Predicting differential treatment outcomes in randomized clinical trials:

A comparison of model-based and machine learning approaches

By

Gloria T. Han

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

August 7th, 2020

Nashville, Tennessee

Approved:

Andrew J. Tomarken, Ph.D.

Lynn S. Walker, Ph.D.

Steven D. Hollon, Ph.D.

David J. Lubinski, Ph.D.

To Charlotte,

whose constant warmth and companionship

during the year of 2020, in the midst of a worldwide pandemic,

I could not repay,

even with a lifetime of dog treats and belly rubs.

ACKNOWLEDGEMENTS

This work would not be possible without the faithful mentorship of my advisor, Andrew J. Tomarken, PhD. Almost 7 years ago, I began my graduate studies with eagerness to combine dual interests in quantitative methods and clinical psychological science, and quickly realized that this endeavor would not be possible without support and guidance from a mentor with unparalleled patience and expertise. Over the years, Dr. Tomarken has enthusiastically encouraged me to tackle challenging statistical topics in clinical psychology research and has persisted, even when my confidence has wavered. This journey has taken on many functional forms. Rarely linear, the trajectory is best characterized as piecewise with moments of stochastic insight. I am thoroughly grateful for the countless weekly meetings and insightful discussions that have strengthened my appreciation for critical and creative thinking in psychopathology research. Most importantly, Dr. Tomarken has taught me to never give up and “when you fall off the bike, you must get back up and ride.”

I am also thoroughly grateful for and indebted to the wonderful mentorship and generosity of Dr. Lynn Walker, who welcomed me to her lab and allowed me to apply my quantitative interests to her exciting program of research, and supported my travel to England and Switzerland to learn from experts in the field. I would also like to thank Drs. Steven Hollon and David Lubinski who have given their time and expertise to shape this dissertation in helpful ways. Last but not least, I am whole-heartedly grateful to my family, friends, and colleagues for providing constant support and cheering me on along the way.

TABLE OF CONTENTS

DEDICATIONii

ACKNOWLEDGEMENTSiii

LIST OF TABLESvi

LIST OF FIGURESvii

Chapter Page

1. Introduction1

Assessing Moderators Using Model-Based Regression2

Limitations of Model-Based Approaches2

Identifying Patient Predictors Using Statistical Learning3

Shift Toward Machine Learning4

A Tale of Two Data Analytic Cultures4

Current Study7

1. Motivating Dataset: A Randomized Clinical Trial (RCT) for Youth with Chronic Abdominal Pain10

Study Background10

Chronic Pain as a Biopsychosocial Phenomenon11

Study Design and Overview of Methods11

Model-Based Analysis and Evidence of Treatment x Subgroup Interaction18

Rationale for Using the Dataset for a Comparison Study21

1. Overview of Selected Modeling Approaches23

Model-Based Approaches: Linear regression and its Extensions25

Machine Learning Approaches: Recursive Partitioning29

Assessing Model Performance34

1. Methods Comparison Study: Data Analytic Plan38

Modeling Approaches39

Model Performance Indices40

Model Comparison Procedure41

Comparison of Significant Predictors42

1. Methods Comparison Study: Results44

Model Performance44

Predictor Performance44

Follow-Up Analysis: A Forest of Stumps *versus* Trees50

Probing Underlying Interactions Using Conditional Inference Trees54

1. Discussion59

Methodological Insights: Model Performance and Selection59

Substantive Insights: Personalizing Interventions for Chronic Abdominal Pain61

Limitations and Future Directions65

Conclusions66

REFERENCES68

LIST OF TABLES

Table Page

1. Strengths and limitations of model-based and machine learning approaches9
2. Demographic factors and clinical characteristics by pain phenotype 16
3. Summary of measures, collection timepoints, informant, and variable types17
4. Estimated slopes from pre- to post- treatment20
5. Definitions of model performance indices used in regression42
6. Model performance results comparing four models across three outcome variables46
7. Model performance indices for conditional random forest models48
8. Model performance indices examining conditional random forest specifying a forest of stumps *versus* a forest of trees for all outcome variables51

LIST OF FIGURES

Figure Page

1. Consolidated Standards of Reporting Trials (CONSORT) chart14
2. Overview of Web-MAP treatment content for patient and parent/caregiver15
3. Effect of treatment on GI symptoms and abdominal pain by pain phenotype21
4. Conceptual overlap between inferential statistics, statistical learning, and machine learning24
5. Representation of a hypothetical regression tree32
6. Comparison of correlation between observed and predicted values47
7. Variable importance plots for final conditional random forest models predicting GI symptoms at post-treatment48
8. Variable importance plots for final conditional random forest models predicting abdominal pain at post-treatment49
9. Variable importance plots for final conditional random forest models predicting pain interference at post-treatment50
10. Distributions of correlations between observed and predicted values for the conditional random forest algorithm allowing interactions *versus* prohibiting interactions53
11. Representative conditional inference tree predicting GI symptoms at post-treatment55
12. Representative conditional inference tree predicting abdominal pain at post-treatment56
13. Representative conditional inference tree predicting pain interference at post-treatment57

CHAPTER 1

INTRODUCTION

Despite ongoing and significant advancements in clinical psychology research, heterogeneity and complexity remain constant themes in our understanding of the diagnostic categories and underlying etiologies of psychopathology (Lahey et al., 2017). Consequently, longitudinal prognoses of psychological conditions remain unclear and results of treatment studies are marked by significant heterogeneity in patient response (Hofmann et al., 2012; Rush et al., 2006). As a result, clinicians often make treatment decisions under prognostic uncertainty and our best psychopharmacologic and psychotherapeutic interventions are only effective in 30 to 50% of patients (Kessler et al., 2016; Kessler, 2018; Rush et al., 2006).

One factor that negatively impacts our ability to effectively allocate treatments for psychopathology is the dearth of clinical decision-making tools that personalize treatments to the individual. Such tools are common in other areas of medicine, such as oncology (Pfeiffer et al., 2013), where algorithmic prediction tools called “nomograms” (for examples, see Memorial Sloan Kettering Cancer Center: <https://www.mskcc.org/nomograms>) serve critical roles in medical decision-making and outperform clinician judgement (Kattan et al., 2013). These prediction tools are consistent with the National Research Council’s charge towards “precision medicine,” (2011) which refers to the practice of improving health outcomes by subgrouping patients and tailoring care to biological, environmental, and lifestyle characteristics rather than treating patients as a homogeneous group.

In clinical research, “gold standard” randomized clinical trials (RCTs) are utilized to compare two or more treatments by examining differences in the *average* treatment effects in a target clinical population. Even in well-characterized clinical samples, such emphasis on average effects inadequately accounts for heterogeneity in treatment effects across the trial population, resulting in imprecise, “one size fits all” treatment recommendations (Subramanian et al., 2018). Acknowledging this shortcoming, researchers have attempted to clarify “for whom and under what conditions” to allocate treatments by testing specific patient characteristics or identifying patient subgroups as potential moderators of treatment outcome using traditional, model-based regression approaches (DeRubeis et al., 2014; Fournier et al., 2009; Kraemer, 2013; Kraemer et al., 2002; Petkova et al., 2020).

**Assessing Moderators Using Model-Based Regression**

Assessment of moderators using model-based regression approaches requires the *a priori* specification of interaction effects that are denoted by product terms in the regression model. Thus, these studies tend to focus on testing one or a few potential moderators at a time. This approach is common for clinical researchers, in part because graduate psychological statistics curricula include and emphasize more traditional regression methods for identifying treatment modifiers in lieu of more advanced regression or exploratory techniques (Aiken et al., 2008). Further, methods to detect differential treatment outcome across levels of a single moderator are well-established and relatively straight-forward (Aiken, West, & Reno, 1991; Preacher, Curran, & Bauer, 2006). This traditional approach for identifying treatment moderators may be particularly useful in cases where potential moderators are well-established from previous empirical work. However, problems arise when there are many potential predictors and moderators of treatment response and the researcher does not have sufficient knowledge about the phenomenon under study to propose definitive hypotheses regarding which predictors to test (Hollon et al., 2019). This issue is especially relevant to cases with complex, higher-order interactions and is particularly relevant to clinical psychology research where theories of psychopathology posit the presence of complex interactions among biological, psychological, and social variables (Engel, 1977, 1981).

**Limitations of Model-Based Approaches: Characterizing Complexity and Underlying Interactions**

To account for the potential complexity among predictors, researchers may test models with many potential predictors of treatment outcome in a multiple regression framework. However, the standard implementation of multiple regression can quickly become confusing, inefficient, and uninterpretable when there are many predictors and the researcher is interested in assessing not only main effects, but also additional interactions among predictors (e.g., with 10 predictors, there would be 45 possible 2-way interactions and 120 possible 3-way interactions). This hinders the detection of potential higher-order interactions because they are never tested by the researcher. In the case that they *are* specified, studies are often underpowered for detecting these interaction effects (Marshall, 2001). Multiple regression also encounters problems with multicollinearity when there are dependencies and linear relationships among two or more independent variables (e.g., correlation between two predictors reflects a linear relationship between two variables) (Farrar & Glauber, 1967; Mansfield & Helms, 1982). The presence of multicollinearity results in estimation of regression coefficients with large standard errors, which hinders hypothesis testing, estimation, and predictive accuracy (Alin, 2010). Model-based methods also impose many assumptions on the relationship between predictors and outcome variables, including a pre-specified functional form and distributional assumptions that can be too restrictive or violated by psychological variables (Erceg-Hurn & Mirosevich, 2008).

**Identifying Patient Predictors Using Statistical Learning**

When researchers are not privy to patient characteristics and interaction effects that predict differential treatment outcomes but have access to many potential candidate predictors, one solution is to use statistical learning methods that embed learning algorithms into traditional statistical models to perform variable selection. A common approach used in precision psychiatry is regularized or penalized regression, which can handle the inclusion of a large number of predictors by imposing a penalty on the regression model and reducing coefficient values with minor contribution to the dependent variable to near or close to zero (Eilers & Marx, 1996; Ruppert & Carroll, 1999; Tibshirani, 1996). As a result, this “penalization” process reduces the model complexity, mitigates issues with multicollinearity, minimizes prediction error, and facilitates model interpretation. Importantly, despite the advantages of penalized regression, the focus is to fit the data *to a model* with *pre-specified assumptions* to formally test hypotheses at the *group level*. Recent work has cautioned that these group-level summaries of effects may not provide information at the level of the individual (Fisher et al., 2018; Molenaar & Campbell, 2009), which is a primary goal of precision medicine efforts.

**Shift Toward Machine Learning**

Acknowledging the limitations of model-based approaches, researchers have turned to machine learning or data mining approaches (e.g., recursive partitioning approaches, random forests, bagging, boosting, support vector machines, neural networks, and deep learning) to flexibly identify patterns of information in clinical datasets that can be used to predict *individualized* treatment outcomes and guide treatment decisions (Chekroud et al., 2016; Dwyer et al., 2018; Passos & Mwangi, 2018). Machine learning is a data-driven approach that does not make assumptions about the relationships between predictors and outcomes, thereby shifting the focus from testing one (or few) predictors of interest at a time to identifying patterns of information in the data that may be useful for predicting outcomes based on individual patient characteristics. Machine learning algorithms do not use traditionally identified statistical parameters and do not impose a pre-specified structure on the relationships between predictors and outcomes (Bzdok & Meyer-Lindenberg, 2018). In this way, machine learning methods offer a “data-driven” approach that is well-suited for detecting complex and high-dimensional interactions without having pre-specified hypotheses about the relations among variables. These higher order interactions may also more adequately reflect the multivariate and interactive nature of predictors of psychopathology (Dwyer et al., 2018).

**A Tale of Two Data Analytic Cultures**

The differences between model-based and machine learning approaches reflect two “cultures” of data analytics. Whereas model-based approaches may offer interpretability, they may be overly simplistic and inadequately capture complex phenomena related to psychopathology and treatment. Though data-driven and flexible machine learning alternatives are attractive for addressing this shortcoming, too much flexibility hinders both interpretability and the advancement of theoretical frameworks representing the phenomenon under study. A summary of the strengths, limitations, and distinctions between model-based and machine learning approaches is organized in Table 1. To advance precision psychiatry, a dialectical stance that acknowledges and balances the strengths and limitations of both approaches is warranted. In fact, the brief mention of statistical learning methods that incorporate learning algorithms from the machine learning lineage, to model-based regression approaches originating from the field of inferential statistics, hints that the two frameworks are not strictly distinct. In efforts to bridge the illusory gap between analytic cultures, two dialectics are considered: 1) The *bias-variance tradeoff* and 2) Emphasis on *causal explanation versus prediction*.

*The Bias-Variance Tradeoff: Balancing interpretability* versus *flexibility*

Ultimately, models are useful when they are generalizable and lead to reproducible findings in future datasets and clinical populations. Model generalizability is hindered by either underfitting or overfitting the model to the data. A model that underfits the data may not include the adequate explanatory variables or insufficiently reflect important relations among variables, thus yielding biased and inaccurate predictions in future datasets. Conversely, an overfitted model may include an extraneous number of predictors and model complexities that are idiosyncratic to the data, resulting in the modeling of noise and random variation instead of meaningful and systematic relationships between predictors and outcomes. Modeling such noise improves model performance in the dataset used for modeling but yields unstable, inaccurate, and highly variable predictions in future datasets. The prediction error associated with underfitting and overfitting are referred to as “bias” and “variance,” respectively. Whereas traditional inferential statistics may increase bias due to the specification of overly-simplistic and underfitted models, complex machine learning algorithms risk introducing undesirable variance and prediction error due to model overfitting. Thus, model optimization involves balancing increased interpretability but potentially greater bias, with enhanced flexibility at the cost of greater variance and prediction error in future datasets.

Described in more detail in Chapter 3, there are well-documented strategies for managing the bias-variance tradeoff (e.g., regularization procedures that help to reduce the dimensionality of the dataset, cross-validation procedures, improving data quality and sample size). Beyond computational strategies, optimizing bias and variance is particularly challenging in clinical psychology and psychiatry. Datasets are susceptible to overfitting due to limited sample sizes and measurement error that is common in behavioral and psychological variables, and underfitting due to limited access to patient predictors across levels of explanation (e.g., biological, genetic, behavioral, and social) (Bzdok & Meyer-Lindenberg, 2018; Yarkoni & Westfall, 2017). As sample sizes in psychosocial intervention studies also tend to be smaller than clinical trials in other fields (e.g., psychotherapy trial vs. drug trial) (Chekroud et al., 2016; Koutsouleris et al., 2016), future work is warranted to understand how machine learning algorithms perform under sample size and measurement constraints.

