ADVANCES IN DATA-DRIVEN RESEARCH METHODOLOGY FOR PRECISION PUBLIC HEALTH

Michael T. Lawson

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2019

Approved by:

Michael R. Kosorok

Eric B. Bair

Michael I. Love

Elizabeth B. Mayer-Davis

Donglin Zeng

ABSTRACT

Michael T. Lawson: Advances in Data-driven Research Methodology for Precision Public Health
(Under the direction of Michael R. Kosorok)

The rise of precision medicine has ushered in manifold opportunities and challenges, many of them linked. For instance: precision medicine offers an avenue to revisit assumption-rich, knowledge-driven research practices, but requires careful and creative thinking to replace them. In this manuscript, we turn our attention to three such areas of interest: subgroup determination, modeling of dynamical systems, and accounting for measurement error. In each case, we construct a statistical and machine learning framework for the problem at hand, develop methodology to address it, and present theoretical and numerical justifications for the methodology.

In the first chapter, we develop a data-driven method for subgroup determination in a clinical trial of treatment or intervention, where subgroups are based on predicted efficacy of treatment and not based on a limited number of a priori-specified markers. The proposed subgroup determination method is illustrated in a trial of a lifestyle intervention in type 1 diabetes, where we use it to determine subgroups who are expected to benefit from intervention and from control conditions. In the second chapter, we formulate a fully nonparametric stochastic differential equation model that performs model selection for factors affecting both the mean and variability of a dynamic process. The model is applied to data arising from a type 1 diabetes trial. In the third chapter, we turn our attention to developing model-agnostic influence statistics to assess the impact of observations' mismeasurement on analysis results. This method is illustrated in detail in three different settings, one of which is a study of water quality with a complex mechanism of measurement error.

"He would have to look up quickening dark & say: *Me. I do. It's mine.*"

For those who inspire.

ACKNOWLEDGEMENTS

We gratefully acknowledge the following contributors to this research: Michael Kosorok, Anna Kahkoska, Elizabeth Mayer-Davis, Jamie Crandell, David Maahs, Michael Seid, and David Holcomb, and the following funding sources: NIEHS T32 ES007018 and the Royster Society of Fellows.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BMIz	Body Mass Index Z-score
CI	Confidence Interval
HbA1c	Hemoglobin A1c
ITR	Individualized Treatment Rule
LASSO	Least Absolute Shrinkage and Selection Operator
MICE	Multiple Imputation using Chained Equations
ODE	Ordinary Differential Equation
OWL	Outcome Weighted Learning
QoL	Quality of Life
RLT	Reinforcement Learning Trees
SD	Standard Deviation
SDE	Stochastic Differential Equation
T1D	Type 1 Diabetes

CHAPTER 1: INTRODUCTION

The rise of big data, "-omics," and precision approaches have revolutionized many areas of healthcare and health research. New technologies have engendered powerful and innovative data, whose size and complexity dwarf those of traditional health research data. High-dimensional and complex data have inspired novel methods capable of utilizing them. But perhaps the most potentially impactful shift is one of mindset: the precision medicine framework offers health research the opportunity to trade knowledge-driven, assumption-rich tools for data-driven, assumption-light approaches. In this research, we examine three areas of health research and propose new data-driven tools for use in those areas.

Simmons et al. (2011) coined the term "researcher degrees of freedom" to refer to a number of related breakdowns in the scientific method that can lead to misleading, non-reproducible, or even outright incorrect conclusions from a study or analysis that is, for the most part, carried out proficiently and correctly. Examples include *post hoc* modifications to inclusion and exclusion criteria, procedures for handling missing data, flexibility in choice of analysis method, and so on. While there is no solitary correct answer to the question raised by researcher degrees of freedom, data-driven research methodology presents a principled approach to many of these issues. The aim of this research is to provide data-driven tools to use in areas of research where they may currently be lacking, and ultimately provide additional support to conducting scientifically rigorous and reproducible research.

Our first area of focus is subgroup analysis of clinical trials, which seeks to clarify the results of a trial by dividing a patient population with heterogeneous intervention response into smaller segments in which intervention response is more homogeneous. Standard methods in subgroup

analysis involve specifying the variables which subgroups can be based on *a priori*, and many involve purely descriptive subgroups, where only a patient's overall prognosis is considered. In Chapter 2, we propose a data-driven method for subgroup determination that is prescriptive in nature, *i.e.* based on the predicted efficacy of treatment for patients within subgroups.

Our second area of focus is the analysis of continuous-time data arising from dynamical processes. The differential equation models used in many applications involving dynamical processes can involve strong assumptions on the functional forms of the terms involved, or even stronger assumptions on the degree of the derivative involved. Additionally, they perform inference only on factors affecting the mean of the process of interest, when the variability of the process may be of direct biological interest. In Chapter 3, we propose a flexible nonparametric first-order stochastic differential equation model that makes minimal assumptions on the functional forms of its covariates to recover the true support for both the mean and variability of a process of interest with multiple covariate processes.

Our final area of focus is the handling of measurement error. In studies with imperfect measurements that are expensive to collect, while it is ideal to correct errors in measurement before their effects can propagate downstream, it is usually not possible, much less efficient, to catch all measurement errors. In Chapter 4, we propose a method for determining the influence of potential mismeasurement in the outcome variable on the results of a study. The proposed measure is model-agnostic, working in a wide variety of study settings, and extensible to the case of simultaneous mismeasurements.

The remainder of this document proceeds as follows. We first carry out a thorough review of the literature surrounding our research topics. Chapter 2 presents our new method for data-driven subgroup determination. Chapter 3 explores our method for learning the mean- and variance-level dynamics of a system. Chapter 4 lays out our method for determining the measurement error-based influence of observations in a study. Chapter 5 discusses directions for future research in these areas. Technical details such as proofs are deferred to the appendices.

CHAPTER 2: PRECISION MEDICINE SUBGROUP ANALYSIS

2.1 Introduction

The Flexible Lifestyles Empowering Change trial (FLEX), an NIH-funded 18-month randomized trial, tested the efficacy of an adaptive behavioral intervention to promote self-management and improve measures of blood glucose control in 258 youth ages 13-16 with type 1 diabetes (T1D). The goal of the FLEX intervention was to increase adherence to type 1 diabetes self-management, including testing blood sugar levels throughout the day, counting carbs, and calculating and delivering insulin doses. Motivational interviewing and problem-solving skills training tailored to participants and their families were integrated into the interventions counseling (Mayer-Davis et al., 2018b). Despite high retention and fidelity, the FLEX study did not show efficacy with respect to the primary outcome of HbA1c at 18 months post-randomization (Mayer-Davis et al., 2018b). However, the intervention was associated with improvements in several secondary psychosocial outcomes, including motivation, problem solving skills, diabetes self-management, and health-related and general quality of life (Mayer-Davis et al., 2018b).

Within any clinical trial, the average treatment effect across all participants may mask important heterogeneous treatment effects visible across different study subjects or subgroups of subjects (Baum et al., 2017). Heterogeneity in response can also obfuscate whether some strategies helped some participants while harming others (Chakraborty and Murphy, 2014; VanderWeele and Knol, 2011). For this reason, statistical analyses that estimate aggregate effects over time for all patients are limited, since they do not account for the fact that treatment 'responders' and 'non-responders' can exhibit vastly divergent patterns of response (Gueorguieva et al., 2011).

In settings where heterogeneity in participant profiles reliably predicts differential response to the efficacy of treatment, the precision medicine approach offers promise (Burton et al., 2012). The precision medicine approach seeks to develop an individualized treatment rule (ITR), a mathematical function that gives recommendations for whether a patient should receive intervention or not. In the FLEX trial, treatments were assigned at baseline, so the ITR based its recommendations solely on patient characteristics available at baseline. As the goal of the FLEX trial was to optimize a patient's improvement over the full 18-month course of the study, the ITR was estimated based on those 18-month improvements in outcome, also called clinical rewards. Once an ITR is estimated, it can be used to target intervention to those patients whom it estimates will benefit most from intervention. An ITR can be summarized based on its value, the average expected reward that results from applying the ITR. That is, the value of an ITR represents the average reward the patient population would have received if that ITR were followed, rather than the observed randomization scheme. ITRs that deliver the best achievable reward are termed optimal. Estimating and applying optimal ITRs may lead to increases in efficiency of prevention and treatment while simultaneously reducing costs of care (Burton et al., 2012; Trusheim et al., 2007). As such, gaining a deeper understanding of the subgroups defined by an optimal ITR—understanding which patients receive improved outcomes under an intervention and which do not—is critical to inform future tailoring of interventions.

Here, we describe a method for the analysis of randomized trials that leverages the full dataset to identify subgroups defined by an estimated optimal ITR. We demonstrate how this method may be applied to data from the FLEX trial to quantify and describe the subgroups where key clinical outcomes were improved on intervention, the subgroups where outcomes were improved on usual care, and the subgroups where outcomes were the same between intervention and usual care. To be consistent with the primary and secondary outcomes of the parent study, we characterized the effect of the FLEX intervention in terms of 18-month changes in HbA1c (primary outcome), perceived quality of life (QoL), and body mass index z-score (BMIz). We focused on baseline predictors, including sociodemographic characteristics, clinical variables, or

psychosocial and behavioral measures, as these can serve as markers to physicians in the future to guide optimal treatment recommendations with regards to the FLEX intervention using the data available at that time.

2.2 Methods

Study sample

We analyzed data from the baseline visit of the Flexible Lifestyles Empowering Change randomized trial (FLEX). FLEX was a randomized clinical trial testing an adaptive, 18 month intervention which includes behavioral skills and problem solving for youth with T1D, with respect to HbA1c (primary outcome), glycemic variability, CVD risk factors, health-related quality of life, and cost effectiveness (Mayer-Davis et al., 2018b; Kichler et al., 2018). Eligible participants were youth ages 13-16 years with type 1 diabetes for ≥ 1 year, literacy in English, HbA1c 8.0-13.0%, and ≥ 1 primary caregiver with no other serious medical conditions or pregnancy (Kichler et al., 2018). Detailed considerations of the FLEX design and baseline participant characteristics have been described elsewhere (Kichler et al., 2018).

Inclusion Criteria

FLEX enrolled 258 adolescents with T1D who were instructed to wear blinded CGM systems for 7 days at baseline. Participants were excluded from the present analysis if they did not have complete CGM data at baseline ($n = 40$) or were missing the outcomes of HbA1c, QoL, or BMIz at baseline or the 18-month measurement visit ($n = 2$).

Measures

All data collection was standardized as per FLEX study protocol, and FLEX assessment staff were trained and certified to perform all study procedures. The full set of study measurements was obtained at baseline and 6 and 18 months post-randomization; a limited set of measurements

was obtained at 3 and 9 months post-randomization (Kichler et al., 2018). Standardized measurements, laboratory data, clinical measures, and questionnaires from the FLEX study are described in detail in Section 2.5.

Outcome Measures

Univariate Outcomes. To assess the intervention's efficacy for our outcomes of interest individually, we considered three univariate outcomes: change in HbA1c, change in self-reported QoL measured by the PedsQL^{TM} score (QoL) (Varni et al., 2001), and constrained change in BMIz. For each univariate outcome, we considered changes between baseline and the 18-month study visit. For HbA1c and QoL, this change was directly equal to the difference between 18-month and baseline outcomes. Change in BMIz was constrained to reward patients who completed the study with a healthy BMIz or who improved their BMIz over the course of the study. For full mathematical definitions of the univariate outcomes, see Section 2.5.

Composite Outcome. To assess the intervention's effect on all outcomes of interest simultaneously, we considered a composite outcome of change in HbA1c, QoL, and BMIz between baseline and 18-mo. The composite outcome is an approximation of constrained optimization based on a hierarchy of the univariate outcomes, HbA1c prioritized the highest and BMIz prioritized the lowest. In essence, patients with an unacceptably high HbA1c will receive a low composite outcome, regardless of their quality of life and BMIz; patients with an acceptable HbA1c but an unacceptably low quality of life will receive slightly higher composite outcomes, regardless of their BMIz; and patients will receive the highest composite outcomes if they have acceptable HbA1c and quality of life, with the magnitude determined by their BMIz. For a full discussion of the composite outcome's definition and properties, see Sections 2.5 and 2.6.

Analysis

Imputation. Missing data in non-CGM covariates were imputed via Multiple Imputation by Chained Equations (MICE) (White et al., 2011). A flexible imputation method, MICE can ac-

count for mixed data types with minimal assumptions when paired with random forests (Stekhoven and Bühlmann, 2011). We generated eleven imputed datasets with MICE, with which we employed a modified version of multiple imputation. We chose the number eleven as the smallest odd number, precluding the possibility of ties in a majority vote, larger than 10. As a sensitivity analysis, we performed all analyses on the subset of patients with complete cases in all covariates and outcomes ($n = 197$; see Section 2.7).

ITR Estimation. We estimated the optimal ITR in our sample with Reinforcement Learning Trees (RLT). An extension of Breiman's Random Forest model, RLT uses reinforcement learning to better discriminate between signal and noise variables among the covariates (Breiman, 2001; Zhu et al., 2015). An important aspect of RLT is its ability to mute covariates, *i.e.* set their effect identically equal to zero, in subsets of the covariate space. The details of how and why muting occurs are best left for the technical discussion in (Zhu et al., 2015); for this analysis, it suffices to state that the predicted outcome under the different values of a binary variable, such as intervention status, can be exactly equal for some patients but different for other patients.

Using RLT allowed us to pose a nonparametric model between the baseline covariates X and the observed clinical rewards R within each imputed dataset. Using this model, we obtained the expected reward for a given patient under both FLEX intervention and usual care. The imputed dataset-specific ITR assigned a patient to one of three groups. If the expected reward was higher under FLEX intervention, the intervention was expected to benefit that patient, and they were assigned to the Intervention Group. If the expected reward was higher under usual care, then usual care was expected to benefit that patient, and they were assigned to the Usual Care Group. And if the expected reward was identical under usual care and FLEX intervention, then intervention status was expected to have no effect for that patient, and they were assigned to the Muted Group. For each patient, we obtained eleven assignments to Intervention, Control, or Muted Group, one assignment per imputed dataset. The estimated optimal ITR assigned each patient to the group designated by a plurality vote of these 11. Once the groups were defined by the ITR, we examined their baseline demographic, clinical, and psychological/social characteristics.

To test the robustness of our modeling assumptions in the FLEX dataset, mild as they were, we estimated the optimal ITR via Outcome Weighted Learning (OWL) (Zhao et al., 2012), a non-model-based approach, and fully characterized the OWL ITR-assigned subgroups in additional exploratory analyses. A comparison of OWL's performance to RLT and a full discussion of OWL can be found in Sections 2.7 and 2.8, respectively.

ITR Evaluation. Once estimated, each ITR was evaluated on the basis of its value V, the expected reward resulting from applying the ITR to the sample rather than the observed randomization scheme. To facilitate comparisons between model-based and non-model-based ITRs, we used the following definition of V:

$$V(\pi) = \frac{\sum_{i=1}^{n} R_i I\left\{A_i = \pi(X_i)\right\}}{\sum_{i=1}^{n} I\left\{A_i = \pi(X_i)\right\}} \tag{2.1}$$

where $I\{E\}$ is the indicator function that takes the value 1 when event E is true and 0 otherwise, i indexes patients, A is the vector of observed intervention assignments, and π is the ITR whose value is being obtained. Point estimates of ITR values were computed using this formula, and confidence intervals for ITR values and differences in ITR values were computed via bootstrapping, as described in Section 2.9.

Statistical Considerations

Descriptive data are presented as mean (standard deviation (SD)), n (%), or median (interquartile range) for variables that are not normally distributed, such as measures of hypoglycemia. Characteristics of individuals in each ITR-defined subgroup were compared using chi-square or ANOVA (Fisher's Exact and Wilcoxon-Mann-Whitney where appropriate). Pairwise comparisons were performed using chi-squared or t-tests (Fischer's exact or Wilcoxon signed-rank test where appropriate). A two-sided p-value of < 0.05 was considered statistically significant. These analyses were exploratory, and the results of this study are not intended to deterministically guide future intervention assignments; as such, p-values were not adjusted for

multiple comparisons. Imputation and ITR estimation were carried out in R, version 3.4.1, using the packages missForest, RLT, and DTRlearn. Descriptive analyses were conducted using SAS, version 9.4.

2.3 Results

The final study sample included 216 adolescents with T1D in the FLEX trial. The sample was 77% non-Hispanic White and 50% female with a mean (SD) age of 14.9 (1.1) years and mean (SD) type 1 diabetes duration of 6.3 (3.7) years at baseline of the trial. At baseline, the mean (SD) HbA1c was 9.6% (1.2%), mean (SD) BMIz was 0.73 (0.91), and the mean (SD) QOL measure was 81.2 (12.4).

Table 2.1 depicts two measures of interest for evaluating the RLT ITR. The first measure is the estimated value of V across the composite outcome and each univariate outcome. The second is the comparison between the value of the estimated optimal ITR and the fixed treatment effects for both intervention and usual care, which are computed as V with $\pi(X)$ assigning intervention or usual care to all patients, respectively. Note that each column of this table has a different natural scale due to the particular distribution of outcomes in question. All estimates of fixed treatment comparisons lie above zero, and all but one of the 95% confidence intervals lie entirely above zero, indicating the estimated optimal ITR achieved higher expected rewards than blanket assignment of treatment or usual care.

Estimate	HbA1c	QoL	BMIz	Composite
$\hat{V}_{\text{opt}}$	0.6738	0.6739	0.9737	2.6985
$\hat{V}_{\text{opt}} - \hat{V}_{\text{trt}}$	0.0109	0.0152	0.0233	1.2085
CI $\hat{V}_{\text{opt}} - \hat{V}_{\text{trt}}$	(-0.0028, 0.0372)	(0.0004, 0.0766)	(0.0018, 0.0376)	(1.1058, 1.3716)
$\hat{V}_{\text{opt}} - \hat{V}_{\text{ctrl}}$	0.0171	0.0225	0.0234	1.0067
CI $\hat{V}_{\text{opt}} - \hat{V}_{\text{ctrl}}$	(0.0033, 0.0433)	(0.0077, 0.0839)	(0.0019, 0.0377)	(0.9040, 1.1698)

Table 2.1: Estimated Value (Bootstrap 95% Confidence Interval) of RLT Imputed ITR by Outcome Variable

Table 3A-D also depicts the characteristics found to be significantly different across FLEX participants in the subgroups assigned to Intervention and Usual Care for the composite outcome and each univariate outcome. For full descriptive tables, please see Table S8. With the exception of the composite outcome, a large number of participants were assigned to the Muted Group.

Regarding the composite outcome, 91 participants (42%) were assigned to the Intervention, while the remaining 125 participants (58%) were assigned to the Control Group. Individuals assigned to intervention subgroup were less likely to have private health insurance (60% in the Intervention Group versus 78% in the Control Group, $P = 0.01$) (Table 3A).

Regarding the HbA1c univariate outcome, 105 participants (49%) were assigned to the Muted Group, 54 participants (25%) were assigned to the Intervention Group, and 57 participants (26%) were assigned to the Control Group. Individuals assigned to the Intervention Group did not have a significantly higher Hba1c than those assigned to Usual Care (9.4% versus 9.2%; $P = 0.44$), but individuals in the Muted Group had higher mean HbA1c at baseline than those assigned to Intervention or Control (9.9%; $P = 0.02$ and $P < 0.01$, respectively). Individuals in the Muted group also had a higher incidence of clinical and clinically serious hypoglycemia ($P < 0.01$), with no significant differences between the Intervention and Control Group (Table 3B).

Regarding the QoL univariate outcome, 63 participants (29%) were assigned to the Muted Group, 89 participants (41%) were assigned to the Intervention Group, and 64 participants (30%) were assigned to the Control Group. Individuals in the Intervention Group were more likely to have an elevated HbA1c at baseline compared to the Muted Group (75% versus 54%, $P = 0.01$) but not the Control Group (61%; $P = 0.08$). Individuals in the Intervention Group also had higher significantly higher depressive symptoms at baseline compared to those in the Muted Group (mean (SD) CESD 9.8 (8.5) versus mean (SD) CESD score 6.9 (5.4); $P < 0.01$), with no significant differences from the Control Group ($P = 0.44$) (Table 3C).

