Improving Tuberculosis Treatment Adherence Support: The Case for Targeted Behavioral Interventions

Improving TB Treatment Adherence Support: The Case for Targeted Behavioral Interventions

Justin J. Boutilier
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, j.boutilier@wisc.edu

Jónas Oddur Jónasson
MIT Sloan School of Management, joj@mit.edu

Erez Yoeli
Keheala, eyoeli@keheala.com

Problem definition: Lack of patient adherence to treatment protocols is a main barrier to reducing the global disease burden of tuberculosis (TB). We study the operational design of a treatment adherence support (TAS) platform that requires patients to verify their treatment adherence on a daily basis.

Academic / Practical Relevance: Experimental results on the effectiveness of TAS programs have been mixed and rigorous research is needed on how to structure these motivational programs, particularly in resource-limited settings. Our analysis establishes that patient engagement can be increased by personal sponsor outreach and that patient behavior data can be used to identify at-risk patients for targeted outreach.

Methodology: We partner with a TB TAS provider and use data from a completed randomized controlled trial. We use administrative variation in the timing of peer sponsor outreach to evaluate the impact of personal messages on subsequent patient verification behavior. We then develop a rolling-horizon machine learning (ML) framework to generate dynamic risk predictions for patients enrolled on the platform.

Results: We find that, on average, sponsor outreach to patients increases the odds ratio of next-day treatment adherence verification by 35%. Furthermore, patients' prior verification behavior can be used to accurately predict short-term (treatment adherence verification) and long-term (successful treatment completion) outcomes. These results allow the provider to target and implement behavioral interventions to at-risk patients.

Managerial Implications: Our results indicate that, compared to a benchmark policy, the TAS platform could reach the same number of at-risk patients with 6%-40% less capacity or reach 2%-20% more at-risk patients with the same capacity, by using various ML-based prioritization policies that leverage patient engagement data. Personal sponsor outreach to all patients is likely to be very costly, so targeted TAS may substantially improve the cost-effectiveness of TAS programs.

1. Introduction

Lack of patient adherence to prescribed medication is a significant barrier to improving public health, particularly for longer-term treatments and in developing countries (Eskås et al. 2016, Kalichman et al. 2016). Various behavioral interventions aimed at improving adherence (often in the form of reminder systems) have been tested through randomized controlled trials (RCTs), yielding very mixed results (DiMatteo et al. 2012).

For Tuberculosis (TB)—the deadliest communicable disease in the world with 10 million new cases and 1.8 million deaths in 2018 (WHO 2019)—poor medication adherence not only endangers the infected individual, but also poses a public health risk by not curbing the infection rate and by

increasing the likelihood of emergent drug resistant strains of the virus (Volmink and Garner 2007). As a result, poor TB medication adherence has been identified as a major barrier to achieving the sustainable development goal of eliminating the TB epidemic by 2030 (WHO 2019).

In this paper, we partner with *Keheala*, a mobile health start-up that currently operates a treatment adherence support (TAS) platform[1] for patients on TB treatment in Kenya. A key aspect of the platform design is that patients are required to verify their treatment adherence on a daily basis through a mobile phone interface. To motivate treatment adherence verification, the platform provides a combination of automated reminders, motivational messages, and personal outreach to patients from peer sponsors who have experience of overcoming TB. For patients enrolled on the platform, the ultimate long-term objective of Keheala is to reduce unsuccessful treatment outcomes. The efficacy of the platform to achieve this long-term objective was confirmed through a randomized controlled trial (RCT) (Yoeli et al. 2019). In light of this success, the main, practical short-term objective of Keheala is to ensure that enrolled patients remain engaged with the platform through adherence verification.[2] However, multiple questions remain, about the efficacy and cost-effectiveness of the platform's interventions. We use data collected during the RCT to address four main research questions:

1. **Does sponsor outreach improve engagement?** We use operational variation in the timing of sponsor outreach to non-verifying patients to quantify the average impact of sponsor outreach on future patient verification behavior (§4).
2. **Is behavioral engagement data useful for predicting patient outcomes?** We develop a rolling-horizon machine learning (ML) framework that uses past verification behavior of patients to generate dynamic and personalized predictions of their risk of verification cessation or an unsuccessful treatment completion (§5).
3. **Does sponsor outreach improve engagement of *at-risk* patients?** We compare the impact estimates for *at-risk* patients with the rest of the patient population by repeating the first step analysis but allowing for heterogeneous treatment effects of sponsor outreach across patient risk groups. Understanding whether sponsor outreach affects the behavior of *at-risk* patients is important because it is a prerequisite for targeting outreach strategies to that population (§6).

[1] We adopt the *platform* term, which Keheala uses to describe its TAS service, while acknowledging that it is not a two-sided platform, in the sense often used in the operations management literature.

[2] We note that the original RCT, our source of data, only collected data on *self-verification* of treatment adherence, not *actual* treatment adherence. Establishing a link between the two remains future work (see discussion in §8). However, increasing patient engagement is a stated objective of Keheala, partly based on summary statistics (included in §A2) that suggest a reduced rate of unsuccessful outcomes among patients who remain engaged with the platform.

4. **What are the managerial implications for targeting outreach to *at-risk* patients?** We compare the predictive accuracy and capacity requirements of the patient prioritization heuristic currently used by Keheala to a range of ML based policies that prioritize patients based on the ML risk stratification (§7).

For the first question, our results indicate that sponsor outreach to patients does increase future engagement with the platform. Using a conditional logistic regression specification (which conditions on patient fixed effects as well as other control variables), we find that receiving a personalized outreach message from a sponsor increases the patients' odds-ratio of next-day treatment adherence verification by 35% ($p < 0.01$), on average.

For the second question, we find that patients' prior verification behaviors are useful predictors for short-term and long-term outcomes. Specifically, for day d of a given patient's course of treatment we combine 35 socio-demographic features with 44 features describing their interactions with the platform prior to day d and train machine learning algorithms (we explore three families of models) to predict outcomes subsequent to day d. The two target outcomes are next-day non-verification of treatment adherence and unsuccessful treatment completion (see detailed definitions in §3.3). We find that predicting these outcomes using only socio-demographic features (i.e., the information available at day $d = 0$) is difficult, obtaining a maximum area under the receiver operating characteristic curve (AUC) of 0.56 and 0.59, for next-day non-verification and unsuccessful treatment outcome, respectively. However, as d increases and we include more behavioral features, the prediction accuracy increases. For example, on day $d = 120$ the AUC scores for the two outcomes increased to 0.9 and 0.79, respectively. Based on these results, we can dynamically identify patients who are either *at-verification-risk* and/or *at-outcome-risk* on a given day.

For the third question, we explore whether the patients who are classified as *at-verification-risk* or *at-outcome-risk* on day d are more or less responsive to sponsor outreach. By exploiting the dynamic risk classification and the administrative variation in which patients receive sponsor messages on a given day, we obtain separate estimates of the effect of sponsor outreach for the *at-verification-risk*, *at-outcome-risk*, *at-both-risks* (defined as the set of patients who are simultaneously in both the aforementioned risk groups), and *not-at-risk* patient groups. We find that patients who are either *at-verification-risk* or *at-both-risks* are responsive to sponsor outreach—with an approximately 70% ($p < 0.01$) increase in the next-day verification odds ratio following peer sponsor outreach. However, the verification behavior of patients who are classified as being *at-outcome-risk* is not significantly affected by sponsor outreach. This result highlights that patients who are *at-outcome-risk* but not *at-verification-risk* might benefit more from messaging strategies that are not focused on treatment adherence verification.

For the fourth question, we explore various ways to implement targeted sponsor outreach, depending on whether Keheala wishes to focus their outreach efforts on patients *at-verification-risk*, *at-outcome-risk* or a combination (union or intersection) of the two groups. We compare our ML-based policies with a benchmark policy based on Keheala's heuristic for identifying *at-risk* patients. We find that regardless of the prioritization, our ML-based policies are either able to achieve much better predictive accuracy with the same capacity as a benchmark policy, or achieve the same predictive accuracy as the benchmark policy with much less capacity. Specifically, Keheala can reach the same number of *at-risk* patients with 6%-40% (depending on the risk definition) less capacity than the benchmark. Alternatively, the platform can reduce the number of false negative (false positive) predictions by 6%-37% (2%-11%) using the same capacity required by the benchmark heuristic. These results highlight the potential cost implications of using ML-based policies for targeting behavioral interventions.

Our rolling-horizon ML framework is being implemented at Keheala to prioritize patients for sponsor outreach. In addition, we make three main contributions, beyond the setting at hand. First, we contribute to the operational design of TAS programs. Our analysis establishes that patient engagement can be increased by personal sponsor outreach. This is important because experimental results on the effectiveness of TAS services has been mixed and there have been calls for rigorous research on how to structure these programs, particularly in resource-limited settings. Second, we illustrate that patient behavior data is useful for predicting short-term engagement as well as long-term treatment outcomes on a rolling basis. In the context of TB, health care providers have long struggled with being unable to predict which patients are likely to become non-adherent and as a a result, have been incapable of timing behavioral interventions for maximum effectiveness. From an operational perspective, these insights are particularly useful because providing personal sponsor outreach to all TB patients is likely to be very costly, so targeted TAS may substantially improve the cost-effectiveness of TAS programs. Third, we contribute to the growing literature on global health operations. To our knowledge, our work is the first to develop empirical evidence to support the operational design of targeted behavioral support systems in resource-limited settings.

2. Literature Review

Our work is broadly related to four streams of existing research. First, the expansive literature on patient treatment adherence and how it can be incorporated in treatment planning and support, specifically for TB (§2.1). Second, the growing literature on using predictive models to support behavioral interventions (§2.2). Third, the operations management (OM) literature on connected healthcare (§2.3). Fourth, the expanding literature on OM in the context of global health (§2.4).

2.1. Patient Treatment Adherence

Substantial evidence exists describing the extent of non-adherence to various medical treatments and its impact on clinical outcomes (DiMatteo et al. 2002). Moreover, a number of TAS interventions have been studied for various clinical conditions, with mixed or modest success. Nieuwlaat et al. (2014), Kardas et al. (2013), and DiMatteo et al. (2012) provide more comprehensive reviews of interventions to enhance medication adherence.

A stream of literature has explored the reasons for lackluster adherence to TB treatment. Munro et al. (2007) conduct a systematic review of qualitative research on patient adherence to TB treatment. They summarize various factors which affect an individual's propensity to adhere to treatment. Among the personal factors identified are personal motivation and personal agency, both of which the Keheala intervention addresses. They also identify environmental factors such as the influence of family, community, and other household members, which speaks to the societal stigma associated with being TB infected. Kardas et al. (2013) review the evidence on the determinants of patient adherence and its link to treatment outcomes, concluding that predicting individual non-adherence is challenging. They highlight that multifaceted TAS is likely needed to successfully improve treatment adherence. More prescriptively, multiple interventions have been introduced to improve TB adherence. Research suggests that cash-based incentives and adherence coaching to promote treatment adherence are likely to be effective in resource-limited settings (Richterman et al. 2018, Hovell et al. 2003), while food-based incentives and directly observed (by health care workers) treatment may not be (Martins et al. 2009, Walley et al. 2001).

