Uncovering the Varied Impact of Behavioral Change Messages on Population Groups

Jiaai Xu JIAAIXU@UMICH.EDU
Computer Science
University of Illinois Urbana-Champaign

Rada Mihalcea MIHALCEA@UMICH.EDU
Computer Science and Engineering
University of Michigan, Ann Arbor

Elena Frank EMFRANK@UMICH.EDU
Michigan Neuroscience Institute
University of Michigan, Ann Arbor

Srijan Sen SRIJAN@MED.UMICH.EDU
Michigan Neuroscience Institute
University of Michigan, Ann Arbor

Maggie Makar MMAKAR@UMICH.EDU
Computer Science and Engineering
University of Michigan, Ann Arbor

Abstract

Individuals such as medical interns who work in high-stress environments often face mental health challenges including depression and anxiety. These challenges are exacerbated by the limited access to traditional mental health services due to demanding work schedules. In this context, mobile health interventions such as push notifications targeting behavioral modification to improve mental health outcomes could deliver much needed support. In this work, we study the effectiveness of these interventions on subgroups, by studying the conditional average causal effect of these interventions. We design a two step approach for estimating the conditional average causal effect of interventions and identifying specific subgroups of the population who respond positively or negatively to the interventions. The first step of our approach follows existing causal effect estimation approaches, while the second step involves a novel tree-based approach to identify subgroups who respond to the treatment. The novelty in the second step stems from a pruning approach that deploys hypothesis testing to identify subgroups experiencing a statistically significant positive or negative causal effect. Using a semi-simulated dataset, we show that our approach retrieves affected subpopulations with a higher precision than alternatives while maintaining the same recall and accuracy. Using a real dataset with randomized push interventions among the medical intern population at a large hospital, we show how our approach can be used to identify subgroups who might benefit the most from interventions.

1. Introduction

Mental health issues, including depression, are among the leading causes for disease-related disability worldwide (Friedrich, 2017), and have been exacerbated by the COVID-19 pandemic (Pfefferbaum and North, 2020). In particular, stressful work environments have been often associated with mental health issues (Stansfeld and Candy, 2006; Tennant, 2001). The persistence of stress in these environments, along with the shortage of readily available mental health care, makes it difficult to reduce these challenges. Among these, physician training through medical internships is known to be a highly stressful experience, with an average of 27% of individuals undertaking such training reporting various forms of depression (Rotenstein et al., 2016; Fang et al., 2022; Hughes et al., 2022; Meeks et al., 2022).

Challenges to the mental health of medical interns are often exacerbated by the fact that the demands of their training schedule leave little room to access traditional mental health care services. Against this backdrop, mobile health interventions (such as push notifications) that encourage behavioral changes with the purpose of improving mental health outcomes are important. Whether or not such interventions are effective is still an open question, with the majority of studies focusing on effectiveness on average (NeCamp et al., 2020) as opposed to effectiveness among different subgroups.

In this paper, we explore the effectiveness of behavioral change message interventions targeting depression among medical interns. We focus on identifying subgroups who experience a meaningful change in outcomes as a response to the intervention. Using data collected from 1,565 medical interns NeCamp et al. (2020), we aim to understand the causal effects of messaging interventions targeting changes in mood, sleep, and step count, and uncover the impact that different messages have for different population groups (e.g., gender, age, profession), thus making these interventions centered on the patient. The identification of these subgroups can help target future interventions towards groups who might benefit the most.

Subgroup discovery in the context of causal analysis is difficult because, unlike supervised learning, we never observe the true variable that we wish to cluster upon, which is the causal effect of the intervention. To address this challenge, we study a two-step process where in the first step we estimate the conditional average treatment effect (CATE) of the intervention. In the second step, we utilize a tree-based approach to identify different subgroups with heterogeneous responses to the treatment. Our work builds upon previous work by Makar et al. (2019), and addresses two main limitations of that prior work. First, the previous approach can lead to redundant subgroups, which arise when two groups are deemed as having different responses to the treatment, when in fact they respond similarly. Second, the previous work might lead to subgroups with highly variable response to the intervention, making it difficult to reliably target the groups who benefit the most from the intervention. This becomes a problem especially when the size of the subgroups is allowed to be small.

The main contribution of this paper is a pruning mechanism that enables identifying meaningful subgroups of the population who have a significant positive (or negative) response to the intervention. This pruning approach addresses the two limitations in the causal subgroup discovery approach suggested in previous work by recasting the problem as one of hypothesis testing. Specifically, our pruning strategy relies on testing the hypotheses that (1) a subgroup is meaningfully different from its neighbors, and (2) a subgroup has a meaningful positive or negative response to the intervention. Using semi-simulated data, we show that our approach is able to identify meaningful subgroups more accurately than previous approaches. Using the real data, we show that different intern characteristics are associated with different responses to interventions.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work has significant implications in the specific context of studying intern mental health, as well as the broader context of understanding the causal effect of messaging interventions on patient health outcomes.

In our specific context focusing on medical interns, our work is expected to lead to insights about designing targeted behavioral message interventions that aim to improve intern mental and behavioral health. By designing methods that identify subpopulations of interns who are likely to experience significant improvements because of the message interventions, our approach can lead to significant reductions in depression rates and stress levels among physicians in training. Such an improvement in intern mental health can lead to improvements in quality of patient care, as well as reduction in medical errors (Fahrenkopf et al. 2008; West et al. 2009; West et al. 2006).

In the broader context, our work will lead to the development of tools and analysis techniques to understand the causal effect of message interventions on patient health. Recently, there has been an increased interest in leveraging text-based interventions to help patients manage chronic illnesses (de Jongh et al. 2012) such as diabetes (Arambepola and RicciCabello 2016), asthma and HIV (Horvath et al. 2012), improve maternal and infant health (Poorman et al. 2015), as well as promote positive behavioral change (Armanasco et al. 2017). Careful analysis of the impact of these interventions on patient outcomes is necessary to identify the characteristics of the subpopulations that are positively impacted by the interventions as well as the characteristics of the interventions that lead to positive outcomes. The work described in this paper is a step forward toward the development of methods to serve that goal, and can be used to estimate the causal effect of message intervention on patient outcomes.