Regarding the BMIz univariate outcome, 136 participants (63%) were assigned to the Muted Group, 44 participants (20%) were assigned to the Intervention Group, and 36 participants (17%)

were assigned to the Control Group. Mean BMIz at baseline of individuals assigned to the Intervention Group was higher than that of those assigned to the Control Group ($P < 0.01$); this group also had a higher proportion of under- or normal weight individuals using weight status cut-offs (54.6% versus 30.6%; $P < 0.01$). Mean BMIz was not significantly different between the Intervention Group and the Control Group ($P = 0.06$; Table 3D).

2.4 Discussion

In this study, we present a method to identify subgroups of participants in a clinical trial for whom the studied intervention would have been beneficial, for whom the usual care condition would have been beneficial, and for whom the intervention did not make a difference, with regards to key clinical outcomes. We then apply this method to re-analyze data from the FLEX trial, which showed no effects of the intervention on the primary study outcome, to demonstrate that there are distinct subgroups with different optimal treatment assignments. We focus the discussion first on the findings from the *post hoc* analysis of the FLEX trial, and then turn to a more general discussion of the method itself.

The application of a method to find distinct subgroups within a single randomized trial sample is appropriate given previous reports of heterogeneity in response to behavioral interventions (Hampson et al., 2000), including heterogeneity of response to the same intervention in different samples of youths with T1D (Channon et al., 2007; Wang et al., 2010). The relative proportions of these subgroups, especially with regards to the large muted group for HbA1c as a univariate outcome, highlight the challenges with glycemic control in this age range (Mayer-Davis et al., 2018b). By contrast, a larger group was estimated to benefit from the FLEX intervention with regards to QoL, which agrees with the main trial's findings that the FLEX intervention had a positive aggregate effect on multiple measures of psychosocial well-being (Mayer-Davis et al., 2018b).

The results also reveal interesting dynamics in the interplay between the three univariate outcomes. Clinical markers that link interventions to subgroups of patients they are likely to ben-

efit take a central role in precision application of interventions (Kichler et al., 2018); as such, it would be ideal to have a large variety of markers corresponding to differential estimated response patterns (Trusheim et al., 2007; Khoury et al., 2012). Although we considered a range of participant characteristics, in the FLEX study, only a limited subset of characteristics emerged to distinguish the ITR-assigned subgroups. Furthermore, the markers were not consistent across the three univariate outcomes. For example, patients expected to be indifferent to intervention when optimizing for 18-month improvement in HbA1c had higher baseline HbA1c, while patients expected to benefit more from intervention when optimizing for 18-month improvement in QoL had higher HbA1c at baseline. We believe these antagonistic effects may contribute to the paucity of reliable predictors for the subgroups governed by the composite outcome, even among covariates that helped predict subgroups for subgroups governed by univariate outcomes.

The results suggest that RLT estimated an optimal ITR that performed well for the FLEX data. As Table 2.1 shows, the estimated optimal ITR achieved a high V for each outcome. For each outcome, the estimated optimal ITR nominally outperformed the two fixed treatment effects. All but one confidence interval corresponding to these fixed treatment comparisons lay entirely above zero, suggesting that these improvements in value were reliable. The size of the differences in Table 2.1 suggests the treatment effect is generally small in this trial, though the natural scale present in this table may make it appear misleadingly so—the interquartile range for HbA1c rewards, for instance, is 0.108 units, and the interquartile range for BMIz rewards is a mere 0.041 units. While we chose RLT to estimate the optimal ITR for the FLEX data based on its performance, an additional philosophical advantage of using RLT is its specification of a Muted Group. Conceptually, this group represents a phenomenon that is distinct from the other two assignments: this subgroup can be thought to contain "true" non-responders, since they do not show a response to either treatment or usual care conditions. In this study, the large size of the Muted Group for each univariate outcome is likely a reflection of a marginal treatment effect, combined with noise in the data. More work is needed to understand the subgroup of adolescents with T1D who fall

into this category, as these represent the population who may be the most difficult to reach via intervention work and may be at the highest risk of long-term complications of the disease.

This analysis is conceptually and analytically distinct from standard subgroup analysis methods, representing a novel approach to subgroup determination. There are several advantages to this method for *post hoc* analysis of randomized trial data. First, in contrast to descriptive methods such as effect modification analysis, which show how treatment response differs across levels of a third modifying variable, this method is prescriptive, determining patients for whom the intervention is expected to be most beneficial. Second, unlike common approaches that start with *a priori* defined subgroups and identify optimal treatment rules for each group, this analysis identifies an optimal treatment strategy across the entire study sample and uses it to determine subgroups of interest. As these subgroups are not specified *a priori* based on hypothesized mechanism of disease, they may represent previously uncharacterized subgroups that are nevertheless relevant to the optimal delivery of intervention. Moreover, the data-driven nature of this method may help remove "researcher degrees of freedom" that can hinder reproducibility (Ioannidis, 2005; Simmons et al., 2011). Third, estimating an optimal ITR pools information from the entire study sample, not just the arm randomized to intervention. Finally, RLT allows us to model the intervention effect with a remarkably small number of assumptions, and its ability to handle high dimensionality allows us to consider a broad range of participant characteristics as suitable clinical markers, including aspects of clinical care, sociodemographic characteristics, and behavioral measures at baseline that may reinforce or challenge the efficacy of a given therapy over time (Khoury et al., 2012).

Limitations of this analysis include the small sample size and high degree of noise in the data, especially in relation to the small effect size of the intervention. Future work may explore the application of ITR-based subgroups to other trial data of different sample sizes with a range of intervention effect magnitudes. However, these results are important groundwork to expand the available tools for matching individuals and subgroups of individuals to existing and newly studied interventions. There remains a striking lack of consensus about the best approach to

promote adherence and improve glycemic control among youth with T1D. The precision delivery of interventions, based on a diverse breadth of data, as modeled in this study, offers a promising road forward.

2.5 Measures

In this section, we present details surrounding the measurement of key variables in the FLEX trial. We first give the details of measurement methodology for several covariates and outcomes. We then present the mathematical definitions of the four outcome variables used for our ITR estimation methods.

2.5.1 Measurement Methodology

Standardized Measurements

All data collection was standardized as per FLEX study protocol, and FLEX assessment staff were trained and certified to perform all study procedures. Adolescents and participating caregivers could choose to complete questionnaires online, through the secure FLEX study website, or during in-person study measurement visits. The full set of study measurements was obtained at baseline and 6 and 18 months post-randomization; a limited set of measurements was obtained at 3 and 9 months post-randomization (Mayer-Davis et al., 2018b).

Laboratory data

A central laboratory (Northwest Lipid Metabolism and Diabetes Research Laboratories, Seattle, WA, USA) provided oversight and conducted all assays. At all timepoints, hemoglobin A1c (HbA1c) was measured in whole blood by using an automated nonporous ion exchange HPLC system (model G-7; Tosoh Bioscience). Measurements of plasma cholesterol, triglycerides, and HDL cholesterol concentrations were performed on a Hitachi 917 autoanalyser (Boehringer Mannheim Diagnostics) at the full measurement visits, after the patient had fasted for at least

eight hours. LDL cholesterol was calculated by the Friedewald equation for those with triglycerides <4.52 mmol/l and by the beta-quantification procedure for those with triglycerides >4.52 mmol/l.

Clinical measures

At baseline and at 6- and 18-months post-randomization, patients wore a blinded CGM (iPro®2 Professional CGM; Medtronic Diabetes, Northridge, CA) for a seven-day period to measure interstitial glucose levels in real time throughout the day and night. Cut-points for glucose used to describe hypoglycemia were established according to recommended values (Danne et al., 2017). Height was measured using a stadiometer, and weight was measured to the nearest 0.1 kg using an electronic scale. Body mass index (BMI, weight (kg) / height^2 (m^2)) was calculated and converted to an age- and sex-specific BMI z-score (BMIz) according to Centers for Disease Control and Prevention growth charts. Blood pressure was measured after five minutes of rest using an aneroid manometer. The second and third of three measures were averaged for systolic and diastolic pressures.

Questionnaires

Patients self-reported race, highest level of parental education, duration of diabetes, insulin delivery method (pump versus multiple daily injections (MDI)), and past use of CGM outside the study in standardized questionnaires. Self-reported race and ethnicity was classified as non-Hispanic white, non-Hispanic Black, Black, and other including Asian/Pacific Islander, Native American, or unknown.

The Diabetes Self-Management Profile Self Report (DSMP-SR) (Wysocki et al., 2012) was used to assess usual practices of diabetes management during the preceding three months, across five domains: exercise, management of hypoglycemia, diet, blood glucose testing, and insulin administration and dose adjustment. Higher scores indicated more diabetes self-management behaviors. The DSMP-SR was modified for the present study to allow a single questionnaire

to be administered regardless of insulin regimen. Symptoms of depression were assessed using the Centers for Epidemiologic Study Depression Scale (CES-D), with higher scores reflecting more depressive symptoms (Radloff, 1977). The composite Pediatric Quality of Life Inventory™ Generic Core Scales (PedsQL™) was used to assess quality of life (QoL) across four domains (physical, emotional, social, and school functioning) during the previous month, with higher scores reflecting better QoL (Varni et al., 2001). Fear of hypoglycemia was completed by both the adolescent and parents and measured three domains (Shepard et al., 2014): maintaining high blood sugar, helplessness/worry about low blood sugar, and worry about negative social consequences.

2.5.2 Outcome Variables

We first introduce notation that will be helpful in our mathematical definitions. Let $R_{1,0}$ and $R_{1,1}$ denote the vector of patient HbA1c at baseline and 18 months, respectively. Let $R_{2,0}$ and $R_{2,1}$ denote the vector of patient quality of life, as determined by the PedsQL Generic scale, at baseline and 18 months, respectively. Finally, let $R_{3,0}$ and $R_{3,1}$ denote the vector of patient BMI Z-score at baseline and 18 months respectively. Let $i = 1, \ldots, n$ index patients, such that $R_{1,0,i}$ denotes patient i's HbA1c at baseline, and so on.

We will define the three univariate outcomes before presenting the definition of the composite outcome.

2.5.2.1 HbA1c univariate outcome

The raw univariate outcome vector for HbA1c is simply given by $R_{1,\text{raw}} = R_{1,1} - R_{1,0}$, i.e. the diffence between HbA1c at baseline and at 18 months. Elevated HbA1c is related to the risk for long-term complications of type 1 diabetes (Nathan et al., 2014). All participants in the FLEX trial had an elevated HbA1c at baseline, meaning reductions in HbA1c are expected to reduce the risk of long-term complications. As such, the raw univariate outcome is scaled such that more negative values are preferable, as they correspond to the greatest reductions in HbA1c.

By convention, the clinical reward in ITR estimation settings is strictly positive, with larger values corresponding to better rewards. As such, we define the scaled univariate HbA1c outcome as follows:

$$R_1 = \frac{\max(R_{1,\text{raw}}) - R_{1,\text{raw}}}{\max(R_{1,\text{raw}}) - \min(R_{1,\text{raw}})}. \tag{2.2}$$

Note that by definition R_1 is restricted between 1 and 0, with larger values corresponding to better outcomes (i.e. greater reductions in HbA1c).

2.5.2.2 QoL univariate outcome

The raw univariate outcome vector for quality of life is given by $R_{2,\text{raw}} = R_{2,1} - R_{2,0}$, i.e. the diffence between QoL scores at baseline and at 18 months. The raw univariate outcome is scaled such that more positive values are preferable, as they correspond to the largest increases in QoL.

By convention, the clinical reward in ITR estimation settings is strictly positive, with larger values corresponding to better rewards. As such, we define the scaled univariate QoL outcome as follows:

$$R_2 = \frac{R_{2,\text{raw}} - \min(R_{2,\text{raw}})}{\max(R_{2,\text{raw}}) - \min(R_{2,\text{raw}})}. \tag{2.3}$$

Note that by definition R_2 is restricted between 1 and 0, with larger values corresponding to better outcomes (i.e. larger increases in QoL).

2.5.2.3 BMIz univariate outcome

While the raw univariate outcome for BMIz, $R_{3,\text{raw}} = R_{3,1} - R_{3,0}$, offers computational simplicity, it is not preferred because it does not take into account the patient's starting BMIz. Poor glycemic control can result in glucose purging and weight loss (Group et al., 1988; Carlson and Campbell, 1993). Given the elevated HbA1c levels at baseline in the FLEX study, it was expected that some participants might gain weight if glycemic control, the primary endpoint, were improved. Moreover, a substantial portion of the patients ended the trial with a BMIz in the healthy range below 1.04, some of whose BMIz did increase over the course of the 18 month

period. Giving these patients a poor clinical reward is inappropriate given the relationship between glycemic control and body weight and the goals of the study. As such, we define the BMIz outcome to selectively penalize weight gain that results in excess body weight in relation to sex- and age-specific BMI percentiles. Let $R_3^{\neq 1}$ be the subvector of $R_{3,\text{raw}}$ corresponding to all i such that $R_{3,1,i} \geq 1.04$ and $R_{3,1,i} > R_{3,0,i}$ for each i. As such, we consider the following constrained BMIz outcome:

$$R_3 = \begin{cases} 1, & \text{if } R_{3,1} < 1.04 \\ 1, & \text{if } R_{3,1} - R_{3,0} < 0 \\ \dfrac{\max(R_3^{\neq 1}) - R_{3,\text{raw}}}{\max(R_3^{\neq 1}) - \min(R_3^{\neq 1})}, & \text{otherwise.} \end{cases} \tag{2.4}$$

By definition, R_3 is constrained to lie between 0 and 1, with larger values corresponding to better outcomes.

2.5.2.4 Composite Outcome

We introduce the composite outcome, a combination of the three univariate outcomes into a single outcome. The composite outcome is an approximation of constrained optimization, and corresponds to a hierarchy of the univariate outcomes. Befitting the goals of the FLEX intervention, we prioritized HbA1c highest, QoL next, and BMIz third. Heuristically, the composite outcome is defined as follows. We specify thresholds for "failure" for HbA1c and QoL, such that a patient is considered to have unacceptable values of that outcome if they fall on the wrong side of that threshold at the end of the study. Patients who fail HbA1c, regardless of their QoL and BMIz, are placed into the first category and take the lowest numerical reward values, between 0 and 1, with the magnitude determined by how poor their HbA1c is. Patients who have an acceptable HbA1c but fail QoL are placed into the second category, with numerical reward values falling between 1 and 2 depending on how poor their QoL is. Finally, patients who end the trial with acceptable HbA1c and QoL are placed into the third category and given the highest numerical values, with values ranging from 2 to 3 depending on how much their BMIz improved. Both fail-

ure criteria can be circumvented by strong enough improvement—for instance, a patient whose HbA1c at 18 months is 9.0 but whose HbA1c fell by 0.7 over the course of the intervention is not considered to have failed HbA1c for the purposes of the composite outcome.

We define the mutually exclusive outcome threshold events E_1, E_2, E_3 to simplify notation. E_1 is the indicator that $R_{1,1} > 8.5$ and $R_{1,1} - R_{1,0} > 0.5$. E_2 is the indicator that $E_1 = 0$, $R_{2,1} < 60$, and $R_{2,1} - R_{2,0} < 10$. E_3 is the indicator that both E_1 and E_2 equal 0. The thresholds for HbA1c were chosen based on clinical cut-points, and the thresholds for QoL were chosen based on sample quantiles of QoL in the sample. The composite outcome is defined as follows:

$$R = \begin{cases} \dfrac{R_{1,1} - 8.5}{\max(R_{1,1}) - 8.5}, & E_1 = 1 \\ 1 + \dfrac{60 - R_{2,1}}{60 - \min(R_{2,1})}, & E_2 = 1 \\ 2 + R_3, & E_3 = 1. \end{cases} \tag{2.5}$$

2.6 Properties of composite outcome

In this section, we explore the properties of the composite outcome. The composite outcome is a method of approximating constrained optimization over multiple outcome variables, and to our knowledge represents a novel approach to doing so. Constrained problems are of frequent interest in medicine, particularly in complex diseases where it is unlikely for one outcome variable to dominate all others. Type 1 diabetes alone, for instance, presents many sets of outcome variables that lend themselves to a constrained approach. Investigators may wish to enforce ideal glycemic control while discouraging weight gain, a known side effect of intensive insulin therapies (Carlson and Campbell, 1993); they may wish to lower long-term measures of glycemic control such as HbA1c while simultaneously keeping patients out of the acutely dangerous hypoglycemic range measured by a continuous glucose monitor; or they may have more complex sets of goals corresponding to several outcome variables, as is the case in the FLEX trial.

As the mathematical definition presented in 2.5.2.4 suggests, the distribution of the composite outcome depends directly upon an explicit hierarchy of the outcome variables in a trial, as well as meaningful regional thresholds that define failure events for those outcome variables. Both the hierarchy and the thresholds rely on domain knowledge. In the FLEX trial, for instance, domain knowledge suggested the regions of interest for one of our outcome variables—HbA1c below 8.5 is considered a significant improvement for this population, which was recruited with high baseline HbA1c (Association, 2016)—and informed the other, while the priorities of the trial dictated the order of the hierarchy. HbA1c was prioritized highest, as it was the outcome of primary interest and directly related to long-term complications of diabetes. QoL was prioritized second, as it was an outcome of secondary interest and represents an important patient-oriented outcome. Due to natural growth in this age range complicating BMI-based outcomes, and due to the complicated relationship between body weight and glycemic control alluded to in Section 2.5.2.3, BMIz was prioritized after both HbA1c and QoL (Mayer-Davis et al., 2018b). Note that the shape of the regions need not always be as simple as those presented in the FLEX trial—failure thresholds for raw BMI, for example, would likely penalize measurements that are too high as well as those that are too low.

This approach offers one main advantage over constrained optimization: ease of use. Constrained optimization can be difficult to implement, especially when the regions of interest are complex, and the burden introduced by considering additional outcome variables can be far from trivial. So long as the hierarchy and the failure regions remain clear, it is simple to add outcome variables into a composite outcome constructed as in 2.5.2.4. Any analyst equipped to carry out ITR estimation via RLT or OWL should be equipped to create a composite outcome and carry out an almost exactly analogous analysis with it; the same cannot be said of coding constrained optimization by hand.

2.6.1 Numerical Experiments

We examine the finite-sample performance of the composite outcome for ITR-based subgroup determination in a trial with heterogeneous treatment effects through a brief set of numerical experiments. In the first, we examine a very simple model to explore the properties of the method and the composite outcome, particularly regarding the muted group. In the second, we examine the performance of the method in settings more akin to those likely to be observed in real studies, paying special attention to comparing the performance of RLT and OWL-based subgroups.