We contribute to this literature in two ways. First, by unpacking which aspects of the Keheala platform are effective for increasing patient engagement. Second, by demonstrating how behavioral data can be leveraged to predict engagement and treatment outcomes, which has been highlighted as a key difficulty in designing effective TAS services (Kardas et al. 2013).

2.2. Analytics for Behavioral Health

In recent years, the use of behavioral interventions, often referred to as behavioral *nudges* or *boosts*, to improve health has grown significantly (Ruggeri et al. 2020). Such interventions have been used to successfully reduce no-shows for scheduled appointments (e.g., through reminder systems, Hasvold and Wootton (2011)), to unsuccessfully attempt to reduce overprescription of drugs (Sacarny et al. 2016), and to improve various other aspects of healthcare delivery (see Mills (2020) for a recent review). To date, however, much of the focus has been on homogeneous interventions as opposed to personalized interventions (Mills 2020). Furthermore, limited evidence exists about the performance of personalized behavioral interventions in resource-limited settings (Ruggeri et al. 2020).

The only other work, in this vein, that focuses on TB treatment adherence is that of Killian et al. (2019). Using TAS data from 99DOTS in Mumbai—which requires patients to make daily calls to

toll-free numbers as means of verifying adherence—they develop ML models to leverage data on prior behavior to predict risk of non-adherence and unfavorable outcomes (defined in the same way as in our work). The objective of these models is to support the prioritization of in-person visits from health workers to patients in need. A key methodological challenge in their analysis is that they do not observe their TAS intervention of interest and thus have to develop a framework to overcome this. In contrast, we are able to evaluate the impact of sponsor outreach, as we observe it in our data. Then, like Killian et al. (2019), we develop an ML framework for identifying *at-risk* patients, which significantly outperforms current practice. However, due to our unique dataset, we are subsequently able to evaluate the effectiveness of sponsor outreach in increasing the engagement of the identified *at-risk* group.

Another part of this literature focuses on developing multi-armed bandit models to determine how to best deliver various mHealth interventions with application to weight-loss (Aswani et al. 2019), physical and dietary behavior (Rabbi et al. 2018), HIV awareness (Yadav et al. 2018), as well as drug addiction and mental health (Bidargaddi et al. 2020). Our work differs from this literature in two important aspects. First, since two of our research questions are causal, our approach to estimate the effects of sponsor outreach is conducted post-hoc (rather than online), using plausibly exogenous variation in sponsor outreach within an observational dataset for identification. Second, the above cited bandit models are usually set up to learn the responsiveness of patients to an intervention. This is fundamentally different from identifying which patients are *at-risk*. A key benefit of separating our causal analysis from our prediction analysis is that we can separately establish the effect of sponsor outreach on engagement for patients who are *at-risk* and those who are not *at-risk*. This distinction is practically important, since Keheala is not simply interested in identifying patients who might be responsive to intervention, but those who are also *at-risk*.

2.3. Operations of Connected Healthcare

The advent of information technology solutions to support healthcare delivery—see Kvedar et al. (2014) for an overview—has sparked a stream of research in operations management. Recent theoretical work has examined optimal control policies to manage referrals to second-level decision agents in telemedical triage systems (Saghafian et al. 2018), how telemedical appointment systems affect physician's scheduling decisions (Bavafa et al. 2019), as well as the impact of telemedicine technologies specialists' productivity and social welfare (Rajan et al. 2019). Empirical work has examined the integration of multiple information technologies at hospitals (Angst et al. 2011), whether offering e-visits increases or reduces physical visits (Bavafa et al. 2018), and how telemedicine can complement traditional delivery channels by increasing demand for healthcare services in resource-limited settings (Delana et al. 2020).

Most related to our work is a recent paper by Lekwijit et al. (2019). They also use data from a large-scale RCT, that tested the effectiveness of a multi-pronged TAS service, to evaluate the impact of specific TAS interventions on short-term adherence and long-term outcomes. Our work complements theirs in developing an evidence base for the operational design of TAS services, since there are significant differences in the TAS service, setting, and disease condition under study. Despite significant differences in the context, our first result, that personalized messages are effective to increase subsequent engagement, is similar to the key result of their paper that personal outreach is effective at increasing adherence. Beyond this first result, our approach differs from theirs as we focus on developing an analytics-driven approach to improve the cost-effectiveness of the TAS service through targeting *at-risk* patients and demonstrating its effectiveness.

2.4. Global Health Operations

Finally, our work contributes to the literature on global health OM in resource-limited settings. This stream of work includes theoretical research on how to structure incentives (Taylor and Xiao 2014, Levi et al. 2017, Zhang et al. 2020) as well as more applied work on how to improve supply chains (De Boeck et al. 2019, 2020, Parvin et al. 2018), HIV diagnostic systems (Jónasson et al. 2017, Deo and Sohoni 2015), emergency medical services (Boutilier and Chan 2020), and inventory management (Gallien et al. 2017, Leung et al. 2016, Natarajan and Swaminathan 2014).

Most related to our work are recent papers developing strategies to support the global battle against TB. Suen et al. (2018) use a partially observable Markov Decision Process framework for optimizing the timing of drug sensitivity testing among TB patients on first line treatment and Suen et al. (2020) develop a contract theory model for optimizing the timing and quantity of monetary incentives to improve treatment adherence and find that the cost-effectiveness of such programs (at least for the case of India) can be increased substantially.

3. Setting and Data Sources

3.1. Global TB and the Case of Kenya

The global disease burden of TB has been considered a global health emergency by the WHO since 1993 (Grange and Zumla 2002). While progress has been made, with 7 million new patients receiving TB care in 2018 (up by 600,000 from 2017), TB remains the deadliest communicable disease in the world with 10 million cases and 1.8 million deaths in 2018 (WHO 2019). Furthermore, TB cases are disproportionately concentrated in resource limited settings, with 24% of global cases in Africa (WHO 2019, Sullivan et al. 2017). Although an effective treatment for TB has been available for about seventy years (Marshall et al. 1948), patient adherence remains a major challenge (Fox et al. 1958, Addington 1979). Lack of adherence is partly due to the duration and complexity of the standard TB treatment regimen, as well as the potential for serious side effects.

Moreover, lack of adherence contributes to the emergence of drug resistant strains of the virus, with about half a million new such cases (78% were multi-drug resistant) in 2018 (WHO 2019).

Like many other countries in the region, Kenya has a long history of battling with TB. The national TB program was originally established in 1980, organizing specialist health workers at national, regional, and district levels. The first National Health Strategic Plan, initiated in 1990 included multiple initiatives to improve TB control. Since the development of the second National Health Strategic Plan (for the period 2000-2010), steps have been taken to improve the program, including the doubling of test facilities and dedicated staff (WHO 2009). Despite these efforts, TB remains an important challenge to public health in Kenya, with the country listed among the 30 high burden TB states (as well as the 30 high burden drug resistant TB states) in the most recent WHO TB report (WHO 2019). In fact, recent estimates indicate a higher prevalence than was previously thought, or 558 infections per 100,000 adult population, since only about 40% of cases are notified to the relevant authorities (Enos et al. 2018).

3.2. Keheala

We partner with the mobile health start-up *Keheala*, who have developed a mobile phone service to motivate and increase adherence amongst patients on TB treatment. For patients enrolled on the platform, Keheala has two main objectives. The ultimate long-term objective is to reduce unsuccessful treatment outcomes. The efficacy of the platform to achieve this long-term objective was confirmed through a randomized controlled trial (RCT) (Yoeli et al. 2019). In light of this success, the main, practical short-term objective of Keheala is to ensure that enrolled patients remain engaged with the platform through adherence verification.

The service comprises four main types of interactions with patients. Each of the patient interactions is carefully designed, based on behavioral principles, to overcome well known barriers to treatment adherence (Sullivan et al. 2017, Yoeli et al. 2019). First, patients are required to verify their medication adherence on a daily basis through a USSD-based[3] mobile phone interface. Second, patients receive automated reminders each day, which are repeated twice, at hourly intervals, if the patient does not verify treatment adherence. Third, during sequences of non-verification, patients are contacted by peer sponsors (former TB patients who have successfully completed treatment) that enquire if they need any assistance and encourage them to adhere to their medication. Since TB patients are required to take their medication at the same time, every day, the support sponsors usually attempt to reach out shortly before that time. The policy of Keheala is that sponsors should contact patients using a personalized message after one day of non-verification and call

[3] Unstructured Supplementary Service Data (USSD) is a communications protocol which is available to all SIM-based feature phones (no smart capabilities or cellular data needed).

their mobile phone after two days of non-verification. However, prior to the trial implementation, Keheala had no data to estimate the amount of sponsor outreach that would be required to follow that policy. The data from the trial shows that due to operational reasons (i.e., staffing schedules), the timing of sponsor outreach during non-verification sequences varied substantially (see discussion of identification strategy in §4.4). Fourth, patients receive automated motivational messages following sustained (over multiple days) verification sequences. Finally, the USSD interface includes menu options that the patients can access to obtain educational information about TB as well as a messaging option for patients to reach out to their support sponsors.

3.3. Data Source

The effectiveness of Keheala was tested in a RCT[4] during 2016. Seventeen health facilities around Nairobi, Kenya participated, randomly enrolling 569 out of 1104 patients into the Keheala service and providing the standard of care to the remaining patients.

As part of the RCT, health workers collected socio-demographic information from all patients (treatment and control arms). This information includes the age, gender, language preferences, housing situation, education, employment, distance from health facility, as well as limited clinical history (see a full list of variables and summary statistics in Table A1 in §A1). In addition, Keheala collected engagement data about each patient during their enrollment in the service. This includes whether a patient verified on a given day, how many reminders they received, whether they were contacted by a peer sponsor, and whether they accessed any of the educational material on the USSD system. The outcome of interest in the RCT was the proportion of patients who had an unsuccessful treatment completion, defined as the patient dying, the treatment failing, or the patient being lost to follow-up for two or more months during treatment (a definition that is widely used in the TB literature (Killian et al. 2019)). The results of the RCT were reported in Yoeli et al. (2019), where Keheala is estimated to reduce the proportion of unsuccessful outcomes by approximately two-thirds (from 13.1% in the control group, to 4.2% in the treatment group).

The aforementioned results strongly indicate the *effectiveness* of Keheala in reducing unsuccessful treatment outcomes. In this paper, we explore the *efficiency* of the service by investigating whether the costly intervention of personalized sponsor outreach increases patient engagement and whether the data on patient behavior can be used to target these interventions towards *at-risk* patients.

[4] The trial was approved by the institutional review board of Kenyatta National Hospital and the University of Nairobi. Participants provided written informed consent. The protocol and statistical plan were registered at ClinicalTrials.gov (ID: NCT03135366).

4. Average Impact of Sponsor Outreach

Our first objective is to evaluate the average effect of peer sponsor outreach on patient engagement through the Keheala platform. Our unit of analysis is a patient-day and the number of observations is the sum of all days during which a patient was enrolled on the Keheala platform (a total of 80,931 patient-days for 569 patients).