*Causal Explanation* versus *Prediction*

Another fundamental conceptual difference between model-based and machine learning approaches is the distinction between statistical inference and predictive modeling. Inferential statistics operates under a hypothesis testing framework with the goal of fitting data to a specified model to *infer* properties of a population and test *causal explanations*. In this context, researchers start with hypotheses about how the variables affect the outcome(s) and draw conclusions about these hypotheses based on the results of the statistical model. Machine learning methods seek to identifying patterns in the data to output a model that optimally *predicts* new or future observations by capturing the *associations* between input and output variables (Shmueli, 2010). Whereas the goal of model-based approaches is to characterize and estimate the relationships between predictors and outcomes at a group level, machine learning focuses on generating accurate predictions of treatment outcomes for the individual. The emphasis on individualized prediction is appropriately aligned with precision medicine efforts aimed at tailoring and selecting interventions for individual patients instead of relying on group-based estimates. However, enhanced individualized prediction does not necessarily prioritize the advancement of a theoretical framework or causal explanation to characterize the clinical phenomenon or population under study. Thus, improved prediction of individualized treatment outcome may overlook the identification “active ingredients” or mechanisms of change for the intervention under study.

The two cultures of statistical modeling (Breiman, 2001) align closely with explanation-focused or prediction-focused approaches to scientific inquiry (Douglas, 2009). Traditionally, psychology researchers have been trained in explanation-focused (i.e., model-based) approaches to identify causal mechanisms and build theories of psychopathology. Though the flexibility of machine learning is attractive and promising, it is unclear whether the flexibility and reliance on fitting a model *to the data* can allow for the identification of causal mechanisms that inform theories of psychological and behavioral processes. Careful implementation and comparison of these approaches is warranted to assess their utility for advancing the field of precision psychiatry, especially in the context of psychosocial intervention studies.

**Current Study**

The present study aims to compare the performance of model-based regression techniques and machine learning methods in the context of a randomized clinical trial (RCT) assessing the effectiveness of internet-delivered cognitive behavior therapy for youth with chronic abdominal pain. The machine learning algorithms used in this study reflect extensions of the classification and regression tree (CART; Breiman, Friedman, Olshen, & Stone, 1984) algorithm, a commonly-used machine learning algorithm intended to identify distinct population subgroups by recursively partitioning the dataset into homogeneous groups based on a designated outcome variable. The primary output of the CART algorithm is a decision tree. Tree-based diagrams are a commonly used schematic in medical practice that translate clinical knowledge to inform practitioner decision-making. For example, tree-based diagrams may be used to formulate medical diagnoses, allocate treatments, and designate clinical practice guidelines (Podgorelec et al., 2002; Shiffman, 1997).

The motivating dataset reflects a substantial sample (*n* = 300) of patients and is an ideal setting for this model comparison study. Prior to randomization into the two treatment arms, participants in this study were first stratified into evidence-based subgroups reflecting patient characteristics that were hypothesized to predict differential response to treatment. Because evidence of treatment x subgroup moderation was found using traditional model-based regression methods, this dataset reflects an optimal setting to examine whether machine learning approaches offer convergent validity and/or incremental utility compared to model-based approaches.

Emphasis is given to comparison of model performance and identifying the unique strengths and limitations of the approaches in accomplishing two tasks: 1) Identifying patient predictors of differential treatment outcome, and 2) Detecting and modeling underlying interaction effects and complex relations among variables as they relate to predicting differential treatment outcomes. In addition to comparing various modeling approaches, the study aims to provide an example for howclinical prediction tools may be implemented in clinical research and practice. Results serve as a springboard for empirical studies investigating the efficacy of psychosocial interventions for youth with chronic pain, and methodological studies assessing the performance of different analytic methods for optimizing personalized prediction of treatment outcomes in psychosocial intervention contexts more broadly.

Table 1. Summary of potential strengths and limitations of model-based and machine learning approaches.

|  |  |
| --- | --- |
| **Model-Based Approaches** | **Machine Learning Approaches** |
| Strengths   * **Theory-informed**, requires user to have *a priori* hypotheses to prevent data dredging * Results are **easily interpretable**/can be more readily framed under existing theories * **Recent developments in regression methods**, such as regularization procedures, do allow for the inclusion of many predictor variables | Strengths   * **Data-driven**, allows for flexible modeling of **complex** relations among variables * **Minimal assumptions or constraints** on data structure * Can be used to understand relations among variables in **observational data** (does not require data collection under controlled conditions) * Can accommodate **many different variable types** * Readily allows for **non-linear functional forms** between predictors and outcomes |
| Potential Limitations   * Requires the user to specify the model, including potential interactions; current **theories may not inform which interactions to test** * Inability to acknowledge the hypothesized **complexity** of underlying causes of psychopathology deems significant findings to have attenuated effects sizes and may hinder clinical significance * **Reliance on statistical significance** based on calculation of *p*-values may be limited based on recent developments in data science | Potential Limitations   * **Overfitting**, especially in high-dimensional datasets or limited sample sizes * Lack of **reproducibility and generalizability** * **Opacity/Interpretability**: Prediction tools and the algorithms themselves may be difficult to understand/interpret; does not inform mechanisms underlying differential treatment response * **Data Quality**: Behavioral and self-report data often collected in treatment studies for psychiatric conditions may not have enough phenotypic variability or granularity |

CHAPTER 2

MOTIVATING DATASET: AN RCT FOR YOUTH WITH CHRONIC PAIN

This section describes the motivating dataset used in the current methods comparison study. The selected dataset if part of a longitudinal RCT assessing the efficacy of internet-delivered CBT for children and adolescents with chronic abdominal pain. A comprehensive report of the full study is currently in preparation (Walker, Stone, Han, et al., in preparation). The goal of this chapter is to provide adequate background regarding the substantive motivations of the RCT and establish its relevance for precision medicine initiatives. First, study procedures and characterization of the sample are presented. Next, a portion of the findings from the main RCT study is outlined to justify the use of the dataset for a comparison of model-based and machine learning approaches. As described below, inclusion of an empirically-based subgrouping variable allowed for the explicit testing of treatment x subgroup interaction using a model-based approach. Evidence of statistical moderation motivates the examination of potentially convergent findings using a series of model-based and non-parametric machine learning approaches outlined in Chapter 3.

**Study Background**

Consistent with the goals of precision medicine, the Institute of Medicine blueprint for the transformation of pain research and care (2011) and the Federal Pain Research Strategy (2017) proposed that treatment outcomes may be improved by subgrouping pain patients and tailoring care to subgroup characteristics rather than treating pain patients as a homogeneous group. Specifically, functional abdominal pain disorders (FAPDs) comprise a common pediatric pain condition characterized by long-lasting intermittent or constant abdominal pain in the absence of an underlying organic cause (Turk & Okifuji, 2002). FAPDs are among the most prevalent chronic pediatric pain problems with estimated worldwide prevalence rates of 13.5% (Korterink et al., 2015). Investigating and optimizing treatment of FAPDs is important because FAPDs are often a precursor to other forms of chronic pain, psychopathologies, and high healthcare utilization in young adulthood (Walker et al., 2010).

**Chronic Pain as a Biopsychosocial Phenomenon**

Conceptualized as a biopsychosocial phenomenon (Engel, 1977, 1981), chronic pain is associated with aberrant sensory processes, cognitive processes (e.g., cognitive appraisal of pain), behavioral responses (e.g., coping styles), and social learning mechanisms (e.g., parent-child communication about pain behaviors) (Howard, 2003; Levy et al., 2010; Stone & Walker, 2017; Walker et al., 2005; Wiech, 2016). Gatchel and colleagues (2007) provide a comprehensive overview of basic neuroscience processes and psychosocial aspects underlying the etiology and treatment of chronic pain and emphasize that “psychological and social factors interact with brain processes to affect health and illness” status. Thus, psychosocial interventions, specifically cognitive behavior therapy, are promising for improving health outcomes in youth with chronic pain with evidence from multiple RCTs indicating improvements in pain intensity, functional disability, and emotional functioning (Eccleston et al., 2002; Levy et al., 2010; Palermo et al., 2009, 2010, 2016). However, only a minority of FAP patients are referred for psychological services and, of these, only a subset has access to these interventions. To address these barriers, internet-delivered CBT may be an alternative, efficient, and more easily delivered approach for treating youth with CAP (Palermo et al., 2009).

**Study Design and Overview of Methods**

The primary aim of the RCT study is to assess the effectiveness of internet-delivered cognitive behavioral therapy for youth with chronic abdominal pain. The RCT includes two treatment arms: 1) internet-delivered cognitive behavioral therapy (CBT) or 2) online education (EDU), which serves as an attention control condition. Data on hypothesized predictors and outcomes of interest were collected at five timepoints: pre-treatment (baseline), mid-treatment (about 6 weeks after baseline), post-treatment (about 12 weeks after baseline), 6-month follow-up (6 months after baseline), and 12-month follow-up (12 months after baseline). Detailed study procedures and results of this RCT (ClinicalTrials.gov: NCT02327377) are presented elsewhere (Walker, Stone, Han et al., in preparation). In this section, a sufficient level of background is provided to justify the use of the RCT dataset for this methods comparison study.

*Participants and Procedures*

Patients (*n* = 300) between 11-17 years old and their parent/caregivers were recruited from new patients referred to Vanderbilt’s Pediatric Gastroenterology Service for evaluation of abdominal pain of at least two months and no prior diagnosis of explanatory organic disease. Eligibility criteria included (1) episodic or chronic abdominal pain ≥ 2 months duration consistent with a FAPD, (2) absence of a chronic disease (e.g., inflammatory bowel disease, diabetes), (3) no hospitalizations within the month prior to enrollment, and (4) both patient and parent able to access the Internet and to read English at the sixth-grade level. The final sample comprised 278 parent-adolescent dyads. Youth were predominately female (66.2%, n = 184) and Caucasian (86.0%, n = 239) with a mean age of 14.62 years (SD = 1.88). Parent participants were primarily mothers or maternal guardians (95.3%, n = 265).

Study procedures were approved by the Vanderbilt University Medical Center Institutional Review Board. Recruitment was from November 2014, through February 2018. Parents of consecutive new patients scheduled for evaluation of abdominal pain at the pediatric gastroenterology clinic of Monroe Carell Jr. Children’s Hospital of Vanderbilt University in Nashville, Tennessee met with a study recruiter when they arrived at the clinic. See Figure 1 for the Consolidated Standards of Reporting Trials (CONSORT) chart.

*Treatment Conditions: Cognitive Behavior Therapy (CBT) and Online Education (EDU)*

Patients randomized to the cognitive behavior therapy (CBT) treatment condition participated in an adapted version of the Web-based Management of Adolescent Pain (Web-MAP) program (Palermo et al., 2009, 2016). Web-MAP is a protocol designed for adolescents with mixed chronic pain conditions. Thus, the eight Web-MAP modules were adapted to be relevant to FAPDs. The online protocol consists of two separate websites, one for the child and one for the parent. Children and parents are each asked to log on once per week to the website to complete a module and assignment, designed to take about 30 minutes to complete. The child modules contain skills-based content in identifying stress, applying deep breathing and progressive muscle relaxation, and modifying cognitions about pain and functional disability. Parent modules provide skills in adaptive communication and interaction patterns. See Figure 2 for an overview of treatment content.

Figure 1. Consolidated Standards of Reporting Trials (CONSORT) chart.

A screenshot of text

Description automatically generated

Figure 2. Overview of Web-MAP treatment content for patient and parent/caregiver.

Children’s Modules

1: Education about chronic pain

2: Recognizing stress and negative emotions

3: Deep breathing and relaxation

4: Implementing coping skills at school

5: Cognitive skills (e.g., reducing negative thoughts)

6: Sleep hygiene and lifestyle

7: Staying active (e.g., activity pacing, pleasant activity scheduling)

8: Relapse prevention

4: Operant strategies II (using reward to increase positive coping; strategies to support school goals)

Parents’ Modules

1: Education about chronic pain

2: Recognizing stress and negative emotions

3: Operant strategies I (using attention and praise to increase positive coping)

5: Modeling of coping behaviors

6: Sleep hygiene and lifestyle

7: Parent-child communication

8: Relapse prevention

Table 2. Demographic factors and clinical characteristics by pain phenotype

|  |  |  |  |
| --- | --- | --- | --- |
| Demographic Factor | High Pain Dysfunctional  (n = 109) | High Pain Adaptive  (n = 114) | Low Pain Adaptive  (n = 55) |
| Adolescent Age, *M ± SD* | 15.08 ± 1.89a | 14.48 ± 1.82b | 14.01 ± 1.79b |
| Adolescent Sex, *% (n)* |  |  |  |
| Female | 79.8% (87)a | 66.7% (76)a | 38.2% (21)b |
| Male | 20.2% (22)a | 33.3% (38)a | 61.8% (34)b |
| Adolescent Race, *% (n)* |  |  |  |
| Caucasian | 80.7% (89) | 88.6% (101) | 92.6% (50) |
| Minority Group\* | 19.3% (21) | 11.4% (13) | 7.4% (4) |
| Child Rome III classification, *% (n)*  Irritable bowel syndrome (IBS)  Functional dyspepsia (FD)  Either IBS or FD | 67.0% (73)a  15.6% (17)a  77.1% (84)a | 62.3% (71)a  19.3% (22)a  78.1% (89)a | 32.7% (18)b  9.1% (5)b  40.0% (22)b |
| Parent who participated in study, *% (n)* |  |  |  |
| Mother or grandmother | 93.6% (102) | 95.6% (109) | 96.4% (53) |
| Father | 6.4% (7) | 4.4% (5) | 3.6% (2) |
| Parent employment, *% (n)* |  |  |  |
| Employed | 66.7% (72) | 67.5% (77) | 61.8% (34) |
| Unemployed | 33.3% (36) | 32.5% (37) | 38.2% (21) |
| Parent education, *% (n)* |  |  |  |
| High school or less | 26.6% (29) | 16.7% (19) | 21.8% (12) |
| Vocational school or some college | 32.1% (35) | 36.8% (42) | 40.0% (22) |
| Four-year college | 31.2% (34) | 32.5% (37) | 25.5% (14) |
| Graduate or professional school | 10.1% (11) | 14.0% (16) | 12.7% (7) |
| Parent marital status, *% (n)* |  |  |  |
| Married or partnered | 69.7% (76) | 71.9% (82) | 78.2% (43) |
| Single, Divorced, or Separated | 30.3% (33) | 21.1% (32) | 21.8% (12) |
| Baseline GI symptom severity (CSSI),  *M ± SD* | 2.05 ± .67a | 1.50 ± .62b | .92 ± .56c |
| Baseline abdominal pain severity (API),  *M ± SD* | 2.75 ± .75a | 2.24 ± .75b | 1.33 ± .48c |
| Baseline pain interference (PROMIS),  *M ± SD* | 57.16 ± 7.24a | 50.92 ± 6.79b | 43.21 ± 5.82c |