2. Related Work

Messaging interventions in healthcare. There is a significant amount of research on message interventions in healthcare, particularly in their effectiveness for improving outcomes for both patients and healthcare workers. One such study by Burner et al. (2014) examined the impact of mobile health interventions on Latino diabetic patients. The study found that text messages served as behavioral triggers to enhance diabetes management, and suggested personalized messages can further improved the effectiveness of the intervention. Fiol-DeRoque et al. (2021) evaluated the efficacy of mobile health interventions on the mental health of healthcare workers during the COVID-19 pandemic and showed significant positive improvements in mental health for those taking psychotherapy or pharmacological treatments. Most previous works have focused solely on the effectiveness of text messaging interventions in addressing existing health problems and there has been limited research on recommended intervention characteristics (Hall et al., 2015). Our study aims to address this gap by investigating the effectiveness of text messaging interventions on a broader population of medical interns and examining how different subgroups respond to these interventions. By doing so, we hope to provide more insight into the future design of intervention characteristics.

Decision tree pruning. Some traditional and widely used post-pruning method for regression trees include minimal cost-complexity pruning and reduced error pruning (Breiman, 2017; Quinlan, 1987). In recent years, several novel pruning methods have been proposed. For example, (Ahmed et al., 2018) introduced Pruning with Bayes Minimum Risk, which prunes the tree by comparing the risk-rate of each node with its children. Another approach proposed by Jeon and Lee (2014) is based on the estimation error and exclude nodes with a small number of the highest class at both node and tree levels. These methods primarily aims to reduce the tree size by estimating the error or cost and pruning the nodes contribute little to the accuracy. In this paper, we focus more on hypotheses testing to exclude meaningless nodes.

3. Intern Health Study Data

Our analysis leverages data from the Intern Health Study (IHS) (NeCamp et al., 2020). IHS is a micro-randomized trial (MRT) which followed medical interns for six months. At the beginning of their intern year and prior to randomization, participants completed a survey collecting basic personal information and characteristics like the Patient Health Questionnaire score which assessed depression. The interns were provided wearable devices (Fitbits) to record their activities and sleep data throughout the study. They were also instructed to download a study app from which they could self report their daily mood scores in response to the question “on a scale of 1-10 how was your mood today?”. Table 1 shows summary statistics of the intern population.

Each week, every intern was randomized to one of four possible messaging interventions: a week of mood notifications, activity notifications, sleep notifications, or no notifications. Examples of the notifications are included in table 2. 75% of the interns receive at least one notification over the duration of the study. The notification text is pulled randomly from one pool of messages without replacement and the pool for each subject would be refilled after the messages had been sent. The outcomes were average daily self-reported mood valence, average daily steps (as a proxy for activity), and average daily sleep duration. Additional details about the study design and collected data can be found in (NeCamp et al., 2020).

3.1. Data Extraction

We focus on estimating the causal effect of the first push notifications. To do so, we collect *pre-intervention variables* which are the intern characteristics at baseline, prior to randomization. These baseline characteristics were collected at the start of the residents' intern year. We define the *intervention period* as the first week of enrollment in the study, i.e., between July 1st, 2019 and July 7th, 2019. We extract all interventions that took place within that one week. We construct three intervention variables: T_{Mood} which takes on a value of 1 if the intern received a notification about mood and 0 otherwise, T_{Step}, T_{Sleep}, are similarly defined with respect to step and sleep. We define the *outcome period* as the

	Intervention N=1198	Nonintervention N=410	Difference
Age, mean	27.74	27.59	0.15
Gender			
Female %	54.76	54.63	0.13
Male %	45.24	45.37	-0.13
Race			
Caucasian %	56.93	56.59	0.34
Asian %	23.12	22.19	0.93
Mixed %	9.02	5.61	3.41
African American %	5.18	7.07	-1.89
Latino %	3.59	6.10	-2.51
Arab/Middle Eastern %	1.50	1.95	-0.45
Native American %	0.17	0	0.17
Other %	0.50	0.49	0.01
Marital status			
Single %	65.19	60.73	4.46
Married %	23.62	26.59	-2.97
Engaged %	10.52	11.22	-0.70
Separated/Divorced %	0.67	1.46	-0.79
Specialty			
Internal Medicine %	22.62	27.56	-4.94
Surgery %	15.19	12.20	2.99
Pediatrics %	11.60	12.44	-0.84
Emergency Medicine %	10.60	9.51	1.09
Family Practice %	7.76	8.29	-0.53
Psychiatry %	7.26	7.32	-0.06
Ob/Gyn %	6.09	9.02	-2.93
Anesthesiology %	5.18	2.68	2.49
Med/Peds %	2.92	1.95	0.97
Transitional %	2.75	3.41	-0.66
Neurology %	2.59	2.93	-0.34
Otolaryngology %	1.42	0.49	0.93
Other %	4.01	2.20	1.81

Table 1: Summary statistics describing the population of interns in the Intern Health Study

week following the intervention, i.e., July 8th, 2019 to July 14th, 2019. During the outcome period, we collect the outcomes corresponding to the interventions: Y_{Mood}, the average mood score submitted by the interns over the outcome period, Y_{Step}, and Y_{Sleep} are the daily step count, and sleep duration in minutes respectively. The latter three outcomes are automatically logged through the Fitbit. Figure 1 shows the full inclusion and exclusion criteria used in our analysis.

Notification type	Example messages
Mood	Whenever you feel down, remember how much you've helped other people since you started internship. That's all you.
Steps	Your mean daily steps of 10,315 last week was above average. Squeeze in even more with a 10 min walk around your building after you eat!
Sleep	Getting enough deep & REM sleep can positively impact your memory & mood. Prioritize sleep when you can to help you feel more refreshed.