4.1. Dependent Variables

For this part of the analysis, the outcome of interest is the future treatment adherence verification of each patient. We define two dependent variables to measure this. First, a binary indicator variable for whether patient i verified treatment adherence on the day after day d, denoted by $Next\_Day\_Verifier_{i,d}$. Second, to understand if there are lasting effects of personalized sponsor messages, we define a binary indicator variable for whether patient i verified treatment adherence for five out of the seven days following day d, denoted by $Next\_Week\_Verifier_{i,d}$.[5]

4.2. Main Independent Variables

Our aim is to evaluate the impact of personal peer sponsor messages to patients, as these require staffing and therefore contribute significantly to the cost-effectiveness of the Keheala service. We define an indicator variable, $Sponsor\_Contact_{i,d}$, that takes the value 1 if a sponsor writes a personal message to patient i on day d.

4.3. Control Variables

Our empirical analysis includes a number of control variables. First, we include patient fixed effects, denoted by $Patient_i$, which control for time-invariant features of each patient. This includes the socio-demographic characteristic of patients which might affect their overall propensity to engage with the platform and verify medication adherence. Second, we include controls for each patient's verification behavior prior to receiving the sponsor message. We define $Last\_Day\_Verifier_{i,d}$ and $Last\_Week\_Verifier_{i,d}$ analogously to our dependent variables, but describing the verification pattern of patient i during the day and week prior to day d, respectively. In addition, we define a consecutive streak of days during which a patient does not verify treatment adherence as a non-verification sequence. We include binary indicators (denoted by $X\_Days\_From\_Verification_{i,d}$ for $X \in \{1, \ldots, 10\}$) which capture the duration of a non-verification sequence if a patient is currently in one. This controls for any systematic interventions that health workers might take during non-verification sequences, but are not included in our data.[6] Third, to capture any general propensity to

[5] We note that since this outcome describes patient engagement behavior during the week subsequent to sponsor contact, it is possible that other interactions with the TAS platform took place during that period. As a result, we consider $Next\_Day\_Verifier_{i,d}$ our primary outcome.

[6] We are not aware of systematic interventions and our robustness checks (see §A7) reveal that our results are robust to excluding these variables.

Table 1 Total number of sponsor messages during non-verification sequences (% of patients).

		Length of non-verification streak				
		1	2	3	4	5
Sum of messages	0	69%	44%	40%	39%	39%
	1	31%	53%	52%	46%	48%
	2		3%	8%	14%	12%
	3			0%	1%	1%
	4				0%	0%
	5					0%
Patient-Days		4,984	1,207	456	188	121
Patients		513	359	220	131	86

Notes: The columns correspond to mutually exclusive non-verification sequences of a certain length. E.g., there were precisely 456 occurrences in which one of 220 patients did not verify adherence for three consecutive days but verified adherence before and after those three days.

increase or decrease adherence verification as patients spend more time on the platform we include a control for the number of days patient i has been enrolled with Keheala prior to day d, denoted by $Days\_On\_Platform_{i,d}$. Fourth, we include binary indicators for how many reminders patient i received on day d, denoted by $One\_Reminder_{i,d}$, $Two\_Reminders_{i,d}$, and $Three\_Reminders_{i,d}$. Finally, we include binary indicators for weekdays to account for patients who might be more or less prone to adherence verification during weekends.

4.4. Empirical Approach

We are able to identify the impact of sponsor outreach by leveraging operational variation in its timing. Specifically, we observe that Keheala's support sponsors did not consistently reach out to all patients following a single day of non-verification. Table 1 summarizes the total number of messages patients received during non-verification sequences of duration of one to five days. We observe that almost 70% of patients who had a single day of non-verification at some point (but verified the previous day and the following day) did not receive any personal messages from a peer sponsor. During longer non-verification sequences, approximately 40% of patients did not receive any messages. Table 2 summarizes at what point during non-verification sequences the patients who received personal messages from peer sponsors were contacted, demonstrating that regardless of the ultimate length of a non-verification sequence, only about a third of patients were contacted on the first day of non-adherence and the majority of those contacted on the second day of non-adherence were being contacted for the first time.

Table 2 Timing of sponsor outreach during non-verification sequences.

Day of non-verification	Contacted by sponsor (%)					First contact during non-verification (%)				
	Length of non-verification streak					Length of non-verification streak				
	1	2	3	4	5	1	2	3	4	5
1	31%	32%	31%	31%	30%	31%	32%	31%	31%	30%
2		28%	25%	27%	18%		25%	22%	23%	16%
3			12%	11%	13%			6%	4%	7%
4				9%	7%				3%	6%
5					7%					2%

Notes: Each row corresponds to a given day in a non-verification sequence. E.g., for patients who had a non-verification sequence of three days at some point, 25% of them were contacted by a sponsor on the second day of non-verification and 6% were contacted for the first time on the third day of non-verification.

The reason for the variability in the timing of sponsor outreach is administrative. Specifically, it is due to variation in the number of available support sponsors (e.g., due to staffing schedules), who are responsible for writing personal messages to non-verifying patients, relative to the daily fluctuations in the number of patients that need outreach on any given day.[7] In summary, the average number of patients enrolled on the platform during each day of the trial was 260, the average number of patients who did not verify adherence on the previous day was 97, but the average number of patients who were contacted with a personalized sponsor message was 24.

We leverage this variation to identify the impact of sponsor outreach on the probability of next-day treatment verification using the following fixed effect logistic specification;[8]

$$\ln\left[\frac{\Pr(Next\_Day\_Verifier_{i,d}=1)}{1-\Pr(Next\_Day\_Verifier_{i,d}=1)}\right] = \beta_1\ Sponsor\_Contact_{i,d} + \lambda_{\mathbf{i,d}}\ \mathbf{X}_{i,d} + \gamma_i\ Patient_i + \epsilon_{i,d}, \quad (1)$$

where β_1 captures the impact of peer sponsor contact on the log-odds of next day verification and the matrix $\mathbf{X}$ includes all the time-varying controls introduced in §4.3. The specification includes fixed effects ($Patient_i$) for each patient, to control for time-invariant differences in patients' propensity to verify treatment adherence. This ensures that β_i is estimated using within patient variation in adherence verification. However, since estimating non-linear models with a high number of fixed effects using maximum likelihood can result in inconsistent parameter estimates (due to the incidental parameters problem (Neyman and Scott 1948)), we estimate the above model using the conditional logistic approach (Chamberlain 1980). This approach conditions on a sufficient statistic

[7] Peer sponsors were not provided with any information beyond what is captured in the dataset. Furthermore, the peer sponsors worked on a rotating shift schedule, so patients were not assigned to specific sponsors.

[8] We repeat the analysis using a linear probability model, as part of our robustness checks in §A7. This does not affect our results.

Table 3 The impact of sponsor contact on future verification (odds ratios).

	(1) Next_Day_Verifier	(2) Next_Week_Verifier
Sponsor_Contact	1.353***	1.343***
	(0.060)	(0.067)
Last_Day_Verifier	2.591***	2.762***
	(0.129)	(0.146)
Last_Week_Verifier	2.352***	2.371***
	(0.111)	(0.158)
Days_On_Platform	0.999	0.996***
	(0.001)	(0.001)
Prior non-verification controls	✓	✓
Reminder controls	✓	✓
Weekday controls	✓	✓
Observations	75,237	63,907
Pseudo R^2	0.108	0.135

Notes: Prior non-verification controls include dummy variables describing the number of days since the last verification (for one to ten days), *Reminder controls* include dummy variables for the number of reminders (one to three) received by the patient on a day, and *Weekday controls* include dummy variables for the weekday at hand. Full coefficient estimates are included in §A3. Standard errors are robust and clustered at the patient level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

for each patient (see details in Chamberlain (1980)) which results in the fixed effects cancelling out in the maximum likelihood function and allows for consistent estimates of the other model coefficients.[9] Finally, we report robust standard errors, clustered by individual patients.

4.5. Results

To evaluate the average impact of personalized sponsor outreach, we estimate the conditional logit regression (1) using our full dataset. The results are shown in Table 3, where columns (1) and (2) display the coefficients (log-odds have been transformed into odds ratios) of our main variables of interest for the two outcomes; $Next\_Day\_Verifier$ and $Next\_Week\_Verifier$, respectively.

Since operational constraints at Keheala provide plausibly exogenous variation in which patients get contacted and when, the first coefficient in column (1) describes the impact of personalized sponsor contact on the ratio of verification probability over non-verification probability (odds ratio) for the average patient on the day following outreach. The coefficient estimate indicates a 35% ($p < 0.01$) increase in this odds ratio, all other factors fixed. The result in column (2) indicates a longer-term effect of sponsor contact of a similar magnitude with the odds ratio of patients verifying at least five out of the next seven days increasing by 34%. While the second result is

[9] In general, this approach has two main drawbacks. First, that patients who have no variation in the outcome variable are excluded from the analysis. This is similar to any fixed effect model, since patients with no variation in the outcome provide no explanatory power. Second, since the fixed effects cancel out during the maximum likelihood estimation, we do not obtain estimates for their coefficients. In our case, this does not pose a problem as we are interested in the effect of sponsor outreach.

encouraging, we note that since this outcome describes behavior during the subsequent week, it is possible that other interactions with the TAS platform took place during that period, which might also affect engagement. Despite that caveat, we include the result as it indicates that personalized sponsor outreach has some lasting effect. Finally, we observe from Table 3 that prior engagement ($Last\_Day\_Verifier$ and $Last\_Week\_Verifier$) are strong predictors of future engagement and that patients' likelihood of engaging with the platform does not seem to change significantly as they spend more time on the platform ($Days\_On\_Platform$ not significant).

To summarize, the results in Table 3 suggest that, on average, peer sponsor messages to patients on TB treatment are effective for increasing the odds of next-day treatment adherence verification. This is in line with previous observations that personalized reminders are more effective than automated reminders (Lekwijit et al. 2019). However, recruiting peer sponsors is costly and will impact the cost-effectiveness of Keheala at scale. This motivates the question of whether the sponsor outreach, which seems effective, could be better targeted at patients who are *at-risk* of ceasing engagement with the platform or having a bad outcome at the end of their treatment regimen. This question motivates the analysis of the subsequent sections §5 and §6.

5. Identifying *At-Risk* Patients

Our second objective is to examine the value of past engagement data for identifying at-risk patients by predicting their future engagement and treatment outcomes. To do this, we develop a framework for generating personalized real-time predictions that identify which patients on the Keheala platform are *at-risk* of either having an unsuccessful treatment outcome or ceasing adherence verification. Such a system will allow Keheala to proactively target the sponsor outreach towards patients who require the most TAS. To this end, we develop a rolling horizon framework to mimic the real time predictions that Keheala could practically obtain for patients currently enrolled on the platform. As before, we index variables describing properties of patient i on day d, where d is the number of days the patient has been on treatment.

5.1. Outcome Variables

In line with Keheala's two main objectives (see §3.2), we generate personalized predictions for two primary outcomes that inform the risk stratification of patients. First, we predict the probability of patient i having an unsuccessful treatment outcome, denoted as the event $\{Bad\_Outcome_i = 1\}$.[10] Second, we use data prior to day d to predict treatment verification cessation on the next day. For this we use the same variable as before, $Next\_Day\_Verifier_{i,d}$, but predict the probability of it taking a value of 0, denoted as the event $\{Next\_Day\_Verifier_{i,d} = 0\}$. With these definitions, a higher probability of either event signifies higher risk.