Several features are similar across experiments. In both cases, the covariates X are drawn i.i.d. from $U(-1, 1)$, with a corresponding Gaussian $p + 1$-vector β that governs the baseline link between covariates and clinical reward R. The observed treatment A is chosen independent of covariates and rewards so that patients are randomized equally to intervention ($A_i = 1$) and control ($A_i = -1$). Each experiment has two clinical rewards, R_1 and R_2. We specify true subgroups for each patient, stored in the n-vector S, where $S_i \in \{-1, 0, 1\}$ denotes whether the patient benefits from control, has identical reward under control and treatment, or benefits from treatment, respectively. We assume that these subgroups apply to each outcome for the sake of simplifying comparisons to the gold standard; the details of how each subgroup is determined varies between experiments. R_1 and R_2 are constructed similarly across experiments. Given fixed treatment effects δ_1 and δ_2 for R_1 and R_2 respectively, each experiment considers two settings. In the first, termed *synergistic*, the treatment effects point in the same direction for both outcomes:

$$R_j = X\beta + \epsilon + \delta_j AS, \tag{2.6}$$

for $j = 1, 2$. In the reverse scenario, termed *antagonistic*, the treatment effects point in opposite ways for the two outcomes:

$$R_1 = X\beta + \epsilon + \delta_1 AS \tag{2.7}$$
$$R_2 = X\beta + \epsilon - \delta_2 AS.$$

The threshold q defines the failure event for R_1: patients with $R_1 < q$ have "failed" R_1 and receive composite outcome $R \in [0, 1]$ based on the magnitude of their R_1, while patients with $R_1 \geq q$ have "acceptable" R_1 and receive $R \in [1, 2]$ based on the magnitude of their R_2, in the manner described in Section 2.5.2.4. We set q to the first quartile of R_1 in the sequel. In each simulation, once the data were generated, we estimated the optimal ITR in the sample using RLT (and, in the second experiment, OWL), then used it to obtain subgroup estimates $\hat{S}$. One set of evaluation metrics for the method is the subgroup recovery sensitivity and specificity, defined as

$$\begin{aligned} \text{sens}_j &= \frac{\sum_{i=1}^{n}(\hat{S}_i = j \cap S_i = j)}{\sum_{i=1}^{n} S_i = j} \\ \text{spec}_j &= \frac{\sum_{i=1}^{n}(\hat{S}_i \neq j \cap S_i \neq j)}{\sum_{i=1}^{n} S_i \neq j}, \end{aligned} \tag{2.8}$$

where $j = -1, 0, 1$. Another evaluation metric is available for the RLT ITR due to the simulated nature of the data. Let R^1 and R^{-1} denote the true value of R under intervention and control. Since we know the magnitude and direction of the true treatment effect for both R_1 and R_2, we can calculate the difference in R obtained by switching a patient on intervention to control or vice versa. For instance, in the synergistic setting, a patient with $S_i = 1$ and $A_i = 1$ would have $R_{i1}^{-1} = R_{i1} - \delta_1$ and $R_{i2}^{-1} = R_{i2} - \delta_2$. Then R_i^{-1} could be obtained by recalculating R in the manner described in (2.5) with these perturbed values. Let $\hat{Q}^1(X)$ and $\hat{Q}^{-1}(X)$ denote the RLT-predicted values of R under treatment and control, respectively. Then we can examine the mean squared

error in treatment effect, defined as

$$\text{MSE} = n^{-1} \sum_{i=1}^{n} \left[\left(R_i^1 - R_i^{-1} \right) - \left(\hat{Q}^1(X_i) - \hat{Q}^{-1}(X_i) \right) \right]^2. \tag{2.9}$$

2.6.1.1 Experiment in Simple Conditions

We first explored a basic model to assess the performance of the method when when all factors are straightforward, particularly with regards to the muted group. In this model, we set $n = 200$ and $p = 1$.

As briefly discussed in Section 2.6.1, we intentionally built a muted group into our simulated data. In particular, we allotted the true subgroup membership according to

$$S = \begin{cases} 1, & X > 0.5 \\ -1, & X < -0.5 \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

In this experiment, we considered both the synergistic and antagonistic setting, and we considered three values for each δ_j: 1, 3, and 10.

As our goal in this simulation was to examine the behavior of the method surrounding the muted group, we used only RLT to estimate the optimal ITR.

Table 2.2 gives the subgroup recovery sensitivity and specificity in the synergistic setting by δ_1, δ_2, and value of S. Overall, we see high sensitivity for both the intervention and control group, especially when the treatment effect for either outcome variable is large. Specificity is less impressive in these groups for large and small treatment effects alike. This is attributable to a curious phenomenon: in none of these simulations did the proposed method recover a muted group. While this seems an indictment of the method, examining the treatment effect MSE in the synergistic setting, displayed in Table 2.3, reveals a different story: the treatment effect MSE is small even when the treatment effects are modest, and very small when treatment effects are

large, which suggests the method is performing well. Figure 2.1 illustrates the disconnect between the conclusions drawn from these evaluation metrics. In particular, when $X_i \in [-0.5, 0.5]$, we know $S_i = 0$, so a method that performs perfectly would have $\hat{Q}^1(X_i) - \hat{Q}^{-1}(X_i)$ within this range. The method does not set any of these differences identically equal to zero for any combination of (δ_1, δ_2); however, for each (δ_1, δ_2), the estimated differences are smaller in this range than outside it, with this trend increasing as the true treatment effect increases. We will return to this observation in our discussion in Section 2.6.1.3.

Table 2.4 gives the treatment effect MSE in the antagonistic setting. Again, MSE is low throughout, suggesting the method performs well even in this conceptually more difficult setting. Figure 2.2 illustrates the predictions and estimated vs. true treatment effects plotted against the true splitting variable X in the antagonistic setting. Again, the predicted differences between interention and control are smaller in magnitude inside the range of the true muted group than outside it, on average, though none are set identically to zero. We also note the somewhat unintuitive behavior of the true treatment effects: although the treatment effect is truly positive for one of the outcome variables, nearly all the true treatment effects are negative. Optimizing on either outcome in a univariate manner would fail to discover this interesting trend, as Figure 2.3 demonstrates: the trend for the univariate outcome appears similar to that exhibited in the synergistic setting.

2.6.1.2 Experiment in Trial-Like Conditions

We briefly explore the performance of the method in conditions more similar to those observed in the FLEX trial. In this experiment, we set $n = 200$ and $p = 30$.

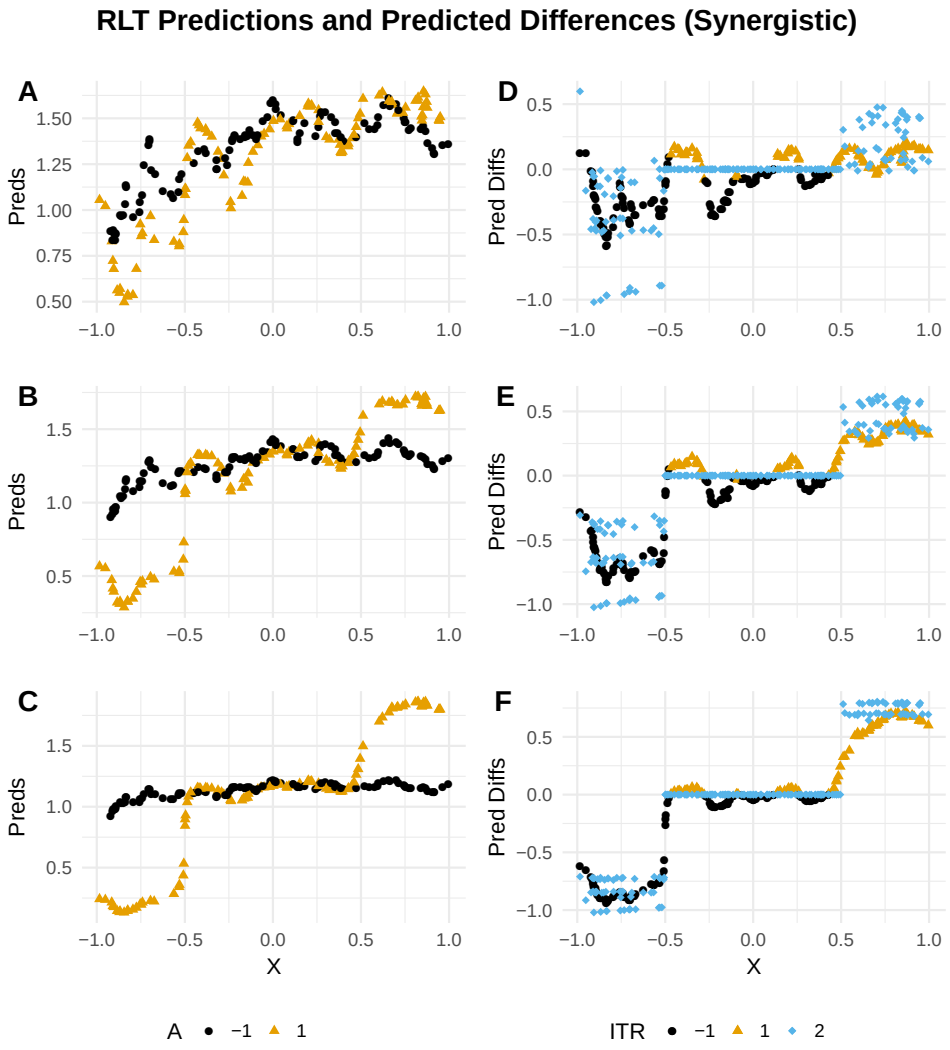

Figure 2.1: (A-C). RLT-predicted values of R by true treatment effect, intervention status, and true splitting variable X for the simple numerical experiment of Section 2.6.1.1, synergistic setting. The true treatment effects for R_1 and R_2 are set to $\delta_1 = \delta_2 = 1, 3, 10$ in A, B, and C, respectively. (D-F). Differences in predicted composite reward R between intervention and control by true treatment effect, RLT ITR assignment, and true splitting variable X, as well as the true treatment effect for R, for the same numerical experiment. ITR=2 denotes the true treatment effect for R. The true treatment effects for R_1 and R_2 are set to $\delta_1 = \delta_2 = 1, 3, 10$ in D, E, and F, respectively.

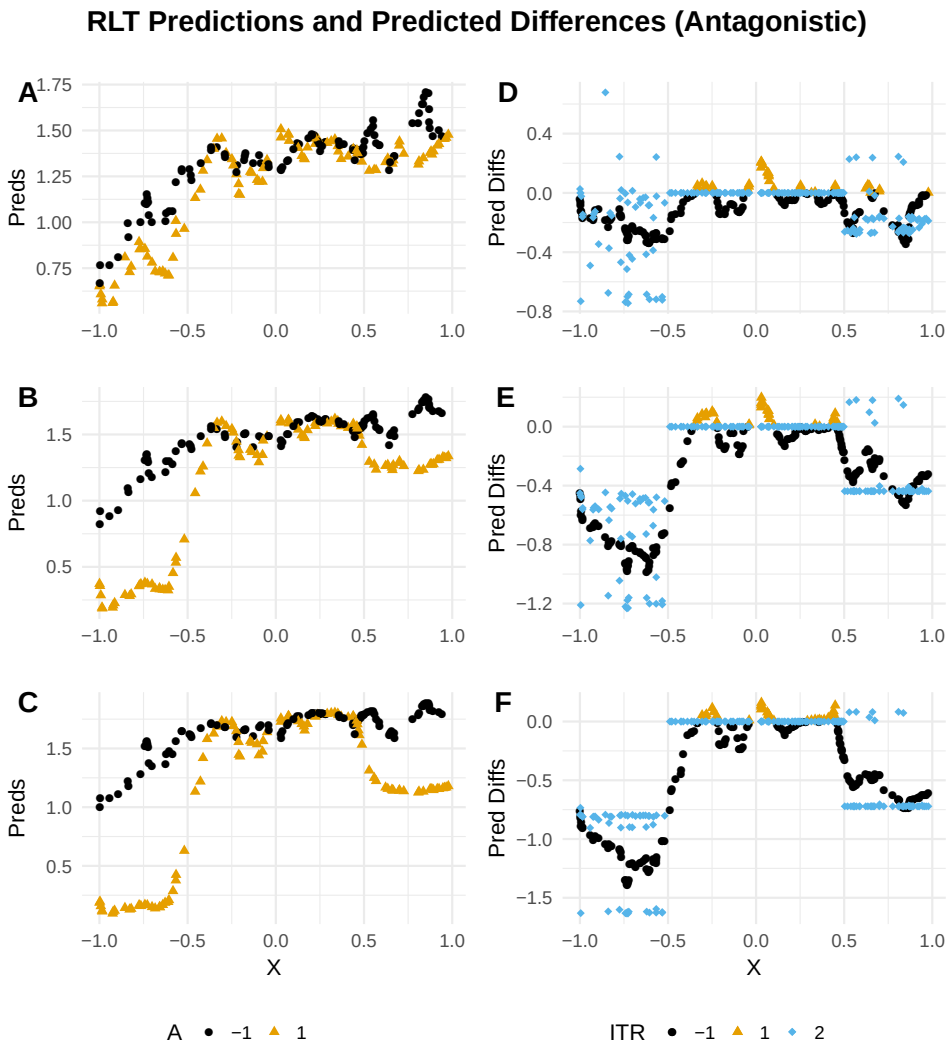

Figure 2.2: (A-C). RLT-predicted values of R by true treatment effect, intervention status, and true splitting variable X for the simple numerical experiment of Section 2.6.1.1, antagonistic setting. The true treatment effects for R_1 and R_2 are set to $\delta_1 = \delta_2 = 1, 3, 10$ in A, B, and C, respectively. (D-F). Differences in predicted composite reward R between intervention and control by true treatment effect, RLT ITR assignment, and true splitting variable X, as well as the true treatment effect for R, for the same numerical experiment. ITR=2 denotes the true treatment effect for R. The true treatment effects for R_1 and R_2 are set to $\delta_1 = \delta_2 = 1, 3, 10$ in D, E, and F, respectively. Note the differences in shape between this panel and Figure 2.1.

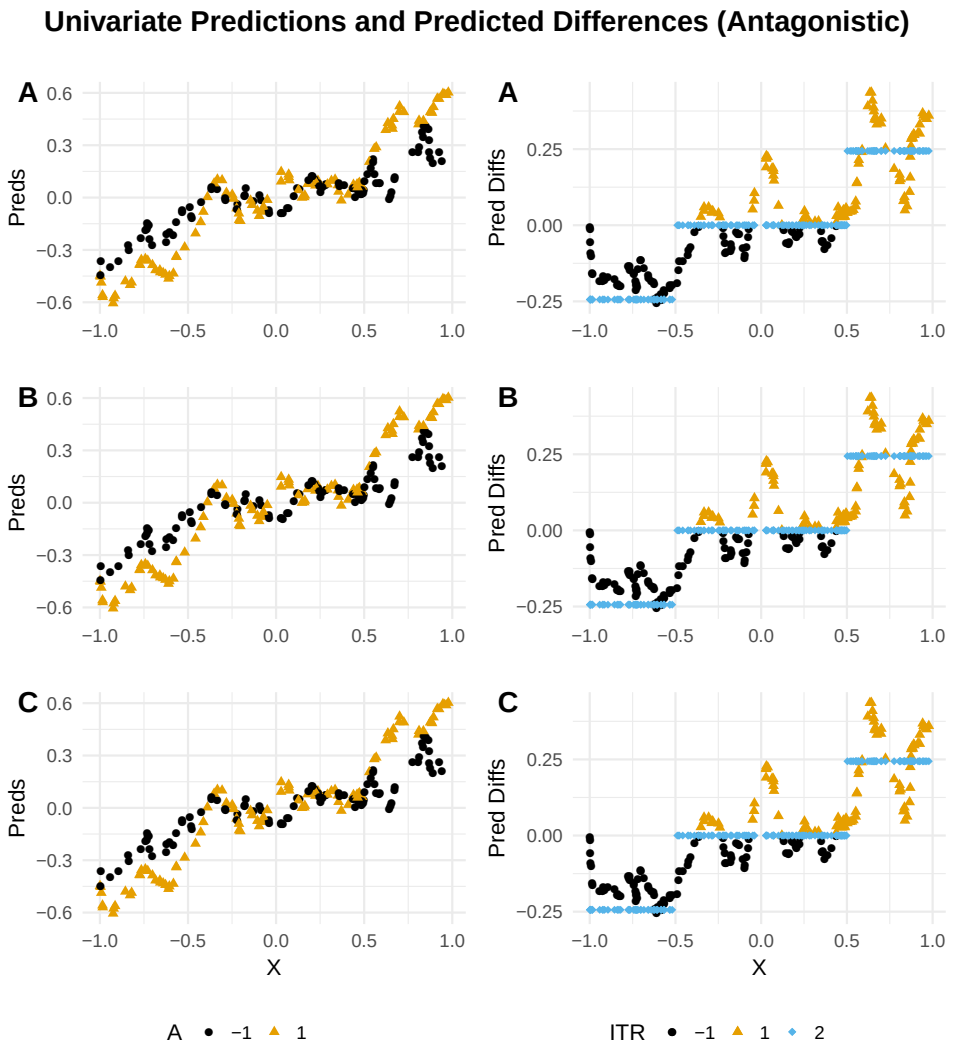

Figure 2.3: (A-C). RLT-predicted values of R_1 by true treatment effect, intervention status, and true splitting variable X for the simple numerical experiment of Section 2.6.1.1, antagonistic setting. The true treatment effect for R_1 is set to $\delta_1 = 1, 3, 10$ in A, B, and C, respectively. (D-F). Differences in predicted reward R_1 between intervention and control by true treatment effect, RLT ITR assignment, and true splitting variable X, as well as the true treatment effect for R_1, for the same numerical experiment. ITR=2 denotes the true treatment effect for R_1. The true treatment effect for R_1 and R_2 is set to $\delta_1 = 1, 3, 10$ in D, E, and F, respectively. Note the similarities in shape between this panel and Figure 2.1.

δ_1	δ_2	Measure	$S=-1$	$S=0$	$S=1$
1	1	Sens	0.962	0.000	0.917
		Spec	0.541	1.000	0.750
	3	Sens	0.981	0.000	1.000
		Spec	0.581	1.000	0.743
	10	Sens	1.000	0.000	1.000
		Spec	0.669	1.000	0.664
3	1	Sens	1.000	0.000	0.917
		Spec	0.547	1.000	0.757
	3	Sens	1.000	0.000	1.000
		Spec	0.581	1.000	0.750
	10	Sens	1.000	0.000	1.000
		Spec	0.581	1.000	0.750
10	1	Sens	1.000	0.000	0.938
		Spec	0.534	1.000	0.776
	3	Sens	1.000	0.000	1.000
		Spec	0.601	1.000	0.730
	10	Sens	1.000	0.000	1.000
		Spec	0.588	1.000	0.743

Table 2.2: Subgroup Recovery Sensitivity and Specificity by Treatment Effect (Simple Experiment, Synergistic Setting)

	$\delta_2=1$	$\delta_2=3$	$\delta_2=10$
$\delta_1=1$	0.054	0.035	0.053
$\delta_1=3$	0.040	0.027	0.015
$\delta_1=10$	0.036	0.022	0.012

Table 2.3: Treatment Effect MSE by True Treatment Effect (Simple Experiment, Synergistic Setting)

As in the simple simulation of Section 2.6.1.1, we divide the sample into a true intervention, control, and muted group. The subgroup divisions are given by

$$S = \begin{cases} -1, & a^t X_{k,\text{int}} < c_1 \\ 1, & a^t X_{k,\text{int}} > c_2 \\ 0, & \text{otherwise}, \end{cases} \tag{2.11}$$

where $X_{k,\text{int}}$ is a matrix composed of the first k columns of X and all their pairwise interactions, a is a random Gaussian vector of the corresponding length, and c_1 and c_2 are specified thresholds.

	$\delta_2 = 1$	$\delta_2 = 3$	$\delta_2 = 10$
$\delta_1 = 1$	0.035	0.049	0.057
$\delta_1 = 3$	0.027	0.040	0.069
$\delta_1 = 10$	0.030	0.043	0.066

Table 2.4: Treatment Effect MSE by True Treatment Effect (Simple Experiment, Antagonistic Setting)

In the sequel, we choose $k = 2$ and set c_1 and c_2 to the first and third quartiles of $a^T X_{k,j}$, respectively. In this experiment, we considered both the synergistic and antagonistic setting, and we considered three values for the true univariate treatment effects $\delta_j, j = 1, 2$: 1, 3, and 10. Additionally, we considered both RLT and OWL to estimate the optimal ITR. We ran 100 simulations for each combination of ITR method, δ_1, δ_2, and setting.

Table 2.5 gives the average subgroup recovery sensitivity and specificity over the 100 simulations for both RLT and OWL for the synergistic setting, divided by true subgroup status and true value of the univariate treatment effects δ_1 and δ_2. Both methods show improvements in sensitivity for the intervention and control groups as the true treatment effect sizes increase. RLT once again struggles to pick up a muted group in any setting, failing to do so entirely when effect sizes are large. Unlike in the numerical experiment of Section 2.6.1.1, RLT does succeed in discovering a muted group with high specificity when effect sizes are small, albeit with modest sensitivity. As noted before, OWL is incapable of discovering a muted group; its sensitivity and specificity remain admirably high in the groups it can recover.