Table 2: Examples of messaging interventions received by interns in the Intern Health Study

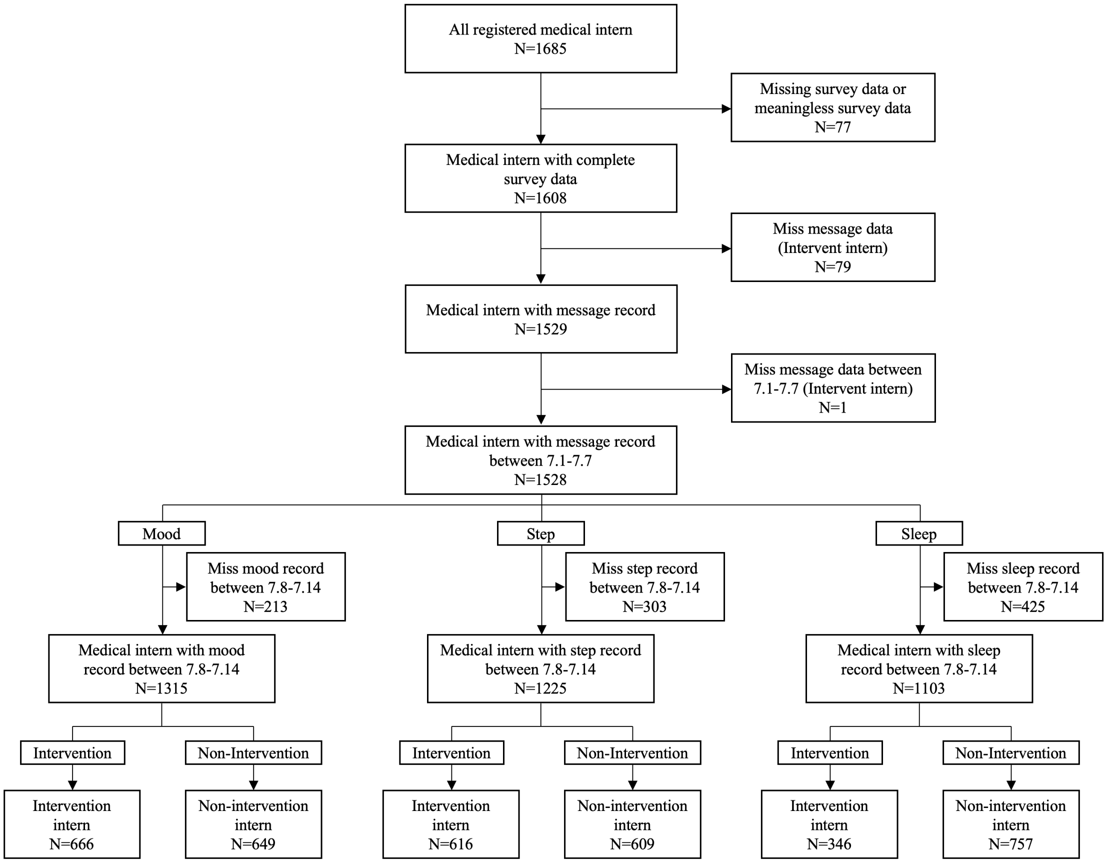
All registered medical intern N=1685
Missing survey data or meaningless survey data N=77
Medical intern with complete survey data N=1608
Miss message data (Intervent intern) N=79
Medical intern with message record N=1529
Miss message data between 7.1-7.7 (Intervent intern) N=1
Medical intern with message record between 7.1-7.7 N=1528
Mood
Step
Sleep
Miss mood record between 7.8-7.14 N=213
Miss step record between 7.8-7.14 N=303
Miss sleep record between 7.8-7.14 N=425
Medical intern with mood record between 7.8-7.14 N=1315
Medical intern with step record between 7.8-7.14 N=1225
Medical intern with sleep record between 7.8-7.14 N=1103
Intervention
Non-Intervention
Intervention
Non-Intervention
Intervention
Non-Intervention
Intervention intern N=666
Non-intervention intern N=649
Intervention intern N=616
Non-intervention intern N=609
Intervention intern N=346
Non-intervention intern N=757

Figure 1: Inclusion and exclusion criteria and demographics of study populations.

4. Technical Background and Preliminaries

We adopt the Neyman-Rubin potential outcomes framework (Rubin, 2005). Throughout, we will use capital letters to denote variables and small letters to denote their value. For each intern i, we observe a set of features $X_i \in \mathcal{X}$, with $\mathcal{X}$ a bounded subset of $\mathbb{R}^d$, a text message intervention $T_i \in \{0,1\}$ and an outcome $Y_i \in \mathbb{R}$. We observe these variables through samples $(x_1, t_1, y_1), ..., (x_n, t_n, y_n) \overset{i.i.d.}{\sim} p(X, T, Y)$. The observed outcome is one of the two *potential outcomes*, Y_0 and Y_1, under control $(T = 0)$ and treatment $(T = 1)$, respectively. Here the treatment takes on a value of 1 if the intern received a messaging intervention, and 0 otherwise. Y_t can be $\in \{0,1\}$ or in a bounded subset of $\mathbb{R}^d$. For mood, $Y_t \in [1, 10]$; for steps and sleep, $Y_t \in \mathbb{R}$. While these interventions are text messages and hence not treatments per se, we stick to this treatment nomenclature to conform with the rest of the causal inference literature.

We define the conditional average treatment effect (CATE) as the expected difference in the outcome under treatment and non-treatment. Specifically, we define CATE as:

$$\tau(x) = \mathbb{E}[Y_1 - Y_0 | X = x].$$

Our goal is to identify subgroups defined by X that experience a meaningful change in their mental and behavior health because of the treatment.

5. Approach

Our main goal is to identify distinct subgroups of interns who experience a significant change to their behavior and mental health as a result of the mobile health intervention. Our estimation strategy follows three steps. First, we estimate the causal effect of the intervention on the interns as outlined in section 5.1. Second, we identify subgroups with heterogeneous treatment effects as outlined in section 5.2. These two steps follow previous work by Makar et al. (2019). Third, we design a novel tree-pruning method targeted towards identifying subgroups that experience a statistically significant change (improvement or decline) in their behavioral and mental health as a result of the intervention, and pruning out ones who do not. This approach is outlined in detail in section 5.3.

We describe those three steps in detail below.