[10] This exercise predicts the probability of the event $\{Bad\_Outcome_i = 1\}$ in the presence of the TAS provided by Keheala during the RCT, in which roughly a third of non-verifying patients received sponsor outreach every day.

5.2. Feature Generation

For prediction purposes we use two sets of patient features; static and dynamic. The static (time-invariant) features comprise the socio-demographic data (described in §3.3 and §A1) for each patient, which is available to Keheala at $d = 0$. We model gender, language preferences, whether patients are slum dwellers, educational background, employment situation, and prior medical history using a set of binary indicators. The other variables—including patient age, the number of other household members, and the travel time to a TB clinic—take integer values. These 35 features inform the prediction at $d = 0$, which illustrates the prediction accuracy that is achievable in the absence of behavioral data (see discussion of results in §5.5). The dynamic (time-variant) features describe the recent and cumulative engagement of patients with the Keahala platform (prior to day d), including their verification history, the number of reminders received, the number of personal messages sent or received, and the amount of time the patient has spent on the platform. Full details are provided in Appendix A5.

Importantly, each one of the above features can be readily calculated by Keheala on any day for any patient on the platform (see discussion of current implementation in §8). We believe that generating a relatively high number of features (44 in total) to describe patient behavior is warranted since the objective of this component of our analysis is prediction accuracy (as opposed to evaluating the impact of each individual feature). In particular, we consider it important to include features for cumulative behavior (from enrollment) and recent behavior (during last week), since the adherence motivation of TB patients is known to oscillate (Munro et al. 2007).

5.3. Models

To evaluate how prediction accuracy would evolve in practice, as Keheala collects more information on patients' behavior, we conduct 177 separate prediction exercises for each outcome. The first corresponds to day 0, in which we predict the outcomes of interest ($\{Bad\_Outcome_i = 1\}$ or $\{Next\_Day\_Verifier_{i,0} = 0\}$) using only the socio-demographic data available on day 0. The subsequent 176 correspond to each day from $d = 10$ to $d = 185$[11], augmenting the socio-demographic data with the dynamic behavioral features defined above.

For each prediction exercise we compare the prediction accuracy of three models; K-Nearest Neighbours (KNN), Regularized Logistic Regression (RLR), and Random Forests (RF). For each algorithm on each day d, we use leave-one-out cross validation to evaluate model performance (in terms of AUC) and predict the out-of-sample outcome of interest for each patient. To compare models, we calculate the number of days that each of the algorithms had the best prediction.

[11] This range covers 95% of our observations, which reflects the fact that the most common course of treatment requires six months of medication. Beyond $d = 185$, we have too few patients to calculate meaningful predictions.

We find that the RF and KNN models have the highest predictive accuracy for engagement and outcomes, respectively (see discussion of prediction results in §5.5). For clarity of exposition, we omit the results from the other models; see §A6.

5.4. *At-risk* Classification

The final step of the ML analysis is to generate distinct patient groups based on the predictions. On each day we assign each patient to an *at-verification-risk* and/or an *at-outcome-risk* group if their predicted probability of the events $\{Next\_Day\_Verifier_{i,0} = 0\}$ or $\{Bad\_Outcome_i = 1\}$ is greater than thresholds $\tau_{verifier}$ or $\tau_{outcome}$, respectively.[12] The choice of the τ thresholds determines a point on the trade-off curve between the number of false positives and false negatives for each outcome. In practice, Keheala could set the thresholds to either emphasize *sensitivity* (i.e., the proportion of truly *at-risk* patients accurately identified) or *specificity* (i.e., proportion of *not-at-risk* patients accurately identified). For our main analysis, we select τ values to match the number of false negatives of a benchmark policy that considers non-verifying patients as being *at-risk* for each outcome, which is the intended heuristic of Keheala. Following this approach, we set $\tau_{verifier} = 0.443$ and $\tau_{outcome} = 0.067$, which results in enrolled patients being classified as *at-verification-risk* and *at-outcome-risk* on 33% and 21% of patient-days, respectively. By matching the number of false negatives of our models to the benchmark policy, we implicitly focus on improving the cost-effectiveness of the platform as our models will identify the same proportion of truly *at-risk* patients but minimize the proportion of patients who are mis-classified as *at-risk*, allowing for more targeted allocation of resources (see more detailed discussion of managerial implications in §7).

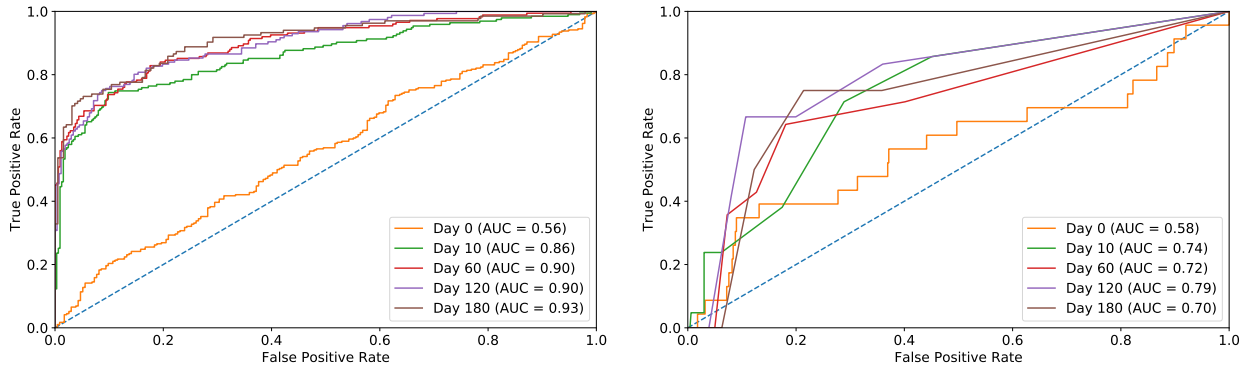
5.5. Results

As we describe in §5.3, we consider three models as part of our rolling-horizon ML framework for identifying *at-risk* patients in real time (i.e., on each day). For clarity of exposition, we focus our discussion of the results on the best performing models for predicting the two adverse events; $\{Next\_Day\_Verifier_{i,d} = 0\}$ and $\{Bad\_Outcome_i = 1\}$. For the former, a RF model provided the highest AUC on 102 out of 177 days (57.6%), whereas a KNN model performed best for the latter outcome, with the highest AUC on 156 out of 177 days (88.1%). Based on the results for those two sets of models, we make four main observations.

First, our models predict both outcomes consistently and with good accuracy. The average AUC (across days $d \in \{0, 10, 11, \ldots, 185\}$) of the RF model for predicting engagement cessation

[12] Another approach would be to define separate thresholds $\tau_{outcome,d}$ and $\tau_{verifier,d}$ for each day d (i.e., the number of days a patient has spent on the Keheala platform). However, using consistent thresholds throughout the entire period is a superior approach since Keheala provides TAS to patients who are at various stages of their treatment regimen. Using different thresholds for different days could result in the undesirable situation of Keheala prioritizing patients who are at a lower risk, simply because the risk distribution for a given day is different from another day. Furthermore, using consistent threshold definitions is easier to implement in practice.

Figure 1 **Predictive performance of rolling-horizon algorithms for patients at different days of enrollment.**



(a) Prediction of $\{Next\_Day\_Verifier_{i,d} = 0\}$ (b) Prediction of $\{Bad\_Outcome_i = 1\}$

Notes: Our rolling-horizong ML framework generates predictions for each patient on each day they spend on the platform. For clarity, the figures above present prediction results for a subset of days. We note that as d increases, we have fewer observations to train our models (since only 20% of patients spend 180 days or more on the platform).

($\{Next\_Day\_Verifier_{i,d} = 0\}$) is 0.883 with a standard deviation of 0.033 across days. Similarly, the average AUC of the KNN model for predicting a bad outcome ($\{Bad\_Outcome_i = 1\}$) is 0.751 with a standard deviation of 0.036 across days.

Second, for both outcomes, we observe that the behavioral data describing individuals' engagement with the TAS platform is valuable for predictions. Figure 1 (a) compares the predictive accuracy of our best performing RF model for $\{Next\_Day\_Verifier_{i,d} = 0\}$ on the subset of days; $d \in \{0, 10, 60, 120, 180\}$. Importantly, the algorithm has no behavioral information to inform the predictions at $d = 0$ and thus relies on socio-demographic data only—with very modest success (AUC=0.56). However, from $d = 10$ to $d = 185$ the model leverages behavioral data in addition to the socio-demographic data and predicts engagement with very high accuracy (AUC $\sim$ 0.9), as illustrated in the ROC curves corresponding to the days $d \geq 10$. Figure 1 (b) provides the accuracy of the best KNN model for predicting $\{Bad\_Outcome_i = 1\}$ for the same subset of days. We observe a milder but similar phenomenon in that outcomes cannot be predicted accurately using socio-demographic data only (AUC=0.59 at day $d = 0$) but improve substantially once behavioral engagement data becomes available (AUC $\sim$ 0.75). To further demonstrate the value of patient behavior data, we repeat our rolling-horizon prediction analysis excluding all features describing the platforms actions, thereby focusing on the predictive power of the features describing patient behavior. This does not affect the predictive accuracy of either model and further supports that our observed predictive accuracy is driven by patient behavior (see Appendix A6 for details).

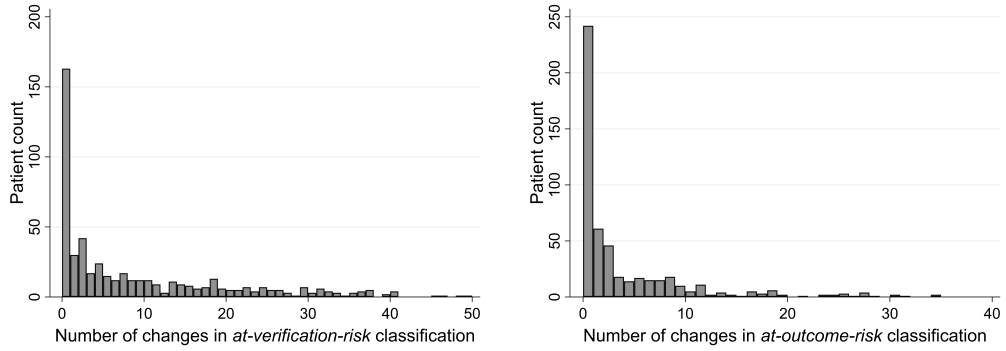
Third, this component of our analysis is focused on prediction and we make no attempt at a causal interpretation. However, it is of interest to understand which variables provide the most

predictive power. We highlight only key observations here and refer to §A6 for details on how these results were obtained. For the KNN model that predicts $\{Bad\_Outcome_i = 1\}$, we find that longest non-verification streak and TB type are most important, followed by indicator variables for location and the number of patient initiated messages to the platform. For the RF model that predicts $\{Next\_Day\_Verifier_{i,d} = 0\}$, we find 12 features with a relative importance[13] of at least 0.025. The four most importance features represent recent behavior (recent verifications, automated motivational verification messages sent from the platform, the number of educational options accessed, and seconds spent on the platform in the last week), but some cumulative behavioral features are also predictive (including the total number of prior verifications, total number of motivational messages, and the length of the longest non-verification streak of the patient).