Note. Adolescent age and sex significantly differed at p < .05 level. Within rows, differing superscripts indicate significant differences between subgroups at p < .05 level. \*Due to the low frequency of some racial groups, races typically identified by the National Institutes of Health as minority groups (Black or African American, American Indian or Alaska Native, Asian, and Mixed Race) were collapsed into a single category. Percentages do not always add up to 100%

Table 3. Summary of measures, collection timepoints, informant, and variable types

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Construct and Measure** |  | **Assessment (Months)** | | | | | **Informant**  **C = Child**  **P = Parent** | **Variable Type** |
|  | **0** | **1** | **2** | **6** | **12** |  |  |
| **Primary Child Moderator** | | | | | | | | |
| Pain Phenotype |  | x |  |  |  | x | C | Categorical/Nominal |
| **Child Measures (potential predictors and/or moderators of treatment response)** | | | | | | | | |
| Pain Catastrophizing |  | x | x | x | x | x | C | Continuous |
| Perceived pain coping efficacy |  | x | x | x | x | x | C | Continuous |
| Anxiety |  | x | x | x | x | x | C | Continuous |
| Depression |  | x | x | x | x | x | C | Continuous |
| **Parent Measures (potential predictors and/or moderators of treatment response)** | | | | | | | | |
| Protectiveness |  | x | x | x | x | x | C, P | Continuous |
| Modeling pain behavior |  | x | x | x | x | x | C, P | Continuous |
| Catastrophizing about child pain |  | x | x | x | x | x | P | Continuous |
| **Child outcomes** | | | | | | | | |
| Abdominal Pain |  | x | x | x | x | x | C | Continuous |
| GI symptoms |  | x | x | x | x | x | C | Discrete/Count |
| Non-GI symptoms |  | x | x | x | x | x | C | Discrete/Count |
| **Other** | | | | | | | | |
| Child trait anxiety |  | x |  |  |  | x | C | Continuous |
| Parent chronic pain |  | x |  |  |  | x | P | Continuous |
| Functional GI disorder symptom criteria |  | x |  |  |  | x | C, P | Categorical/Nominal |
| Family demographics |  | x |  |  |  |  | P | Categorical/Nominal |
| Health service utilization |  |  |  |  |  |  |  |  |
| Outpatient Visits |  | x |  |  |  |  |  | Discrete/Count |
| Emergency Room Visits |  | x |  |  |  |  |  | Discrete/Count |
| Treatment expectancies |  | x |  |  |  |  | C, P | Continuous |
| Treatment engagement |  | x | x | x | x | x | C, P | Continuous/Ordinal |
| Treatment satisfaction |  |  |  |  |  | x | C, P | Continuous |

*Outcome Measures*

The primary outcome measure was youth report of GI symptoms (e.g., pain, upset stomach, constipation, diarrhea, bloating) on the GI Symptom Subscale of the Children’s Somatic Symptoms Inventory (CSSI-24; Walker et al., 2009). Secondary outcome measures were: (1) youth report of abdominal pain severity on the Abdominal Pain Index (API; Laird et al., 2015) a multidimensional measure of pain with items assessing the frequency, duration, and intensity of abdominal pain, and (2) youth report of activity impairment on the PROMIS Pain Interference scale (Varni et al., 2010), a measure assessing the extent to which pain hinders engagement with social, physical, and emotional domains.

**Model-Based Analysis and Evidence of Treatment x Subgroup Interaction**

To test the hypothesis that pain phenotype is a moderator of treatment outcome, a primary focus of the analysis used in the original RCT was a model-based approach to explicitly test the Treatment x FAP Phenotype interaction effect. A portion of the findings are presented here to provide support for the use of this dataset in the current methods comparison study.

Data analysis based on intention-to-treat[[1]](#footnote-1) was conducted using R Version 3.5.3 and R Studio Version 1.0.143. Outcome variables included youth report of GI symptoms (CSSI-24), abdominal pain (API), and pain interference (PROMIS Pain Interference). Longitudinal data were analyzed using linear mixed effects models (Raudenbush & Bryk, 2002), with 5 time points nested within each individual (Pre-treatment, Mid-treatment, Post-treatment, 6-month follow-up, 12-month follow-up). To capture changes in outcomes during the treatment and follow-up periods, models were based on a piecewise analysis of time where Piece 1 examined changes during the treatment period (Pre-, Mid-, to Post-treatment), and Piece 2 modeled changes during the follow-up period (Post-, 6-month follow-up, to 12-month follow-up). A piecewise approach also models typical trends in treatment studies where the greatest effects occur by the end of the treatment period and change levels off after follow-up (Palermo et al., 2016).

Separate piecewise models were specified for each outcome variable. To test whether pain subgroups moderated the effect of treatment condition on outcomes, models included the three-way Treatment x Subgroup x Time interaction during the treatment (Piece 1) and follow-up (Piece 2) periods. Age and sex were entered as covariates.

*Evidence of Treatment Effect Moderation by FAP Phenotype*

Pain subgroup significantly moderated the effect of treatment on GI symptoms (*t*(853) = -2.94, *p* = 0.003) and abdominal pain (*t*(844) = -2.13, *p* = 0.03), but not pain interference, during the treatment period. Figure 3 provides a visual representation of the Treatment x FAPD pain subgroup x Time interactions for GI Symptoms (Panel A) and abdominal pain (Panel B). As shown in Figure 3 and presented in more detail in Table 4, the significant three-way interaction indicates that change in GI symptoms and abdominal pain during the treatment period depended on both treatment condition and pain subgroup. Specifically, within the HPD subgroup, youth assigned to CBT demonstrated significantly greater reductions in GI symptoms compared to youth in EDU (t(853) = -3.24, p=0.02, d = -0.31). Within the CBT treatment condition, HPD youth demonstrated significantly greater reduction in GI symptoms (t(853) = -3.89, p = 0.002, d = -0.41) and abdominal pain (t(844) = -3.63, p = 0.004, d = -0.45) compared to LPA youth.

Regarding covariates, higher age was associated with significantly higher GI symptoms (t(270) = 3.30, p = 0.001, d = 40) and female sex was associated with significantly higher abdominal pain (t(270) = 1.97, p=0.05, d = 0.24) and pain interference (t(270) = 2.90, p = 0.004, d = 0.35). Interaction effects of treatment condition and pain subgroup with age and gender were not significant (all p’s > .05).

Table 4. Estimated slopes from pre- to post-treatment and effect sizes (Cohen’s d) for pairwise slope comparisons between Pain Subgroup and within Treatment Condition

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Measure | Treatment Condition | FAP Subgroup | Slope from Pre- to Post-Treatment (SE)  [95% CI] | Pairwise Contrast (SE)  Effect Size (Cohen’s d) | | |
| HPD vs.  LPA | HPD vs.  HPA | HPA vs.  LPA |
| GI Symptoms | CBT | HPD | -0.06 (0.01)  [-0.07, -0.04] | -0.04 (0.01)  ***d* = -0.41\*\*\*** | -0.02 (0.01)  ***d* = -0.24\*** | -0.02 (0.01)  *d* = -0.20 |
| HPA | -0.03 (0.01)  [-0.05, -0.02] |
| LPA | -0.01 (0.01)  [-0.03, 0.01] |
| EDU | HPD | -0.02 (0.01)  [-0.04, -0.01] | 0.003 (0.01)  *d* = 0.03 | 0.003 (0.01)  *d* = 0.03 | 0.0002 (0.01)  *d* = 0.001 |
| HPA | -0.03 (0.01)  [-0.04, -0.02] |
| LPA | -0.03 (0.01)  [-0.05, -0.01] |
| Abdominal Pain | CBT | HPD | -0.07 (0.01)  [-0.09, -0.06] | -0.06 (0.02)  ***d* = -0.45\*\*\*** | -0.01 (0.01)  *d* = -0.11 | -0.04 (0.02)  ***d* = -0.35\*\*** |
| HPA | -0.06 (0.01)  [-0.08, -0.04] |
| LPA | -0.02 (0.01)  [-0.04, 0.01] |
| EDU | HPD | -0.05 (0.01)  [-0.07, -0.03] | -0.009 (0.02)  *d* = -0.07 | -0.002 (0.01)  *d* = -0.02 | -0.007  (0.02)  *d* = -0.06 |
| HPA | -0.05 (0.01)  [-0.06, -0.03] |
| LPA | -0.04 (0.01)  [-0.07, -0.02] |

\*p<0.05; \*\*p<0.01; \*\*\*p<0.001

Effect sizes were calculated using the difference of pairwise estimates, divided by the pooled standard deviation of the population providing the context for the effects.

Figure 3. Effect of treatment on GI symptoms and abdominal pain by pain phenotype

**A close up of a map

Description automatically generated**

Note. A) Three-way interaction for GI Symptoms; B) Three-way interaction for Abdominal Pain.

**Rationale for Using the Dataset for a Comparison Study**

The RCT described above represents an ideal exemplar dataset for several reasons. First, the biopsychosocial conceptualization of pain suggests the presence of complex interactions, and the multitude of predictors collected in the study allows for the examination of these effects. Second, the dataset is from a well-designed RCT study that meets “gold standard” recommendations. Specifically, it reflects a substantial sample (*n* = 300) in the context of psychosocial intervention studies. Third, the study includes the empirically-supported construct of FAP phenotype (Walker et al., 2010, 2012) as a potential moderator. As described, FAP phenotype is yielded by a previously-validated clustering algorithm and reflects the patient’s pain severity and pain-related psychological processes (e.g., pain threat, catastrophizing, pain coping efficacy beliefs). In this way, pain phenotype can be considered a “composite moderator” (Kraemer, 2013) that can be easily specified as a moderator in a traditional “top-down” model-based framework (i.e., product term specifying the FAP Phenotype x Treatment interaction) while also acknowledging the complex biopsychosocial aspects of pain. This will facilitate comparison of the *substantive* conclusions that can be drawn from “top-down” model-based approaches and “bottom-up” machine learning methods. For example, identification of patient characteristics used for phenotyping as the most important predictors of treatment outcome in a machine learning framework contributes additional validation of FAP phenotype as a clinically useful construct for predicting individualized response to treatment.

*Using a Pre-Post Framework for Model Comparison*

Results of the piecewise linear mixed effects model analysis indicated the presence of a treatment x subgroup interaction from pre- to post-treatment. In longitudinal RCTs, this is expected given that any evidence of a treatment effect will be most prominent at post-treatment due to temporal proximity to the active treatment period, compared to 6-month follow-up and 12-month follow-up time points, when general regression to the mean is inevitable. Thus, the current model-comparison study focuses on using baseline predictors to predict differential treatment outcome at *post-treatment*. Narrowing the scope of the current study to a pre-post framework additionally facilitates a detailed and in-depth comparison of modeling approaches.

CHAPTER 3

OVERVIEW OF SELECTED MODELING APPROACHES

The conceptual goals of precision medicine correspond to two data analytic tasks: 1) Variable selection to identify patient predictors associated with differential treatment outcome, and 2) Modeling complex and meaningful interactions and relationships between variables, including interactions between treatment condition and patient characteristics. To address the first task, all candidate modeling approaches used in this methods comparison study were chosen based on their ability to select, from a set of many patient variables, the most important subset of baseline characteristics for predicting treatment outcomes. The model-based approaches differed from machine learning approaches in their ability to detect underlying interactions and complex relations among variables without pre-specification by the researcher.

The selected models can be organized on a continuum from traditional, parametric, model-based approaches to data-driven, nonparametric, machine learning approaches. Specifically, they include multiple linear regression (Aiken et al., 2012), linear regression with feature selection (Hocking, 1976), linear regression with elastic net regularization (Zou & Hastie, 2005), and random forest analysis (Ishwaran, 2007; Strobl et al., 2007).

Figure 4. Conceptual overlap between inferential statistics, statistical learning, and machine learning

**A close up of text on a black background

Description automatically generated**

Referencing Figure 4, traditional least squares regression is situated in the yellow part of the diagram, as predictors of interest are explicitly included to test dependency of the outcome variable. As a common modeling approach used in *inferential statistics*, linear regression is used to *confirm* a plausible hypothesis regarding the relationship between predictor(s) and outcome (e.g., “Is there a relationship between pain catastrophizing and chronic pain?”). Linear regression with feature selection and regularization represent extensions of the traditional linear regression approach and are useful when the researcher intends to *explore* which predictors are most relevant in predicting the outcome of interest (e.g., What patient characteristics are most associated with chronic pain?”). By extending upon the classical linear regression model, the additional feature selection and regularization techniques shift the focus from *statistical inference* to *statistical learning*, represented by the green portion of the diagram in Figure 4. This is achieved by embedding learning algorithms that serve to iteratively identify the optimal set of predictors or features for optimizing model performance. Finally, random forest analysis represents a nonparametric machine learning algorithm that identifies important variables while allowing for nonlinearities and interactions to be learned from the data without user pre-specification. By learning from the data, machine learning algorithms (represented in the blue portion of the diagram in Figure 4) are focused on generating a *predictive* model that can reliably predict outcomes for hypothetical, unseen datasets, or to predict future outcomes for an individual (e.g., “How much pain will this patient have after completing the 8-week intervention?”).