Tables 2.6 and 2.7 give the average treatment effect MSE over the 100 simulations by true value of univariate treatment effects δ_1 and δ_2 for the synergistic and antagonistic settings, respectively. As in Section 2.6.1.1, MSE values are generally low in this numerical experiment, especially in the synergistic setting, suggesting that the method performs well at estimating the true treatment effect in trial-like settings even when its performance in identifying members of the muted group appears to suffer.

δ_1	δ_2	Measure	Method=RLT			Method=OWL		
			$S=-1$	$S=0$	$S=1$	$S=-1$	$S=0$	$S=1$
1	1	Sens	0.445	0.293	0.427	0.722	0.000	0.719
		Spec	0.678	0.705	0.698	0.610	1.000	0.609
	3	Sens	0.719	0.067	0.801	0.810	0.000	0.824
		Spec	0.706	0.934	0.650	0.667	1.000	0.646
	10	Sens	0.928	0.000	0.941	0.866	0.000	0.876
		Spec	0.740	1.000	0.690	0.693	1.000	0.674
3	1	Sens	0.964	0.009	0.718	0.929	0.000	0.727
		Spec	0.521	0.992	0.830	0.550	1.000	0.775
	3	Sens	0.986	0.000	0.934	0.938	0.000	0.868
		Spec	0.678	1.000	0.778	0.654	1.000	0.744
	10	Sens	0.991	0.000	0.983	0.945	0.000	0.919
		Spec	0.723	1.000	0.759	0.697	1.000	0.730
10	1	Sens	0.999	0.000	0.832	0.991	0.000	0.624
		Spec	0.568	1.000	0.843	0.436	1.000	0.868
	3	Sens	0.999	0.000	0.972	0.984	0.000	0.835
		Spec	0.671	1.000	0.809	0.596	1.000	0.809
	10	Sens	0.999	0.000	0.998	0.974	0.000	0.939
		Spec	0.715	1.000	0.778	0.702	1.000	0.750

Table 2.5: Subgroup Recovery Sensitivity and Specificity by Method and Treatment Effect (Trial-Like Conditions, Synergistic Setting)

	$\delta_2=1$	$\delta_2=3$	$\delta_2=10$
$\delta_1=1$	0.109	0.129	0.139
$\delta_1=3$	0.153	0.137	0.119
$\delta_1=10$	0.091	0.080	0.055

Table 2.6: Treatment Effect MSE by True Treatment Effect (Trial-Like Conditions, Synergistic Setting)

2.6.1.3 Discussion of Numerical Experiments

One salient take-away from the simple numerical experiment of Section 2.6.1.1, especially the disconnect between subgroup recovery sensitivity and specificity and treatment effect MSE illustrated by Figure 2.1, is that slight alterations to this method may lead to improvements in its performance as measured by subgroup recovery sensitivity and specificity. In real data, we will not know the true treatment effect—the method will, for better or worse, be judged on its efficacy in determining the correct subgroup label. In this experiment, the method reliably estimated

	$\delta_2 = 1$	$\delta_2 = 3$	$\delta_2 = 10$
$\delta_1 = 1$	0.099	0.107	0.123
$\delta_1 = 3$	0.150	0.146	0.137
$\delta_1 = 10$	0.094	0.104	0.141

Table 2.7: Treatment Effect MSE by True Treatment Effect (Trial-Like Conditions, Antagonistic Setting)

much smaller treatment effects in the range defined by the muted group than the range defined by the intervention and control groups, but failed to set these effects identically equal to zero. An "ϵ-insensitive" version of the method, in which predicted values under treatment and control must differ by at least ϵ to be placed into either the control or intervention group, might offer improved performance with the same attractive interpretability. As discussed more fully in Section 2.8, the implementation of OWL used here cannot create a muted group, but an ϵ-insensitive extension of OWL might offer the same attractive features. For either potential ϵ-insensitive method, challenges prove to arise from determining the best automated procedure for choosing ϵ.

Another notable trend observed in the numerical experiments in both Sections 2.6.1.1 and 2.6.1.2 is the high specificity for the muted group even when effect sizes were small. This fact may increase our confidence in the existence of the muted group assigned by RLT in the FLEX trial, while the satisfactory sensitivity and specificity for both the intervention and control groups in numerical experiments is promising for those groups as well.

2.7 Sensitivity Analyses

In this section, we present sensitivity analyses pertaining to two of our methodological decisions: the decision to impute missing covariates with MICE, and the decision to use RLT to compute the estimated optimal ITR. We address the former by carrying out the method on the subset of patients with complete cases in all covariates and outcomes; we address the latter by estimating the optimal ITR with OWL. As such, we ultimately explored four settings: complete cases and RLT; complete cases and OWL; imputed cases and RLT; and imputed cases and OWL.

Table 2.8 gives $\hat{V}_{\text{opt}}$ for each combination of dataset, ITR estimation method, and outcome trained upon. Each column of table 2.8 has its own natural numerical scale, due to the distribution of the underlying rewards. The combination of method and dataset that performed the best for each outcome, in terms of value, is shown in bold. 95% confidence intervals for each value estimate are given by the bootstrap, described in fuller detail in Section 2.9. For the three univariate outcomes, the different ITRs show only minor differences in value. OWL in the imputed dataset performs nominally best for HbA1c and QoL, while OWL in the complete cases performs nominally best for BMIz. Of these comparisons, only one seems to indicate a significant difference: RLT in the imputed dataset appears to perform worse than OWL in the imputed dataset for the univariate HbA1c outcome, with the former's 95% CI lying entirely above the latter's. For the composite outcome, however, RLT in the imputed dataset provides a substantial improvement in value over all other estimated ITRs, with its 95% CI lying well above all other ITR's.

Dataset	Method	HbA1c	QoL	BMIz	Composite
Complete cases	RLT	0.6786	0.6434	0.9670	2.1359
		(0.660,0.702)	(0.643,0.702)	(0.954,0.984)	(2.030,2.272)
	OWL	0.6905	0.7006	**0.9857**	2.0734
		(0.693,0.707)	(0.698,0.722)	(0.984,0.989)	(2.060,2.167)
Imputed	RLT	0.6738	0.6739	0.9737	**2.6985**
		(0.661,0.689)	(0.669,0.722)	(0.960,0.988)	(2.696,2.858)
	OWL	**0.7044**	**0.7021**	0.9855	1.9867
		(0.702,0.705)	(0.701,0.703)	(0.985,0.986)	(1.959,2.070)

Table 2.8: Estimated Value (Bootstrap 95% Confidence Interval) of ITR by Method, Dataset, and Outcome Variable

This improvement in value is the primary reasons we chose RLT in the imputed dataset as the primary ITR estimation method of interest. The others are more philosophical in nature. RLT's creation of the muted group, as described in the main text, is another argument in favor of RLT—the muted group may prove to be an important consideration for real-life decisions about targeting inteventions. In particular, the fact that the majority of adolescents are "indifferent" to the FLEX intervention with regards to HbA1c, as Table 3 in the main text demonstrates, is a reflection of the challenges in controlling glycemia within the age range studied by the trial, as

well as an indication that future work is needed to better tailor the FLEX intervention toward the specific needs of adolescents. Additionally, the arguments that typically apply to imputation—it allows us to use more data, and the assumptions required to impute are least impactful when the amount of total observations filled in by imputation is small—apply to the FLEX trial as well.

For completeness, we present Table S8, the "full" version of Table 3 in the main text with all covariates that were considered instead of only those that had significant differences between subgroups for at least one ITR. We also present Table S9, the analogue of Table S8 for the subgroups defined by OWL in the imputed dataset. For conciseness, we do not present the analogous tables for the ITRs estimated in the complete cases dataset.

Table S9 depicts the characteristics of FLEX participants in the subgroups assigned by the OWL ITR to intervention and usual care for the composite outcome and each univariate outcome. Regarding the composite outcome, 101 participants (47%) were assigned to intervention, while the remaining 115 participants (53%) were assigned to usual care. Individuals assigned to the intervention subgroup were more likely to be female (57% versus 44%; P= 0.04), more likely to be non-Hispanic White race/ethnicity (85% versus 70%; P=0.05), and less likely to have private health insurance (57% versus 82%; $P < 0.01$). Participants assigned to intervention had a longer disease duration at baseline (7.3 (3.7) years versus 5.5 (3.5) years; $P < 0.01$) and were less likely to use an insulin pump (57% versus 83.5%; $P < 0.01$). They also reported lower problem-solving abilities at baseline (SPSI score of 103.2 (13.0) versus 108.4 (12.5); $P < 0.01$).

Regarding the HbA1c univariate outcome, 118 participants (55%) were assigned to intervention and 98 participants (45%) were assigned to Usual Care. Individuals assigned to intervention were more likely to be non-Hispanic white race/ethnicity (91% versus 60%; $P < 0.01$), less likely to use an insulin pump (64% versus 80%; $P < 0.01$) and experienced more clinically serious hypoglycemia at baseline. They also reported higher motivation at baseline ($P = 0.03$) and lower diabetes-related family conflict ($P = 0.04$).

Regarding the QoL univariate outcome, 123 participants (57%) were assigned to intervention and 93 participants (43%) were assigned to usual care. Individuals assigned to inter-

vention showed higher glycemic variability (coefficient of variation of 41.5% (8.0%) versus 37.9% (7.4%) and experienced more clinical and clinically serious hypoglycemia at baseline (all $P < 0.01$).

Regarding the BMIz univariate outcome, 116 participants (54%) were assigned to intervention and 100 participants (46%) were assigned to usual care. Individuals assigned to intervention showed a longer disease duration at baseline (7.8 (3.8) months versus 4.7 (3.0) months; $P < 0.01$). They also showed a higher frequency of clinical hypoglycemia ($P < 0.01$), higher diabetes adherence (DSMP score of 57.0 (12.0) versus 53.9 (10.7); $P = 0.05$), and lower fear of hypoglycemia as measured by the helplessness/worry ($P = 0.01$) and worry about negative social consequences ($P = 0.02$) subscales.

Compared to RLT (depicted in Table S8), the OWL ITR (depicted in Table S9) assigned a larger proportion of the FLEX sample to the intervention group, ranging from 47% for the composite outcome to 57% for quality of life. Characteristics that were significantly different across RLT-assigned subgroups were not consistent with the characteristics that were significantly different across the OWL-assigned subgroups. In some cases, the OWL-assigned subgroups showed significant differences in additional characteristics. For example, individuals assigned to intervention for the composite outcome were less likely to have public health insurance in the subgroups defined by both RLT and OWL, but those in the OWL-defined group showed additional differences in sex, race/ethnicity, disease duration, and insulin pump use versus the participants that OWL assigned to receive usual care. In other cases, differences across the RLT-assigned subgroups were not replicated across the OWL-assigned subgroups. One compelling explanation for this fact is the ability of RLT to form a muted group, in contrast to OWL, which results in patients from the OWL-assigned intervention and usual care groups being reassigned to the muted group. For example, the individuals muted by RLT for the HbA1c univariate outcome had significantly higher HbA1c at baseline compared to the intervention and usual care groups. By contrast, the OWL-assigned subgroups showed no differences in HbA1c, although both groups showed consistent differences in hypoglycemia at baseline.

2.8 Outcome weighted learning (OWL)

Outcome weighted learning (OWL) is a precision medicine method for estimating the optimal ITR in a sample. Unlike RLT, which poses a model between X and R and inverts it to find the ITR, OWL considers a class of functions Π to which all estimated ITRs can belong, then directly estimates the optimal ITR $\hat{\pi}_{\text{opt}}$ by minimizing a loss function applied to this class (Zhao et al., 2012). The challenge in implementing OWL comes from specifying a class Π that is robust enough to allow sufficiently close estimation of the true optimal ITR π^*_{opt} but also computationally tractable. For details on the specific implementation of OWL we employed, Residual Weighted Learning with a linear kernel, see Zhou et al. (Zhou et al., 2017).

One key difference between RLT and OWL is that OWL does not create a muted group. Residual weighted learning ultimately relies on support vector machines, which are strict classifiers—all observations are assigned to either treatment or control.

This implies that OWL and RLT may be suited to different scenarios. In scenarios where the true muted group is small, OWL's focus on direct estimation of the ITR may lead to improvements in overall value. But in scenarios where the true muted group is large, OWL will classify them into either intervention or control regardless of the estimated treatment effect. This may impact value estimates, but more importantly misrepresents the efficacy of both intervention and control. And naturally, any attempt to characterize the ITR-based treatment or control group through their covariates will be highly influenced by having additional group members under OWL, especially group members whose differential treatment response suggests they may increase the heterogeneity of the group.

2.9 Imputation Bootstrapping Procedure

In this section we describe the bootstrapping procedure we employed to estimate the variability of an ITR which was estimated from multiple imputed datasets. We begin by restating, and introducing notation for, our multiple imputation procedure.

Let $k = 1, \ldots, K$ index the K datasets imputed by MICE, $X_{\text{imp},k}$. As noted in the main text, we used $K = 11$ to preclude the possibility of ties on a majority vote for OWL, but we leave K general here. As before, let $i = 1, \ldots, n$ index patients and $j = 1, \ldots, p$ index covariates, such that each $X_{\text{imp},k}$ is an $n \times p$ matrix. Let $\hat{\pi}_{\text{opt},k}$ denote the estimated optimal ITR estimated from the data $X_{\text{imp},k}$, and let $\hat{V}_{\text{opt},k}$ denote its estimated value. Let the estimated optimal ITR for the whole sample be denoted $\hat{\pi}_{\text{opt}}$, with estimated value $\hat{V}_{\text{opt}}$. Our multiple imputation procedure follows the steps outlined in Algorithm 1.

Algorithm 1 Multiple imputation procedure for $\hat{\pi}_{\text{opt}}$

1. Generate $X_{\text{imp},k}, k = 1, \ldots, K$, via MICE
2. For $k = 1, \ldots, K$, compute $\hat{\pi}_{\text{opt},k}$ via the method of interest
3. Set $\hat{\pi}_{\text{opt}}$ to the plurality vote of the $\hat{\pi}_{\text{opt},k}$
4. Obtain the estimated value $\hat{V} = \left(\sum_i R_i I\left\{A_i = \hat{\pi}_{\text{opt}}(X_i)\right\}\right) / \left(\sum_i I\left\{A_i = \hat{\pi}_{\text{opt}}(X_i)\right\}\right)$

We can use Algorithm 1 to obtain point estimates of V_{opt}. However, due to multiple imputation, large-sample theory no longer provides a simple form for an estimate of the variability of V_{opt}. Although the bootstrap is not guaranteed to be valid in all precision medicine settings, under reasonable regularity conditions for the estimated ITR, bootstrap validity will hold.

The first and second steps of Algorithm 1 are, unsurprisingly, its most computationally intensive. To avoid replicating as many computationally intensive steps as possible when bootstrapping, we adopt the bootstrapping procedure described in Algorithm 2.

Strictly speaking, the bootstrapping procedure in Algorithm 2 does not capture all of the variability that arises from MICE, in the sense that we do not impute via MICE after generating each bootstrap sample. As the proportion of missing data in any one column in this sample was quite low, however, we did not believe capturing this particular source of variance was crucial given the computational costs it would incur.

Bootstrap procedures to estimate the confidence interval of a single ITR's value, or the difference in estimated value between two ITRs, proceed in a manner analogous to Algorithm 2, following the usual bootstrapping recommendations that B should be higher to estimate a con-

Algorithm 2 Bootstrapping to estimate the variability of V_{opt}

1. Sample the indices $i = 1, \ldots, n$ B times with replacement to generate the bootstrapping index vectors, $I_b, b = 1, \ldots, B$
2. For $b = 1, \ldots, B$:
   (a) For $k = 1, \ldots, K$:
      i. Generate the bth bootstrap sample of the kth imputed dataset $\tilde{X}^b_{\text{imp},k}$, by stacking the rows of $X_{\text{imp},k}$ corresponding to I_b
      ii. Compute $\hat{\pi}^b_{\text{opt},k}$ via the method of interest
   (b) Set $\hat{\pi}^b_{\text{opt}}$ to the plurality vote of the $\hat{\pi}^b_{\text{opt},k}$
   (c) Compute $\hat{V}^b_{\text{opt}} = \left(\sum_i R_i I\left\{A_i = \hat{\pi}^b_{\text{opt}}(X_i)\right\}\right) / \left(\sum_i I\left\{A_i = \hat{\pi}^b_{\text{opt}}(X_i)\right\}\right)$
3. Compute $\widehat{SE}(\hat{V}_{\text{opt}}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{V}^b_{\text{opt}} - \bar{V}_{\text{opt}}\right)$

fidence interval than a variance. As the procedure is entirely analogous, we do not explicitly outline it here. For the sensitivity analyses presented in Section 2.7, we calculated 95% CIs based on $B = 1000$ bootstrap replicates.

CHAPTER 3: VOLATILITY LEARNING IN DYNAMICAL SYSTEMS

3.1 Introduction

Ordinary differential equations (ODEs) offer an attractive avenue for modeling random processes that vary in continuous-time. They have been employed in a variety of health science research settings, including metabolic modeling in type 1 diabetes (Lehmann and Deutsch, 1992), multi-stage modeling in cancer (Spencer et al., 2004), and recovery of gene regulatory networks in several diseases (Wu et al., 2014b; Song et al., 2018). We consider the setting where the dynamics of one process are of particular interest, but several other history or covariate processes vary along with the process of interest. We can represent such a setting as

$$dX(t;\theta) = f(Z(t;\theta),\theta)dt; \qquad t \in [0,1], \tag{3.1}$$

where the process of interest $X(t;\theta)$ and the history processes $H_1(t;\theta),\ldots,H_{p-1}(t;\theta)$ are collected in $Z(t;\theta) \equiv (X(t), H_1(t),\ldots,H_{p-1}(t))^T$, and the functional form of f may be known or unknown. The index t is over time, and is condensed to $[0,1]$ without loss of generality. Typically, an ODE of the form given in (3.1) will include a scalar initial condition $X(0;\theta) = C$. Together with an initial condition, (3.1) describes the mean motion of the process of interest in a general way.

In some applications, however, merely describing the mean motion of the process is not sufficient. In type 1 diabetes, for instance, the variability of serum glucose may serve as a better predictor of long-term complications (Monnier et al., 2008) and certain comorbidities (Saisho, 2014) than exposure to sustained hyperglycemia; in cancer epigenomics, the volatility of methyla-

tion may be linked to informative changes in the genome (Wagner et al., 2014). In such settings, an attractive approach is to model the dynamical system with a Stochastic Differential Equation (SDE), taking the form

$$dX(t;\theta) = \mu(Z(t;\theta),\theta)dt + \sigma(Z(t;\theta),\theta)dW(t); \qquad t \in [0,1], \tag{3.2}$$

where $W(t)$ is standard Brownian motion and the functional forms μ and σ may be known or unknown. The form (3.2) contains a *drift* term governed by μ that describes the motion of the drift of $X(t;\theta)$ and a *diffusion* term governed by σ that describes the volatility of $X(t;\theta)$. In practice, we will assume the processes are observed at $n+1$ discrete time points $0 = t_0 < t_1 < \cdots < t_n = 1$, and that the observations Y_{ji} are subject to measurement error exogenous to the process, such that

$$Y_{ji} = Z_j(t_i;\theta^*) + \epsilon_i, \qquad j = 1,\ldots,p,\ i = 1,\ldots,n, \tag{3.3}$$

where θ^* is a true parameter vector and the ϵ represent independent measurement errors. For ease of expression, we will frequently suppress the dependence of $Z(t;\theta)$ on θ.