Fourth, we examine the consistency of the prediction results. Specifically, we observe the risk classification of each patient throughout their time on the platform and count how frequently they are reassigned (we will refer to those as *switches*) from *not-at-risk* (for one or both of the main outcomes) to *at-risk*, and vice versa. Qualitatively, we want to know whether the risk predictions for a given patient are mostly constant across days—which might be accurate but not provide much useful information for dynamic prioritization. Similarly, we check whether the risk classification of patients change very frequently—which would indicate some noise in the classification and make it less valuable for resource allocation. Figure 2 demonstrates the number of switches (using the baseline thresholds $\tau_{outcome}$ and $\tau_{verifier}$ introduced in §5.4) of a patient throughout the course of the trial, for *at-verificaton-risk* in (a) and for *at-outcome-risk* in (b). We observe from (a) that 163 patients have no switches in their *at-verification-risk* classification for their duration on the platform. Specifically, 107 patients are never considered *at-verification-risk* and 56 patients are always considered *at-verification-risk*. The remaining 353 patients switch groups once or more during their time on the platform. The average number (standard deviation) of switches for *at-verification-risk* was 9 (11) across all patients (12 (11) for the 353 patients with one or more switches). We observe slightly more consistent predictions for the *at-outcome-risk* classification, with 203 patients who are never classified as at risk for a bad outcome, 39 patients who are always classified as *at-outcome-risk*, and the remaining 274 switching between groups. The average number (standard deviation) of switches for this outcome is 4 (6) for the whole population, but 7 (7) for the patients who switch at least once. Notably, the rate of unsuccessful treatment outcomes in the population which never enters the *at-outcome-risk* group is 1.5%, as compared to 23% among the patients who are always classified as *at-outcome-risk* (and a rate of 4% among all patients).

[13] The importance of each feature is computed as the total reduction of the Gini importance metric (i.e., the model objective function) brought by that feature. These values are then normalized to sum to one across all features.

Figure 2 Number of switches between risk groups.



(a) (b)

Notes: The graphs, above, summarize the number of times each patient switches to or from being classified as *at-risk* for each of the two outcomes $\{Next\_Day\_Verifier_{i,d}=0\}$ (a) and $\{Bad\_Outcome_i=1\}$ (b).

6. Impact of Sponsor Outreach for *At-Risk* Patients

Our third objective is to evaluate the effect of sponsor outreach on *at-risk* patients. The aim of the first empirical analysis, described in §4, is to evaluate the population average impact of sponsor outreach on the odds of next-day treatment adherence verification. If it is possible to identify patients who are *at-risk* of verification cessation or unsuccessful treatment outcomes, using the methods described in §5, the ultimate question is whether peer sponsor outreach is effective for increasing the engagement of this population. A positive answer to this question indicates that Keheala should focus its outreach effort on this *at-risk* population.

6.1. Empirical Approach

We explore this question by repeating the analysis described in §4, but allowing for heterogeneous treatment effects based on whether patients are *at-risk* or *not-at-risk*, at the time they receive the sponsor message. Using the *at-risk* definitions introduced in the previous subsection, we extend the regression specification (1) as follows;

$$\begin{aligned}\ln\left[\frac{\Pr(Next\_Day\_Verifier_{i,d}=1)}{1-\Pr(Next\_Day\_Verifier_{i,d}=1)}\right] &= \alpha_1\ Sponsor\_Contact_{i,d} * Not\_At\_Risk_{i,d} \\ &+ \alpha_2\ Sponsor\_Contact_{i,d} * At\_Verificaton\_Risk_{i,d} \\ &+ \alpha_3\ Sponsor\_Contact_{i,d} * At\_Outcome\_Risk_{i,d} \\ &+ \alpha_4\ Sponsor\_Contact_{i,d} * At\_Both\_Risks_{i,d} \\ &+ \alpha_5\ At\_Verificaton\_Risk_{i,d} + \alpha_6\ At\_Outcome\_Risk_{i,d} \\ &+ \lambda_{\mathbf{i,d}}\ \mathbf{X}_{i,d} + \gamma_i\ Patient_i + \epsilon_{i,d}.\end{aligned} \tag{2}$$

In the above specification, the impact of sponsor outreach is estimated separately for each risk group, relative to no outreach. We capture the effect of sponsor outreach for patients classified as *at-verification-risk*, *at-outcome-risk*, or *at-both-risks* (defined as the set of patients, on a given day,

who are predicted to be at risk for both adverse outcomes) using α_2, α_3, and α_4, respectively. For comparison, we also evaluate the impact of sponsor outreach to patients who belong to neither risk group on a given day (α_1). As before, we include the control variables introduced in §4.3 (including patient fixed effects), estimate the coefficients using a conditional logistic approach, and cluster standard errors at the patient level. Similar to our first empirical analysis, identification is enabled by the fact that for administrative reasons, only a subset of patients belonging to a given risk group on a given day, receive sponsor outreach on that day.[14]

6.2. Results

We know from §4.5 that peer sponsor outreach is effective for increasing patient verification on average. In this section, we explore heterogeneous effects, by estimating (2), which focuses on whether sponsor outreach is effective for the *at-risk* population identified in §5.5. As before, identification is enabled by the operational constraints at Keheala resulting in plausible exogenous variation in which patients are contacted and which are not.

The results of this analysis are included in Table 4. Columns (1) and (3) report results for the analysis which corresponds to our main analysis in Table 3. We include these estimates here because our analysis of heterogeneous effects relies on a smaller sample than the previous analysis, since we only have *at-risk* predictions for patients until they reach 185 days on the platform. These results correspond closely (in direction, magnitude, and significance) to the results presented in Table 3.

We make two sets of main observations about our results (reported as odds-ratios) for the heterogeneous effects (across risk groups) of sponsor outreach, on the likelihood of next day verification, reported in column (2). First, unsurprisingly the odds ratio associated with being *at-verification-risk* is significant and less than one, confirming that patients classified in this risk group are less likely to verify treatment adherence on the following day. In contrast, the patients classified as *at-outcome-risk* on a given day are not necessarily less likely to verify treatment adherence on the following day. Including these controls is important as it ensures that the interaction effects ($Sponsor\_Contact * At\_Verification\_Risk$ and $Sponsor\_Contact * At\_Outcome\_Risk$) reflect the impact on the verification odds-ratio for patients in each of the *at-risk* groups.

Second, the impact of sponsor outreach on subsequent verification behavior is not uniform across risk groups. For patients who are not considered *at-verification-risk* or *at-outcome-risk* on a given day, sponsor outreach has a statistically significant ($p < 0.05$) but modest ($\alpha_1 = 1.157$) effect to increase the odds of next day treatment verification. For patients, who are classified as

[14] Specifically, §A4 shows that patients received sponsor outreach on 8.8% (9.1%) of patient-days when they were classified as *at-verification-risk* (*at-outcome-risk*). This is comparable to the percentage (9.1%) of patient-days for which patients who were identified by the intended benchmark heuristic of Keheala (of reaching out to patients who did not verify on the previous day) received sponsor outreach.

Table 4 The heterogeneous impact of sponsor contact on future verification, by risk classification (odds ratios)

	(1) Next_Day_Verifier	(2) Next_Day_Verifier	(3) Next_Week_Verifier	(4) Next_Week_Verifier
Sponsor_Contact	1.361***		1.340***	
	(0.064)		(0.073)	
Sponsor_Contact*Not_At_Risk		1.157**		1.129*
		(0.068)		(0.072)
Sponsor_Contact*At_Verification_Risk		1.719***		1.790***
		(0.139)		(0.173)
Sponsor_Contact*At_Outcome_Risk		0.972		1.118
		(0.126)		(0.168)
Sponsor_Contact*At_Both_Risks		1.707***		1.557***
		(0.174)		(0.190)
At_Verification_Risk		0.540***		0.516***
		(0.031)		(0.035)
At_Outcome_Risk		1.011		1.073
		(0.066)		(0.127)
Last_Day_Verifier	2.581***	2.319***	2.776***	2.473***
	(0.139)	(0.114)	(0.161)	(0.131)
Last_Week_Verifier	2.348***	1.801***	2.372***	1.795***
	(0.120)	(0.085)	(0.171)	(0.127)
Days_On_Platform	1.000	1.000	0.996***	0.996***
	(0.001)	(0.001)	(0.001)	(0.001)
Prior non-verification controls	✓	✓	✓	✓
Reminder controls	✓	✓	✓	✓
Weekday controls	✓	✓	✓	✓
Observations	64,098	64,098	50,661	50,661
Pseudo R^2	0.103	0.108	0.128	0.133

Notes: Prior non-verification controls include dummy variables describing the number of days since he last verification (for one to ten days), *Reminder controls* include dummy variables for the number of reminders (one to three) received by the patient on a day, and *Weekday controls* include dummy variables for the weekday at hand. Full coefficient estimates are included in §A3. Standard errors are robust and clustered at the patient level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

at-verification-risk on a given day, the impact of sponsor outreach is more substantial ($\alpha_2 = 1.719$, $p < 0.01$). This confirms that sponsor outreach is effective for improving the engagement of patients who are at risk of verification cessation. In contrast, the verification behavior of patients who are classified as being *at-outcome-risk* is not significantly affected by sponsor outreach. This could reflect the fact that some patients can have a high likelihood of a bad outcome, despite being diligent in verifying treatment adherence. (This also highlights that patients who are *at-outcome-risk* might benefit from a different type of sponsor outreach, which is not focused on adherence verification. See further discussion of managerial implications in §7.). Most importantly, the interaction $Sponsor\_Contact * At\_Both\_Risks$ is positive ($\alpha_4 = 1.707$) and significant ($p < 0.01$), indicating that sponsor outreach to patients who are both *at-outcome-risk* and *at-verification-risk* can be motivated to continue treatment adherence verification. This is encouraging for Keheala since it is a priority to reach patients who are likely to stop adherence verification (i.e., are *at-verification-risk*) and are likely to need TAS to successfully complete their treatment (i.e., are *at-outcome-risk*).

The results in column (4) of Table 4, which show the impact of sponsor outreach on treatment adherence verification during the subsequent week, are similar to our results for next day verification (in column (2)). As before, we note the caveat that when we use $Next\_Week\_Verifier$ as the outcome variable, it is possible that patients were contacted by a sponsor on another occasion during the subsequent week (regardless of whether they were contacted on the day at hand or not). However, these results suggest some lasting effect of sponsor outreach, specifically for patients who are *at-verification-risk* or *at-both-risks*.

In summary, the implication of the results in Table 4 is that sponsor outreach can be effective for improving patient engagement (through treatment adherence verification). This is important for Keheala since it confirms that sponsor outreach is effective for increasing the engagement of patients who would otherwise be at risk of adherence verification cessation. Furthermore, the results are quite robust. As part of our robustness checks in §A7, we repeat this analysis on a matched sample, using a dynamic binary panel model (Bartolucci and Nigro 2010), using various classification thresholds $\tau_{outcome}$ and $\tau_{verifier}$, as well as estimating a linear probability model, obtaining qualitatively similar results in all cases.

