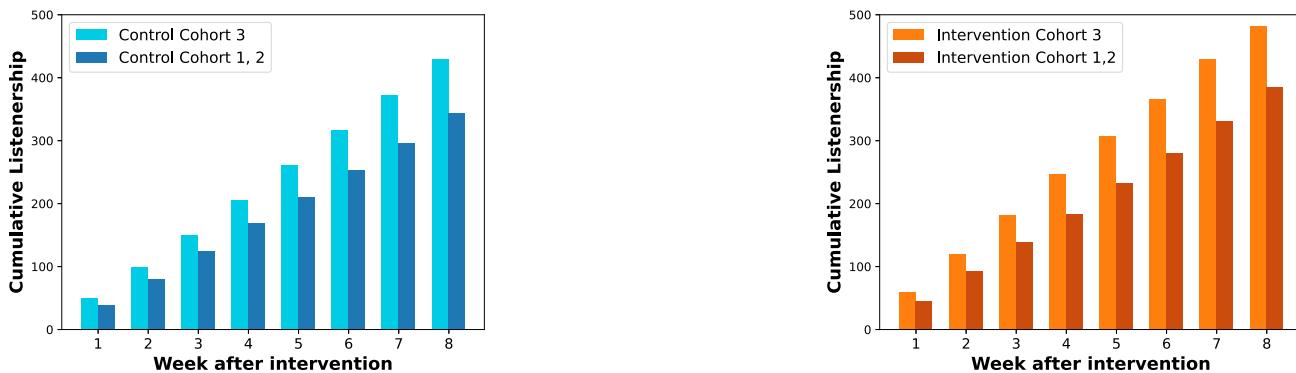
Furthermore, as seen in figures attached in the Appendix, we observe that towards the end of Cohort 2, the cumulative score differences between the intervened and control groups increase, with the intervened group performing better. These results become more evident as we also observe a significant decrease in the p-values for Cumulative Weeks 6 and 7. Table 1 further shows the overall improvement on these three questions and the corresponding p-value

Question	Correct Answer (s)	Control Score	Intervention Score	p-Value
If yes what was the baby's weight at birth? (Follow up to knowing weight)	'It was 2.5' or 'It was more than 2.5'	$0.766 \pm 0.019$	$0.836 \pm 0.018$	<b>0.007991</b>
If yes, what did the husband do? (Follow up to telling husband about anxiety)	'The husband took her to the doctor'	$0.059 \pm 0.011$	$0.074 \pm 0.012$	0.367425
When should the baby be fed the first milk after birth?	'Within 1 hour'	$0.633 \pm 0.022$	$0.645 \pm 0.023$	0.709437
What to do to dry baby's umbilical cord after birth?	'It should be kept clean and dry' or 'Nothing should be applied'	$0.545 \pm 0.023$	$0.575 \pm 0.023$	0.353922
If yes, what made you want to listen to the call? (Follow up to question about if they listen to mMitra calls)	'Doubts clarified', 'Endorsement', 'Guidance', 'Knowledge of physiological changes', 'Fills gaps left by medical advice'	$0.32 \pm 0.013$	$0.331 \pm 0.014$	0.596208
Does anyone in your family know that you are getting informative calls from mMitra?	'Husband', 'In-laws', 'Mother', 'Relative'	$0.844 \pm 0.016$	$0.856 \pm 0.017$	0.611126
Do they hear these calls too?	'Yes , always', 'Yes, I put the mobile on speaker so that everyone in the family can hear the call', 'yes , sometimes' (partial points) 'No , but I tell them what the call was about' (partial points)	$0.302 \pm 0.014$	$0.293 \pm 0.015$	0.663034
Do you pick up mMitra calls even when other people are around?	'Yes', 'Always', 'Sometimes' (partial points)	$0.538 \pm 0.02$	$0.524 \pm 0.021$	0.630469
Do you pick up mMitra call even when you are busy at work?	'Yes', 'Always', 'Sometimes' (partial points)	$0.46 \pm 0.019$	$0.467 \pm 0.021$	0.795084

**Table 5:** Scores on single correct questions along with expected answers, standard error and p-values for cohorts 1 and 2.