If the number of covariate processes, p, is high, it is frequently of scientific and practical interest to discover which of the processes in $Z(t)$ inform the dynamics of $X(t)$ in some sense that is significant. The structure in (3.2) provides a convenient way to express this concept: if μ is a function of Z_j, then we say Z_j *regulates* the drift of X, while if σ is a function of Z_j, we say Z_j regulates the volatility of X. Note that these conditions are neither disjoint nor necessarily co-occurring. Several well-studied statistical tools are suited to the task of inducing sparsity, and thus estimating regulators. Examples include the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) and the group LASSO (Yuan and Lin, 2006), which have been applied and adapted to a variety of research settings. Recent work provides substantive insight into the questions of parameter estimation and model selection in the context of high-dimensional ODEs (Lu et al., 2011; Henderson and Michailidis, 2014; Wu et al., 2014a; Chen

et al., 2017), including several that do not make strong assumptions on the form of f in (3.1). Chen et al. (2017) in particular provides theoretical justifications for a model selection scheme in the context of ODEs. These tasks remain relatively unexplored in the context of SDEs, however, especially with similarly weak assumptions on functional forms for μ and σ.

In this paper, we propose a nonparametric method for estimating the unknown functions μ and σ and recovering the true regulators of both the drift and volatility of X in (3.2). The remainder of the paper is laid out as follows. In Section 3.2, we propose our procedure. In Section 3.3, we explore its theoretical properties. In Section 3.4, we study its performance in finite samples by means of numerical experiments. In Section 3.5, we apply it to a dataset arising from a clinical study of youth with type 1 diabetes. We conclude with a discussion and directions for future research in this vein in Section 3.6. Details of proofs are provided in the supplementary material.

3.2 Methods

In this section, we propose a method for parameter estimation and model selection at both the drift and volatility level for a process of interest characterized by an ODE.

3.2.1 Notation

Let t be the time index of the dynamical system. Without loss of generality, assume we observe the system at $n+1$ times $0 = t_0 < t_1 < \cdots <, t_n = 1$. As before, we let $X \in \mathcal{X}(\cdot)$ denote the process of interest, $H_1, H_2, \ldots, H_{p-1} \in \mathcal{X}(\cdot)$ denote the covariate or history processes, and $Z(t) = (X(t), H_1(t), \ldots, H_{p-1}(t))^T$ denote the entire p-vector of the process at time t.

Let Y_{ij} denote the observation of the jth process at the ith time point, where the numbering of j is the same as that of Z. Let $\mathcal{X}(h)$ denote a nonparametric class of functions defined on $[0, 1]$ and indexed by smoothing parameters h, and let $\mathcal{X}^p(h)$ denote the pth Cartesian product of $\mathcal{X}(h)$. Let $Q(\cdot)$ refer to an arbitrary function from $\mathcal{X}(\cdot)$. We will refer to the ℓ_2-norm of a vector or matrix as $\|\cdot\|_2$ and the ℓ_2-norm of a function f on the interval $[0, 1]$ as $\|f\|_{2,[0,1]}$, i.e. $\|f\|^2_{2,[0,1]} \equiv \int_0^1 f^2(t)dt$. The minimum and maximum eigenvalues of a square matrix A will be

denoted $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$, respectively. True values of processes and parameters will be identified by an asterisk: for instance, α^* denotes the true value of α.

3.2.2 Proposed Model

As in Chen et al. (2017), we allow components of the ODE to be high-dimensional, although we assume the main process of interest X is univariate. We assume that the ODE for the process of interest describes the behavior of the process drift and volatility together. That is, we assume its derivative takes the form

$$dX(t) = \mu\left(Z(t)\right) dt + \sigma\left(Z(t)\right) dW(t), \tag{3.4}$$

where $\mu : \mathcal{X}^p(\cdot) \to \mathbb{R}$ and $\sigma : \mathcal{X}^p(\cdot) \to \mathbb{R}$ are unknown functional forms, and $W(\cdot)$ is standard Brownian motion.

We assume that the true form of μ and σ is additive. That is, there exist functions $f_1, f_2, \ldots, f_p$ and $g_1, g_2, \ldots, g_p$ such that (3.4) can be written

$$dX(t) = \left[\sum_{j=1}^{p} f_j\left(Z_j(t)\right)\right] dt + \left[\sum_{j=1}^{p} g_j\left(Z_j(t)\right)\right] dW(t). \tag{3.5}$$

If the true function f_j^* in (3.5) is nonzero, we refer to Z_j^* as a true regulator of the drift of X^*. Similarly, if the true function g_j^* is nonzero, we refer to Z_j^* as a true regulator of the volatility of X^*. Let $S_\mu \equiv \{j : \|f_j^*\|_2 \neq 0, j = 1, \ldots, p\}$ denote the set of true drift regulators of X^*, and similarly let $S_\sigma \equiv \{j : \|g_j^*\|_2 \neq 0, j = 1, \ldots, p\}$ denote the set of true volatility regulators of X^*.

We will use finite and known bases of functions, $\phi(\cdot) \equiv (\phi_1(\cdot), \phi_2(\cdot), \ldots, \phi_{M_1}(\cdot))^T$ and $\psi(\cdot) \equiv (\psi_1(\cdot), \psi_2(\cdot), \ldots, \psi_{M_2}(\cdot))^T$, to approximate f and g, respectively. Note that M_1 and M_2 are not necessarily equal, but ϕ and ψ in general will be of the same form (e.g. cubic spline or

Fourier functions). We assume that the basis ϕ approximates f by

$$f_j(x) = \phi(x)^T \alpha_j + \delta_{\mu,j}(x), \alpha_j \in \mathbb{R}^{M_1} \tag{3.6}$$
$$g_j(x) = \psi(x)^T \beta_j + \delta_{\sigma,j}(x), \beta_j \in \mathbb{R}^{M_2},$$

where $\delta_{\mu,j}$ and $\delta_{\sigma,j}$ are residuals. Using these basis functions, we can express the ODE for the process of interest in the form

$$\begin{aligned} dX(t) = &\left[\alpha_0 + \sum_{j=1}^{p} \phi(Z_j(t))^T \alpha_j + \sum_{j=1}^{p} \delta_{\mu,j}(Z_j(t))\right] dt \\ &+ \left[\beta_0 + \sum_{j=1}^{p} \psi(Z_j(t))^T \beta_j + \sum_{j=1}^{p} \delta_{\sigma,j}(Z_j(t))\right] dW(t). \end{aligned} \tag{3.7}$$

3.2.3 Estimating the Process Mean

To estimate α, as in (Chen et al., 2017), we proceed by integrating both sides of (3.7) from 0 to t, which yields

$$\begin{aligned} X(t) - X(0) = &\left[\alpha_0 t + \sum_{j=1}^{p} \Phi_j(t)^T \alpha_j + \sum_{j=1}^{p} \int_0^t \delta_{\mu,j}(Z_j(s)) ds\right] \\ &+ \left[\beta_0 (W(t) - W(0)) + \sum_{j=1}^{p} \left\{\int_0^t \psi(Z_j(s)) dW(s)\right\}^T \beta_j + \sum_{j=1}^{p} \int_0^t \delta_{\sigma,j}(Z_j(s)) dW(s)\right], \end{aligned} \tag{3.8}$$

where the integrated basis $\Phi(\cdot)$ is defined as

$$\Phi_j(t) = (\Phi_{j1}(t), \Phi_{j2}(t), \ldots, \Phi_{jM_1}(t))^T = \int_0^t \phi(Z_j(s)) ds, \qquad j = 1, 2, \ldots, p, \tag{3.9}$$

and $\Phi_0(t) = t$.

The proposed method solves for α by taking the expectation of both sides of (3.8), yielding a form similar to that in (Chen et al., 2017):

$$\begin{aligned}\hat{\alpha} = & \underset{C_0\in\mathbb{R},\alpha_0\in\mathbb{R},\alpha_1,\ldots,\alpha_p\in\mathbb{R}^{M_1}}{\arg\min} \frac{1}{2n} \\ & \times \sum_{i=1}^{n}\left[Y_{i1} - C_0 - \alpha_0\widehat{\Phi}_0(t_i) - \sum_{j=1}^{p}\alpha_j^T\widehat{\Phi}_j(t_i)\right]^2 \\ & + \chi_n\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{p}\left\{\alpha_j^T\widehat{\Phi}_j(t_i)\right\}^2\right]^{1/2},\end{aligned} \tag{3.10}$$

where

$$\hat{Z}(\cdot;h) = \underset{Q(\cdot)\in\mathcal{X}(h)}{\arg\min}\sum_{i=1}^{n}\|Y_i - Q(t_i)\|_2^2, \tag{3.11}$$

and

$$\hat{\Phi}_0(t) = t,\ \hat{\Phi}_j(t) = \int_0^t \phi(\hat{Z}_j(s;h))ds, j = 1,\ldots,p. \tag{3.12}$$

In (3.10), χ_n is a nonnegative and group sparsity-inducing tuning parameter. We will estimate S_μ with the estimated set of drift regulators, $\hat{S}_\mu \equiv \{j : \|\hat{\alpha}_j\|_2 \neq 0, j = 1,\ldots,p\}$.

3.2.4 Estimating the Process Volatility

We first note that rearranging, integrating over the interval $(0, t_i)$, and squaring both sides of (3.4) gives us an estimator of the volatility of the process of interest in that interval, defined as

$$V(t_i) \equiv \left[X(t_i) - X(0) - \int_0^{t_i}\mu\left(Z(s)\right)ds\right]^2 = \left[\int_0^{t_i}\sigma\left(Z(s)\right)dW(s)\right]^2. \tag{3.13}$$

We will sometimes use the shorthand V_i to denote $V(t_i)$. Once we obtain $\hat{\alpha}$, we can estimate $\mu(\cdot)$ with

$$\hat{\mu}\left(Z(t)\right) \equiv \hat{\alpha}_0 + \sum_{j=1}^{p}\phi(\hat{Z}_j.(t))^T\hat{\alpha}_j, \tag{3.14}$$

which in turn allows us to estimate $V(t_i)$ with

$$\hat{V}(t_i) \equiv \left[\hat{X}(t_i) - \hat{X}(0) - \int_0^{t_i} \hat{\mu}\left(Z(s)\right) ds\right]^2. \tag{3.15}$$

We proceed by taking the expectation of both sides of (3.13). After some simplification, this yields

$$\mathbb{E}\left[V(t_i)\right] = \beta^T \left(\int_0^{t_i} \left[2(t_i - s)\tilde{U}(s)\right] ds\right) \beta = \beta^T U_i \beta, \tag{3.16}$$

where $\beta = (\beta_0, \beta_1^T, \beta_2^T, \ldots, \beta_p^T)^T$, $\tilde{U}(s)$ is an $M_2 p + 1 \times M_2 p + 1$ symmetric matrix with

$$\tilde{U}(s) \equiv \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & \Psi_{11}(s) & \Psi_{12}(s) & \ldots & \Psi_{1p}(s) \\ 0 & \Psi_{21}(s) & \Psi_{22}(s) & \ldots & \Psi_{2p}(s) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \Psi_{p1}(s) & \Psi_{p2}(s) & \ldots & \Psi_{pp}(s) \end{pmatrix}, \tag{3.17}$$

where

$$\Psi_{jk}(s) = \psi(Z_j(s))\psi(Z_k(s))^T, j = 1, \ldots, p, k = 1, \ldots, p, \tag{3.18}$$

and

$$U_i \equiv \int_0^{t_i} \left[2(t_i - s)\tilde{U}(s)\right] ds. \tag{3.19}$$

For each (j, k), we will estimate $\Psi_{jk}(s)$ with

$$\hat{\Psi}_{jk}(s) = \psi\left(\hat{Z}_j(s)\right) \psi\left(\hat{Z}_k(s)\right)^T, \tag{3.20}$$

and the estimators of $\tilde{U}(s)$ and U_i, $\hat{\tilde{U}}(s)$ and $\hat{U}_i$ respectively, follow analogously.

One approach is to directly find an estimator $\tilde{\beta}$ that satisfies

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(V_i - \beta^T U_i \beta\right)^2 + \lambda_n J(\beta), \tag{3.21}$$

where $J(\beta)$ is a penalty term added in accordance with bias-variance tradeoff logic to control the complexity of $\tilde{\beta}$. It is quite challenging to find appropriate solutions from (3.21), however, especially ones that enforce group sparsity in a satisfactory manner through $J(\beta)$. Consider beginning with an estimator $\tilde{\beta}$ from (3.21), then enacting a one-step update Δ to reach a new estimator $\hat{\beta} \equiv \tilde{\beta} + \Delta$. First note that for any $M_2p+1 \times M_2p+1$ positive definite symmetric matrix A,

$$\begin{aligned} \hat{\beta}^T A\hat{\beta} - \tilde{\beta}^T A\tilde{\beta} &= (\tilde{\beta}+\Delta)^T A(\tilde{\beta}+\Delta) - \tilde{\beta}^T A\tilde{\beta} \\ &= 2\tilde{\beta}^T A\tilde{\beta} + 2\tilde{\beta}^T A\Delta + \Delta^T A\Delta - 2\tilde{\beta}^T A\tilde{\beta} \\ &= 2\tilde{\beta}^T A\hat{\beta} - 2\tilde{\beta}^T A\tilde{\beta} + \Delta^T A\Delta. \end{aligned} \tag{3.22}$$

Then we can say

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(V_i - \hat{\beta}^T\hat{U}_i\hat{\beta})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(V_i - \tilde{\beta}^T\hat{U}_i\tilde{\beta} - (\hat{\beta}^T\hat{U}_i\hat{\beta} - \tilde{\beta}^T\hat{U}_i\tilde{\beta})\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - 2\tilde{\beta}^T U_i\hat{\beta} + \Delta^T\hat{U}_i\Delta\right)^2, \end{aligned} \tag{3.23}$$

where the last equality follows from (3.22). As we will discuss in further detail in Section 3.3, provided $\tilde{\beta}$ is consistent, we may ignore the final term inside the parentheses in (3.23), giving us the objective function

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{M_2p+1}}{\arg\min}\left(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - 2\tilde{\beta}^T U_i\beta\right)^2 + \lambda_n\left[\frac{1}{n}\sum_{i=1}^{n}\left(\beta^T\hat{U}_i\beta\right)^2\right]^{1/2}. \tag{3.24}$$

Thus we arrive at Algorithm 3.

In practice, the group LASSO of Yuan and Lin (2006) assists in inducing group sparsity as desired in Algorithm 3. The groups are defined by the groups of basis functions in the same way as the rows of $\tilde{U}$ in (3.17), i.e., a single group of length 1 followed by p groups of length M_2.

Algorithm 3 Estimation of locally linear $\hat{\beta}$

1. Find a consistent estimator $\tilde{\beta}'$ from (3.21).
2. Rescale to $\tilde{\beta} = \tilde{\gamma}^{1/2}\tilde{\beta}$, where $\tilde{\gamma} = \arg\min_{\gamma} \sum_{i=1}^{n} \left(V_i - \gamma \tilde{\beta}'^T \hat{U}_i \tilde{\beta}' \right)$.
3. While $\epsilon > \epsilon_{\max}$, where $\epsilon_{\max}$ is a pre-specified threshold:
   (a) Find $\hat{\beta}'$ using (3.24).
   (b) Rescale to $\hat{\beta} = \hat{\gamma}^{1/2}\hat{\beta}'$, where $\hat{\gamma} = \arg\min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \left(\hat{V}_i - \gamma \hat{\beta}'^T \hat{U}_i \hat{\beta}' \right)^2$.
   (c) Compute $\epsilon = \|\tilde{\beta} - \hat{\beta}\| / \|\hat{\beta}\|$.
   (d) Set $\tilde{\beta} = \hat{\beta}$.
4. Return $\hat{\beta}$.

Let $\hat{\beta}$ denote the final estimate provided by Algorithm 3 after convergence, and let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1^T, \hat{\beta}_2^T, \ldots, \hat{\beta}_p^T)^T$ as before. We can estimate S_σ with the estimated set of volatility regulators implied by $\hat{\beta}$, $\hat{S}_\sigma \equiv \{j : \|\hat{\beta}_j\|_2 \neq 0, j = 1, 2, \ldots, p\}$.

3.3 Theoretical Properties

In this section, we establish consistency for the proposed method's variable selection. In this section, we provide the main assumptions and theoretical results required to do so; detailed proofs are deferred to section A.1 of the supplementary material. In this section, we will let $|S|$ denote the cardinality of set S. For ease of notation, we will let $S^0 = 0 \cup S$.

A crucial first step in establishing variable selection consistency is bounding the error introduced by using the smoothed estimates $\hat{Z}(\cdot; h)$ instead of the true trajectories $Z^*(\cdot)$. The smoothed estimates are obtained from local polynomial regression.

For ease of presentation, we have assumed the measurement errors in (3.3) are normally distributed; as explored in Tsybakov (2009), generalizations to bounded or sub-Gaussian errors are possible, though not explored further in this paper.

Assumption 3.1. *The measurement errors in* (3.3) *in are independent, and* $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n, j = 1, \ldots, p$.

We assume that the true trajectories $Z_j^*, j = 1, \ldots, p$, are smooth. Note that the form of our model in (3.4), in particular the Brownian motion term, bounds the amount of smoothness we can assume for X^*—namely, the ℓth derivative of X_j^* will not exist for any $\ell \geq 1/2$. For the sake of simplicity, we have extended this limitation to the rest of the Z_j^*.

Assumption 3.2. *Assume that the solutions* $Z_j^*, j = 1, \ldots, p$, *belong to the Hölder class* $\Sigma(\tau_1, L_1)$, *where* $0 < \tau_1 < \frac{1}{2}$. *That is, for* $\ell \in (0, \tau_1)$,

$$|Z_j^{*(\ell)}(t) - Z_j^{*(\ell)}(t')| \leq L_1 |t - t'|^{\tau_1 - \ell}, \qquad \forall t, t' \in [0, 1], j = 1, \ldots, p.$$

Using these assumptions, we can obtain a concentration inequality for $\|\hat{Z}_j - Z_j^*\|$.

Theorem 3.1. *Suppose that Assumptions 1-2 are satisfied. Let $\hat{W}_j$ be the local polynomial regression estimator of order $\ell \geq 3$ with bandwidth*

$$h_n \propto n^{(\upsilon-1)/(2\tau_1+1)} \tag{3.25}$$

for some positive $\upsilon < 1$. Then, for each $j = 1, \ldots, p$,

$$\|\hat{Z}_j - Z_j^*\|^2 \leq C_2 n^{\frac{2\tau_1}{2\tau_1+1}(\upsilon-1)} \tag{3.26}$$

holds with probability at least $1 - 2\exp\{-n^\upsilon/(2C_3\sigma^2)\}$ for some constants C_2 and C_3.

The concentration inequality in Theorem 3.1 relies on the concentration bounds for Gaussian errors established in Boucheron et al. (2013). Note that our rate is slower than that established in Chen et al. (2017). This slower rate is the cost of including Brownian motion in (3.4), as discussed above, and thereby indirectly the cost of being able to model the volatility of X.

Note that, as the bound in (3.26) holds uniformly for $j = 1, \ldots, p$ with probability at least $1 - 2p\exp\{-n^\upsilon/(2C_3\sigma^2)\}$, the bound will hold uniformly for $j = 1, \ldots, p$ with probability converging to 1 if $p = o(\exp\{n^\upsilon/(2C_3\sigma^2)\}$.

In Section 3.2.4, we argued that the final term in the final equality of (3.23) could be ignored. We state our reasoning more clearly here. We must make one assumption about the estimator $\tilde{\beta}$.

Condition 3.1. *The estimator $\tilde{\beta}$ is almost surely consistent. That is,$\tilde{\beta} \underset{as}{\to} \beta^*$.*

Note that the almost sure convergence in Condition 3.1 is stronger than we need for our present argument; however, it will assist with further theoretical results later in this section. We also briefly define $o_{as}(\cdot)$ notation for completeness:

Definition 3.1. We say $X_n = o_{as}(Y_n)$ if there exists a set A such that $P(A) = 1$ and $\forall \omega \in A$, $\frac{X_n(\omega)}{Y_n(\omega)} \to 0$.