5.1. CATE Estimation

We divide the training data into three subgroups $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{D}_3$. The first two will be utilized to identify all subgroups of interns with heterogeneous CATE, while the last will be used for our novel pruning approach, described in the next section. In order to identify all subgroups of interns experience heterogeneous effects, we first estimate their CATE. We do so be estimating the nuisance parameter, $m_t(x) := \mathbb{E}[Y = y \mid X = x, T = t]$, which is a mapping from the interns' pre-treatment characteristics to the expected value of Y under some treatment decision t, the observed outcomes learned by minimizing:

$$\hat{m}_t = \min_{m_t \in \mathcal{M}_t} \frac{1}{|\mathcal{D}_i^t|} \sum_{i \in \mathcal{D}_i^t} \mathcal{L}(m_t(x), y), \quad t \in \{0,1\} \tag{1}$$

where $\mathcal{D}_i^t$ is the set of interns in $\mathcal{D}_1$ who received treatment t, $|A|$ is the cardinality of the set A, $\mathcal{L}$ is some loss function such as the logistic loss or mean squared error, $\mathcal{M}_t$ is an appropriately chosen function class such as linear models, Bayesian additive regression trees, random forests or deep neural networks. Different estimators of the nuisance parameters can be utilized. For example, instead of estimating two different mappings for each treatment group (i.e., T-learners), it is possible to estimate a single function $m(x,t)$ (i.e., S-learners) (Künzel et al., 2019). Since our analysis is based on data from a randomized control trial, we do not need to estimate other nuisance parameters such as the propensity to treat in order to adjust for confounding. However, our suggested approach is trivially extendable to situations where the data is collected from observational data with a biased treatment assignment. Such an extension would require estimating the propensity score $e(x,t) := \mathbb{E}[T = t \mid X = x]$ to re-weight the cohort using importance weighting, i.e., $w_t(x) := \mathbb{E}[T = t]/e(x,t)$. In that case, equation (1) is replaced with:

$$\hat{m}_t = \min_{m_t \in \mathcal{M}_t} \sum_{i \in \mathcal{D}_i^t} \tilde{w}_t \mathcal{L}(m_t(x), y), \quad t \in \{0, 1\}, \tag{2}$$

where $\tilde{w}_t$ is the normalized version of w_t such that $\sum_{i \in \mathcal{D}_i^t} \tilde{w}_t = 1$.

Without loss of generality, we assume that a T-learner is used to estimate the conditional outcomes under treatment and non-treatment. The CATE can then be imputed as follows for all $i \in \mathcal{D}_2$:

$$\hat{\tau}(x_i) = \hat{m}_1(x_i) - \hat{m}_0(x_i) \ . \tag{3}$$

5.2. Identifying All Subgroups

In the second step, we seek to identify all the subgroups that display heterogeneity in the treatment effect. As Makar et al. (2019) show, such a task can be done using decision trees, which, by virtue of their splitting function, partition the input space $\mathcal{X}$, according to the similarity in the outcome. Here the outcome is the imputed CATE, $\tau^*(x)$. To get the subgroup partitioning, we minimize the following objective:

$$\hat{\Pi}, \hat{\boldsymbol{\mu}} = \sum_j \operatorname{argmin} \left\{ \frac{1}{\#(i : i \in \ell_j)} \sum_{i \in \mathcal{D}_2} \Big(\hat{\tau}(x_i) - \mu_j(\ell_j) \Big)^2 \right\}, \tag{4}$$

where Π is a partition over the input space, $\mu_j(\ell_j)$ is the mean of leaf j, $\boldsymbol{\mu} = \{\mu_j\}$ for all j. In other words, each terminal leaf j in the partitioning $\hat{\Pi}$ denotes a distinct subgroup, and the corresponding $\hat{\mu}_j$ is the estimated response of the j^{th} group.

An important question remains: how do we regularize or prune this tree? In the original paper, Makar et al. (2019) were primarily studying settings where the ability to collect data about the interns at test time (i.e., when deciding which treatment to administer to new interns) is limited. In such a case, the tree should only be grown to a depth that is consistent with the number of features that can be collected at test time. In our case, this constraint does not exist. Instead, our focus is to identify subgroups that experience *meaningful* change.

5.3. Identifying Subgroups with Significant Change

We suggest a novel approach to identify subgroups that experience a significant change as a result of the treatment. Our approach is a bottom-up pruning strategy which attempts to identify meaningful subgroups by (1) ensuring that each split corresponds to two *significantly different* subgroups and (2) identifying subgroups who have a *meaningful* response to the treatment. We say that a subgroup's response is meaningful if it is statistically significantly different from zero.

Our approach proceeds by discarding of $\hat{\boldsymbol{\mu}}$, the estimates of the mean CATE at each of the terminal leaves acquired from the previous step. Instead, we use $\mathcal{D}_3$ to estimate $\tilde{\boldsymbol{\mu}}$, which is the estimates of the mean CATE at each of the terminal leaves, as well as $\tilde{\boldsymbol{\sigma}}$, which is the estimate of the standard deviation in each of the terminal leaves. We note that using the a separate, third dataset $\mathcal{D}_3$ that has not been used to train the decision tree (i.e., identify the partitioning rules) is important to avoid issues of bias. By doing so, the confidence intervals estimated in the third step are valid, making this an "honest" approach (Athey and Imbens, 2016). We estimate the confidence interval of each of the terminal leaves by estimating $\tilde{\mu}_j \pm z\frac{\tilde{\sigma}}{|\ell_j|}$, where z is the critical value, and it depends on the required level of confidence. We then identify which of the terminal leaves have a confidence interval that crosses zero. For each terminal leaf, if its corresponding confidence interval crosses zero, we regard the terminal leaf as significant and if it doesn't, we regard the terminal leaf as insignificant. Then if a node is a significant leaf node or it is on the branch of a significant leaf node, we keep it. Otherwise, we prune it. Then for the pruned tree, we only have the significant terminal leaves and their pathways.

We note that issues relating to multiple hypothesis testing arise in our pruning approach outlined in this section and section 5.2. We follow other authors in setting a lower significance level (0.001) as a way to address these issues Zhang et al. (2012). In general, a Bonferroni adjustment could be used to address issues relating to multiple hypothesis testing here.