**Model-Based Approaches: Linear regression and its extensions**

In linear regression, a dependent variable *Y* is modeled as a linear function of an independent variable *X*. When there is more than one independent variable, the natural extension is multiple regression, where we include multiple independent variables into the same linear model. This approach is reasonably robust, especially if the number of observations *n* is distinctly larger than the number of variables *p* (n >> p). As mentioned previously, linear regression faces challenges handling multicollinearity, insufficient power to detect effects of interest, and potential violations of distributional assumptions as the number of predictors increases. The multiple regression framework also deteriorates when there are many predictors and the researcher is interested in assessing not only main effects, but also additional interactions among predictors. In the context of precision medicine, where the primary goal is to understand how many individual patient characteristics differentially predict (i.e., moderate) response to treatment, linear regression may be limited. This is especially true in cases where the researcher has access to “high-dimensional” datasets with many candidate predictors but does not have *a priori* knowledge of the most relevant variables associated with treatment outcomes. Insufficient knowledge of the relevant predictors subsequently hinders the identification of complex, higher-order interaction effects.

To provide clarity on the relevant predictors, many automated variable selection (also termed “feature selection”) methods have been developed. Consistent with the goals of precision medicine, feature selection helps to prevent overfitting and improve model performance by identifying relevant patient characteristics for making accurate predictions of treatment outcome. In inferential statistics, substantive knowledge informs viable hypotheses regarding the most important predictors of treatment outcome. Combining substantive knowledge with a less subjective algorithmic approach can facilitate novel and additional insight into the underlying processes related to individual treatment response. A comprehensive review of feature selection methods is beyond the scope of this text and has been documented elsewhere (e.g., Saeys et al., 2007).

Briefly, automated feature selection methods include filter methods, wrapper methods, and embedded methods. Filter methods assess the “relevance” of features by using univariate statistics to identify predictors that have a strong statistical relationship with the outcome variable. For example, correlations may be used to identify the top 10 features that are most correlated with the outcome. Wrapper methods assess the “usefulness” of the features by algorithmically selecting a subset of the predictors across many iterations, training the model with the selected predictors, and using cross-validation to identify the subset of predictors that yields the best model performance. Finally, embedded methods integrate a feature selection algorithm into the learning algorithm and perform feature selection during the model training process. In the current study, a sequential search algorithm and penalized regression algorithm were used to perform automated feature selection.

*A Wrapper Method for Variable Selection: Sequential search*

One wrapper method for performing variable selection is to specify a search procedure that iteratively considers different combinations of predictors for predicting an outcome, and then finds the set of predictors that optimizes the performance of the predictive model by minimizing the prediction error in the test data. Various search methods can be specified, such as an exhaustive search where all possible combinations of predictors in the dataset are assessed, a random search, and a sequential search. Whereas exhaustive searches can be unfavorably computationally intensive, sequential searches offer an ideal compromise between computational time and probability of finding the best-performing feature combination. There are many sequential search strategies, including forward search (i.e., start with an empty model and sequentially add variables that improve the model performance until additional features no longer improve performance), backward search (i.e., start with a full model with all variables and remove one variable at each step until additional removals no longer improve model performance), and a floating search (i.e., start with either an empty or full model and sequentially add or remove one variable at each step until neither an addition nor removal improves model performance). For the current study, sequential floating backward search was used. This method starts with a full model that includes all predictors and removes or adds one variable at each step until neither an addition nor a removal improves model performance. Floating searches have been documented as an optimal sequential search algorithm for variable selection that balances computational efficiency with overall model performance (Pudil et al., 1994).

*An Embedded Method for Variable Selection in Linear Regression: Regularization*

Another statistical learning approach involves embedding shrinkage or regularization procedures to facilitate search and selection of predictors to construct the linear model. Regularization is a technique that introduces an additional constraint into the optimization of a predictive model that biases the model toward lower complexity to prevent overfitting. Three well-known and commonly used regularization techniques for linear models include ridge regression (Hoerl & Kennard, 1970), least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), and elastic net regression (EN; Zou and Hastie, 2005). To understand linear regression with regularization, it is helpful to first revisit the procedure used to identify the line of best fit in ordinary least squares (OLS) regression without regularization. Recall that residuals for a particular combination of intercept and slope are calculated for each case and squared. These squared residuals are then summed to give the *sum of squares* of the residuals*.* The line that minimizes the sum of squares is used to model the relationship between predictor and outcome. The sum of squares is represented in mathematical notation as the following equation:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 1 |

In Equation 1, is the value of the outcome variable for case , and is its value predicted by the model. This equation represents the sum of vertical distances of each case from the line. Mathematical functions that are minimized by machine learning algorithms to select the best combination of parameters are called *loss functions*. Thus, minimizing the sum of squares, or “least squares,” is the loss function of OLS regression.

Ridge regression is a form of penalized regression that modifies the least squares loss function by including a “penalty” called the L2 norm, which sums the squares of all the model parameters , from the first parameter to the last , and adds this summation to the least squares loss function. As a result, the ridge regression algorithm tempers model complexity by balancing the selection of parameters that minimize the sum of squares while also selecting the parameters that minimize this new penalty. The L2 norm is multiplied by λ, a value that is used to control the degree to which we wish to penalize model complexity. *Lambda* (λ) can take on any value from 0 to infinity; larger λ values impose greater penalty for model complexity. As an aside, λ represents a parameter that affects model performance but is not estimated *from* the data. Such parameters are called “hyperparameters” in the machine learning literature. As described in more detail below, hyperparameters can be optimized or tuned as part of the modeling process through a process called hyperparameter tuning. As outlined in Equation 2, minimizing the L2 loss function serves to “shrink” parameters estimates *toward* 0 to prevent overfitting. The L2 norm is displayed below as Equation 2:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 2 |

The LASSO uses a different loss function, called the L1 norm. Instead of adding the squared parameter values in the loss function, the L1 norm sums the absolute value of the parameter values. As a result, it is able to shrink certain parameter estimates *to* 0, thus removing less relevant predictors and identifying a remaining predictor subset. The L1 norm is displayed below as Equation 3:

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 3 |

Finally, EN regression combines the L1 and L2 penalties in its loss function and finds a combination of parameter estimates somewhere between those yielded by ridge and LASSO procedures. As shown in Equation 4, the loss function used in EN regression includes two hyperparameters, λ and α, that can be tuned during the model optimization process to specify the degree of penalization placed on the model using the L2 *versus* the L1 norm.

|  |  |  |
| --- | --- | --- |
|  |  | Eq. 4 |

Examination of Equation 4 shows that when α is 0, the L1 norm becomes 0, and the model uses the same loss function as ridge regression. When α is 1, the L2 norm becomes 0, and the model uses the same loss function as LASSO regression. When α is between 0 and 1, the loss function reflects a mixture of ridge regression and LASSO. EN regression is often recommended when the decision between ridge regression and LASSO is not entirely clear. For the current study, EN regression was chosen.

**Machine Learning Approaches: Recursive Partitioning**

The approaches outlined above offer solutions to the variable selection problem. However, as extensions of linear regression, they still involve distributional and model-based assumptions. Thus, nonparametric machine learning algorithms may more appropriately model complexity and interactions among predictors. The current study focuses on a subset of algorithms called recursive partitioning approaches that have been both implemented and developed to detect treatment heterogeneity (Loh et al., 2015; Su, Tsai, Wang, Nickerson, & Li, 2009). The goal of recursive partitioning is to explore regression functions for the conditional distribution of the outcome variable on the predictor space (across all predictor variables). In other words, the approach is inherently motivated by the identification of heterogeneous treatment effects as they depend on the covariates in the dataset. In contrast to traditional model-based methods (e.g., linear regression), this approach does not require the pre-specification of a model and the results also do not need to be interpreted as a model. Instead, recursive partitioning is a procedure that iteratively and adaptively defines the homogenous effect of different covariates based on a designated fitting criterion.

*Classification and Regression Trees (CART)*

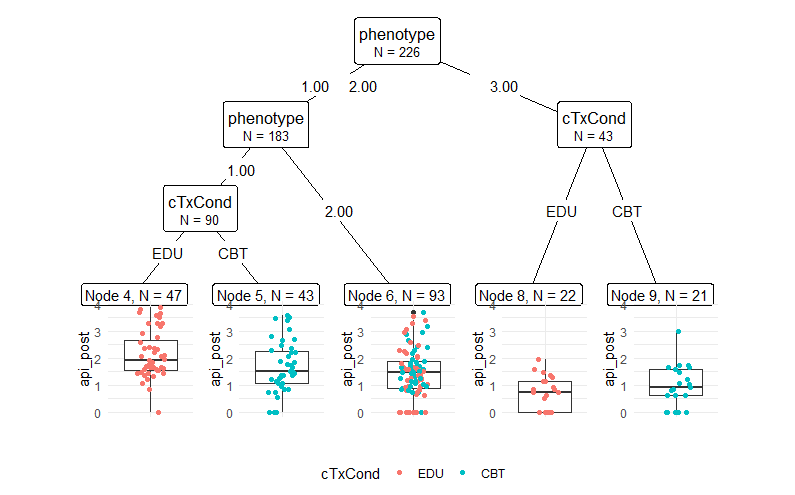
CART is a recursive partitioning method that builds either a classification or regression tree for predicting categorical or continuous outcome variables, respectively. The classic CART algorithm was popularized by Breiman, Friedman, Olshen & Stone (1984). The prediction tree yielded by the CART algorithm defines the predictive model that is used to summarize the relationships among variables and how they predict the outcome of interest. Regression trees are built based on two primary procedures: 1) recursive partitioning of the data space and 2) simple regression modeling in each cell of the data space.

Prediction trees are visually represented by an easily interpretable tree diagram. The node at the top of the tree is the “root node” and contains the complete dataset. The “terminal nodes” or “leaves” of the tree represent the cells of the partition. The classic CART algorithm partitions the data by conducting a “greedy search,” searching over all binary splits of all predictors for the split (i.e., cut point) that will produce the most homogenous subgroups of the data. The data are split from a “parent node” into two “child nodes.” Next, CART continues to partition the data using the best overall split within the child nodes. CART repeats this process for each predictor in the model and iteratively tries out all possible splits, ultimately settling on the split that minimizes the “node impurity,” defined as the total sum of squared deviations from node centers.

Importantly, CART has a tendency to overfit the data because the algorithm may find binary splits that result in completely or nearly completely homogenous subgroups but that have trivial sample sizes (Breiman, Olshen, & Stone, 1984; Hastie, Tibshirani, Friedman, 2001). These nodes likely reflect idiosyncrasies of the dataset that are not generalizable in other datasets. To prevent overfitting, a stopping criterion is specified, usually by setting a minimal sample size for each terminal node or stopping the algorithm if the decrease in node impurity for the iteration would be less than some specified threshold. Another way to prevent overfitting is to grow a very large tree and then “prune” the tree using “cost-complexity pruning” (Breiman et al., 1984). This method limits the number of partitions by adding a parameter that penalizes larger and more unstable trees. This pruning method seeks to identify the nested subtree that minimizes the risk associated with the complexity of the tree.

The splitting approach used in CART is known to be unfair in the presence of predictors of different variable types, categorical variables with different numbers of categories, or differing numbers of missing values (e.g., Breiman, 1984; Shih & Tsai, 2004). To mitigate this variable selection bias, conditional inference trees utilize multiplicity-adjusted conditional tests (Hothorn et al., 2006) instead of maximum impurity reduction as the splitting criterion. For any node to be split, the procedure conducts a global permutation test of the null hypothesis that there is no association between any of the predictors and the outcome within the node. If this null hypothesis is not rejected, the node is not split and becomes a terminal node. Otherwise, individual null hypotheses of no association to the outcome are tested for each predictor, and the variable with the smallest p-value is selected for splitting. To visualize the output of a decision tree, an example of a regression tree is presented in Figure 5.

Figure 5. *Representation of a hypothetical regression tree*



The hypothetical tree presented in Figure 5 depicts an interaction effect between phenotype and treatment condition and shows how response to treatment is moderated by phenotype. As shown, the decision tree algorithm first identified a split at the root node based on pain phenotype. Patients in the Low Pain Adaptive phenotype (phenotype = 3) were separated from the High Pain Adaptive (phenotype = 2) and High Pain Dysfunctional (phenotype = 1) subgroups. For individuals in the Low Pain Adaptive phenotype, individuals in the CBT condition had higher levels of abdominal pain at post-treatment (Node 9) compared to individuals in the EDU condition (Node 8). Individuals in the High Pain Adaptive phenotype had moderate levels of abdominal pain at post treatment that did not differ based treatment condition (Node 6). Finally, for individuals in the High Pain Dysfunctional phenotype, patients assigned to CBT had lower abdominal pain (Node 5) compared to patients assigned to EDU (Node 4).

*Bagging and Boosting*

Although decision trees have notable strengths in their ability to detect complex interactions when many predictors are included in the model, there are several well-documented limitations. First, the predicted trees are rarely stable and using the CART or conditional inference tree algorithm in a new sample often does not yield the same splits and nodes. To improve generalizability of results, bootstrap aggregation, or “bagging” (Breiman, 1996), and boosting (Shapire et al., 1998) are used to fit an ensemble of many trees. Bagging generates the prediction by repeatedly creating new datasets using bootstrap sampling (i.e., random sampling from the data with replacement), fitting the tree algorithm to each bootstrap sample, and aggregating the results to determine the most stable features of the tree. Boosting is another approach that is used to improve the accuracy of the prediction tree. Unlike bagging, which uses a simple averaging of results to obtain an overall prediction tree, boosting uses a weighted average of results obtained from applying the tree algorithm to each successive sample. In boosting, incorrectly predicted cases from the prior step are given increased weight in the next iteration. At the end of this process, a weighted vote is taken to determine the optimal prediction tree.

*Random Forest Analysis*

Random forest analysis is an ensemble machine learning algorithm that fits a forest of decision trees (e.g., *ntree* = 1000). A random forest is random in two ways. First, each tree is based on a random subset of the observations, and secondly, each split within each tree is created based on a random subset of *mtry* candidate variables. This addresses the instability of single trees, as the randomness creates differences in individual tress’ predictions and the overall prediction of the forest is the average of the predictions from the individual teres. As a result, random forests prevent overfitting while addressing potential collinearity issues that may arise if the predictors are highly correlated; the random selection of predictors at each node gives each of the predictors a greater chance of being selected and used for splits in each successive bootstrap tree.

The primary disadvantage of random forest is that it does not produce a single, readily-interpretable tree because it is aggregating over the results of many bootstrap trees. However, it does provide variable importance measures derived from assessing the contribution of each variable for predicting across the aggregated bootstrap trees. Thus, the most important variables can be considered important correlates of the outcome of interest. Given the well-known instability of single trees and the documented limitations of the CART algorithm, this methods comparison study will use conditional random forest (CRF) analysis as the representative machine learning algorithm. CRF analysis fits a forest of conditional inference trees to identify the most important predictors of treatment outcome and can accommodate the detection of underlying and complex interaction effects.