**Figure 1:** Comparison of **Weekwise Listenership Gain** by comparing listenership behaviours post vs pre intervention for both Intervention and Control for all Cohorts. Drop in Intervention Group is significantly lower than Control indicating the impact of interventions in improving listenership behaviours of the Intervention Group.



**Figure 2:** Comparison of the cumulative average listenership of **Cohort 1 and 2 vs Cohort 3** after "intervention" for Control group (left) and Intervention Group (right). Both the plots indicate that the **base listenership for Cohort 3 was higher than Cohort 1 and 2**. achieved by week 7, emphasizing the statistical significance on these findings.

### Significance of these findings

- Calcium and Iron supplementation:** Improved performance on the questions regarding the continued intake of iron and calcium supplements after delivery—is particularly significant as it indicates enhanced knowledge and adherence to crucial postnatal supplementation guidelines. It has been noted that adherence to micronutrient supplementation in the postnatal period is very low compared to the antenatal period, primarily due to lack of counselling by health workers. Hence, this improved understanding is paramount for safeguarding maternal health during the demanding postpartum period [7]. For instance, consistent calcium supplementation is vital as it plays a key role in replenishing maternal bone mineral density, which can be depleted during pregnancy and lactation [8]. Similarly, adequate iron intake is critical; iron deficiency after childbirth not only negatively impacts the mother's cognitive functions, potentially affecting her ability to care for herself and her infant, but can also have detrimental effects on the baby's overall growth [5].
- Knowing the birth weight of the baby:** is significant because this information indicates if the mother has monitored a key health metric from birth. Birth Weight is a precise marker for the infant's health at birth and special home-care measures are required for infants who have been born with low birth weight. Hence, mother's ability to report birth weight is a critical indicator of knowledge [22]. As World Health Organization [29] emphasizes, birth weight is a critical indicator for neonatal survival and for broader public health monitoring. Moreover, when a mother knows the precise birth weight, it facilitates more effective health worker follow-ups, assisting them in monitoring the child's subsequent health and development.

Therefore, higher scores (Table 1) reflecting correct understanding and practice of iron and calcium supplementation and in recalling birth weight demonstrate a meaningful improvement in the mothers' grasp of vital health knowledge and health-promoting behaviors.

### 4.3 Impact of Listenership gain on Survey responses for Cohort 3

The above findings were established for Cohort 1 and 2. For Cohort 3 however, we cannot establish a statistically significant difference

between the intervention and control groups. This is primarily due to the reason that Cohort 3 already has a higher listenership on average (fig. 2) than Cohort 1 and 2, even in the Control Group, suggesting that any marginal further increase in listenership may not significantly impact the knowledge gained by the beneficiaries. This also happens because critical health information is repeated across messages at different times, thereby improving the likelihood of awareness around key topics when overall listenership is already good. A combined evaluation across all cohorts is provided in the Appendix.

## 5 Conclusion

This paper compellingly demonstrates that AI-based interventions, by successfully enhancing engagement with mHealth programs, can lead to statistically significant improvements in crucial health behaviors among beneficiaries, such as adherence to iron or calcium supplementation and improved understanding of critical health topics relevant to pregnancy and infancy. This not only highlights the potential of AI to drive meaningful advancements in maternal and child health but also illuminates a pathway for other public health programs. The evidence presented, which translates previously established listenership differences into tangible gains in knowledge, suggests that similar AI-driven strategies could be effectively adapted to boost outcomes in diverse health initiatives. For instance, programs targeting chronic disease management, vaccination uptake, Diabetes, nutritional education could leverage tailored AI interventions to improve participant engagement, knowledge retention, and ultimately, adoption of healthier behaviors and achievement of better health outcomes across various populations.