6. Experiments

Due to the fundamental problem of causal inference, we cannot directly evaluate our approach using real data because the true causal effect is never observed. In addition, the true subgroups are unknown even if the causal effect is observed. For these reasons, we follow other authors (Hill, 2011; Dorie et al., 2019; Shalit et al., 2017; Shi et al., 2019) in evaluating our approach on semi-simulated data in addition to our analysis on the real data. Specifically, we extract the intern characteristics and their intervention assignment from the real data, but simulate their potential outcomes and hence their causal effect.

6.1. Simulation

Setup. We set up the simulation such that causal effect of the intervention varies by the interns' specialties. The following equations summarize how we simulate the potential outcomes under treatment and non-treatment:

$$Y_0(X) = 0.5X_{\text{Surg}} + 0.8X_{\text{EM}} + 1\min\{X_{\text{IM}}, X_{\text{FM}}, X_{\text{Peds}}\} + 1.5\min\{X_{\text{Psych}}, X_{\text{OBG}}\} \tag{5}$$
$$Y_1(X) = Y_0(X) + 5X_{\text{EM}} + 4.4X_{\text{Peds}} + 3.4X_{\text{OBG}} + 3.1X_{\text{FM}} + 3\min\{X_{\text{IM}}, X_{\text{Psych}}\} + 1X_{\text{Surg}}, \tag{6}$$

where X_{Surg} is a binary variable taking on a value of 1 when the intern speciality is surgery. Similarly with X_{EM} (emergency medicine), X_{IM} (internal medicine), X_{FM} (family medicine), X_{Peds} (pediatrics), X_{Psych} (Psychiatry), and X_{OBG} (obstetrics and gynaecology). The simulation setup and parameters were chosen two reflect there are some subgroups that experience similar yet non-identical causal effect. For example, internal medicine interns have a CATE equal to 2, while family medicine interns have a CATE equal to 2.1.

In this simulation setting, we randomly assign the intervention T by drawing it from a binomial distribution. Specifically, we set $T \sim \text{Binomial}(0.5)$. We set the observed outcome to be the outcome under the randomly assigned intervention with some added noise drawn from a normal distribution with mean equal to 0 and standard deviation equal to 0.5. Specifically:

$$y_i = t_i \cdot Y_1(x_i) + (1 - t_i)Y_0(x_i) + \varepsilon, \text{ with } \varepsilon \sim \text{Normal}(0, 0.5)$$

We split the data into 70% training and validation and 30% held out for testing. We simulate 50 different datasets with different train/test splits and different noise draws. We report the average and standard deviation of all performance metrics over all 50 simulations.

Baseline and Implementation. We compare our approach to **classic-pruned** trees, which follow the same procedure as ours outlined in section 5 but does not identify subgroups with significant change. Specifically, both our approach and the classic-pruned baseline proceed by splitting the training and validation data into two datasets, each comprising of 50% of the training and validation data. The first is used for the CATE estimation step while the second is used for the tree building. In the CATE estimation step we use T-learners, which split the data into two groups based on the observed treatment assignment at train a different model that maps the intern features to the likely value of the outcome Künzel et al. (2019). We fit an L1 regularized model for each of the two treatment groups with 3-fold cross validation to pick the regularization parameter from the following candidate values $\{1e^{-4}, 1e^{-3}, 1e^{-2}, 0.1, 1\}$. The remaining 50% of the training and validation data is used for the tree building step. In the tree building step, we use 3 fold cross validation to pick the maximum depth of the tree, picking from the following possible values $\{5, 6, 7, 8, 9, 10, 20, 50, 100\}$. In addition to those two steps, our approach proceeds with the two steps of identifying significant subgroups as detailed in sections 5.2 and 5.3.

Evaluation. We evaluate the performance of our approach and the baseline on two different fronts. First, is the accuracy of the estimated CATE. We compute the mean squared error (MSE) between the true CATE and the estimated CATE, which is sometimes referred to as the precision of estimation in heterogeneous effects, to evaluate the accuracy of the estimated CATE. In this semi-simulated setting, we have access to the true CATE, which enables us to estimate the MSE. Second, we evaluate the accuracy of identifying relevant subgroups. To do so, we measure the precision and recall of the different subgroups created by the trees. Specifically, the precision and recall of an arbitrary partitioning Π are defined as follows:

$$\text{Precision}(\Pi) = \frac{1}{n} \sum_i \frac{|\{\text{Relevant features}\} \cap \{\text{Features}_i(\Pi)\}|}{|\{\text{Features}_i(\Pi)\}|},$$

$$\text{Recall}(\Pi) = \frac{1}{n} \sum_i \frac{|\{\text{Relevant features}\} \cap \{\text{ Features}_i(\Pi)\}|}{|\{\text{Relevant Features}\}|},$$

where $\text{Features}_i(\Pi)$ are the features that lie on the pathway to the terminal node corresponding to intern i's prediction and the relevant features are the features encoding the interns' speciality, outlined in equations 5 and 6.

Results. Table 3 shows the results of the semi-simulated setting. We see that our approach has a better precision in identifying relevant subgroups, without sacrificing recall or the accuracy of the CATE estimate as measured by the MSE.

	Precision	Recall	MSE
Ours	0.45 (0.04)	0.46 (0.03)	0.03 (0.01)
Classic-pruned	0.39 (0.03)	0.46 (0.03)	0.03 (0.01)

Table 3: Results on the semi-synthetic data. Our approach has a better precision in identifying relevant subgroups, without sacrificing recall or the accuracy of the CATE estimate as measured by the MSE.

6.2. Real Data

We apply our approach to the real data; in this setting, the true CATE and the true relevant subgroups are unknown, which makes quantitative evaluation challenging. Our main aim here is to present the qualitative results obtained by applying our approach to the real data, and discuss the implications of these results for designing targeted interventions.