Define the optimization problem in (3.23) with

$$M_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - 2\tilde{\beta}^T U_i\hat{\beta} + \Delta^T\hat{U}_i\Delta\right)^2. \tag{3.27}$$

Then we have

$$M_n(\hat{\beta}) = \tilde{M}_n(\hat{\beta}) + \frac{2}{n}\sum_{i=1}^{n}\left[(V_i - \tilde{\beta}^T\hat{U}_i\tilde{\beta})\Delta^T\hat{U}_i\Delta\right] + \frac{1}{n}\sum_{i=1}^{n}\left(2\tilde{\beta}^T\hat{U}_i\Delta + \Delta^T\hat{U}_i\Delta\right)^2, \tag{3.28}$$

where

$$\tilde{M}_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - 2\tilde{\beta}^T U_i\hat{\beta}\right)^2. \tag{3.29}$$

The rightmost term in (3.28) is clearly $o_{as}(\|\Delta\|^2)$. Under Condition 3.1, the middle term will be $o_{as}(\|\Delta\|^2)$ provided that $\mathbb{E}[\|U\|]$ is bounded. Hence we have

$$M_n(\hat{\beta}) = \tilde{M}_n(\hat{\beta}) + o_{as}(\|\Delta\|^2),$$

and thus $\tilde{M}_n(\beta)$, the left-hand term in (3.24), locally approximates $M_n(\beta)$.

The proof of variable selection consistency for $\hat{\beta}$ requires four additional assumptions about the true process $X^*(t)$, given here, along with one technical assumption and four technical conditions, which are presented in the supplementary materials.

Assumption 3.3. *The expectation of the derivative of X^* is an additive function of $Z_j^*, j = 1, \ldots, p$. In other words,*

$$\mathbb{E}\left[dX^*(t)\right] = \left[\alpha_0^* + \sum_{j=1}^{p} f_j^*(Z_j^*(t))\right]dt, \qquad \alpha_0^* \in \mathbb{R},$$

where $\int_0^1 f_j^*(Z_j^*(t))dt = 0$ for all j. Additionally, the functions $f_j^*, j = 1, \ldots, p$, belong to a Sobolev class $W(\tau_2, L_2)$ on a finite interval with $\tau_2 \geq 3$. Furthermore, for any $j = 2, \ldots, p$,

assume that the basis function residuals $\delta_{\mu,j}$ satisfy

$$\int_0^1 \delta_{\mu,j}^2(s)ds = \int_0^1 \left[f_j^*(Z_j^*(s)) - \phi^T(Z_j^*(s))\alpha_j^*\right]^2 ds \leq Q_\mu(M_1+1)^{-2\tau_2},$$

where Q_μ is a global constant.

We further assume that the derivative of $X^*(t)$ minus its drift is an additive function of $Z_j^*, j = 1, \ldots, p$. In other words,

$$dX^*(t) - \mu^*(X^*(t)dt) = \left[\beta_0^* + \sum_{j=1}^p g_j^*(Z_j^*(t))\right] dW(t), \qquad \beta_0^* \in \mathbb{R}, \tag{3.30}$$

where $\int_0^1 g_j^*(Z_j^*(t))dt = 0$ for all j. Additionally, the functions $g_j^*, j = 1, \ldots, p$, belong to a Sobolev class $W(\tau_3, L_3)$ on a finite interval with $\tau_3 \geq 3$. Furthermore, for any $j = 2, \ldots, p$, assume that the basis function residuals $\delta_{\sigma,j}$ satisfy

$$\int_0^1 \delta_{\sigma,j}^2(s)ds = \int_0^1 \left[g_j^*(Z_j^*(s)) - \psi^T(Z_j^*(s))\beta_j^*\right]^2 ds \leq Q_\sigma(M_2+1)^{-2\tau_3}, \tag{3.31}$$

where Q_σ is a global constant.

Assumption 3.4. *The eigenvalues of $\int_0^1 \Phi_{S_\mu^0}\Phi_{S_\mu^0}^T dt$ are bounded above by $C_{\max}$ and bounded below by a positive number $C_{\min}$, and for all $k \notin S_\mu^0$, the eigenvalues of $\int_0^1 \Phi_k\Phi_k^T dt$ are bounded below by $C_{\min}$. In other words,*

$$0 < C_{\min} \leq \Lambda_{\min}\left(\int_0^1 \Phi_{S_\mu^0}\Phi_{S_\mu^0}^T dt\right) \leq \Lambda_{\max}\left(\int_0^1 \Phi_{S_\mu^0}\Phi_{S_\mu^0}^T dt\right) \leq C_{\max},$$

and

$$C_{\min} \leq \Lambda_{\min}\left(\int_0^1 \Phi_k\Phi_k^T dt\right) \forall k \notin S_\mu^0.$$

Additionally, assume that the eigenvalues of $\int_0^1 \left(\Psi_{S_\sigma^0}\right)_{ijk}\left(\Psi_{S_\sigma^0}\right)_{ilm} dt$ are bounded above by $D_{\max}$ and bounded below by a positive number $D_{\min}$, and for all $k \notin S_\sigma^0$, the eigenvalues of

$\int_0^1 (\Psi_k)_{ijk} (\Psi_k)_{ilm} \, dt$ *are bounded below by* $D_{\min}$*, where* $(\cdot)_{ijk}(\cdot)_{ilm}$ *denotes a tensor product with Einstein notation.*

Assumption 3.5. *Assume that*

$$\max_{j \notin S_\mu^0} \left\| \left(\int_0^1 \Phi_j \Phi_{S_\mu^0}^T dt \right) \left(\int_0^1 \Phi_{S_\mu^0} \Phi_{S_\mu^0}^T dt \right)^{-1} \right\|_2 \leq \iota,$$

and

$$\max_{j \notin S_\sigma^0} \left\| \left(\int_0^1 (\Psi_j)_{ijk} \left(\Psi_{S_\sigma^0}\right)_{ilm} dt \right) \left(\int_0^1 \left(\Psi_{S_\sigma^0}\right)_{ijk} \left(\Psi_{S_\sigma^0}\right)_{ilm} dt \right)^{-1} \right\|_2 \leq \kappa, \tag{3.32}$$

where $(\cdot)_{ijk}(\cdot)_{ilm}$ *denotes a tensor product with Einstein notation.*

Assumption 3.6. *Assume that*

$$f_{\min} > \chi_n \frac{4\sqrt{2 s_\mu C_{\max}}}{C_{\min}} \text{ and } \iota < \frac{1}{4} \sqrt{\frac{C_{\min}}{s_\mu C_{\max}}},$$

where $f_{\min} = \min_{j \in S_\mu} \left\{ \int_0^1 [f_j^*(Z_j^*(t))]^2 dt \right\}^{1/2}$ *is the minimum regulatory effect for the drift of* X^**.*

Similarly, assume that

$$g_{\min} > \lambda_n \frac{4\sqrt{2 s_\sigma D_{\max}}}{D_{\min}} \text{ and } \kappa < \frac{1}{4} \sqrt{\frac{D_{\min}}{s_\sigma D_{\max}}}, \tag{3.33}$$

where $g_{\min} = \min_{j \in S_\sigma} \left\{ \int_0^1 [g_j^*(Z_j^*(t))]^2 dt \right\}^{1/2}$ *is the minimum regulatory effect for the volatility of* X^**.*

Assumption 3.4 addresses the identifiability of the elements of $\{t, Z_{S_\mu}^*\}$ and $\{t, Z_{S_\sigma}\}$ and the non-degeneracy of the integrated basis functions outside of S_μ^0 and S_σ^0. Assumption 3.5 restricts the amount of interaction allowed between elements in $\{t, Z_{S_\mu}^*\}$ and outside it, and between elements in $\{t, Z_{S_\sigma}\}$ and outside it. These assumptions preclude concurvity, a necessary condition for parameters in an additive model to be identifiable, as explored more deeply in Buja

et al. (1989). Assumption 3.6 relates the quantities in prior assumptions to the sparsity tuning parameter λ_n, as is typical in the LASSO literature.

Given these assumptions, the proof of variable selection consistency for $\hat{\alpha}$ follows Chen et al. (2017). For the sake of brevity, we exclude it here and refer interested readers to Section A of the supplementary materials of that paper.

We are now ready to state the primary theoretical result.

Theorem 3.2. *Suppose that Assumptions 3.1-3.3, Assumption A.1 in the supplementary material, and Conditions A.1-A.4 in the supplementary material hold. Then $\hat{\beta}$, the estimator from Algorithm 3, is asymptotically almost sure consistent for the true value β^*. Furthermore, almost surely for sufficiently large n, the estimator $\hat{\beta}$ from Algorithm 3 recovers the correct support for σ, i.e. $\hat{S}_\sigma = S_\sigma$.*

The full proof is deferred to the supplement, but follows two main steps. We first establish the overall consistency of the estimator $\tilde{\beta}$ in (3.21) when $J(\beta)$ takes a ridge-like form. We then show that, given any consistent estimator $\tilde{\beta}$, the one-step improvement schema proposed in Algorithm 3 will provide estimation consistency as well as support recovery consistency.

3.4 Simulations

We examine the finite-sample performance of the proposed method in simulations. In each setting, data are generated from Gaussian processes, some endogenous (i.e. their future values rely only on their current state) and some exogenous (i.e their drift and/or volatility are allowed to rely on other processes in the dynamical system). In each case, the process of interest is exogenous. Following the convention set by other statistical work in ODE systems, we use smoothing splines with bandwidth specified by GCV to find the smoothing estimates $\hat{Z}$ in (3.11). We consider two choices of basis functions for ϕ and ψ in (3.12) and (3.20): the Fourier basis and cubic splines with two internal knots. We compute the integrals in (3.12) and (3.16) numerically with step size 0.001.

3.4.1 Variable selection in additive ODEs

In this simulation, we examine the performance of the proposed method in simultaneous variable selection for drift-level and volatility-level effects. A subset of size p_{endo} of the $p-1$ history processes are generated endogenous, and the remainder of the history processes and the process of interest are exogenous. We enforce strict unidirectional dependence. That is, the first of the exogenous history processes is allowed to depend only on the p_{endo} endogenous processes, the next is allowed to depend on the p_{endo} endogenous processes and the first exogenous processes, and so on. The process of interest's drift and volatility each depend on a fraction π of the $p-1$ history processes, with the groups not necessarily overlapping. In all cases, the true influence set is chosen randomly, as are the functional forms of the dependence. Due to the unidirectional dependence in this system, we can generate these processes directly rather than solving them via Euler's method, which may be complicated by the stochastic nature of the equations.

After generating the data, we apply the proposed method and determine $\hat{S}_\mu$ and $\hat{S}_\sigma$. We assess the performance of the method with the model selection sensitivity and specificity for X, defined as

$$\begin{aligned} \text{sens}_\mu &= \frac{|\{j : j \in S_\mu \cap \hat{S}_\mu\}|}{|\{j : j \in \hat{S}_\mu\}|} \\ \text{spec}_\mu &= \frac{|\{j : j \notin S_\mu \cup \hat{S}_\mu\}|}{|\{j : j \notin \hat{S}_\mu\}|}, \end{aligned} \tag{3.34}$$

and analogously for sens_σ and spec_σ. We ran $N = 100$ independent simulations for each combination (n, p, π).

Table 3.1 gives the model selection sensitivity plus specificity for $\mu(\cdot)$ and $\sigma(\cdot)$, averaged across $N = 100$ simulations and both values of π, for the specified values of n and p. The results show fairly strong performance in group recovery for $\mu(\cdot)$, especially when p is small or n is large, but more middling performance in group recovery for $\sigma(\cdot)$.

Component	n	$p = 10$	$p = 25$	$p = 50$
μ	50	1.40	1.18	1.10
	100	1.68	1.32	1.25
	200	1.76	1.32	1.27
σ	50	1.14	1.09	1.05
	100	1.15	1.11	1.07
	200	1.15	1.12	1.09

Table 3.1: Average sens + spec for $\hat{S}_\mu$ and $\hat{S}_\sigma$ across $N = 100$ independent simulation runs and various values of n and p.

3.5 Clinical Application

In this section, we illustrate one of our method's applications to clinical data arising from a clinical study of youth with type 1 diabetes (T1D). We first (Section 3.5.1) introduce the clinical study and explain the data we use. We then (Section 3.5.2) apply our method to these data and discuss the results.

3.5.1 CCAT study and data

T1D is the cell-mediated autoimmune destruction of the beta-cells of the pancreas, resulting in an absolute insulin deficiency and hyperglycemia. As a result, patients with type 1 diabetes are tasked with the daily management of blood glucose levels using exogenous insulin replacement in the form of multiple daily injections or continuous infusions (Mayer-Davis et al., 2018a; Association et al., 2018). Levels of blood glucose outside the normal range can lead to adverse consequences: sustained hyperglycemia, or high serum glucose, is associated with increased risk of complications such as cardiovascular disease and stroke (Nathan et al., 2014; Maahs et al., 2014), while acute hypoglycemia, or low serum glucose, invites the risk of coma or even death (Cryer et al., 2003). A growing body of research, however, suggests that due to its effect on oxidative stress, the volatility of serum glucose may be as important as the raw levels, if not more, when it comes to predicting complications of T1D (Monnier et al., 2008; Saisho, 2014).

Hemoglobin A1c is a well-studied and commonly used measure of a patient's glycemic control, representing average exposure to hyperglycemia over the preceding three months. However,

hemoglobin A1c does not capture transient glucose excursions or glycemic variability. With the recent emergence of continuous glucose monitoring (CGM) systems, which provide a reading of blood glucose levels on a minute-to-minute scale, attention has turned to this data to better characterize dysglycemia in the setting of type 1 diabetes. While the physiologic effect of insulin on blood glucose levels is causal and clear, the exact effect of physical activity and dietary intake can be more heterogenous and is less well-characterized, although these factors play important parts in the overall management of diabetes (Wright and Hirsch, 2017; Beck et al., 2017; Monnier et al., 2008; Kilpatrick et al., 2008). Additionally, the exact nature of the dependence between blood glucose volatility and these factors is not well-studied, especially at the resolution that CGM data provide. Due to the density of blood glucose readings that CGM data provide as well as the physiologic and patient-oriented implications of blood glucose variabilty, we chose to apply our method to data arising from CGM. We note that best practices for linking measures of blood glucose process volatility to clinically meaningful thresholds of blood glucose variability remain to be established; however, our approach is at least as likely as existing methods to offer a sufficient resolution for system volatility to do so.

The Carbohydrate Counting in Adolescents with T1D (CCAT) study followed 30 adolescent outpatients with T1D over 5 days with the goal of measuring acute changes to their blood glucose levels as well as key factors known to affect blood glucose levels: insulin, dietary intake, and physical activity (Maahs et al., 2012). Participants wore a CGM and an accelerometer-based tracker of physical activity (PA) for these 5 days. During the entire course of the study, patients' insulin doses were tracked, either by an insulin pump (20 participants) or an insulin pen recording multiple daily injections (10 participants). During day 1 and 3 of observation, participants logged their dietary intake, which was confirmed using time-stamped cell phone photographs. Dietary intake was then divided into a number of macronutrient categories, including carbohydrates, fats, and protein.

3.5.2 Application to CCAT study

We illustrate a proof-of-concept of our method by applying it to data from one CCAT patient. As the primary process of inferential and predictive interest in the CCAT study was blood glucose, we let $X(\cdot)$ represent blood glucose (mg/dL), while $H(\cdot)$ contained PA (counts/min), bolus insulin dose (U), carbohydrates consumed (g), fat consumed (g), and protein consumed (g). As the effects of dietary data were of scientific interest, we limited ourselves to the two days containing dietary data. We averaged the data from this patient over ten-minute intervals, giving us $n = 288$ evenly spaced measurements through the 2-day period.

The results of our CCAT analysis revealed a few interesting trends. First, all explanatory factors in $Z(t)$ were found to have a significant effect on both the patient's drift and volatility of blood glucose—that is, $\hat{S}_\mu = \hat{S}_\sigma = \{1, \ldots, p\}$. Given the posited causal links between blood glucose and the explanatory factors considered in this study, this should perhaps come as little shock for the glycemic drift; the fact that all factors appear to play a significant role in the volatility of blood glucose may represent an interesting finding worth exploring at a larger scale, though. The saturation of the regulator set has large potential implications on future interventions aimed at controlling blood glucose in a patient population similar to the patient analyzed here: namely, it suggests that physical activity, bolus dose, and the macronutrient breakdown are all important factors to consider intervening upon. We note, however, that this statement carries a hefty caveat: this is merely a proof-of-concept analysis. Substantial work must be done to study the properties of this method in already-collected data of this nature before we can make any practical suggestions for the collection of future research data. On the more technical side of things, we found that $M_1 = 3$ and $M_2 = 6$ this patient. This mirrors the trend discussed in Section 3.4.1: $\sigma(\cdot)$ appears to be a more complicated function than $\mu(\cdot)$, as using further basis depth to characterize its shape appears to be justified by cross-validation error.

3.6 Discussion

In this chapter, we present a method for the analysis of a system of continuous-time stochastic processes, performing parameter estimation and model selection for both the drift and the volatility of the process of interest. Our model requires only light assumptions, relying on additivity for its composite functions and a Brownian motion term to govern the process of interest's drift. The algorithm employed in this chapter to fit the chosen model can function in settings where $p > n$ and perform model selection thanks to its underlying reliance on the group LASSO, though other choices of algorithm are possible. We demonstrate several attractive theoretical properties of the algorithm, including model selection consistency, and explore its application to both simulated and clinical data.

We observe several caveats and limitations. One limitation we must address is the middling performance of the method at model detection sensitivity and specificity for $\sigma(\cdot)$ in numerical experiments with modest sample sizes. We believe that this stems from the fact that estimation of $\sigma(\cdot)$ is a more difficult task than estimation of $\mu(\cdot)$, due in part to the fact that the former depends upon the latter. Additionally, we would like to highlight that this is, to our knowledge, a novel approach to the problem. We are not naïve enough to believe our estimation method cannot be improved upon with further research, and improvements in the estimation of $\sigma(\cdot)$ will afford improvements in the operating characteristics studied here. Another limitation is the computational burden imposed by the chosen algorithm—in particular, fitting the group LASSO many times incurs a nontrivial computational cost. Tasks involved in computation are parallelizable, which mitigates this burden. The final limitation pertains to the chosen model: while additivity is a light assumption, among the least restrictive posed for any model of derivatives in a dynamical system, it is still an assumption. It is possible that relaxing strict additivity, for example by including interaction terms, would lead to improved performance without incurring substantial changes to the underlying algorithm.

Dynamic systems pose a unique set of challenges to researchers interested in learning from their data. In cases where it is not enough to simply know where a process will drift, but also to understand how volatile it will be when it does, new approaches are required. This research presents one such approach.

CHAPTER 4: MEASUREMENT INFLUENCE DIAGNOSTICS

4.1 Introduction

In real-world studies, measurements may be measured imperfectly for any of a variety of reasons: instrument failure, errors in sampling methodology or data collection, even the faulty nature of human memory. In many settings, observations can be observed again, or re-measured, potentially more carefully, to obtain a measurement with a higher degree of certainty. In settings where measurements are costly, however, the ideal approach from the perspective of data quality—re-measuring all suspect data points until they are measured reliably well—is unlikely to be feasible, and even less likely to be cost-effective. Additionally, not all data points are created equal. The concept that some observations are especially influential is well-known and accepted when it comes to correctly measured data impacting the results of data analysis, and there is no reason the same logic should not apply to the concept of measurement error. As such, choosing points to re-measure at random may prove inefficient.

In such settings, a reliable way to assess the importance of re-measuring any given observation, which we term *measurement influence*, may provide useful guidelines for real-world practice. In this chapter, we develop a method for determining the measurement influence of observations in error-prone and costly-to-measure data scenarios. We furthermore extend its utility by allowing it to be agnostic to the modeling assumptions employed and by allowing it to assess the influence of k-tuples of observations together.

Our method adapts the concept of leave-one-out influence metrics, such as the jackknife residual and Cook's distance, to the setting of measurement error. Namely, rather than calculating the change in a single-number summary of the model when an observation is removed, we calcu-

late the change in a single-number summary of the model when that observation is mismeasured by up to a pre-specified amount. This formulation of the influence statistic is model-agnostic, relying on modeling assumptions only so far as they are required to determine the single-number summary—in fact, it works for model-free approaches, provided that a suitable single-number summary of performance is available.