## 6 Ethical Considerations

It is important to highlight that the study was designed together with the NGO, relying on their domain expertise and was approved by the ARMMAN Ethics Review Board. It was registered on the Clinical Trials Registry India (CTRI) - CTRI/2023/02/049472. Verbal consent was obtained from the participants. The research team at Google only had restricted access to anonymized data and no PII data was used for the study.

## References

- [1] SDG Target 3.1 Maternal mortality — who.int. <https://www.who.int/data/gho/data/themes/topics/sdg-target-3-1-maternal-mortality>. [Accessed 31-05-2024].
- [2] Armmann Foundation. Armmann Foundation, 2024. URL <https://www.armmann.org/>. Accessed: March 6, 2024.
- [3] Armmann Foundation. mMitra Program, 2024. URL <https://armmann.org/mmitra/>. Accessed: March 6, 2024.
- [4] S. Bagheri and A. Scaglione. The restless multi-armed bandit formulation of the cognitive compressive sensing problem. *IEEE Transactions on Signal Processing*, 63(5):1183–1198, 2015.
- [5] J. L. Beard and J. R. Connor. Iron deficiency alters brain development and functioning in mothers and infants. *Nutrition*, 21(6):564–570, 2005. doi: 10.1016/j.nut.2005.01.019.
- [6] N. Boehmer, Y. Nair, S. Shah, L. Janson, A. Taneja, and M. Tambe. Evaluating the effectiveness of index-based treatment allocation. *arXiv preprint arXiv:2402.11771*, 2024.
- [7] L. L. Brown, B. E. Cohen, E. Edwards, C. E. Gustin, and Z. Noreen. Physiological need for calcium, iron, and folic acid for women of various subpopulations during pregnancy and beyond. *Journal of Women's Health*, 30(2):207–211, 2021.
- [8] G. M. Chan, K. Hoffman, and M. McMurry. Calcium supplementation improves bone mineral density in postpartum women. *American Journal of Obstetrics and Gynecology*, 190(4):1158–1164, 2004. doi: 10.1016/j.ajog.2004.01.028.
- [9] A. Dasgupta, N. Boehmer, N. Madhiwalla, A. Hedge, B. Wilder, M. Tambe, and A. Taneja. Preliminary study of the impact of ai-based interventions on health and behavioral outcomes in maternal health programs. *arXiv preprint arXiv:2407.11973*, 2024.
- [10] K. Gupta, S. Roy, R. C. Poonia, S. R. Nayak, R. Kumar, K. J. Alzahrani, M. M. Almfai, and F. N. Al-Wesabi. Evaluating the usability of mhealth applications on type 2 diabetes mellitus using various mcdm methods. *Healthcare*, 10(1), 2022. ISSN 2227-9032. doi: 10.3390/healthcare10010004. URL <https://www.mdpi.com/2227-9032/10/1/4>.
- [11] A. Hegde and R. Doshi. Assessing the impact of mobile-based intervention on health literacy among pregnant women in urban india. In *American Medical Informatics Association Annual Symposium*, page 1423, 2016.
- [12] J. A. Killian, M. Jain, Y. Jia, J. Amar, E. Huang, and M. Tambe. New approach to equitable intervention planning to improve engagement and outcomes in a digital health program: Simulation study. *JMIR diabetes*, 9(1):e52688, 2024.
- [13] S. Kinsey, J. Wolf, N. Saligram, V. Ramesan, M. Walavalkar, N. Jaswal, S. Ramalingam, A. Sinha, and T. Nguyen. Building a personalized messaging system for health intervention in underprivileged regions using reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*, 2023. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/668. URL <https://doi.org/10.24963/ijcai.2023/668>.
- [14] J. M. Lachin, J. P. Matts, and L. Wei. Randomization in clinical trials: conclusions and recommendations. *Controlled clinical trials*, 9(4):365–374, 1988.