Implementation. Here, we split the data into 2/3 for CATE estimation and 1/3 for tree building and pruning. For the CATE estimation step, we fit an L1 regularized model with 4-fold cross validation to pick the regularization parameter from the following candidate values $\{1e^{-4}, 1e^{-3}, 1e^{-2}, 0.1, 1, 10, 1e^{2}, 1e^{3}\}$. For the tree building step, we used 4-fold cross validation to pick the maximum depth of the tree, picking values between 1 and 11.

Causal Effect of Mood Interventions on Reported Mood Score. Here, we study the effect of mood interventions (T_{Mood}) on the mood score as reported by the interns (Y_{Mood}). Figure 2 shows the decision tree obtained by following our approach. Blue leaves represent subgroups that experience a significant non-zero effect. Examining the tree shows that Anesthesiology interns experience a positive causal effect as a result of mood interventions. This improvement in mood is independent of other demographic characteristics, or baseline depression scores. Non-Anesthesiology interns who have a high PHQ score (measuring baseline depression) experience a negative causal effect corresponding to a deterioration in their mood scores because of mood interventions. This finding points to the idea that mobile health interventions might have a negative effect on some sub-populations who have a high baseline depression score.

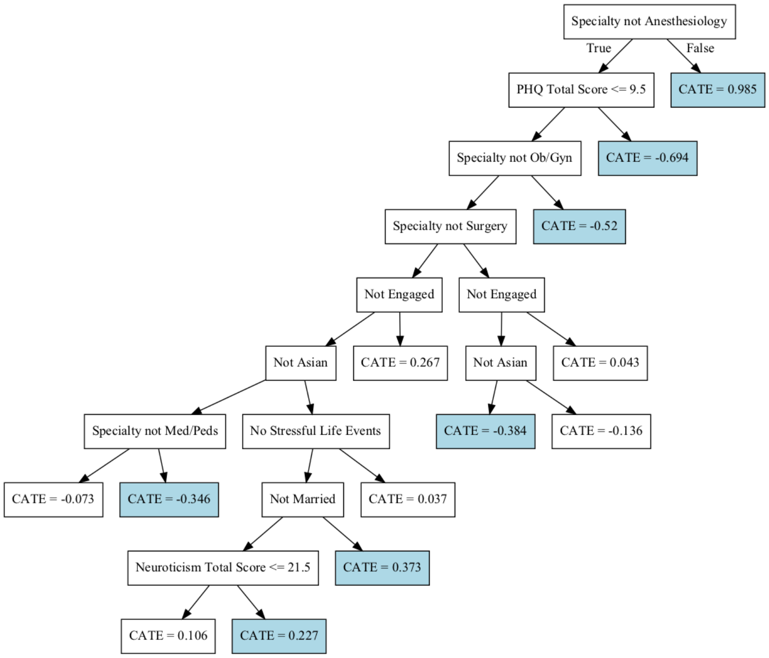

Figure 2: Causal effect of mood interventions on mood outcome. Figure shows the decision tree obtained by following our approach. Blue leaves represent groups that experience a significant non-zero effect.

Causal Effect of Sleep Intervention on Duration of Sleep in Minutes. Here, we study the effect of sleep interventions (T_{Sleep}) on the duration of sleep in minutes as measured by the Fitbit (Y_{Sleep}). Figures 3 and 4 show the decision tree obtained by following our approach, with blue leaves defined similarly as before. To improve readability, we split the tree into two figures representing the right subtree (figure 3) and the left subtree (figure 4). Interestingly, we find that demographic characteristics play a more prominent role in shaping the response to the intervention compared to the mood study. Specifically, sex and race of the intern are the first two characteristics used to split the different subgroups. For example, African American men experience an average of 28 minute increase in sleep because of sleep interventions, whereas Caucasian men have a more heterogeneous response depending on their specialty and their marital status.

Causal Effect of Step intervention on Step count. Next, we study the effect of step interventions (T_{Step}) on the daily step count as measured by the Fitbit (Y_{Step}). Figure 5 and 4 show the decision tree obtained by following our approach, with blue leaves defined similarly as before. To improve readability, we split the tree into two figures representing the right subtree (figure 5) and the left subtree (figure 6). The results here show that the baseline reported personal history of depression is one of the main factors associated

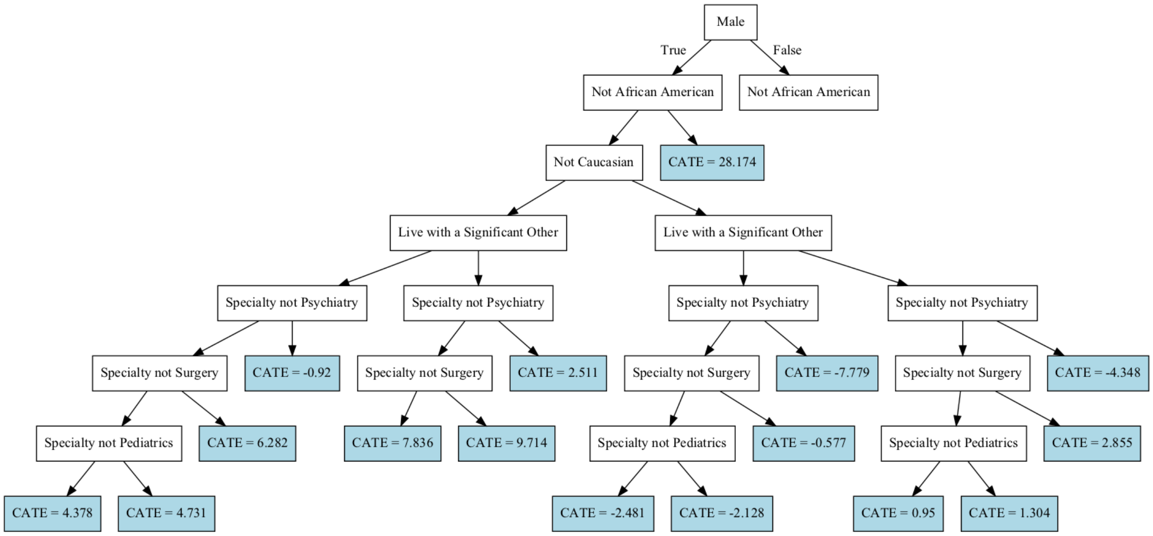

Figure 3: Causal effect of sleep interventions on sleep outcome. Figure shows the left half of the decision tree obtained by following our approach. Blue leaves represent groups that experience a significant non-zero effect.