7. Managerial Implications

We illustrate the managerial implications of our analysis, for Keheala, in two steps. First, we discuss how the ML predictions (developed in §5) can be used to generate ML-based policies to prioritize which patients receive sponsor outreach on which day (§7.1). We then compare the ML-based policies to the benchmark heuristic of Keheala—reaching out to patients who fail to verify medication adherence the previous day—in terms of prediction accuracy and capacity requirements (§7.2). Finally, we discuss managerial implications of our work, beyond the case of Keheala (§7.3).

7.1. Developing ML-based Outreach Strategies

As we discuss in §1, Keheala has two main objectives for patients enrolled on the platform; a long-term objective of increasing the probability of successful treatment completion and a short-term objective of ensuring that patients remain engaged with the platform through adherence verification. The two predictions from our rolling-horizon ML framework—one for the risk of verification cessation and one for the risk of having an unsuccessful treatment outcome—allow Keheala to generate a range of outreach strategies, depending on their prioritization of the two risks.

We explore four such prioritizations, focusing exclusively on each of the two types of risk (the *Verification Risk Prioritization* and the *Outcome Risk Prioritization*) as well as their intersection (the *Verification AND Outcome Risk Prioritization*) and union (the *Verification OR Outcome Risk Prioritization*). For each prioritization, the classification thresholds ($\tau_{verifier}$ and $\tau_{outcome}$) determine how many instances (patient-days) of peer sponsor outreach would be required by following

a corresponding ML-based policy. A high $\tau_{verifier}$ ($\tau_{outcome}$) results in only patients with a high predicted probability of non-verification on the following day (high predicted probability of an ultimately unsuccessful treatment outcome on a given day) being considered *at-risk*. A high $\tau_{verifier}$ is therefore likely to result in a high number of false negatives (e.g., mis-classifying a patient who will not verify treatment adherence on the following day as not *at-verification-risk*), but would require relatively few instances of sponsor outreach. In contrast, a low $\tau_{verifier}$ is likely to result in a high number of false positives (e.g., mis-classifying a patient who will verify on the subsequent day as *at-verification-risk*) and would result in many instances of sponsor outreach.

7.2. Comparison of Accuracy and Capacity Requirements

Table 5 summarizes the main performance metrics for each policy (including the benchmark heuristic). Prediction accuracy is described using the instances (patient-days) in which each classification policy resulted in a true positive (T.P.), false positive (F.P.), false negative (F.N.) or true negative (T.N.) prediction[15]. Capacity refers to the number of sponsor outreach instances the platform would have to conduct to reach out to all patients classified as *at-risk* by a given policy.

Figure 3 provides a visual depiction of which patients would receive outreach according to each of the four ML-based strategies. The scattered dots are the same in each panel, denoting the ML prediction of verification risk and outcome risk for a given patient on a given day. We observe that while the RF-generated prediction of non-verification on the subsequent day (our measure for engagement) is continuously distributed between zero and one, the KNN-generated predictions for the probability of a bad outcome are clustered on ten values (ranging from 0 to 0.3).[16] The long-dash (short-dash) lines depict the $\tau_{verifier}$ and $\tau_{outcome}$ thresholds for each *matching F.N.* (*matching capacity*) policy and the shaded areas demonstrate the prediction characteristics of the patients who would receive sponsor outreach, according to each policy.

7.2.1. Verification Risk Prioritization. The top panel of Table 5 compares the benchmark heuristic to two ML-based policies tailored to the case in which Keheala prioritizes identifying patients who are *at-verification-risk*. We observe that the heuristic accurately predicts the next day treatment verification behavior of patients on 56,724 (18,100 + 38,624) instances. However, on 6,319 occasions a patient is mis-classified as being *at-verification-risk* when they are not (we

[15] We acknowledge that if Keheala would have consistently followed the heuristic during the field trial, this could bias the number of false positives for the heuristic upwards (e.g., if a patient was considered *at-risk* by the heuristic on a given day and therefore received sponsor outreach, they might have subsequently verified treatment adherence and hence been classified as a false positive). However, Table A4 in §A4 demonstrates that the subset of patients who were classified as *at-risk* and received sponsor outreach is consistent in size, regardless of the classification policy. We therefore believe that the statistics for the heuristic are comparable to those for the ML-based policies.

[16] This clustering is a function of the number of neighbours (10) and the relatively fewer data points for the outcome prediction (only one outcome per patient).

Table 5 Prediction accuracy and capacity requirements comparison for the heuristic and ML-based policies.

Verification Risk Prioritization

Objective: Identify patients who will not verify on subsequent day

	T.P.		F.P.		F.N.		T.N.		Capacity	
Benchmark heuristic	18,100		6,319		6,416		38,624		24,419	
ML Policy (matching F.N.)	18,100	(0%)	4,560	(-28%)	6,416	(0%)	40,383	(5%)	22,660	(-7%)
ML Policy (matching capacity)	18,625	(3%)	5,593	(-11%)	5,891	(-8%)	39,350	(2%)	24,218	(-1%)

Outcome Risk Prioritization

Objective: Identify patients who will have a bad outcome

	T.P.		F.P.		F.N.		T.N.		Capacity	
Benchmark heuristic	1,371		23,048		748		44,292		24,419	
ML Policy (matching F.N.)	1,406	(3%)	13,243	(-42%)	713	(-4%)	54,097	(23%)	14,649	(-40%)
ML Policy (matching capacity)	1,649	(20%)	22,370	(-3%)	470	(-37%)	44,970	(2%)	24,019	(-2%)

Verification AND Outcome Risk Prioritization

Objective: Identify patients who will not verify on subsequent day and have a bad outcome

	T.P.		F.P.		F.N.		T.N.		Capacity	
Benchmark heuristic	1181		23238		196		44844		24419	
ML Policy (matching F.N.)	1,181	(0%)	17,142	(-26%)	196	(0%)	50,940	(14%)	18,323	(-25%)
ML Policy (matching capacity)	1,227	(4%)	22,740	(-2%)	150	(-23%)	45,342	(1%)	23,967	(-2%)

Verification OR Outcome Risk Prioritization

Objective: Identify patients who will not verify on subsequent day or have a bad outcome

	T.P.		F.P.		F.N.		T.N.		Capacity	
Benchmark heuristic	18,290		6,129		6,968		38,072		24,419	
ML Policy (matching F.N.)	18,330	(0%)	4,541	(-26%)	6,928	(-1%)	39,660	(4%)	22,871	(-6%)
ML Policy (matching capacity)	18,676	(2%)	5,741	(-6%)	6,582	(-6%)	38,460	(1%)	24,417	(0%)

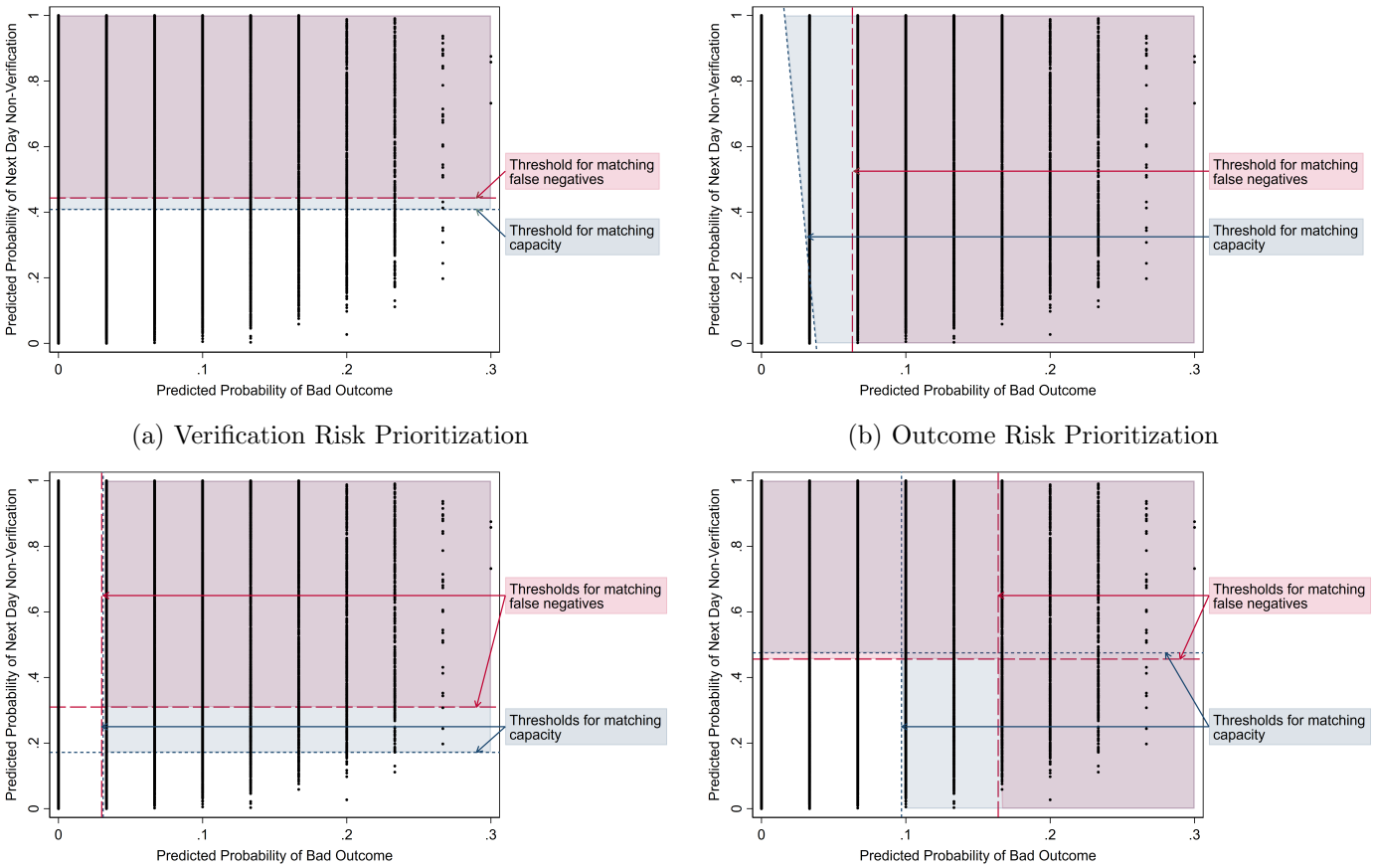
Notes: The units in the table are in patient-days. E.g., 18100, in the top row, corresponds to the sum of all instances when a given patient at a given day was correctly classified as *at-risk* for not verifying treatment adherence on the following day, using the Non-Verifier heuristic. T.P., F.P., F.N., and T.N., denote true positives, false positives, false negatives, and true negatives for the outcome defined in the objective for each policy, respectively. Capacity refers to the number of sponsor outreach instances (patient-days) required to reach out to each patient identified as *at-risk* for a given policy. Numbers in parentheses represent the performance change, relative to the Non-verifier heuristic.

refer to these false positives as *waste*, since it would result in unnecessary sponsor outreach) and on 6,416 occasions a patient is mis-classified as not being *at-verification-risk* when they are (we refer to these false negative instances as *misses*).