- [15] K. Liu and Q. Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.
- [16] A. Mate, L. Madaan, A. Taneja, N. Madhiwalla, S. Verma, G. Singh, A. Hegde, P. Varakantham, and M. Tambe. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 12017–12025. AAAI Press, 2022.
- [17] N. Murthy, S. Chandrasekharan, M. P. Prakash, A. Ganju, J. Peter, N. Kaonga, and P. Mechael. Effects of an mhealth voice message service (mmitra) on maternal health knowledge and practices of low-income women in india: findings from a pseudo-randomized controlled trial. *BMC Public Health*, 20(1):1–10, 2020.
- [18] V. Nair, K. Prakash, M. Wilbur, A. Taneja, C. Namblard, O. Adeyemo, A. Dubey, A. Adereni, M. Tambe, and A. Mukhopadhyay. Adviser: Ai-driven vaccination intervention optimiser for increasing vaccine uptake in nigeria. 2022. URL <https://arxiv.org/pdf/2204.13663.pdf>.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [20] Y. Qian, C. Zhang, B. Krishnamachari, and M. Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 123–131, 2016.
- [21] S. Shah, K. Wang, B. Wilder, A. Perrault, and M. Tambe. Decision-focused learning without decision-making: Learning locally optimized decision losses. *Advances in Neural Information Processing Systems*, 35:1320–1332, 2022.
- [22] M. V. Smitha, P. Indumathi, S. Parichha, S. Kullu, S. Roy, S. Gurjar, and S. Meena. Compliance with iron-folic acid supplementation, associated factors, and barriers among postpartum women in eastern india. *Human Nutrition & Metabolism*, 35:200237, 2024.
- [23] R. S. Tshikomana and M. M. Ramukumba. Implementation of mhealth applications in community-based health care: Insights from ward-based outreach teams in south africa. *PLOS ONE*, 17(1):1–15, 01 2022. doi: 10.1371/journal.pone.0262842. URL <https://doi.org/10.1371/journal.pone.0262842>.
- [24] S. Verma, A. Mate, K. Wang, N. Madhiwalla, A. Hegde, A. Taneja, and M. Tambe. Restless multi-armed bandits for maternal and child health: Results from decision-focused learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1312–1320. ACM, 2023.
- [25] S. Verma, G. Singh, A. Mate, P. Verma, S. Gorantla, N. Madhiwalla, A. Hegde, D. Thakkar, M. Jain, M. Tambe, and A. Taneja. Expanding impact of mobile health programs: SAHEL for maternal and child care. *AI Magazine*, 44(4):363–376, 2023.
- [26] K. Wang, S. Verma, A. S. Mate, S. Shah, A. Taneja, N. Madhiwalla, A. Hegde, and M. S. Tambe. Scalable decision-focused learning in restless multi-armed bandits with application to maternal and child care. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 12138–12146. AAAI Press, 2023.
- [27] P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.
- [28] B. Wilder, L. Onasch-Vera, G. T. DiGuiseppe, R. Petering, C. Hill, A. Yadav, E. Rice, and M. Tambe. Clinical trial of an ai-augmented intervention for hiv prevention in youth experiencing homelessness. *CoRR*, abs/2009.09559, 2020. URL <https://arxiv.org/abs/2009.09559>.
- [29] World Health Organization. *Neonatal and Perinatal Mortality: Country, Regional and Global Estimates*. World Health Organization, 2006. URL <https://apps.who.int/iris/handle/10665/43444>.
- [30] Z. Yu, Y. Xu, and L. Tong. Deadline scheduling as restless bandits. *IEEE Transactions on Automatic Control*, 63(8):2343–2358, 2018.
- [31] Q. Zhao, B. Krishnamachari, and K. Liu. On myopic sensing for multi-channel opportunistic access: structure, optimality, and performance. *IEEE Transactions on Wireless Communications*, 7(12):5431–5440, 2008.