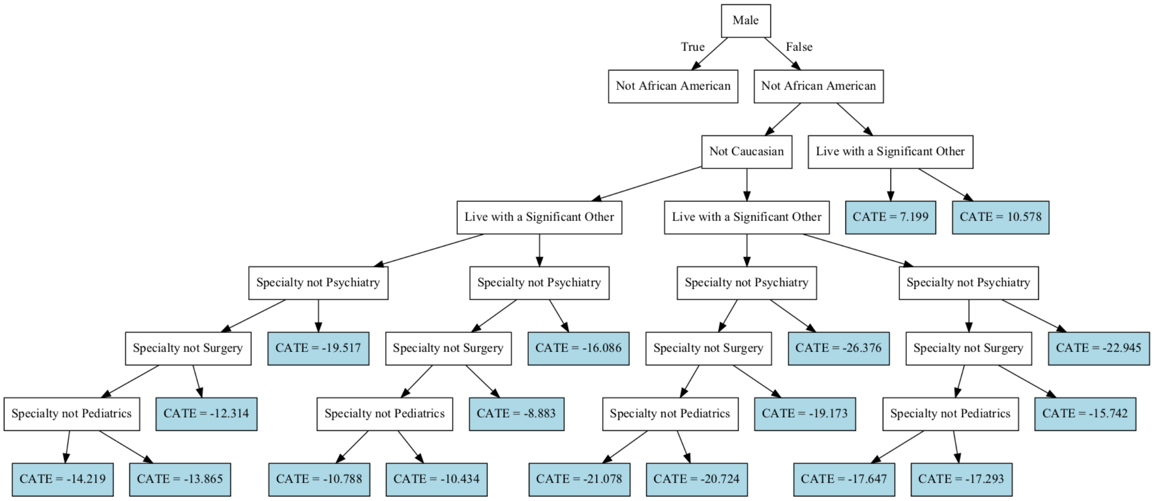

Figure 4: Causal effect of sleep interventions on sleep outcome. Figure shows the right half of the decision tree obtained by following our approach. Blue leaves represent groups that experience a significant non-zero effect.

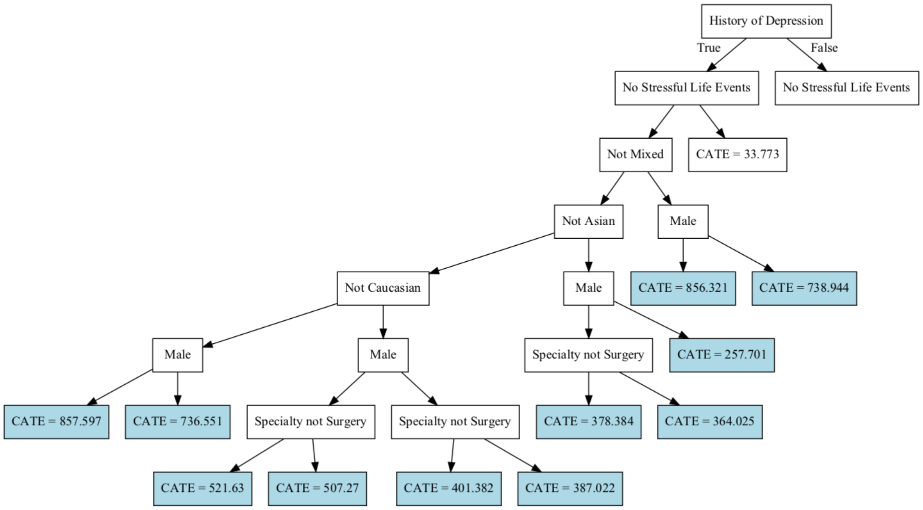

Figure 5: Causal effect of step interventions on step outcome. Figure shows the left half of the decision tree obtained by following our approach. Blue leaves represent groups that experience a significant non-zero effect.

with the effectiveness of the intervention: all subgroups of interns who did not report the presence of a personal history with depression experience an increase in the average step count because of the intervention. The effectiveness among the group of interns who report a personal history of depression is mixed: subgroups who report experiencing a stressful life event in addition to a personal history of depression tend to experience a reduction in their step count because of the intervention.

Limitations. There are a number of limitations to our study. First, we considered a "snapshot" setting where we assess the causal effect of the intervention at a single point in time. A more comprehensive look at the causal effect of these push notifications should take into account the dynamic, time varying nature of the intervention. Future extensions of our work will include developing our subgroup discovery method to accommodate time varying interventions. Second, the subgroups themselves should not be interpreted causally. While the final estimate has a causal interpretation, the subgroups themselves are groups that are associated with a positive or negative causal effect. We stress that the interpretation of our discovery approach should be, for example "Being an African American man is *associated* with a positive *causal* effect for the sleep intervention". In our setting, this is sufficient since the subgroups are defined with respect to variables that we cannot intervene upon (e.g., we cannot change an intern's sex, race or specialty). However, in settings where it is possible to intervene upon and change the characteristics of the subgroups, methods relating to causal discovery (e.g., Zhang et al. (2012)) might be more suitable here.