In contrast, the second line of the panel demonstrates how an ML-based policy can be tuned to achieve the same number of misses, with 28% less waste, which results in the policy requiring 7% fewer sponsor outreaches than the heuristic. Alternatively, an ML-based policy can be tuned to require the same amount of capacity as the heuristic, but resulting in 11% less waste (F.P.) and 8% fewer misses (F.N.).

The inclusion thresholds for the two ML-based policies are demonstrated by the horizontal lines in Figure 3(a). In this case, the threshold for classifying patients as *at-risk* is horizontal, as it relies exclusively on the probability of a patient not verifying treatment adherence on the subsequent day. Intuitively, the threshold for matching the capacity of the heuristic is lower than the threshold for the policy which matches the misses (F.N.), as it allows for more instances of a patient being classified as *at-verification-risk* on a given day.

Figure 3 Risk predictions and inclusion thresholds for ML-based risk classification policies.



(a) Verification Risk Prioritization

(b) Outcome Risk Prioritization

(c) Verification AND Outcome Risk Prioritization

(d) Verification OR Outcome Risk Prioritization

Notes: The dots are the same in each panel, denoting the prediction of verification risk and outcome risk for a given patient on a given day. The shaded areas correspond to the instances (patient-days) when a patient would have been targeted by a given outreach policy. The red long-dash (blue short-dash) thresholds correspond to the inclusion criteria for the ML-based policy which matches the current heuristic on false negatives (capacity requirements).

7.2.2. Outcome Risk Prioritization. The second panel in Table 5 describes the predictive accuracy and capacity requirements of each classification strategy for identifying patients who are *at-outcome-risk*, corresponding to the case in which Keheala prioritizes identifying patients who are *at-risk* of having an unsuccessful treatment outcome, based on patients' behavior on the platform.

As before, we observe that the ML-based policies can outperform the heuristic—either by reducing capacity or by reducing waste. The heuristic correctly classifies patients *at-outcome-risk* on 1,371 instances, while only missing 748 instances where patients eventually have an unsuccessful treatment outcome despite having verified treatment adherence on the previous day (which results in them not being classified as *at-risk* by the heuristic). However, the relatively few misses (F.N.) come at the expense of having significant waste (F.P.). In contrast, an ML-based strategy tuned to achieve the same number of misses is able to do so with 42% lower waste. Alternatively, an

ML-based strategy requiring the same capacity could jointly achieve a 3% and 37% reduction in waste and misses, respectively.

Figure 3(b) demonstrates the classification thresholds for the two ML-based strategies prioritizing outcome risk. The policy which closest matches the misses (F.N.) of the heuristic is the one which classifies all patients who have a predicted probability of a bad outcome exceeding 0.06 as *at-outcome-risk*. The alternative policy, which matches the heuristic on capacity requirements, expands the set of *at-outcome-risk* patients to also include 71% of the instances when a patient had a predicted probability of a bad outcome of 0.03.

7.2.3. Verification AND Outcome Risk Prioritization. The third panel in Table 5 describes the performance of the benchmark heuristic and two ML-based strategies for identifying patients who at a given day are both *at-verification-risk* and *at-outcome-risk*. As before, we observe that using an ML-based classification strategy, in this case based on both prediction models, is able to either achieve the same number of misses with 26% less waste or reduce the waste and misses by 2% and 26%, respectively—using the same capacity. Figure 3(c) demonstrates the thresholds for the ML-based strategies. In this case, only instances when a patient is above both $\tau_{verifier}$ and $\tau_{outcome}$ on a given day result in that patient being classified as at risk. As a result, only patients who are in the shaded upper-right corner of the figure would be considered *at-risk*.

7.2.4. Verification OR Outcome Risk Prioritization. The final potential prioritization approach aims to identify patients who are either *at-verification-risk* or *at-outcome-risk*. In Figure 3(d) this is illustrated by the fact that patients whose predictions on a given day lie above either $\tau_{verifier}$ or $\tau_{outcome}$ would be considered at risk and prioritized for sponsor outreach (i.e., the shaded area includes the highest values on both axes). Similar to before, an ML-based policy can either reduce waste by 26% while maintaining a similar number of misses or reduce both waste and misses by 6% with the same capacity requirements as the benchmark heuristic.

7.3. Broader Managerial Implications

Beyond the practical impact on the operations of Keheala, our work has more general implications for the design of mHealth programs as well as other TB TAS services.

7.3.1. Implications for the Design of mHealth Programs. Fostering patient engagement is a well-known challenge for behavioral mHealth applications (Subbaraman et al. 2018, Druce et al. 2019, Birnbaum et al. 2015). Our work demonstrates that personal outreach from support sponsors is effective for maintaining patient engagement. Going forward, providers of behavioral interventions could benefit from incorporating personal outreach in the design of their platforms, in addition to automated reminders and push notifications.

An obvious drawback of the above recommendation is that personal outreach requires more resources than automated messages and can make or break the cost-effectiveness (i.e., the business model) of behavioral mHealth programs. However, our work demonstrates that requiring patients to provide self-verification of the desired behavior has the side-benefit—beyond motivating the behavior itself—of providing valuable data to the mHealth platform.

In addition, our work is motivated by an mHealth application in a resource-limited setting. Our results demonstrate that even in settings with limited technological infrastructure, personalizing behavioral health interventions is not only feasible but potentially a major driver of operational efficiencies. Even if data is collected through low-tech interactions with patients, sophisticated analytics can be conducted on the back-end, allowing for highly accurate risk scoring of individual patients. Such risk scoring, in turn, enables timely and targeted interventions.

Finally, understanding how to design mHealth applications focused on TAS is useful. The effectiveness of TAS services is inconclusive (Lekwijit et al. 2019, DiMatteo et al. 2012). We study the operations of a service which has been shown to be effective through an RCT, to unpack which elements of its service (peer sponsor outreach, specifically) are effective for maintaining the engagement of patients with the platform. Understanding the specifics of how to structure TAS is relevant for many conditions besides TB, such as cardiovascular health (Chaudhry et al. 2007, 2010).

7.3.2. Implications for TB TAS. Our work is highly relevant for other TAS services focused on TB, including programs that use mHealth platforms, electronic pill bottles, and ingestible sensors. The results from controlled studies on the effectiveness of TAS services are mixed and there have been calls for programs to enhance (rather than limit) personal engagement to improve TB patient treatment adherence (Subbaraman et al. 2018), particularly since predicting when a patient benefits from motivational support can be challenging. Our work establishes that personal outreach can improve engagement and provides evidence for shifting the current treatment adherence paradigm from observational (i.e., monitoring and collecting data on patient adherence) to actionable (i.e., combining patient adherence data with analytics to actively improve patient treatment adherence in a targeted way). Our results indicate that other TB TAS programs should be designed to collect engagement data via active participation from end-users, because that data is useful to provide targeted outreach.

8. Conclusion

The third objective of the U.N. Sustainable Development Goals aims to "ensure healthy lives and promote well-being for all". A key target (3.3) is to end the TB epidemic by 2030 (WHO 2019). Despite the current strains of TB being curable, progress towards the target has been slow (WHO 2019). Reasons include the long duration of state-of-the-art treatment regimens (usually six months

or longer) which—in the presence of community stigma, challenges in access to care, and treatment side effects—drives significant fluctuations in TB patients' motivation to adhere to treatment (Yoeli et al. 2019). Rigorous evidence is lacking for how to motivate diligent and sustained TB treatment adherence and how to identify patients in need of support.

We contribute to developing an evidence-base for TAS operations by conducting a post-hoc analysis of data collected by Keheala in an RCT. The results of the original trial show a significant impact on treatment outcomes for patients enrolled on the Keheala platform (Yoeli et al. 2019). In this paper, we take the next step of exploring whether the most costly intervention of the Keheala platform (direct peer sponsor outreach) is effective for improving patient treatment adherence verification and whether the behavioral data collected by the platform allows for personalized outreach strategies through an ML-based risk stratification of the enrolled patients.

We find that, on average, sponsor outreach to patients increases the odds ratio of next-day treatment adherence verification by 35%. Furthermore, patients' prior verification behavior can be used to accurately predict short-term (treatment adherence verification) and long-term (successful treatment completion) outcomes. These results allow the provider to target and implement behavioral interventions to at-risk patients. Compared to a benchmark policy, the TAS platform could reach the same number of at-risk patients with 6%-40% less capacity or reach 2%-20% more at-risk patients with the same capacity, by using various ML-based prioritization policies that leverage patient engagement data. Personal sponsor outreach to all patients is likely to be very costly, so targeted TAS may substantially improve the cost-effectiveness of TAS programs.

The scope of our work is to estimate the impact of sponsor outreach on patient engagement and evaluate the usefulness of treatment adherence self-verification data for risk prediction. Multiple important research questions remain for future work. Most importantly, our data source does not allow for establishing a clear association between patient *self-verification* of treatment adherence and *actual* treatment adherence (although summary statistics suggest a correlation, see discussion in §A2). Establishing this link is important, since our work demonstrates that sponsor outreach can improve patient self-verification of adherence and it is well known that actual treatment adherence is a prerequisite of successful treatment completion (WHO 2020). While researchers have assumed equivalence between self-verification of adherence and actual adherence (see e.g., Killian et al. 2019), additional field data collection that permits a rigorous demonstration of the link between the two would solidify the intuition that self-verification leads to improved health outcomes.

Acknowledgments

The authors are grateful to Jon Rathauser, founder and CEO of Keheala, for the collaboration. The authors also thank the Keheala Operational Team (Alice, Faith, Edwin, Jacinta, Jill, Lewis, Moreen, Trish). Finally, the authors thank Jiong Wei Lua for excellent research assistance.

References

Addington, W.W. 1979. Patient compliance: the most serious remaining problem in the control of tuberculosis in the united states. *Chest* **76**(6) 741–743.

Angst, C.M., S. Devaraj, C.C. Queenan, B. Greenwood. 2011. Performance effects related to the sequence of integration of healthcare technologies. *Production and Operations Management* **20**(3) 319–333.

Aswani, A., P. Kaminsky, Y. Mintz, E. Flowers, Y. Fukuoka. 2019. Behavioral modeling in weight loss interventions. *European Journal of Operational Research* **272**(3) 1058–1072.

Bartolucci, F., V. Nigro. 2010. A dynamic model for binary panel data with unobserved heterogeneity admitting a n-consistent conditional estimator. *Econometrica* **78**(2) 719–733.

Bartolucci, F., V. Nigro. 2012. Pseudo conditional maximum likelihood estimation of the dynamic logit model for binary panel data. *Journal of Econometrics* **170**(1) 102–116.

Bavafa, H., L.M. Hitt, C. Terwiesch. 2018. The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* **64**(12) 5461–5480.

Bavafa, H., S. Savin, C. Terwiesch. 2019. Redesigning primary care delivery: Customized office revisit intervals and e-visits. *Available at SSRN 2363685* .

Bergstra, J., Y. Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13** 281–305.

Bidargaddi, N, G Schrader, P Klasnja, J Licinio, S Murphy. 2020. Designing m-health interventions for precision mental health support. *Translational psychiatry* **10**(1) 1–8.