This chapter is laid out as follows. In Section 4.2, we present the general mathematical framework and specific data settings we will consider in this paper. In Section 4.3, we describe the method. In Section 4.4, we present numerical experiments examining the performance of the method in finite samples in the specific data settings previously outlined. In section 4.5, we apply the method to datasets arising from clinical and environmental applications. We conclude with a discussion in Section 4.6. Directions for future work are presented in Chaper 5.

4.2 General Framework

We present the most general framework in which the proposed method applies. Let $X \in \mathcal{X} \subset \mathbb{R}^p$ be an $n \times p$ covariate matrix, which is considered measured without error. Note that this assumption is reasonable in many settings, such as settings where the covariates include geographic location, year, and so on. Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be an $n \times 1$ outcome vector, which may be subject to errors in measurement. Let $P \in \mathcal{P}$ denote the possibly nonparametric model used to obtain the performance summary of interest, $\Psi(P) \in \Psi(\mathcal{P})$. Note that $\Psi(P)$ need not be a functional of a model in the most traditional sense—the value of an individualized treatment rule estimated by Outcome Weighted Learning suffices, for instance. In general, it we require that an efficient influence function $D(\cdot)$, as in Van Der Laan and Rubin (2006), can be defined for P. Finally, let $\delta : \Psi(\mathcal{P}) \times \Psi(\mathcal{P}) \to \mathbb{R}$ be a distance metric that quantifies how far apart two realizations of the summary of model performance lie.

We now specify two data settings which we will return to in simulations and data applications in Sections 4.4 and 4.5.

In the regression setting, we will presume Y is continuous in nature, and that the focus of modeling is predictive accuracy. That is, we set $\Psi(P) = \hat{Y}$, the predicted value of Y obtained by predicting at the observed points $X_1, \ldots, X_n$ using P. There are many possible choices of regression model P, each with their own set of required assumptions and benefits regarding predictive accuracy. Choices of P include ordinary least squares (OLS) regression, penalized regression such as the elastic net, and tree-based methods such as random forests (RF). We will primarily illustrate the impact of assumptions in the regression setting by comparing OLS and RF. In this setting, a natural distance metric δ is the Euclidean distance between n-vectors.

In the precision medicine setting, we will presume we observe $\{X_i, A_i,\ Y_i\}, i = 1, \ldots, n$, where $X_i \in \mathcal{X} \subset \mathbb{R}^p$ is a covariate vector, $A_i \in \{-1, 1\}$ is the assigned treatment, and $Y_i \in \mathbb{R}$ is the clinical reward. We will assume that the goal of analysis is to estimate an individualized treatment rule (ITR), a function $\pi : \mathcal{X} \to \{-1, 1\}$ which recommends treatment based on covariate state, that is optimal in the sense of achieving the highest value $V(\pi) = \mathbb{E}_\pi[X]$. As such, the natural summary of model performance is $\Psi(P) = \hat{V}(\pi) = \sum_{i=1}^{n} Y_i I\{A_i = \pi(X_i)\} / \sum_{i=1}^{n} I\{A_i = \pi(X_i)\}$, and the natural distance metric δ is univariate squared distance. Particular choices of model P can be supplied by direct ITR estimation methods, such as outcome weighted learning (OWL), or indirect ITR estimation methods, in which an explicit model is posed between (X, A) and Y and then inverted to obtain π.

4.3 Method

In this section, we describe the proposed method in the most general data setting. We first introduce the *maximal perturbation range* of $Y_i, i = 1, \ldots, n$, denoted $\Gamma(Y_i)$. $\Gamma(Y_i)$ is the range of values Y_i is assumed able to take if it is mismeasured. In many settings, the simple and symmetric form $\Gamma(Y_i) = (Y_i - \gamma, Y_i + \gamma)$ will suffice, where $\gamma > 0$ heuristically represents the "biggest" plausible measurement error. Other settings may suggest other choices for Γ: if Y is nonnegative, for instance, $\Gamma(Y_i) = [\max(0, Y_i - \gamma), Y_i + \gamma)$ is more appropriate. In practice, we will only use a grid of K values from $\Gamma(Y_i)$, which we will denote $\Gamma_k(Y_i), k = 1, \ldots, K$.

As described in Section 4.2, we must also specify a distance metric $\delta : \Psi(\mathcal{P}) \times \Psi(\mathcal{P}) \to \mathbb{R}^+$, which quantifies the distance between two realizations of $\Psi(\cdot)$. In many applications, the choice of δ will be obvious. If $\Psi(P) \in \mathbb{R}\ \forall P \in \mathcal{P}$, for instance, a natural choice is $\delta(\Psi(P_1), \Psi(P_2)) = (\Psi(P_1) - \Psi(P_2))^2$, while the Euclidean norm is a natural choice when $\Psi(P) \in \mathbb{R}^p$.

The method proceeds in a manner similar to leave-one-out methods such as jackknife residuals or Cook's distance. The key difference is that instead of dropping the observation Y_i entirely, we assume it has been measured incorrectly, such that it lies within the range given by $\Gamma(Y_i)$. We then compute the largest change in $\Psi(P)$ resulting from mismeasurement in the set $\Gamma(Y_i)$ and save this value as Δ_i, the *measurement influence* of observation i. Algorithm 4 provides the mathematical details of this procedure.

Algorithm 4 Estimation of Measurement Influence

1. For $i = 1, \ldots, n$:
   (a) For $k = 1, \ldots, K$:
      i. Set $\tilde{Y}_{(ik)} = (Y_1, \ldots, \Gamma_k(Y_i), \ldots, Y_n)^T$.
      ii. Let $\tilde{P}_{(ik)}$ denote the model incorporating $X, \tilde{Y}_{(ik)}$.
      iii. Compute $\Delta_{ik} = \delta(\Psi(P), \Psi(\tilde{P}_{(ik)}))$.
   (b) Compute $\Delta_i = \max_k \Delta_{ik}$.

Algorithm 4 provides stable and easily interpretable results when P is low in modeling uncertainty. For some choices of P, however—especially machine learning models whose relative lack of traditional influence statistics may make our method particularly attractive—this may not be a reasonable assumption. In cases where P is subject to greater modeling uncertainty, we propose a simple adaptation of Algorithm 4: average the results of M runs of the model when computing Δ_{ik}. Algorithm 5 provides the details of this procedure. We explore the rationale behind Algorithm 5 more thoroughly in Section 4.4.1.1.

Algorithm 5 Estimation of Measurement Influence with Modeling Uncertainty

1. For $i = 1, \ldots, n$:
   (a) For $k = 1, \ldots, K$:
      i. For $m = 1, \ldots, M$:
         A. Set $\tilde{Y}_{(ikm)} = (Y_1, \ldots, \Gamma_k(Y_i), \ldots, Y_n)^T$.
         B. Let $\tilde{P}_{(ikm)}$ denote the mth realization of the model incorporating $X, \tilde{Y}_{(ikm)}$.
         C. Compute $\Delta_{ikm} = \delta(\Psi(P), \Psi(\tilde{P}_{(ikm)}))$.
      ii. Compute $\Delta_{ik} = \frac{1}{M} \sum_{m=1}^{M} \Delta_{ikm}$
   (b) Compute $\Delta_i = \max_k \Delta_{ik}$.

4.4 Numerical Experiments

We explore the performance of the proposed measurement influence statistic in several numerical experiments. This section is laid out as follows. First, in Section 4.4.1, we assess our measurement influence statistic in the regression data setting, taking special care to illustrate how the properties vary with different choices of P. We devote Section 4.4.1.1 to an analysis of the impact of modeling uncertainty. In Section 4.4.2, we assess the performance of our measurement influence statistic in the precision medicine setting.

4.4.1 Regression Setting

We first consider the regression setting. As outlined in Section 4.2, in the regression setting, we assume Y is a continuous real-valued outcome variable. In these simulations, we set $Y \in \mathcal{Y} = \mathbb{R}$, and we allow mismeasurement to occur symmetrically up to a distance γ from the observed values, i.e. $\Gamma(Y_i) = [Y_i - \gamma, Y_i + \gamma],\ i = 1, \ldots, n$. Our summary of model performance is $\Psi(P) = \hat{Y} = (\hat{Y}(X_1), \ldots, \hat{Y}(X_n))$, the model-predicted values of the outcome at the n observed points in covariate space, with distances between realizations of $\Psi(P)$ given by the Euclidean norm in $\mathbb{R}^n$.

To facilitate visualization of Δ values, we first consider the case of an underlying linear relationship between X and Y with low-dimensional X. Specifically, we consider $n = 25$ and $p = 1, 2$, with $Y \sim \mathcal{N}(X_{\text{int}}\beta, \sigma^2)$, where X_{int} is X with an intercept column of ones appended.

Figure 4.1 compares the values of Δ given by an OLS and RF model as P, along with X and Y, when $n = 25$ and $p = 1$. We can immediately observe a strong trend in Figure 4.1: Δ values computed using OLS are synonymous with extreme values of X. We do not observe the same trend for Δ values computed using RF. To illustrate this disconnect, we can visualize the prediction curves obtained when the maximal and minimal impact measurement error occurs for both models. Figure 4.2 compares the OLS prediction lines from the observed, minimal impact, and maximal impact measurement errors, along with an analogous plot using the observed, minimal impact, and maximal impact RF prediction curves. The rationale is clear: the OLS prediction line takes a strong predetermined shape, and mismeasurements near the extremes of X change that shape the most. Meanwhile, the RF prediction curve is more local, allowing points with greater impacts on local prediction to attain higher Δ values. Figures 4.3 and 4.4 are analogous to panels (a) and (b) of Figure 4.1, respectively, in the case where $p = 2$, and they reveal similar trends. We note that raw distance from the origin does not govern OLS-based Δ values as strongly as distance normalized by the variability of X. Figures 4.5 and 4.6, analogously to panels (a) and (b) of Figure 4.2, show the observed, minimal impact, and maximal impact prediction planes and surfaces from OLS and RF, respectively. We note a similar difference in the placement and interpretation of high-Δ observations between OLS and RF as in the $p = 1$ case.

We now consider the case where the underlying relationship between X and Y is locally linear, but not linear overall. We begin again with $p = 1$. Figure 4.7 shows the difference in Δ values between OLS and RF in this setting. Note that the OLS Δ values still exclusively consist of values extreme in X, a trend that no longer appears reasonable given the data structure at hand, while the RF Δ values identify points more in line with the local nature of the data. We can again visualize this disconnect with the use of prediction curves, as Figure 4.8 illustrates. When we consider a locally planar model with $p = 2$, the same trends emerge as before: OLS

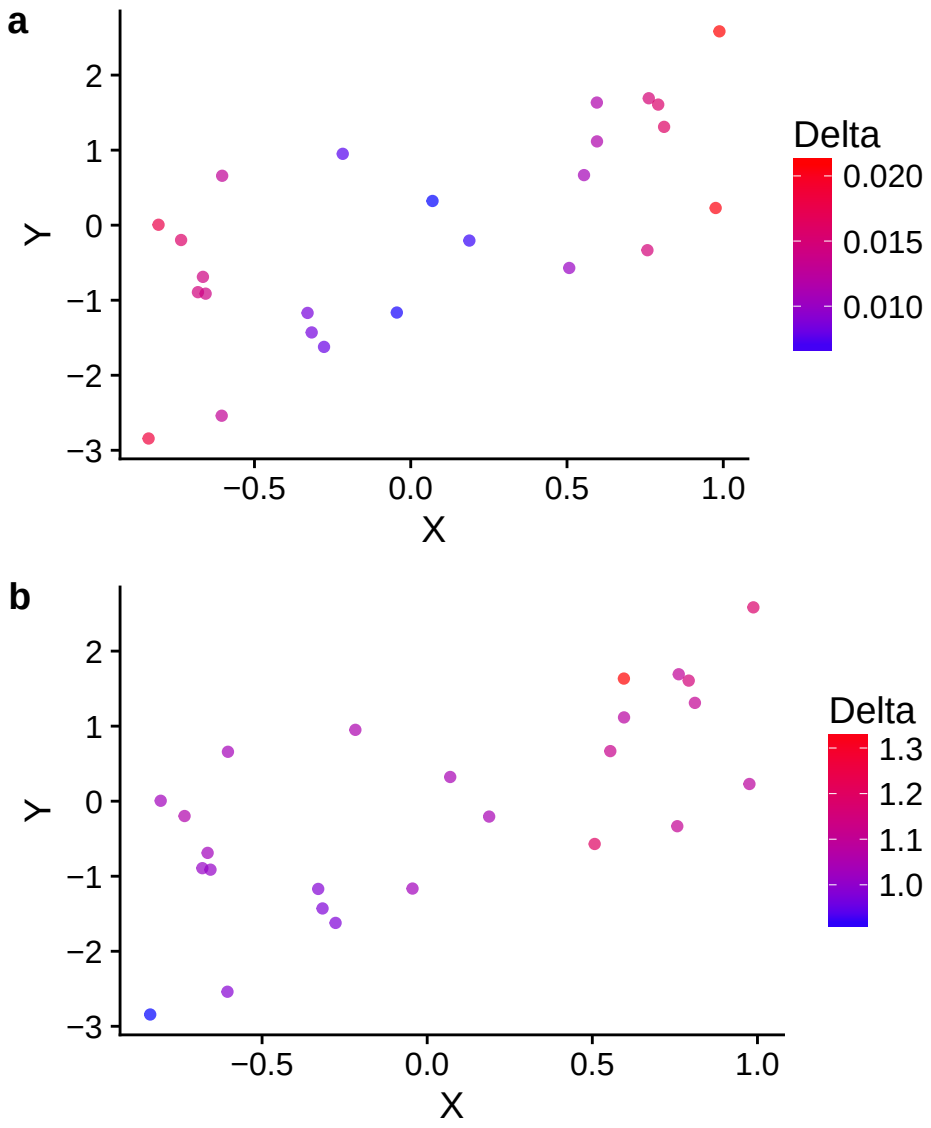

Figure 4.1: Δ values from an (a) OLS and (b) RF model in the regression data setting with a true underlying linear relation between X and Y and $n = 25, p = 1$. Deeper blue points have lower relative Δ, while brighter red points have higher relative Δ. Note the tendency of high Δ values from an OLS model to seek extreme values of X, while Δ values from RF do not exhibit the same trend.

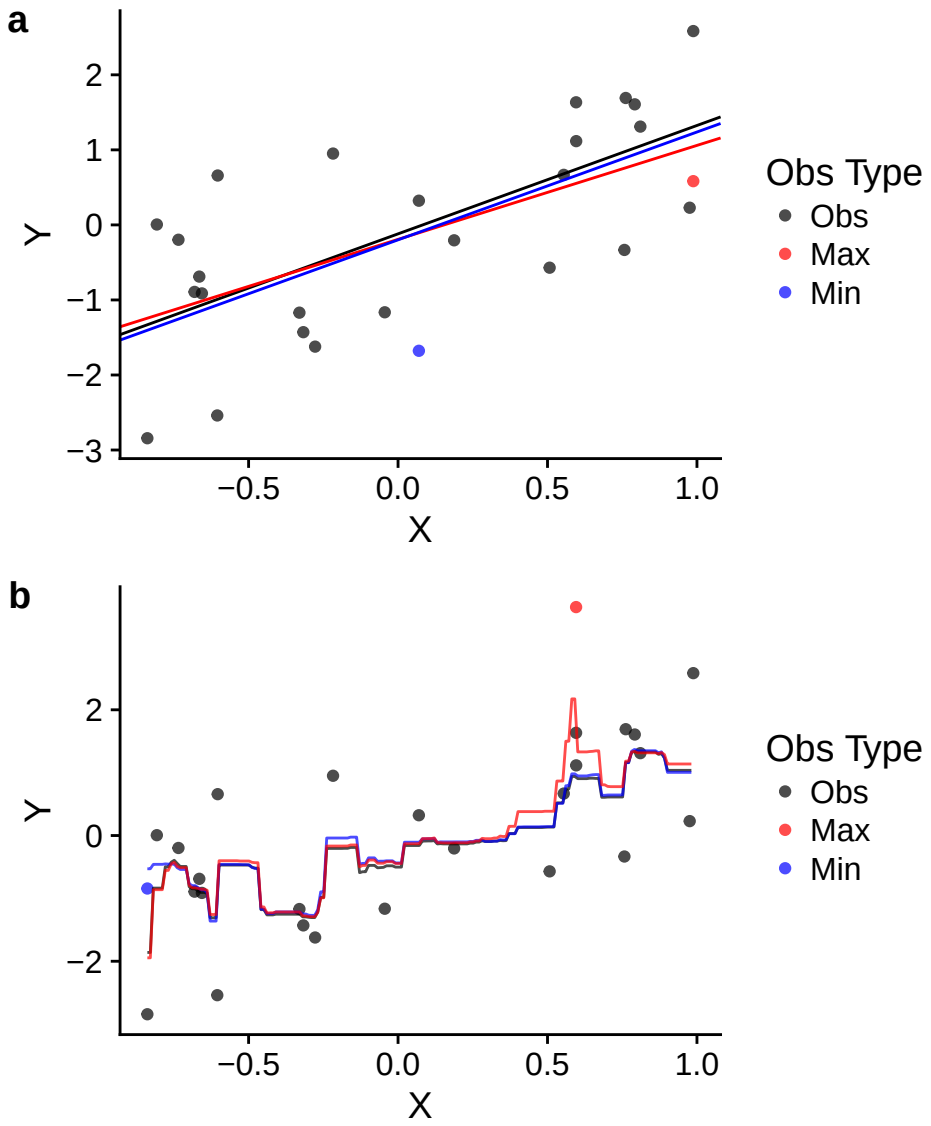

Figure 4.2: Prediction curves incorporating the observed, maximal impact, and minimal impact measurement errors from an (a) OLS and (b) RF model in the regression data setting with a true underlying linear relation between X and Y and $n = 25, p = 1$. The black points represent the observed data, while the red and blue points represent $\Gamma_k(Y_i) : \Delta_{ik} = \Delta_i$ for the observation i with the maximal and minimal values of Δ_i, respectively. Note the increased overall distance between the red and black curves, compared to the blue and black curves.

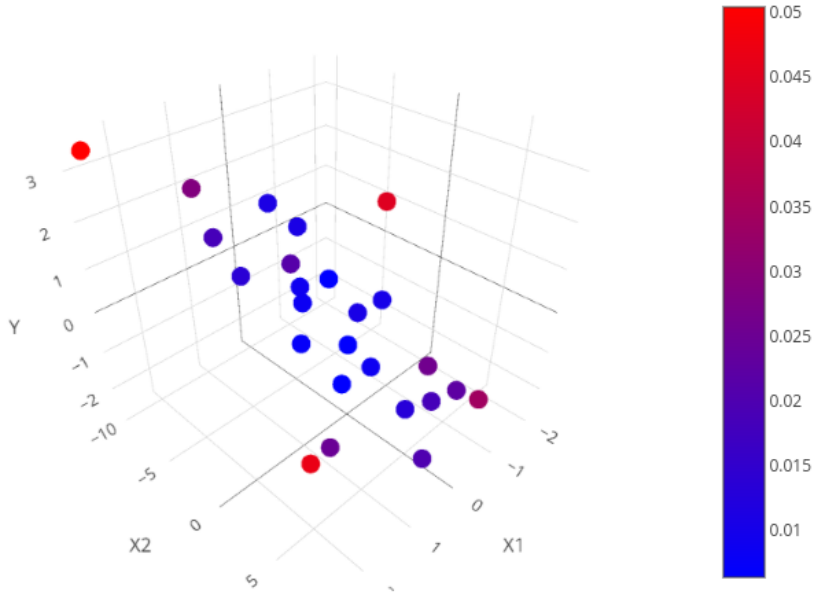

Figure 4.3: Δ values from an OLS model in the regression data setting with a true underlying linear relation between X and Y and $n = 25, p = 2$. Deeper blue points have lower relative Δ, while brighter red points have higher relative Δ. Note the tendency of high Δ values from an OLS model to seek variance-weighted extreme values of X, a tendency that carries over from the $p = 1$ case. A fully interactive version of this plot can be found at https://plot.ly/ mtlawson/19/#/.