Where Next From Here. In addition to considering a time varying, dynamic intervention, we hope to extend our work next to different classes of interventions. For example, while a subset of the messages are based on cognitive therapy, others are based on mindfulness or on motivational interviewing, among others. In future work, we will look at all

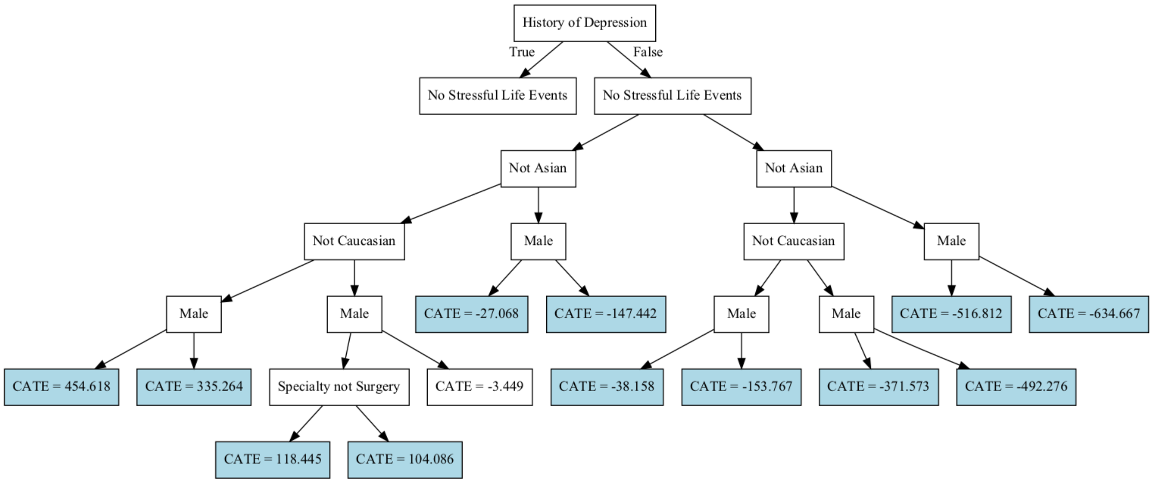

Figure 6: Causal effect of step interventions on step outcome. Figure shows the right half of the decision tree obtained by following our approach. Blue leaves represent groups that experience a significant non-zero effect.

possible interventions as defined with respect to both the type of intervention (sleep, step, mood) as well as its class. Moreover, instead of defining the intervention to be a binary variable, we will consider CATE approaches for high dimensional interventions.

7. Conclusion

In this work, we studied the identification of subgroups of a population who respond positively or negatively to an intervention of interest. We presented a novel approach that relies on pruning trees using a series of hypothesis tests. Using semi-synthetic data, we showed that our approach is able to retrieve meaningful subpopulations with better precision while maintaining the same recall and accuracy. Using a real dataset with interventions aiming to promote mental health among medical interns, we showed that our approach can be used to identify meaningful subgroups who respond differently to push notifications about sleep, exercise and mood patterns.

Acknowledgments

The authors are grateful for feedback from the anonymous reviewers. This work was partially supported by the National Science Foundation under Grant No. 2153083, the University of Michigan Eisenberg Family Depression Center, and the Templeton Foundation (#62256). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Eisenberg Family Depression Center, or the Templeton Foundation.

References

Ahmed Mohamed Ahmed, Ahmet Rizaner, and Ali Hakan Ulusoy. A novel decision tree classification based on post-pruning with bayes minimum risk. *Plos one*, 13(4):e0194168, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Leo Breiman. *Classification and regression trees*. Routledge, 2017.

Elizabeth R Burner, Michael D Menchine, Katrina Kubicek, Marisela Robles, and Sanjay Arora. Perceptions of successful cues to action and opportunities to augment behavioral triggers in diabetes self-management: qualitative analysis of a mobile intervention for low-income latinos with diabetes. *Journal of medical Internet research*, 16(1):e2881, 2014.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. 2019.

Yu Fang, Sara Lodi, Tasha M Hughes, Elena Frank, Srijan Sen, and Amy SB Bohnert. Work hours and depression in us first-year physicians. *New England Journal of Medicine*, 387(16):1522–1524, 2022.

Maria Antònia Fiol-DeRoque, Maria Jesús Serrano-Ripoll, Rafael Jiménez, Rocío Zamanillo-Campos, Aina María Yáñez-Juan, Miquel Bennasar-Veny, Alfonso Leiva, Elena Gervilla, M Esther García-Buades, Mauro García-Toro, et al. A mobile phone–based intervention to reduce mental health problems in health care workers during the covid-19 pandemic (psycovidapp): randomized controlled trial. *JMIR mHealth and uHealth*, 9(5): e27039, 2021.

Mary Jane Friedrich. Depression is the leading cause of disability around the world. *Jama*, 317(15):1517–1517, 2017.

Amanda K Hall, Heather Cole-Lewis, and Jay M Bernhardt. Mobile text messaging for health: a systematic review of reviews. *Annual review of public health*, 36:393–415, 2015.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Tasha M Hughes, Jennifer F Waljee, Yu Fang, Srijan Sen, and Amy Bohnert. New-onset depression among surgical interns. *JAMA surgery*, 157(6):543–545, 2022.

Hae Sook Jeon and Won Don Lee. Performance measurement of decision tree excluding insignificant leaf nodes. In *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 122–127. IEEE, 2014.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Maggie Makar, Adith Swaminathan, and Emre Kıcıman. A distillation approach to data efficient individual treatment effect estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4544–4551, 2019.

Lisa M Meeks, Jennifer Cleary, Adam Horwitz, Karina Pereira-Lima, Zhuo Zhao, Yu Fang, and Srijan Sen. Analysis of depressive symptoms and perceived impairment among physicians across intern year. *JAMA Network Open*, 5(1):e2144919–e2144919, 2022.

Timothy NeCamp, Srijan Sen, Elena Frank, Maureen A Walton, Edward L Ionides, Yu Fang, Ambuj Tewari, and Zhenke Wu. Assessing real-time moderation for developing adaptive mobile health interventions for medical interns: micro-randomized trial. *Journal of medical Internet research*, 22(3):e15033, 2020.

Betty Pfefferbaum and Carol S North. Mental health and the covid-19 pandemic. *New England journal of medicine*, 383(6):510–512, 2020.

J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

Lisa S Rotenstein, Marco A Ramos, Matthew Torre, J Bradley Segal, Michael J Peluso, Constance Guille, Srijan Sen, and Douglas A Mata. Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis. *Jama*, 316(21):2214–2236, 2016.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Stephen Stansfeld and Bridget Candy. Psychosocial work environment and mental health—a meta-analytic review. *Scandinavian journal of work, environment & health*, pages 443–462, 2006.

Christopher Tennant. Work-related stress and depressive disorders. *Journal of psychosomatic research*, 51(5):697–704, 2001.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.