Birnbaum, F., D.M. Lewis, R. Rosen, M.L. Ranney. 2015. Patient engagement and the design of digital health. *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine* **22**(6) 754.

Boutilier, J.J., T.C.Y. Chan. 2020. Ambulance emergency response optimization in developing countries. *Operations Research* **68**(5) 1315–1334.

Chamberlain, G. 1980. Analysis of Covariance with QualitativeData. *The Review of Economic Studies* **47**(1) 225–238.

Chaudhry, S.I., J.A. Mattera, J.P. Curtis, J.A. Spertus, J. Herrin, Z. Lin, C.O. Phillips, B.V. Hodshon, L.S. Cooper, H.M. Krumholz. 2010. Telemonitoring in patients with heart failure. *New England Journal of Medicine* **363**(24) 2301–2309.

Chaudhry, S.I., C.O. Phillips, S.S. Stewart, B. Riegel, J.A. Mattera, A.F. Jerant, H.M. Krumholz. 2007. Telemonitoring for patients with chronic heart failure: a systematic review. *Journal of Cardiac Failure* **13**(1) 56–62.

De Boeck, K., C. Decouttere, J.O. Jónasson, N. Vandaele. 2020. Vaccine supply chains in resource-limited settings: Mitigating the impact of rainy season disruptions *Working Paper*.

De Boeck, K., C. Decouttere, N. Vandaele. 2019. Vaccine distribution chains in low-and middle-income countries: A literature review. *Omega* **97**.

Delana, K., S. Deo, K. Ramdas, G. Babu, T. Ravilla. 2020. Multichannel delivery in healthcare: The impact of telemedicine centers in southern india. *Available at SSRN 3505318* .

Deo, S., M. Sohoni. 2015. Optimal decentralization of early infant diagnosis of hiv in resource-limited settings. *Manufacturing & Service Operations Management* **17**(2) 191–207.

DiMatteo, M.R., P.J. Giordani, H.S. Lepper, T.W. Croghan. 2002. Patient adherence and medical treatment outcomes a meta-analysis. *Medical Care* 794–811.

DiMatteo, M.R., K.B. Haskard-Zolnierek, L.R. Martin. 2012. Improving patient adherence: a three-factor model to guide practice. *Health Psychology Review* **6**(1) 74–91.

Druce, K.L., W.G. Dixon, J. McBeth. 2019. Maximizing engagement in mobile health studies: lessons learned and future directions. *Rheumatic Disease Clinics* **45**(2) 159–172.

Enos, M., J. Sitienei, J. Ong'ang'o, B. Mungai, M. Kamene, J. Wambugu, H. Kipruto, V. Manduku, J. Mburu, D. Nyaboke, et al. 2018. Kenya tuberculosis prevalence survey 2016: Challenges and opportunities of ending tb in kenya. *PloS One* **13**(12) e0209098.

Eskås, P.A., S. Heimark, J. Eek Mariampillai, A.C.K. Larstorp, F.E.M. Fadl Elmula, A. Høieggen. 2016. Adherence to medication and drug monitoring in apparent treatment-resistant hypertension. *Blood Pressure* **25**(4) 199–205.

Fox, W., et al. 1958. The problem of self-administration of drugs; with particular reference to pulmonary tuberculosis. *Tubercle* **39**(5) 269–74.

Gallien, J., I. Rashkova, R. Atun, P. Yadav. 2017. National drug stockout risks and the global fund disbursement process for procurement. *Production and Operations Management* **26**(6) 997–1014.

Grange, J.M., A. Zumla. 2002. The global emergency of tuberculosis: what is the cause? *The journal of the Royal Society for the Promotion of Health* **122**(2) 78–81.

Hasvold, P.E., R. Wootton. 2011. Use of telephone and sms reminders to improve attendance at hospital appointments: a systematic review. *Journal of Telemedicine and Telecare* **17**(7) 358–364.

Hovell, M.F., C.L. Sipan, E.J. Blumberg, C.R. Hofstetter, D. Slymen, L. Friedman, K. Moser, N.J. Kelley, A.Y. Vera. 2003. Increasing latino adolescents' adherence to treatment for latent tuberculosis infection: a controlled trial. *American Journal of Public Health* **93**(11) 1871–1877.

Jónasson, J.O., S. Deo, J. Gallien. 2017. Improving hiv early infant diagnosis supply chains in sub-saharan africa: Models and application to mozambique. *Operations Research* **65**(6) 1479–1493.

Kalichman, S.C., L. Eaton, M.O. Kalichman, T. Grebler, C. Merely, B. Welles. 2016. Race-based medical mistrust, medication beliefs and hiv treatment adherence: test of a mediation model in people living with hiv/aids. *Journal of Behavioral Medicine* **39**(6) 1056–1064.

Kardas, P., P. Lewek, M. Matyjaszczyk. 2013. Determinants of patient adherence: a review of systematic reviews. *Frontiers in Pharmacology* **4** 91.

Killian, J.A., B. Wilder, A. Sharma, V. Choudhary, B. Dilkina, M. Tambe. 2019. Learning to prescribe interventions for tuberculosis patients using digital adherence data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2430–2438.

Kvedar, J., M.J. Coye, W. Everett. 2014. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. *Health Affairs* **33**(2) 194–199.

Lekwijit, S.T., C. Terwiesch, D. Asch, K. Volpp. 2019. Evaluating the efficacy of connected healthcare: An empirical examination of patient engagement systems and their impact on readmission. *Available at SSRN 3420495* .

Leung, N.Z., A. Chen, P. Yadav, J. Gallien. 2016. The impact of inventory management on stock-outs of essential drugs in sub-saharan africa: secondary analysis of a field experiment in zambia. *PloS one* **11**(5).

Levi, R., G. Perakis, G. Romero. 2017. On the effectiveness of uniform subsidies in increasing market consumption. *Management Science* **63**(1) 40–57.

Marshall, G., J.W.S. Blacklock, C. Cameron, N.B. Capon, R. Cruickshank, J.H. Gaddum, F.R.G. Heaf, A.B. Hill, L.E. Houghton, J.C. Hoyle, et al. 1948. Streptomycin treatment of pulmonary tuberculosis: a medical research council investigation. *The BMJ* **2**(4582) 769–82.

Martins, N., P. Morris, P.M. Kelly. 2009. Food incentives to improve completion of tuberculosis treatment: randomised controlled trial in dili, timor-leste. *BMJ* **339** b4248.

Mills, S. 2020. Personalized nudging. *Behavioural Public Policy* 1–10.

Munro, S.A., S.A. Lewin, H.J. Smith, M.E. Engel, A. Fretheim, J. Volmink. 2007. Patient adherence to tuberculosis treatment: a systematic review of qualitative research. *PLoS Medicine* **4**(7).

Natarajan, K.V., J.M. Swaminathan. 2014. Inventory management in humanitarian operations: Impact of amount, schedule, and uncertainty in funding. *Manufacturing & Service Operations Management* **16**(4) 595–603.

Neyman, J., E.L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society* 1–32.

Nieuwlaat, R., N. Wilczynski, T. Navarro, N. Hobson, R. Jeffery, A. Keepanasseril, T. Agoritsas, N. Mistry, A. Iorio, S. Jack, et al. 2014. Interventions for enhancing medication adherence. *Cochrane Database of Systematic Reviews* (11).

Parvin, H., S. Beygi, J.E. Helm, P.S. Larson, M.P. Van Oyen. 2018. Distribution of medication considering information, transshipment, and clustering: Malaria in malawi. *Production and Operations Management* **27**(4) 774–797.

Rabbi, M., M.S.H. Aung, G. Gay, M.C. Reid, T. Choudhury. 2018. Feasibility and acceptability of mobile phone–based auto-personalized physical activity recommendations for chronic pain self-management: pilot study on adults. *Journal of medical Internet research* **20**(10) e10147.

Rajan, B., T. Tezcan, A. Seidmann. 2019. Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science* **65**(3) 1236–1267.

Richterman, A., J. Steer-Massaro, J. Jarolimova, L.B.L. Nguyen, J. Werdenberg, L.C. Ivers. 2018. Cash interventions to improve clinical outcomes for pulmonary tuberculosis: systematic review and meta-analysis. *Bulletin of the World Health Organization* **96**(7) 471.

Ruggeri, K., A. Benzerga, S. Verra, T. Folke. 2020. A behavioral approach to personalizing public health. *Behavioural Public Policy* 1–13.

Sacarny, A., D. Yokum, A. Finkelstein, S. Agrawal. 2016. Medicare letters to curb overprescribing of controlled substances had no detectable effect on providers. *Health Affairs* **35**(3) 471–479.

Saghafian, S., W.J. Hopp, S.M.R. Iravani, Y. Cheng, D. Diermeier. 2018. Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* **64**(11) 5180–5197.

Subbaraman, R., L. de Mondesert, A. Musiimenta, M. Pai, K.H. Mayer, B.E. Thomas, J. Haberer. 2018. Digital adherence technologies for the management of tuberculosis therapy: mapping the landscape and research priorities. *BMJ global health* **3**(5).

Suen, S., M.L. Brandeau, J.D. Goldhaber-Fiebert. 2018. Optimal timing of drug sensitivity testing for patients on first-line tuberculosis treatment. *Health Care Management Science* **21**(4) 632–646.

Suen, S., D. Negoescu, J. Goh. 2020. Design of incentive programs for optimal medication adherence. *Available at SSRN 3308510* .

Sullivan, B.J., B.E. Esmaili, C.K. Cunningham. 2017. Barriers to initiating tuberculosis treatment in sub-saharan africa: a systematic review focused on children and youth. *Global Health Action* **10**(1) 1290317.

Taylor, T.A., W. Xiao. 2014. Subsidizing the distribution channel: Donor funding to improve the availability of malaria drugs. *Management Science* **60**(10) 2461–2477.

Volmink, J., P. Garner. 2007. Directly observed therapy for treating tuberculosis. *Cochrane Database of systematic reviews* (4).

Walley, J.D., M.A. Khan, J.N. Newell, M.H. Khan. 2001. Effectiveness of the direct observation component of dots for tuberculosis: a randomised controlled trial in pakistan. *The Lancet* **357**(9257) 664–669.

WHO. 2009. A brief history of tuberculosis control in kenya. Tech. rep.

WHO. 2019. *Global Tuberculosis Report*. World Health Organization and others.

WHO. 2020. *Global Tuberculosis Report*. World Health Organization and others.

Yadav, A., B. Wilder, E. Rice, R. Petering, J. Craddock, A. Yoshioka-Maxwell, M. Hemler, L. Onasch-Vera, M. Tambe, D. Woo. 2018. Bridging the gap between theory and practice in influence maximization: Raising awareness about hiv among homeless youth. *IJCAI*. 5399–5403.

Yoeli, E., J. Rathauser, S.P. Bhanot, M.K. Kimenye, E. Mailu, E. Masini, P. Owiti, D. Rand. 2019. Digital health support in treatment for tuberculosis. *New England Journal of Medicine* **381**(10) 986–987.

Zhang, C., A. Atasu, T. Ayer, L.B. Toktay. 2020. Truthful mechanisms for medical surplus product allocation. *Manufacturing & Service Operations Management* **22**(4) 735–753.