**Assessing Model Performance**

Machine learning algorithms are evaluated based on their generalizability, which is defined as the ability for the model to make accurate predictions in new and unseen datasets. Recall the discussion of bias-variance tradeoff, in which the goal is to minimize inaccurate predictions (i.e., bias) and optimize variability (i.e., variance) when the model is applied to future datasets. To balance these two sources of prediction error, a key component of the model optimization process involves using both a training dataset to train the model, and a test dataset to evaluate the predictive accuracy of the model. Of note, the implementation of and distinction between training and test sets from machine learning differs from a traditional inferential statistics framework, which focuses on drawing causal explanations by minimizing the training error when fitting a model to the study sample (i.e., “training data”) and does not include validation of the model in a separate sample (i.e., “test data”) (Yarkoni & Westfall, 2017). The real-world analogue to a test set is data collected from an independent replication study. In the absence of a replication study, most conclusions in the psychology literature result from statistical procedures that minimize training error, which is typically an underestimate of the prediction error yielded by hypothetical replication datasets (James, Witten, Hastie, & Tibshirani, 2013). Thus, a potential strength of machine learning is the natural implementation of a train-test framework to optimize the tradeoff between bias and variance and improve the detection of meaningful patterns in the data.

*Resampling Strategies: K-fold cross-validation and out-of-bag bootstrap*

Various resampling strategies have been developed to manage the bias-variance tradeoff and assess the performance of the model. Resampling strategies assess and validate a model by repeatedly splitting the entire dataset into training sets and test sets. The learning algorithm is trained on each training set and predictions are made on the corresponding test set from which the performance measure is calculated. Then, the individual performance values are aggregated (e.g., by calculating their mean). There are various resampling strategies, including cross-validation (e.g., holdout cross-validation, k-fold cross-validation, leave-one-out cross-validation) and bootstrap resampling approaches. Resampling strategies for assessing model performance have been documented in detail in machine learning texts (e.g., Simon, 2007). For the current study, k-fold cross validation and out-of-bag bootstrap resampling were used.

K-fold cross-validation involves randomly dividing the sample into *k* groups or “folds” of approximately equal size. The first fold is used as a validation or test dataset and the model is fit on the remaining folds. The performance metric, such as mean squared error (), is then computed using the observations in the held-out fold to calculate the predictive ability of the model. This process is repeated times and then yields estimates of the test error (. At the end of this procedure, k-fold cross validation yields an aggregated cross-validation estimate of the test error. In practice, researchers often use 5-fold or 10-fold cross validation, though the optimal number of folds is arbitrary and likely dependent on the sample size (Bengio & Grandvalet, 2003). In out-of-bag bootstrap resampling, new data sets are drawn from the data set, , with replacement, each of the same size as the original dataset. In the *-*th iteration of the procedure, is used as the training set and the remaining elements from are used as the test set. When the sample size is large and as the number of bootstrap sample increases, the average bootstrap sample contains approximately 63.2% of the original observations and omits 36.8% (Chernick & LaBudde, 2014). Due to sampling with replacement, individual observations may appear more than once in the test set. Importantly, for each iteration, observations from the training set are not also present in the test set.

*Nested resampling*

As described in the overview of machine learning models used in the current study, some machine learning algorithms involve variables internal to the model that control how the model makes predictions. These variables, called hyperparameters, affect the performance of the model but are not estimated from the data. Three general strategies are used to specify hyperparameter values: 1) Choosing a sensible default value that has worked on similar modeling problems, 2) Manually trying a few different values and identifying the one that yields the best performance, or 3) Using a procedure called hyperparameter tuning to automate the selection process. As hyperparameters are not typically known *a priori,* hyperparameter tuning is the ideal procedure for maximizing the likelihood of selecting the hyperparameter values that balance the bias-variance tradeoff and minimize prediction error of the model. It is recommended to include hyperparameter tuning in the model training and testing process. This is achieved through *nested cross-validation*, where a resampling procedure in the inner loop is used to identify the optimal hyperparameter values, which is then passed to an outer cross-validation loop. In the outer cross-validation loop, the optimal hyperparameters are used for model fitting in each training set and model performance is tested in the outer loop test sets.

The general process of nested cross-validation is described below:

1. Split the data into training and test sets using a specified resampling procedure. This splitting of the dataset occurs in the outer resampling loop.
2. In each training set, another resampling approach is used to test potential hyperparameter values. The resampling approach embedded within each outer loop training set is called the inner loop.
3. The hyperparameter that gives the best cross-validated performance from each inner loop is then used in the outer loop.
4. The model is trained on each training set of the outer loop using the optimized hyperparameters from its inner loop. These models are used to make predictions on their tests sets from the outer loop.
5. In the outer loop, the performance metrics calculated in each outer loop test set are aggregated to estimate how the model will perform on future, unseen data.

The current study used nested resampling to estimate model performance for EN regression and random forest, which both include hyperparameters. Out-of-bag bootstrap resampling with 100 iterations was used in the outer loop and 5-fold cross-validation was used in the inner loop. For each iteration, a bootstrap sample was selected and used as an outer fold training set. Within this training set, 5-fold cross-validation was used. Each of the inner sets cross-validates a single hyperparameter that was randomly chosen from a set of potential hyperparameter values by splitting the bootstrap sample into training and test sets. For each fold in these inner sets, a model was trained using the training set and evaluated on the test set, using the hyperparameter value of the inner loop training set. The hyperparameter from each inner cross-validation loop that yielded the best model performance was then used to train the models in the outer loop. Over 100 outer loop iterations, a distribution of 100 model performance metrics was generated for each model, allowing for between-model comparison of model performance using inferential statistics.

CHAPTER 4

METHODS COMPARISON STUDY: DATA ANALYTIC PLAN

*Sample Characteristics*

Details on study design, participants, and measures for the dataset used in this methods comparison study are presented in Chapter 2. As depicted in the CONSORT chart (Figure 1), *n* = 300 participants were randomized into either the CBT or EDU treatment conditions. The final sample used for analysis included *n* = 278 participants who enrolled in and completed the study.

*Predictors*

A total of 41 baseline variables were used to predict outcomes at post-treatment. Baseline variables include participants’ and parents’ demographic information, physical symptoms, psychological variables, and coping styles. A list of baseline variables is presented in Table 3.

*Outcomes*

The primary outcome measure was youth report of GI symptoms (CSSI-24; Walker et al., 2009). Secondary outcome measures were: (1) youth report of abdominal pain severity on the Abdominal Pain Index (API; Laird et al., 2015) a multidimensional measure of pain with items assessing the frequency, duration, and intensity of abdominal pain and, (2) youth report of activity impairment on the PROMIS Pain Interference scale (Varni et al., 2010), a measure assessing the extent to which pain hinders engagement with social, physical, and emotional domains.

*Missing Data*

All analyses were performed in R (R Development Core Team, 2013). Post-treatment outcome analyses included all participants who completed both baseline and post-treatment assessments. Percent missing at post-treatment was 17% for GI symptoms and pain interference (remaining n = 231 of 278) and 19% for API (remaining n = 226 of 278). Proportion missing was very minimal for baseline predictors: 40 out of 41 predictors had either complete or less than 2.7% missing data. Socio-economic status had the highest proportion of missing data, ranging from 9.9% – 10.2% across the three outcomes. Missing data were imputed using classification and regression trees such that continuous variables were imputed using regression trees and categorical variables were imputed using classification trees using the ‘*rpart’* package in R. An imputed dataset was created for each outcome variable and was used for all subsequent analyses involving that outcome.

**Modeling Approaches**

Analyses were completed using the *glmnet* (Friedman et al., 2010), *party* (Hothorn et al., 2006)*, partykit* (Hothorn & Zeileis, 2015) and *mlr* packages (Bischl et al., 2016) in R (R Development Core Team, 2013). In order from most simple to complex, the final modeling approaches included: 1) linear regression with all 41 baseline predictors; 2) linear regression with a sequential floating backward search (SFBS) algorithm for variable selection; 3) linear regression with elastic net regularization (EN); and 4) conditional random forest analysis (CRF). All three model-based approaches included only main effects of the predictors. The EN regression and CRF modeling approaches required tuning of their hyperparameters as part of the modeling process. For EN regression, *λ* was allowed to vary from 0 to 5 and α varied from 0 to 1. For the CRF parameter set, number of trees, *ntree,* varied from 50 to 500 and *mtry*, the number of variables selected to determine each split, varied from 1 to 41 variables.

As explained previously, it is well-understood that implementing linear regression with 41 predictors (of which many are inter-correlated) would not be advisable. However, this approach was included in the initial comparison of model performances as an ideal “reference” model due to its familiarity to researchers. Given the limitations described previously, standard linear regression is hypothesized to perform very poorly against the other models. The hypothesized poor performance of this model provides a helpful contrast to remind the reader why alternative models from the statistical and machine learning literatures that can account for variable selection are more preferred for precision medicine initiatives.

**Model Performance Indices**

The model performance indices used in the current study include mean squared error (MSE), mean absolute error (MAE), coefficient of determination (*R*2), and the correlation between observed and predicted values (*R*). Each estimate of model performance uses a different formula to capture the discrepancy between observed and predicted values. Definitions of model performance indices are presented in Table 5.

Of note, the coefficient of determination (*R*2) is a commonly used measure of goodness of fit for regression models that is used to estimate the ratio of explained variance of the model to the total variance of the dependent variable. Most researchers are familiar with this measure in the context of OLS regression, in which *r*2 denotes the proportion of variance in the dependent variable explained by the independent variable(s). Importantly, there are several definitions of *R*2, leading to inconsistencies and misspecification of *R*2 throughout the literature, especially when using modeling approaches outside of the standard least squares regression context (e.g., linear regression without an intercept term, nonlinear fitting techniques). Consistent with the recommendations by Kvalseth (1985), the most general definition of the coefficient of determination is: 1 – [(Sum of Squares of Residuals)/(Total Sum of Squares)]. For a model with perfect fit, the modeled values would exactly match the observed values, resulting in a sum of squares of residuals of 0 and a *R*2 of 1. A baseline model, which always predicts the mean of the dependent variable will yield an *R*2 of 0. Models that have worse predictions than the mean will yield negative *R*2 values. Though *R*2 typically ranges from 0 and 1, negative *R*2 values in the current study indicate a) gross model misspecification, or b) data are grossly contaminated with outliers that are not accounted for by the model. In other words, an *R*2 less than 0 denotes a complete lack of fit.

**Model Comparison Procedure**

Out-of-bag bootstrap resampling with 100 iterations was used to assess the performance of each learning algorithm. During resampling, the entire dataset is repeatedly split into training and test sets. The algorithm is trained on each training set and predictions are made on the corresponding test set, yielding the final performance measures. Because EN and CRF include the additional step of hyperparameter tuning, nested resampling was used for these algorithms to first identify the optimal hyperparameters using the inner loop. Then, model performance was estimated using the test set of the outer loop. Specifically, 5-fold cross validation was used on the inner loop to perform hyperparameter tuning for the EN and CRF models to obtain the ideal parameter set. To search the parameter space, a random search algorithm with 500 iterations was specified, which randomly picks parameter combinations 500 times to identify the optimal parameter configuration. This was chosen over other search methods, such as the grid search, which tests all possible parameter configurations and is computationally expensive. After identifying the optimal hyperparameter set using the inner loop, the parameter set was then applied to the training set of the outer bootstrap resampling procedure, and model performance was calculated using the out-of-bag, bootstrapped test sets.

For each model, the bootstrap resampling procedure yielded 100 estimates for each model performance index, allowing for between-model comparison via inferential statistics. Of note, there is no standardized scale to designate descriptive categories for goodness of fit. Thus, model performance indices are best used for between-model comparison for the same regression task (i.e., same outcome variable).

Table 5. Definition of model performance indices used in regression

|  |  |
| --- | --- |
| **Model Performance Index** | **Definition** |
| MSE: Mean Square Error | Sum of the squared absolute residuals between observed values for each case and values predicted by the model, divided by the number of cases (i.e., the mean of squared residuals). The MSE is more sensitive to outliers than the MAE because the size of the squared residual grows quadratically, indicating greater distance from the model prediction. Lower MSE indicates better predictive accuracy. |
| MAE: Mean Absolute Error | Sum of the absolute residuals between observed values for each case and values predicted by the model, divided by the number of cases. The MAE is interpreted as the mean absolute distance of the cases from the predicted values. |
| *R*2: R-squared or Coefficient of Determination | The proportion of variance in the dependent variable that is explained by the model. *R*2 typically ranges from 0 to 1. However, the formula used for *R*2 can yield negative values, indicating that the model is not better than predicting the mean of the dependent variable. Negative values also denote that the modeling approach is likely grossly inappropriate for the modeling task. |
| Pearson’s *R* | Correlation between observed and predicted values. A correlation of 1 would indicate that the model predictions were identical to the actual cases used in the analysis. |

**Comparison of Significant Predictors**

After comparing model performance, the best performing model was selected to examine predictor importance. In EN regression, this is done by examining the predictors that are kept and then rank ordering the coefficients based on magnitude, with larger coefficients denoting greater significance. In CRF, conditional permutation variable importance values and corresponding variable importance plots indicate the most important predictors. Larger variable importance values indicate greater importance of the variable.

The purpose of examining variable importance was twofold: 1) To identify the most important predictors of each outcome, and 2) To validate the FAP phenotype variable by comparing whether the most important predictors correspond to the variables used to phenotype patients into FAP subgroups. To this end, two sets of predictors were used for each outcome variable. The first set included all baseline predictors. The second set included the FAP phenotype variable and excluded all baseline variables that were used for phenotyping. Comparison of variable importance plots yielded by the two predictor sets (all predictors vs. predictors with FAP phenotype) allows for the direct assessment of 1) whether the predictors used for phenotyping are important in predicting treatment outcomes and 2) whether the FAP phenotype variable successfully captures the information from the individual predictors and remains an important variable in the second model.

It is hypothesized that the CRF model will outperform the other models due to its higher complexity and ability to detect underlying and clinically meaningful interaction effects among predictors. To the extent that this is true, follow-up analyses will probe underlying interaction effects by plotting representative conditional inference trees for each outcome variable.

CHAPTER 5

METHODS COMPARISON STUDY: RESULTS

**Model Performance**

Model performance indices for all four models are presented in Table 6. Results indicate that the CRF algorithm significantly outperformed the other modeling approaches. Linear regression with all 41 predictors demonstrated the poorest performance compared to the other three models. Distributions of the 100 bootstrap-sampled correlations between observed and predicted values of linear regression with SFBS, EN regression, and CRF are presented in Figure 6. Results from a one-way analysis of variance (ANOVA) indicate that for GI Symptoms and Pain Interference, CRF significantly outperformed EN regression, which outperformed linear regression with SFBS. For Abdominal Pain, CRF outperformed both EN regression and linear regression with SFBS, but the two model-based approaches were not significantly different from each other in terms of model performance.

**Predictor Importance**

Given the model comparison results, CRF was selected for subsequent analyses to examine predictor performance. Variable importance results are presented by outcome variable. The purpose of examining variable importance was twofold: 1) To identify the most important predictors of each outcome, and 2) To compare whether the most important predictors correspond to the variables used to phenotype patients into FAP subgroups.

The first set of predictors included all baseline variables. To minimize redundancy, the original set of 41 baseline predictors was reduced to 35 baseline variables. This reduction was achieved by excluding baseline variables that were very similar to each other or from the same measure. For example, the Children’s Somatization Inventory yields scores for both number of GI and non-GI symptoms in the form of counts, and a continuous score denoting severity of GI and non-GI symptoms. To reduce redundancy, the continuous severity score was kept while the count variables were omitted. The second predictor set of 27 variables included the three-level FAP phenotype variable (High Pain Dysfunctional, High Pain Adaptive, and Low Pain Adaptive) and excluded the individual baseline predictors that were used to create this subgroup variable. 10-fold cross validation was used for hyperparameter tuning and then variable importance was extracted for each tuned model. Model performance was assessed using nested resampling, as hyperparameters were tuned using 10-fold cross validation in the inner loop and optimal parameters were used in 100 bootstrap samples in the outer loop to estimate model performance. Model performance indices are presented in Table 7.

Variable importance plots do not provide an official “cutoff” or threshold to determine a subset of important predictors. Instead, visual inspection of variable importance plots allows for discrimination of important predictors compared to those that do not add incremental utility.

Table 6. *Model performance results comparing four models across three outcome variables*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Performance | | | | |
| All 41 Baseline Predictors | | | | |
| Model | Mean Squared Error  (*MSE*) | *R*2 | Mean Absolute Error  (*MAE)* | Mean *R* |
| **Linear Regression with all predictors** |  |  |  |  |
| GI Symptoms | 0.65 (0.50) | -0.25 (0.96) | 0.59 (0.06) | 0.64 (0.07) |
| Abdominal Pain | 3.53 (5.91) | -2.82 (6.06) | 0.85 (0.16) | 0.55 (0.17) |
| Pain Interference | 4.22 (6.88) | -11.15 (20.04) | 0.58 (0.18) | 0.52 (0.22) |
| **Linear Regression with sequential search** |  |  |  |  |
| GI Symptoms | 0.36 (0.06) | 0.33 (0.08) | 0.48 (0.04) | 0.63 (0.02) |
| Abdominal Pain | 0.65 (0.08) | 0.30 (0.08) | 0.64 (0.04) | 0.62 (0.02) |
| Pain Interference | 0.25 (0.03) | 0.31 (0.07) | 0.39 (0.03) | 0.62 (0.02) |
| **Linear Regression with elastic net regularization** |  |  |  |  |
| GI Symptoms | 0.42 (0.06) | 0.22 (0.11) | 0.51 (0.04) | 0.65 (0.03) |
| Abdominal Pain | 0.75 (0.12) | 0.17 (0.13) | 0.69 (0.06) | 0.63 (0.03) |
| Pain Interference | 0.29 (0.05) | 0.21 (0.11) | 0.41 (0.03) | 0.66 (0.02) |
| **Conditional random forest** |  |  |  |  |
| GI Symptoms | 0.39 (0.07) | 0.26 (0.07) | 0.50 (0.04) | 0.77 (0.02) |
| Abdominal Pain | 0.71 (0.10) | 0.24 (0.07) | 0.66 (0.05) | 0.76 (0.02) |
| Pain Interference | 0.26 (0.03) | 0.30 (0.06) | 0.40 (0.03) | 0.78 (0.01) |

*MSE* = Mean of Squared Errors, defined as mean((response – truth)2); range = [0, infinity)

*R2* = R-squared, also called “coefficient of determination,” defined as

[1 – (Residual Sum of Squares) / (Total Sum of Squares)]; range = (-infinity, 1]

*MAE* = Mean of Absolute Errors, defined as

mean(abs(response - truth)); range = [0, infinity)

Mean *R* = Mean Pearson’s correlation between observed and predicted values across 100 bootstrap samples

Figure 6. Comparison of correlation between observed and predicted values (*R*) between linear regression with SFBS (LinearSFBS), EN regression (ElasticNet), and CRF (cForest) analysis

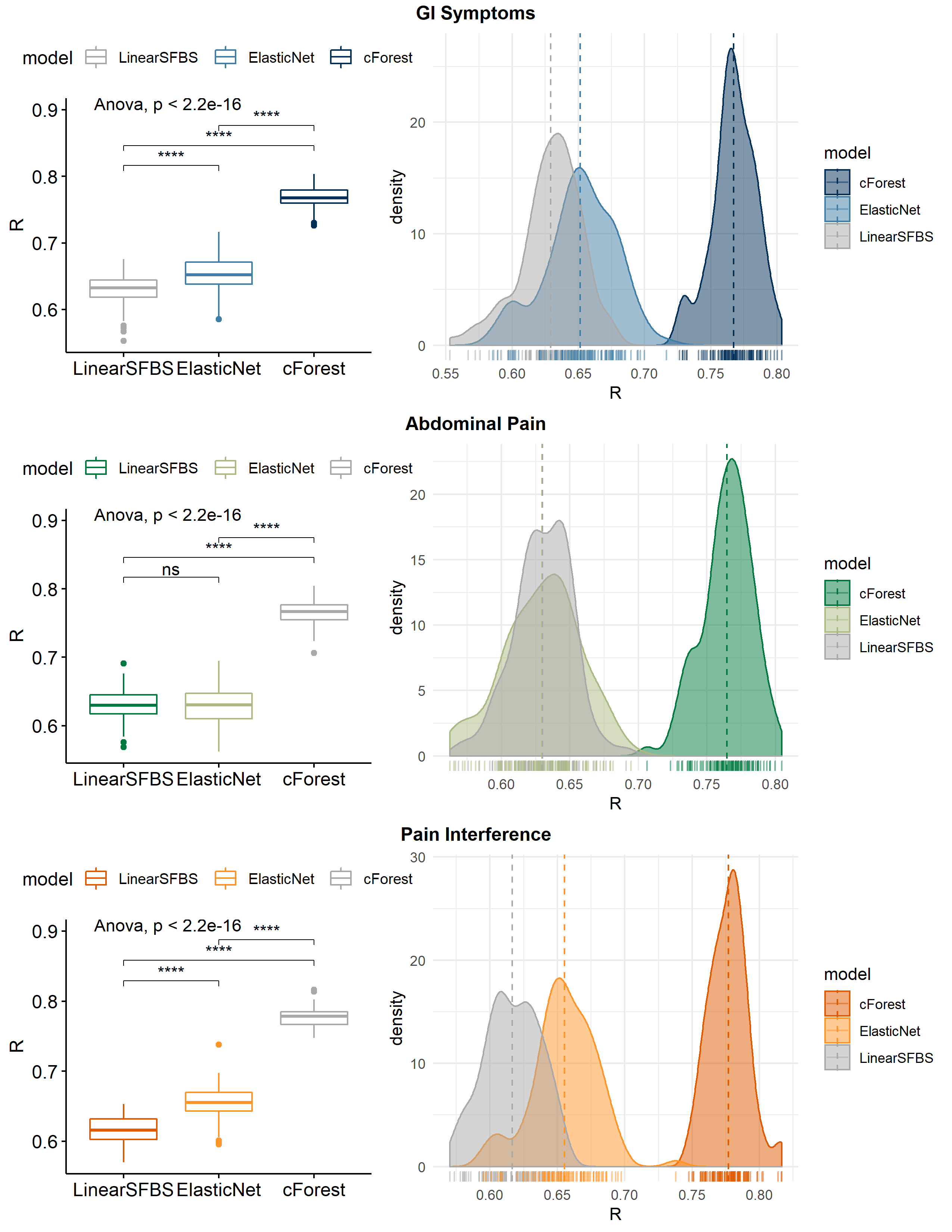
Note*. \*\*\*\* = p < 0.0001; ns = not significant*

Table 7. Model performance indices for CRF models with all predictors (p = 35) and with the FAP phenotype variable (p = 27).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Mean Squared Error  (MSE) | R2 | Mean Absolute Error  (MAE) | R |
| **GI Symptoms** |  |  |  |  |
| All predictors | 0.39 (0.07) | 0.26 (0.07) | 0.50 (0.04) | 0.77 (0.02) |
| w/ Phenotype | 0.41 (0.06) | 0.24 (0.07) | 0.51 (0.04) | 0.75 (0.02) |
| **Abdominal Pain** |  |  |  |  |
| All predictors | 0.71 (0.10) | 0.24 (0.07) | 0.66 (0.05) | 0.76 (0.02) |
| w/ Phenotype | 0.77 (0.10) | 0.17 (0.06) | 0.69 (0.05) | 0.73 (0.02) |
| **Pain Interference** |  |  |  |  |
| All predictors | 0.26 (0.03) | 0.30 (0.06) | 0.40 (0.03) | 0.77 (0.01) |
| w/ Phenotype | 0.26 (0.03) | 0.29 (0.06) | 0.40 (0.03) | 0.76 (0.02) |

A close up of a map

Description automatically generatedFigure 7. Variable importance plots for final CRF models predicting GI symptoms at post-treatment

As displayed in Figure 7, the most important predictors predicting GI symptoms at post-treatment were baseline GI symptoms and level of depressive symptoms, followed by anxiety symptoms and parents’ protectiveness in response to child’s pain. Five of the nine variables used for phenotyping were included in the top 10 variables relevant for predicting GI symptoms at post-treatment. In the model with 27 predictors including FAP phenotype, the phenotype variable was the most important predictor, followed by child’s anxiety symptoms, parents’ protectiveness over child’s pain, and patient’s sleep disturbance.

Figure 8. Variable importance plots for final CRF models predicting abdominal pain at post-treatment

A close up of a map

Description automatically generated

As shown in Figure 8, the most important predictors of abdominal pain at post-treatment included baseline abdominal pain, child’s pain threat appraisal, GI symptoms, depressive symptoms, and functional disability. Of note, all these baseline predictors were used to subgroup patients into their respective FAP phenotypes at baseline. In the model with FAP phenotype, the phenotype variable was again the most important predictor, followed by patient’s sleep disturbance and number of outpatient visits.

Figure 9. Variable importance plots for final CRF models predicting pain interference at post-treatment

A close up of a map

Description automatically generated

Displayed in Figure 9, the most important baseline predictors predicting pain interference at post-treatment included child’s anxiety symptoms, depressive symptoms, functional disability, pain threat appraisal, and sleep disturbance. Three of these variables were used to subgroup patients into their respective FAP phenotypes at baseline. In the model including FAP phenotype, child’s anxiety and sleep disturbance continued to be the most important predictors, followed by FAP phenotype.

**Follow-Up Analysis: A Forest of Stumps *versus* Trees**

Results of model performance indicated that CRF outperformed EN regression. A primary difference between the two approaches is that CRF accommodates the modeling of complex, interactive, and non-linear relationships among predictors, which contrasts EN given that only main effects were included for the EN regression analysis. Thus, to test the hypothesis that there exist underlying complex interactions among the baseline predictors, the CRF algorithm was run with two different parameter sets for hyperparameter tuning. The first parameter set required that the *maxdepth* parameter was set to 1, indicating that the algorithm would only include conditional inference trees without interactions (i.e., a random forest algorithm made up of ‘stumps’). The second parameter set allowed tree depth to vary from 1 to 10. Hyperparameter tuning via 10-fold cross validation was used to identify the optimal tree depth in the random forest model.

Table 8. Model performance indices examining CRF specifying a forest of stumps vs. a forest of trees for all outcome variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Performance | | | | |
| Stumps vs. Trees | | | | |
| Model | Mean Squared Error  (MSE) | *R*2 | Mean Absolute Error  (MAE) | Mean *R* |
| **GI Symptoms** |  |  |  |  |
| CRF with stumps | 0.42 (0.07) | 0.24 (0.05) | 0.52 (0.04) | 0.60 (0.02) |
| CRF with trees | 0.40 (0.06) | 0.25 (0.08) | 0.50 (0.04) | 0.77 (0.02) |
| **Abdominal Pain** |  |  |  |  |
| CRF with stumps | 0.76 (0.08) | 0.17 (0.05) | 0.69 (0.04) | 0.57 (0.03) |
| CRF with trees | 0.71 (0.09) | 0.23 (0.08) | 0.67 (0.05) | 0.77 (0.02) |
| **Pain Interference** |  |  |  |  |
| CRF with stumps | 0.27 (0.03) | 0.26 (0.06) | 0.41 (0.02) | 0.62 (0.02) |
| CRF with trees | 0.25 (0.04) | 0.29 (0.07) | 0.39 (0.03) | 0.78 (0.01) |

Table 8 provides model performance indices comparing the CRF algorithm under two parameter settings. Results indicate that the CRF algorithm allowing for the modeling of interactions (i.e., forest of trees) significantly outperformed the CRF algorithm prohibiting the modeling of interactions among predictors (i.e., a forest of stumps). Figure 10 provides a visual comparison of the two approaches and illustrates the distributions of the correlations between observed and predicted values for each bootstrap sample.

Figure 10. Distributions of correlations between observed and predicted values across 100 bootstrap samples for the CRF algorithm allowing interactions among predictors (forest of decision trees) *vs.* prohibiting interactions (forest of decision stumps)

A close up of a map

Description automatically generated

**Probing Underlying Interactions Using Conditional Inference Trees**

Follow-up analyses support the idea that the CRF algorithm is superior to model-based regression methods due to its ability to model underlying interactions among predictors. Instead of requiring the user to specify underlying, unknown, and complex relations, the CRF algorithm is designed to identify such patterns in the data. As an ensemble algorithm, one limitation of the CRF is that the output of CRF yields the variable importance plots, not an interpretable decision tree. Thus, to probe underlying interactions, a “representative” conditional inference tree was plotted for each outcome using the most important predictors from the CRF analyses that included the FAP phenotype variable (p = 27). Importantly, treatment condition was not an important predictor based on CRF analyses for any of the outcomes. However, given that these data are from an RCT study with the goal of identifying not just differential outcomes, but differential treatment outcome based on FAP phenotype, treatment condition was also included in the plotting of individual conditional inference trees. This specification provides an analogue to the original model-based linear mixed effects model approach used in the longitudinal RCT study, as it allows for the detection of Treatment x FAP subgroup interaction in the tree-based framework.

For GI symptoms, the conditional inference tree analysis included child’s anxiety symptoms, FAP phenotype, and treatment condition. For abdominal pain, FAP phenotype, number of outpatient visits, sleep disturbance, and treatment condition were included as predictors. For pain interference, FAP phenotype, child’s anxiety, child’s sleep disturbance, and treatment condition were included as predictors. The minimum number of individuals at each terminal node was set at n = 15.

Figure 11. Representative conditional inference tree predicting GI Symptoms at post-treatment

A picture containing map, text, small, sitting

Description automatically generated

Figure 11 displays the conditional inference tree predicting GI symptoms at post-treatment. The algorithm identified a three-way interaction between FAP phenotype, child’s anxiety, and treatment condition. As shown, the algorithm first partitioned individuals based on FAP phenotype such that the High Pain Dysfunctional (phenotype = 1) individuals are separated from individuals in the High Pain Adaptive (phenotype = 2) and Low Pain Adaptive (phenotype =3) subgroups. In the High Pain Dysfunctional subgroup, individuals with a T score of ≤ 53.5 on the PROMIS anxiety measure have lower GI symptoms at post-treatment (Node 3), irrespective of treatment condition. For individuals in the High Pain Dysfunctional subgroup with T scores of > 53.5 on the PROMIS anxiety measure, individuals in the CBT group have lower GI symptoms at post-treatment (Nodes 7 and 8) compared to those assigned to EDU (Node 5). Specifically, CBT appeared to be most effect for individuals in the High Pain Dysfunctional subgroup with a moderate level of anxiety symptoms (T scores between 53.5 and 64.5), as shown in the comparison between Node 7 and Node 8. Overall, individuals in the High Pain Adaptive (phenotype = 2) subgroup had greater GI symptoms at post-treatment (Nodes 11 and 12) compared to individuals in the Low Pain Adaptive subgroup (phenotype = 3; Nodes 14 and 15), and this effect was moderated by level of anxiety symptoms, not treatment condition.

Figure 12. Representative conditional inference tree predicting abdominal pain at post-treatment

A close up of a map

Description automatically generated

The conditional inference tree predicting abdominal pain at post-treatment is presented in Figure 12. The representative conditional inference tree depicts three-way interactions between FAP phenotype, sleep disturbance, and treatment condition, and FAP phenotype, number of outpatient visits, and treatment condition. As shown in the tree diagram, the algorithm immediately partitions the sample based on FAP phenotype, separating the Low Pain Adaptive (phenotype = 3) subgroup from the two high pain subgroups (phenotype = 1 and phenotype = 2). For the Low Pain Adaptive subgroup, individuals assigned to EDU had lower abdominal pain at post treatment (Node 16) compared to individuals assigned to CBT (Node 17). For the two high pain subgroups, individuals with greater than six outpatient visits were partitioned out and demonstrated overall high levels of abdominal pain at post-treatment (Node 14), irrespective of treatment condition. For the remaining individuals in the High Pain Dysfunctional subgroup, individuals assigned to CBT had overall lower abdominal pain at post-treatment (Node 8) compared to those assigned to EDU (Nodes 6 and 7). For individuals in the High Pain Adaptive subgroup, individuals with ≤ 2 outpatient visits and assigned to EDU exhibited the lowest levels of abdominal pain at post-treatment (Node 11), while individuals with between two to six outpatient visits demonstrated comparable levels of abdominal pain regardless of treatment condition (Nodes 12 and 13).

Figure 13. Representative conditional inference tree predicting pain interference at post-treatment

A close up of a map

Description automatically generated

The conditional inference tree predicting pain interference at post-treatment is presented in Figure 13. The representative conditional inference tree depicts a four-way interaction between child’s anxiety, FAP phenotype, sleep disturbance, and treatment condition. For individuals with a T score on the PROMIS anxiety measure ≤ 45.1, there was no evidence of moderation based on treatment. Instead, sleep disturbance and FAP phenotype were more relevant predictors of pain interference at post-treatment (Nodes 4, 5, and 6). For individuals with a T score of > 45 on the PROMIS anxiety measure, individuals in either the High Pain Adaptive (phenotype = 2) or High Pain Dysfunctional (phenotype = 1) subgroups demonstrated low levels of pain interference at post-treatment if they had less sleep disturbance (T score ≤ 53.2) and were assigned to CBT (Node 14).

CHAPTER 6

DISCUSSION

*“You cannot answer a question that you cannot ask,*

*and you cannot ask a question that you have no words for*.”

*-Judea Pearl, Book of Why (Pearl & MacKenzie, 2018)*

The present study examined the performance and utility of model-based and machine learning approaches for predicting differential treatment outcomes in the context of a randomized clinical trial (RCT) assessing the effectiveness of internet-delivered cognitive behavior therapy for youth with chronic abdominal pain. In addition to implementing cross-validation to compare model performances, the availability of a well-designed RCT dataset with an empirically-supported subgrouping variable allowed for the “cross-validation” of data analytic findings with substantive knowledge relevant to tailoring treatment for individuals with chronic abdominal pain. Results allow for discussion of methodological insights specific to model performance and selection, and substantive findings regarding the effectiveness of CBT for youth with chronic abdominal pain.

**Methodological Insights: Model Performance and Selection**

Findings support the superior performance of conditional random forest (CRF) analysis, a bottom-up, data-driven, machine learning approach, compared to model-based alternatives, with correlations between observed and predicted values ranging from .76 to .78. The CRF method demonstrated utility in completing two analytic tasks relevant to precision medicine: 1) Identifying baseline patient characteristics relevant to treatment outcomes at post-treatment, and 2) Modeling underlying interactions and complex relations among important variables. Comparing the CRF model to model-based approaches where only main effects were included allowed for an intrinsic test of the presence or absence of underlying interactions among explanatory variables. Superior performance of the CRF model suggested the presence of underlying interactions among variables that were missed in model-based approaches. This possibility was further tested by comparing model performance of the CRF algorithm with the constraint of modeling only a forest of “stumps” (i.e., no interaction effects) versus allowing the algorithm to model a forest of trees (i.e., allowing for the modeling and detection of underlying interactions). Results indicated that allowing for the modeling of interaction effects yielded significantly improved performance and that only allowing main effects yielded model performance that was comparable to the main-effect-only, model-based approaches. This finding corroborates the assertion that standard model-based approaches used in psychosocial intervention trials may underfit the data and omit meaningful and complex higher-order interactions, thus leading to inaccurate and imprecise predictions at the individual level. As such, developing prediction models in datasets with a high number of predictors may be particularly beneficial for advancing precision psychiatry and clinical psychology treatment research (Bzdok & Meyer-Lindenberg, 2018; Dwyer et al., 2018).

By using the most important predictors from the CRF models to fit individual conditional inference trees for each outcome, representative tree models uncovered interaction effects that would be omitted in a traditional model-based framework. Without *a priori* knowledge of interactions or clinically informed hypotheses, researchers would face difficulty justifying the inclusion of interaction effects in the model building process in a traditional inferential statistics framework. In other words, without knowing which interaction effects to test, a researcher could not examine whether they exist. In addition to detecting underlying interactions, the conditional inference trees also revealed unique threshold effects that would be difficult to probe in a model-based framework. For example, when predicting abdominal pain at post-treatment (Figure 12), individuals with > 6 outpatient visits were immediately partitioned out from the individuals in the two high pain subgroups. For individuals with ≤ 6 outpatient visits, those assigned to EDU had significantly lower levels of abdominal pain at post-treatment compared to CBT, but *only if* they also had ≤ 2 outpatient visits prior to study enrollment. The corollary of this effect in a model-based framework would be a three-way FAP Phenotype x Outpatient Visit x Treatment Condition interaction. The tree-based diagram shows that this moderation effect is driven by the subset of 22 participants in the High Pain Adaptive subgroup who also had ≤ 2 outpatient visits. Substantively, this result suggests that healthcare utilization is an important proxy for pain severity and adaptive coping, such that individuals with less reliance on the healthcare system will naturally remit over time and do not require more time- and resource-intensive CBT intervention. Importantly, this level of specificity would not be offered by a significant three-way interaction in a model-based framework.

The well-known instability of single trees may prompt skepticism regarding the generalizability of the tree diagrams presented for each outcome variable. One potential solution is to recode the variables used in the tree analysis based on clinical cutoffs. Though discretization of variables is typically not recommended, building a tree-based diagram using established clinical cutoffs more closely simulates clinical decision-making. Such visual representations of a predictive model also facilitate real-world application of data-driven, algorithmic findings. Just as clinical neuroscientists seek to translate neurobiological findings “from bench to bedside,” clinical data scientists derive tree-based diagrams as readily interpretable tools for distilling complex theories of psychopathology to help clinicians optimize and tailor care to the individual.

**Substantive Insights: Personalizing Interventions for Chronic Abdominal Pain**

*Validation of FAP subgroups*

A primary strength of the motivating dataset chosen for this study was the presence of a hypothesized subgrouping variable that was incorporated in the original RCT study for detecting possible treatment x subgroup interaction. The FAP phenotype variable stratified patients based on many baseline patient characteristics, thus serving as a composite moderator of treatment outcome (Kraemer, 2013). The availability of this subgrouping variable allowed for the explicit testing of treatment x subgroup interaction in a traditional model-based framework. Results of the original study (Walker, Stone, Han, et al., in preparation) indicated evidence for treatment x subgroup interaction from pre- to post-treatment, thus providing an ideal setting to pursue possible convergent findings using data-driven, machine learning approaches. Variable importance plots from the CRF analysis incorporating all baseline variables showed that, of the nine variables used to stratify patients by FAP phenotype, five out of the nine variables were consistently in the top 10 predictors for all three outcome variables. Furthermore, variable importance results from the CRF model that included FAP phenotype but excluded variables used for phenotyping demonstrated that the FAP phenotype variable was the most important predictor of GI symptoms and abdominal pain, and the third most important predictor for pain interference, at post-treatment. These results suggest that, out of all the patient characteristics collected at baseline, the variables used to create the FAP phenotype are indeed important predictors of treatment outcome. Furthermore, the FAP phenotype variable adequately encodes the explanatory information in each of the variables used in the phenotype clustering algorithm. As such, results of the current study provide additional validation of the FAP phenotype variable as a useful clinical stratification tool that is predictive of differential longitudinal outcomes.

*Moderators of treatment outcomes*

For GI symptoms, conditional inference trees demonstrated a three-way FAP Phenotype x Treatment x Anxiety Symptoms interaction. Consistent with the treatment x subgroup interaction identified in the original study, individuals in the High Pain Dysfunctional phenotype demonstrated greater improvement in GI symptoms when assigned to CBT compared to EDU. However, this effect was particularly pronounced for individuals with a moderate level of anxiety symptoms (53.5 < PROMIS T ≤ 64.5). Individuals in the HPD subgroup with sub-clinical levels of anxiety symptoms (T Score ≤ 53.5) had low levels of GI symptoms at post-treatment, irrespective of treatment condition, while individuals with a high level of anxiety (T Score > 64.5) maintained moderate to high levels of GI symptoms at post-treatment. Results suggest that individuals with clinical levels of anxiety symptoms may require additional intervention beyond what is targeted in the Web-MAP protocol. Overall, levels of GI symptoms for individuals in the High Pain Adaptive and Low Pain Adaptive subgroups were lower than individuals in the High Pain Dysfunctional subgroup at post-treatment. Within the two adaptive subgroups, higher levels of anxiety symptoms at baseline were also related to higher levels of GI symptoms at post-treatment. These findings are consistent with previous work demonstrating the relationship between anxiety and self-reported GI symptoms for youth with chronic abdominal pain (Walker & Green, 1989).

For abdominal pain, the conditional inference tree highlighted the moderating roles of sleep disturbance and outpatient visits (i.e., health service utilization) to further characterize treatment x subgroup interaction. Specifically, individuals in the High Pain Dysfunctional subgroup demonstrated greater improvement when assigned to CBT compared to EDU. However, for HPD individuals assigned to EDU, greater baseline sleep disturbance was associated with higher abdominal pain at post-treatment. This finding suggests that the sleep hygiene modules of the Web-MAP protocol may be particularly beneficial and that maintaining healthy sleep behaviors is a key mechanism of therapeutic change for patients with high levels of abdominal pain at baseline. The tree diagram also indicated that number of outpatient visits may be an important indicator of impairment such that individuals who had > 6 outpatient visits prior to the intervention study maintain high levels of abdominal pain at post-treatment, regardless of treatment condition. Furthermore, individuals with ≤ 2 outpatient visits prior to the intervention demonstrate low levels of abdominal pain at post treatment when assigned to the EDU condition. This suggests that individuals with low health service utilization who already possess positive coping strategies for managing pain may be particularly suited for the less intensive EDU treatment condition. Importantly, the regression tree results provide additional nuance to relevant patient characteristics and help to identify which modules of the Web-MAP intervention, such as sleep hygiene, may be particularly effective.

For pain interference, individuals were immediately partitioned based level of anxiety symptoms. Individuals in the High Pain Dysfunctional subgroup who also had greater sleep disturbance (T score > 55.8) demonstrated the highest level of pain interference at post-treatment. Evidence of FAP Phenotype x Treatment Condition moderation was seen for individuals in the two adaptive subgroups (High Pain Adaptive + Low Pain Adaptive). For these individuals, individuals assigned to CBT demonstrated lower pain interference at post-treatment compared to those assigned to EDU. Interestingly, results from the tree diagrams predicting abdominal pain and pain interference at post-treatment partially corroborate a recently published study from the current RCT dataset investigating the role of sleep in adolescents with chronic abdominal pain (Murphy, Palermo, Tham, Stone, Han, Bruehl, Garber & Walker, 2020). Murphy and colleagues (2020) investigated the relationship between sleep disturbance and pain-related variables, including functional disability and healthcare utilization, at baseline. Model-based linear regression showed that, controlling for age and sex, sleep disturbance predicted additional variance in functional disability, lending support to the role of sleep disturbance in daily pain interference. Furthermore, their study used model-based negative binomial regression to show that sleep disturbance predicted number of emergency room visits, but not outpatient visits. The CRF model with 27 baseline predictors predicting abdominal pain at post-treatment (Figure 8) showed that number of outpatient visits was the third most important variable and ER visits was ranked fourteenth. This extends findings by Murphy and colleagues (2020) by suggesting that healthcare utilization in the form of outpatient visits may be particularly relevant to a patient’s *response to treatment*. In the current study, the tree-based model also revealed a threshold of > 6 outpatient visits as an indicator of treatment resistance at post-treatment.

Taken together, results from the conditional inference tree analysis suggest a similar pattern of results across outcome variables, such that evidence of Treatment x FAP Phenotype interaction is most pronounced for individuals in the High Pain Dysfunctional subgroup. If resources were limited, patient improvement would be maximized by assigning CBT to these individuals. Anxiety symptoms, sleep disturbance, and healthcare utilization in the form of outpatient visits were shown to additionally moderate response to treatment. Specifically, irrespective of treatment condition, high levels of anxiety and sleep disturbance may explain non-remission of symptoms and reflect important targets for additional intervention (e.g., exposure therapy or CBT for insomnia).

**Limitations and Future Directions**

The present study should be interpreted in light of study limitations. First, the implementation of machine learning algorithms in clinical psychology and psychiatry, and specifically for treatment of chronic pain, is still in a preliminary stage. Though nested cross-validation was used to prevent overfitting and estimate accurate model performance indices, external validation from independent samples is warranted to provide additional support for the identification of important variables and interaction effects. Precision medicine also advocates for the inclusion of variables across levels of explanation. Therefore, future work may prioritize the inclusion of physiological, biological, or genetic predictors of differential experience of chronic pain. In the current study, the primary explanation for the superior performance of the CRF analysis is its ability to detect underlying interaction effects among predictors for predicting treatment outcome. However, other explanations are plausible and could be tested with the inclusion of additional models.

*Other potential comparison models*

The recursive partitioning approach utilized in tree-based models identifies patient subgroups by using patient predictors to partition individuals into homogenous subgroups. In contrast to model-based approaches, partitioning the dataset into homogenous subgroups provides a non-parametric definition of moderation that does not rely on explicit specification or modeling of functional relations between predictor and outcome. In fact, previous work has shown that when using model-based approaches, the presence of non-linear functional relations between predictors and outcomes may lead to the detection of spurious moderation effects, especially when such relations are not pre-specified in the model (Lubinski & Humphreys, 1990). Thus, in addition to EN regression, assessing the performance of model-based approaches that allow for non-linear functional forms is warranted to decipher whether improved model performance of CRF was due to its ability to accommodate non-linearities, detect underlying interaction effects, or both. For example, penalized spline regression (Yu & Ruppert, 2002) is a statistical learning technique that allows for variable selection and models non-linearities in the data. Inclusion of this model in the comparison study would provide a stepwise comparison of main-effect-only linear models, main-effect-only non-linear models, and non-parametric models agnostic to underlying functional forms.

To test if superior CRF performance is primarily due to the detection of moderation effects, future work may also consider using more complex model-based approaches from the statistical learning literature that can select two-way and/or higher-order interaction effects (e.g., stepwise regression approaches such as *stepAIC* or *bootStepAIC* in R; Zhang, 2016). Finally, the original motivating dataset included outcome measurements at multiple longitudinal timepoints. Recent developments in tree-based and random forest approaches extend these algorithms to a longitudinal context (Kundu & Harezlak, 2019). Given the growing availability of such resources, longitudinal extensions may be particularly appropriate for the current motivating dataset.

**Conclusions**

The current study represents a thorough demonstration of the utility of machine learning methods in precision psychiatry. Results showed that data-driven, machine learning approaches (i.e., conditional random forest) resulted in superior prediction of treatment outcomes due, in part, to its capability in accommodating and identifying underlying and complex interaction effects among patient variables. The comparison of model-based and machine learning approaches in the context of a substantive clinical area (i.e., chronic pain) revitalizes a key debate in our field, previously named the “clinical-statistical controversy” (Grove & Meehl, 1996). In their article, Grove and Meehl (1996) entertain arguments against actuarial prediction tools and provide rebuttals for each one. In the end, they conclude that since it can be shown that actuarial prediction is superior to the clinical judgement of even the most skilled clinicians, researchers and practitioners should prioritize statistical prediction. The current methods comparison study extends this important discourse by examining statistical judgement from both *model-based* and *data-based* frameworks. Results strongly support the need for more widespread acknowledgement and incorporation of complex modeling approaches, such as machine learning methods, in clinical psychology and psychiatry to allow analytic approaches to match the known complexity of psychopathology and improve prospects for precision psychiatry.

REFERENCES

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno’s (1990) survey of PhD programs in North America. *American Psychologist*, *63*(1), 32–50.

Aiken, L. S., West, S. G., Pitts, S. C., Baraldi, A. N., & Wurpts, I. C. (2012). Multiple linear regression. In *Handbook of Psychology, Second Edition*. American Cancer Society.

Aiken, L., West, S., & Reno, R. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.

Alin, A. (2010). Multicollinearity. *WIREs Computational Statistics*, *2*(3), 370–374.

Bengio, Y., & Grandvalet, Y. (2003). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research*, *5*, 1089–1105.

Bischl, B., Lang, M., Kotthoﬀ, L., Schiﬀner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). mlr: Machine Learning in R. *The Journal of Machine Learning Research, 17*(1), 5938-5942.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadworth International Group.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250.

Chernick, M. R., & LaBudde, R. A. (2014). *An introduction to bootstrap methods with applications to R*. John Wiley & Sons.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE*, *9*(1), e83875.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science*, *76*(4), 444–463.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118.

Eccleston, C., Morley, S., Williams, A., Yorke, L., & Mastroyannopoulou, K. (2002). Systematic review of randomized controlled trials of psychological therapy for chronic pain in children and adolescents, with a subset meta-analysis of pain relief: *Pain*, *99*(1), 157–165.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible Smoothing with $B$-splines and Penalties. *Statistical Science*, *11*(2), 89–102.

Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science*, *196*(4286), 129–136.

Engel, G. L. (1981). The clinical application of the biopsychosocial model. *The Journal of Medicine and Philosophy*, *6*(2), 101–124.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, *49*(1), 92–107.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106-E6115.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., & Gallop, R. (2009). Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. *Journal of Consulting and Clinical Psychology*, *77*(4), 775–787.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gatchel, R. J., Peng, Y. B., Peters, M. L., Fuchs, P. N., & Turk, D. C. (2007). The biopsychosocial approach to chronic pain: Scientific advances and future directions. *Psychological Bulletin*, *133*(4), 581–624.

Grove, W., & Meehl, P. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293.

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440.

Hollon, S. D., Cohen, Z. D., Singla, D. R., & Andrews, P. W. (2019). Recent developments in the treatment of depression. *Behavior Therapy*, *50*(2), 257–269.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research, 16*(1), 3905-3909.

Howard, R. F. (2003). Current status of pain management in children. *JAMA*, *290*(18), 2464–2469.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, *1*, 519–537.

Kattan, M. W., Yu, C., Stephenson, A. J., Sartor, O., & Tombal, B. (2013). Clinicians versus nomogram: Predicting future technetium-99m bone scan positivity in patients with rising prostate-specific antigen after radical prostatectomy for prostate cancer. *Urology*, *81*(5), 956–961.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., de Jonge, P., Nierenberg, A. A., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2016). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*(1), 22–36.

Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current Opinion in Psychiatry*, *31*(1), 32.

Korterink, J. J., Diederen, K., Benninga, M. A., & Tabbers, M. M. (2015). Epidemiology of pediatric functional abdominal pain disorders: A meta-analysis. *PLOS ONE*, *10*(5), e0126982.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *The Lancet Psychiatry*, *3*(10), 935–946.

Kraemer, H. (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: A parametric approach. *Statistics in Medicine*, *32*(11), 1964–1973.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*(10), 877.

Kundu, M. G., & Harezlak, J. (2019). Regression trees for longitudinal data with baseline covariates. *Biostatistics & Epidemiology*, *3*(1), 1–22.

Kvalseth, T. O. (1985). Cautionary note about R2. *The American Statistician*, *39*(4), 279–285. JSTOR.

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin, Psychological Bulletin*, *143, 143*(2, 2), 142, 142–186.

Laird, K. T., Sherman, A. L., Smith, C. A., & Walker, L. S. (2015). Validation of the Abdominal Pain Index using a revised scoring method. *Journal of Pediatric Psychology*, *40*(5), 517–525.

Levy, R. L., Langer, S. L., Walker, L. S., Romano, J. M., Christie, D. L., Youssef, N., DuPen, M. M., Feld, A. D., Ballard, S. A., Welsh, E. M., Jeffery, R. W., Young, M., Coffey, M. J., & Whitehead, W. E. (2010). Cognitive behavioral therapy for children with functional abdominal pain and their parents decreases pain and other symptoms. *The American Journal of Gastroenterology*, *105*(4), 946–956.

Loh, W.-Y., He, X., & Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, *34*(11), 1818–1833.

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious" moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological bulletin*, *107*(3), 385.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician*, *36*(3), 158–160.

Marshall, R. J. (2001). The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology*, *54*(6), 603–609.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117.

Palermo, T. M., Eccleston, C., Lewandowski, A. S., Williams, A. C. de C., & Morley, S. (2010). Randomized controlled trials of psychological therapies for management of chronic pain in children and adolescents: An updated meta-analytic review. *PAIN®*, *148*(3), 387–397.

Palermo, T. M., Law, E. F., Fales, J., Bromberg, M. H., Jessen-Fiddick, T., & Tai, G. (2016). Internet-delivered cognitive-behavioral treatment for adolescents with chronic pain and their parents: A randomized controlled multicenter trial. *Pain*, *157*(1), 174–185.

Palermo, T. M., Wilson, A. C., Peters, M., Lewandowski, A., & Somhegyi, H. (2009). Randomized controlled trial of an Internet-delivered family cognitive–behavioral therapy intervention for children and adolescents with chronic pain. *PAIN*, *146*(1), 205–213.

Passos, I. C., & Mwangi, B. (2018). Machine learning-guided intervention trials to predict treatment response at an individual patient level: An important second step following randomized clinical trials. *Molecular Psychiatry*.

Petkova, E., Park, H., Ciarleglio, A., Todd Ogden, R., & Tarpey, T. (2020). Optimizing treatment decision rules through generated effect modifiers: A precision medicine tutorial. *BJPsych Open*, *6*(1).

Pfeiffer, R. M., Park, Y., Kreimer, A. R., Lacey, J. V., Pee, D., Greenlee, R. T., Buys, S. S., Hollenbeck, A., Rosner, B., Gail, M. H., & Hartge, P. (2013). Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: Derivation and validation from population-based cohort studies. *PLoS Medicine*, *10*(7), e1001492.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. *Journal of medical systems*, *26*(5), 445-463.

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, *15*(11), 1119–1125.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Ruppert, D., & Carroll, R. J. (1999). *Penalized regression splines*. Cornell University Operations Research and Industrial Engineering.

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., ... & McGrath, P. J. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR\* D report. *American Journal of Psychiatry*, *163*(11), 1905-1917.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Shiffman, R. N. (1997). Representation of clinical practice guidelines in conventional and augmented decision tables. *Journal of the American Medical Informatics Association*, *4*(5), 382–393.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Simon, R. (2007). Resampling strategies for model assessment and selection. In *Fundamentals of data mining in genomics and proteomics* (pp. 173-186). Springer, Boston, MA.

Stone, A. L., & Walker, L. S. (2017). Adolescents’ observations of parent pain behaviors: Preliminary measure validation and test of social learning theory in pediatric chronic pain. *Journal of Pediatric Psychology*, *42*(1), 65–74.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, *10*(2).

Subramanian, S. V., Kim, R., & Christakis, N. A. (2018). The “average” treatment effect: A construct ripe for retirement. A commentary on Deaton and Cartwright. *Social Science & Medicine*, *210*, 77–82.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Turk, D. C., & Okifuji, A. (2002). Psychological factors in chronic pain: Evolution and revolution. *Journal of Consulting and Clinical Psychology*, *70*(3), 678–690.

Varni, J. W., Stucky, B. D., Thissen, D., DeWitt, E. M., Irwin, D. E., Lai, J.-S., Yeatts, K., & DeWalt, D. A. (2010). PROMIS Pediatric Pain Interference Scale: An item response theory analysis of the pediatric pain item bank. *The Journal of Pain: Official Journal of the American Pain Society*, *11*(11), 1109–1119.

Walker, L. S., Beck, J. E., Garber, J., & Lambert, W. (2009). Children’s Somatization Inventory: Psychometric properties of the revised form (CSI-24). *Journal of Pediatric Psychology*, *34*(4), 430–440.

Walker, L. S., Dengler-Crish, C. M., Rippel, S., & Bruehl, S. (2010). Functional abdominal pain in childhood and adolescence increases risk for chronic pain in adulthood. *PAIN*, *150*(3), 568–572.

Walker, L. S., Sherman, A. L., Bruehl, S., Garber, J., & Smith, C. A. (2012). Functional abdominal pain patient subtypes in childhood predict functional gastrointestinal disorders with chronic pain and psychiatric comorbidities in adolescence and adulthood. *PAIN®*, *153*(9), 1798–1806.

Walker, L. S., Smith, C. A., Garber, J., & Claar, R. L. (2005). Testing a model of pain appraisal and coping in children with chronic abdominal pain. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *24*(4), 364–374.

Wiech, K. (2016). Deconstructing the sensation of pain: The influence of cognitive processes on pain perception. *Science*, *354*(6312), 584–587.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, *97*(460), 1042–1054.

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, *4*(7).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

1. All participants were included in analysis in the group in which they were randomized, regardless of whether they had received the intervention, e.g., due to noncompliance or drop-out. [↑](#footnote-ref-1)