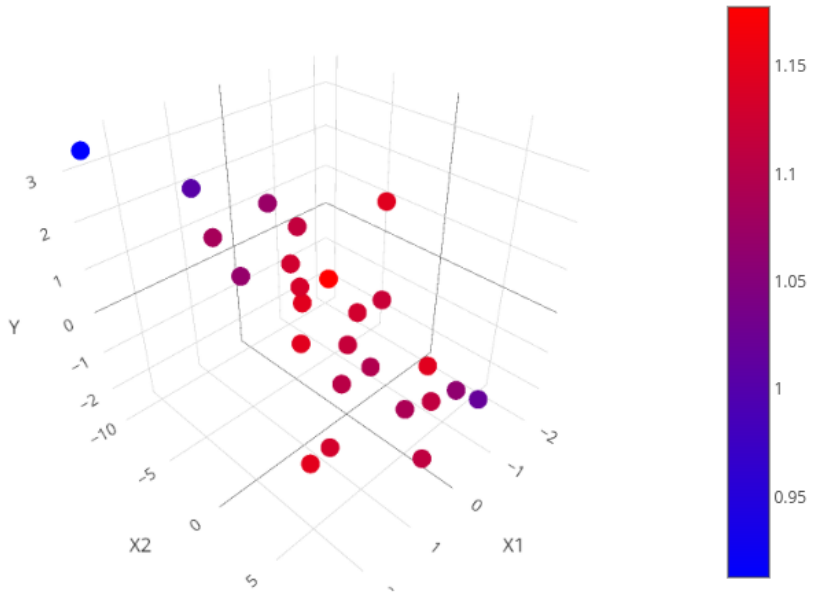

Figure 4.4: Δ values from an RF model in the regression data setting with a true underlying linear relation between X and Y and $n = 25, p = 2$. Deeper blue points have lower relative Δ, while brighter red points have higher relative Δ. Note that high Δ values are no longer restricted to variance-weighted extreme values of X, a tendency that carries over from the $p = 1$ case. A fully interactive version of this plot can be found at https://plot.ly/ mtlawson/21/#/.

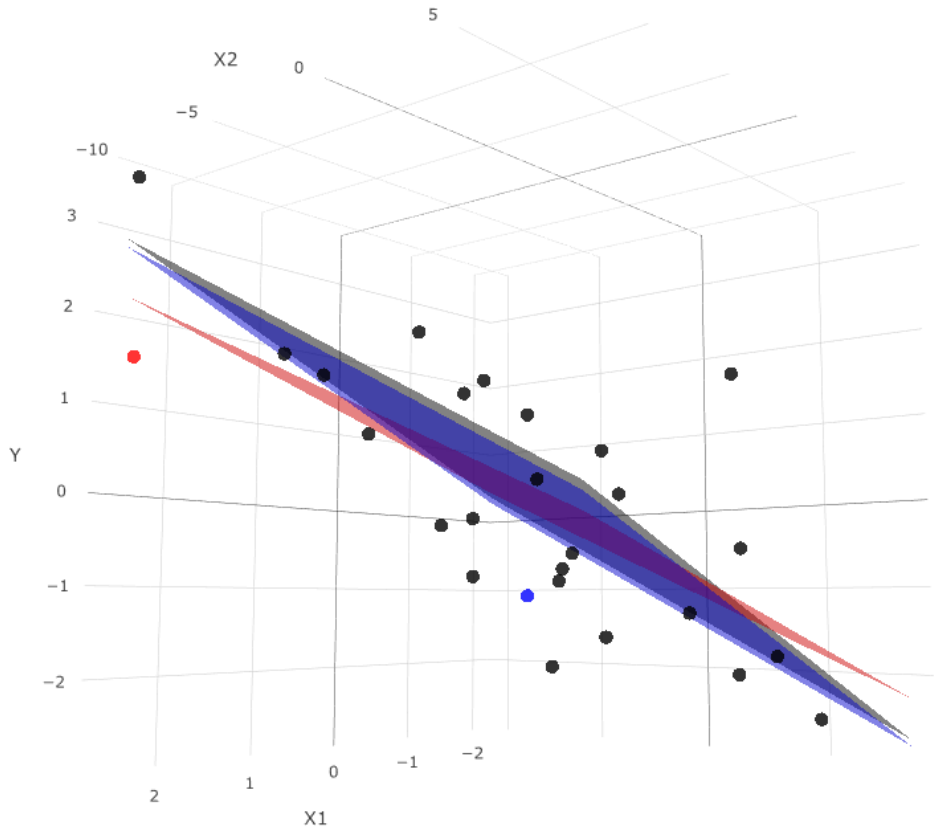

Figure 4.5: OLS prediction surfaces incorporating the observed data, minimal impact mismeasurement, and maximal impact mismeasurement, based on Δ values from an OLS model. The black points and black plane correspond to the observed data and the prediction surface from them, the blue point and plane correspond to the minimum-Δ observation after mismeasurement and the prediction surface after incorporating this point, and the red point and plane correspond to the maximum-Δ observation after mismeasurement and the prediction surface after incorporating this point. Note the increased distance between the red and black planes, relative to the red and blue planes.

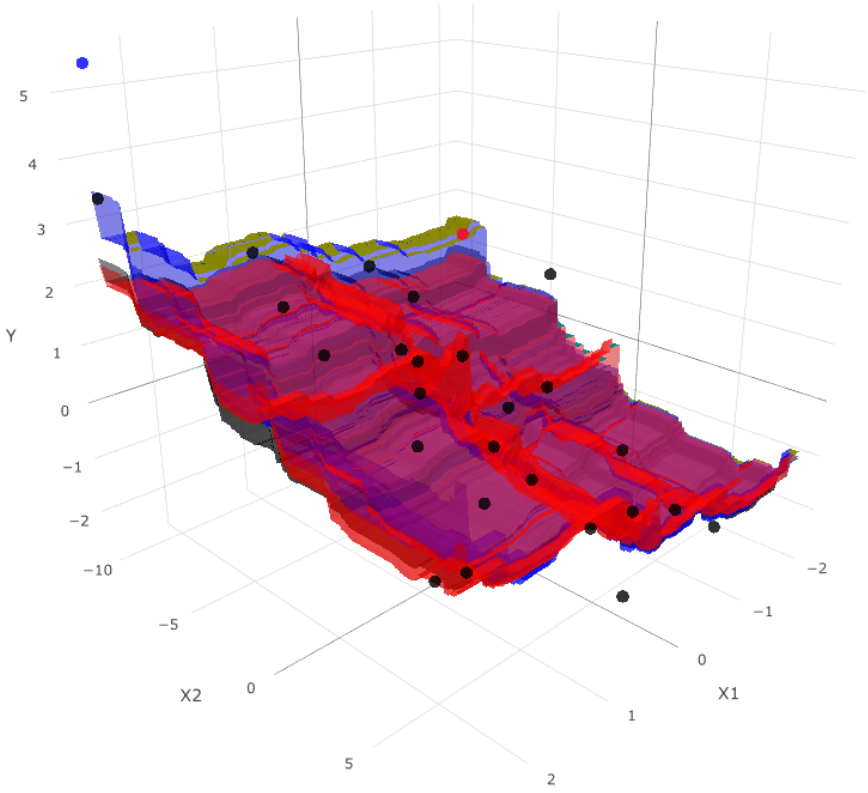

Figure 4.6: RF prediction surfaces incorporating the observed data, minimal impact mismeasurement, and maximal impact mismeasurement, based on Δ values from an RF model. The black points and black surface correspond to the observed data and the prediction surface from them, the blue point and surface correspond to the minimum-Δ observation after mismeasurement and the prediction surface after incorporating this point, and the red point and surface correspond to the maximum-Δ observation after mismeasurement and the prediction surface after incorporating this point. Note the differences in how the blue and red surfaces depart from the black. The blue surface largely departs from the black in the margin of lowest X_2 values, where few points lie, with the rest adheres closely to the observed prediction surface. The red surface, meanwhile, has a large ridge through the central body of the points, where many observations lie, separated from the black surface, while again it adheres closely in other regions.

continues to ignore the local structure and seek extreme X values (Figure 4.9), while RF does not solely identify extreme values of X (Figure 4.10). The impact on prediction is most evident when we examine the predictive surfaces: only extreme values of X have a large impact on the OLS predictive plane when mismeasured (Figure 4.11), while mismeasured values that have large impact on the RF predictive surface appear to take local structures into account (Figure 4.12).

4.4.1.1 Impact of Modeling Uncertainty

In Section 4.3, we alluded to the need to account for modeling uncertainty when calculating Δ. We explore this concept briefly. Consider the regression setting with a true locally linear model, analogous to that described in Section 4.4.1, with $n = 100$ and $p = 10$. Fix a value γ and allow $\Gamma(Y_i) = [Y_i - \gamma, Y_i + \gamma],\ i = 1, \ldots, n$. Suppose we choose P to be the random forest model. Then, holding X, Y, and γ constant, suppose we fit M independent runs of the random forest model, $P_1, \ldots, P_M$, and computed Δ_i for each observation i as in Algorithm 4. Let Δ_{im} denote the Δ value for the ith observation obtained by considering the mth model run. If modeling uncertainty has low overall importance, we would expect $\Delta_{i1}, \ldots, \Delta_{iM}$, to be very close.

Figure 4.13 shows side-by-side boxplots across $M = 100$ model runs for each of the $n = 100$ observations in this simulated dataset. Clearly, modeling uncertainty has a nonzero effect in this setting. High and low Δ values appear to be distinguishable from each other: if we compare the mean Δ values across the $M = 100$ runs between each of the $\binom{100}{2} = 4950$ pairs of observations using pairwise two-sample t-tests of equality, 574 or 11.6% of pairwise comparisons remain significant after Bonferroni correction. However, modeling uncertainty does appear to threaten the ability of Algorithm 4's approach to distinguish between more similar Δ values. Familiar statistical arguments suggest that the mean of M model runs will provide a more precise estimate of the Δ values for each observation. While we refrain from providing a blanket statement on what values of M are preferable, we suggest that M should be chosen to reflect the tradeoff between increased precision and increased computational burden.

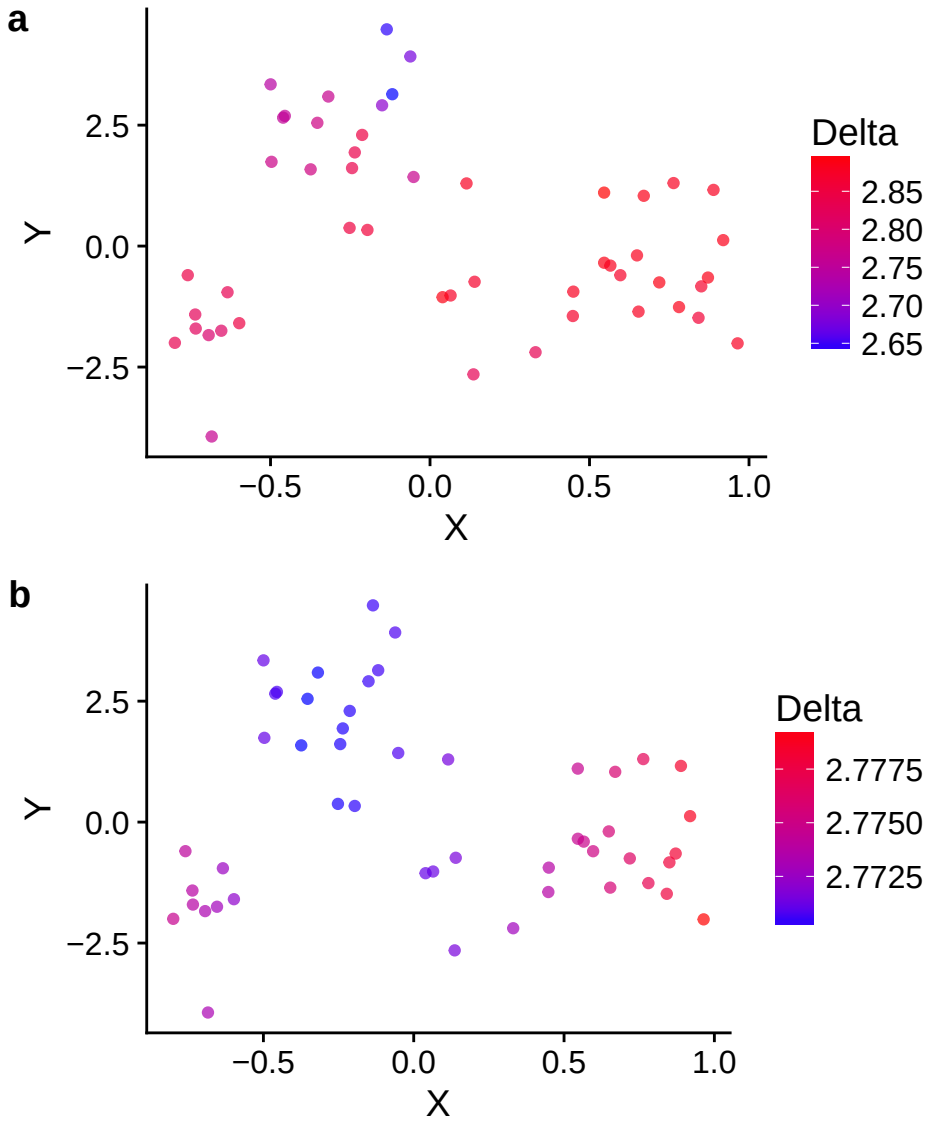

Figure 4.7: Δ values from an (a) OLS and (b) RF model in the regression data setting with a true underlying locally linear relation between X and Y and $n = 25$, $p = 1$. Deeper blue points have lower relative Δ, while brighter red points have higher relative Δ. Note the tendency of Δ values from an OLS model to seek extreme values of X—a tendency that is no longer attractive for this data setup—while Δ values from RF do not exhibit the same behavior.

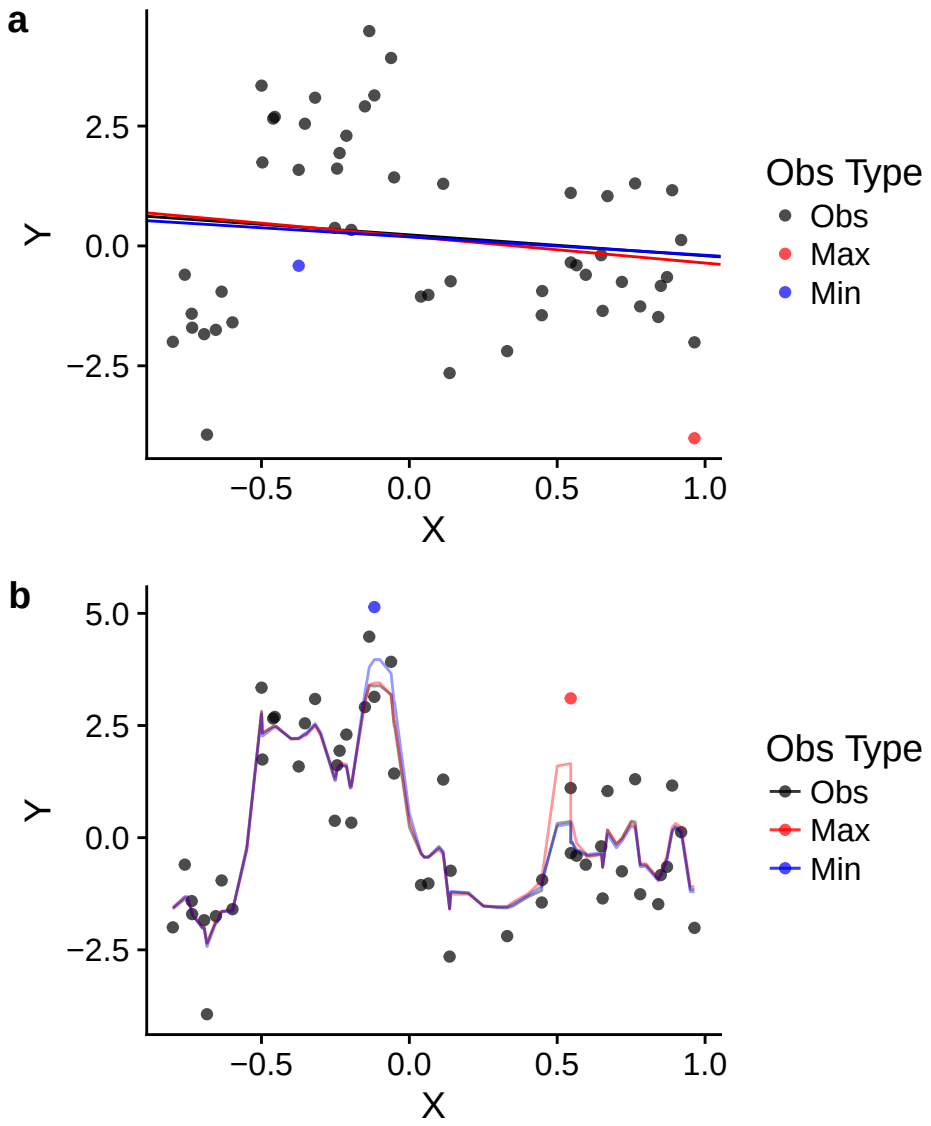

Figure 4.8: Prediction curves incorporating the observed, maximal impact, and minimal impact measurement errors from an (a) OLS and (b) RF model in the regression data setting with a true underlying locally linear relation between X and Y and $n = 25, p = 1$. The black points represent the observed data, while the red and blue points represent $\Gamma_k(Y_i) : \Delta_{ik} = \Delta_i$ for the observation i with the maximal and minimal values of Δ_i, respectively. Note the increased overall distance between the red and black curves, compared to the blue and black curves.

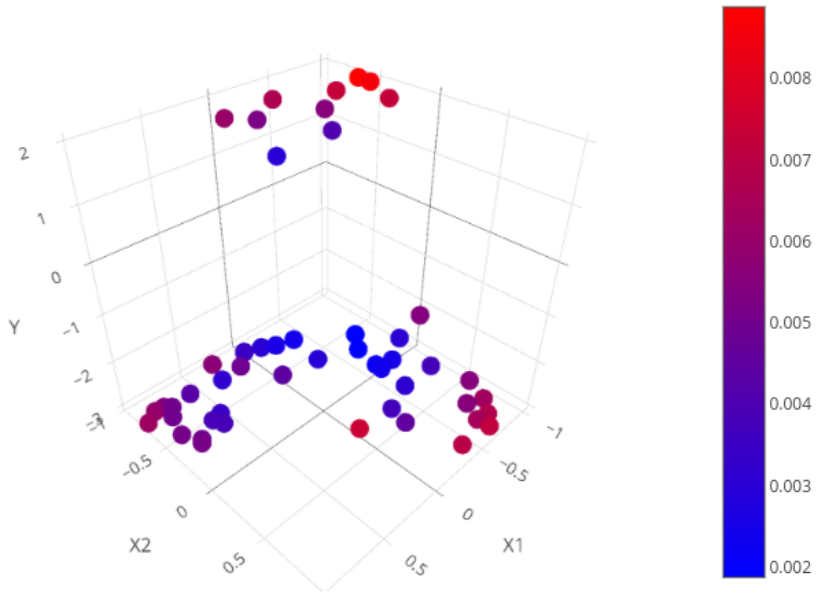

Figure 4.9: Δ values from an OLS model in the regression data setting with a true underlying local relation between X and Y and $n = 25, p = 2$. Deeper blue points have lower relative Δ, while brighter red points have higher relative Δ. Note the tendency of high Δ values from an OLS model to seek extreme values of X, a tendency that carries over from the $p = 1$ case, and which does not take into account the full trends present in these data. A fully interactive version of this plot can be found at https://plot.ly/ mtlawson/23/#/.

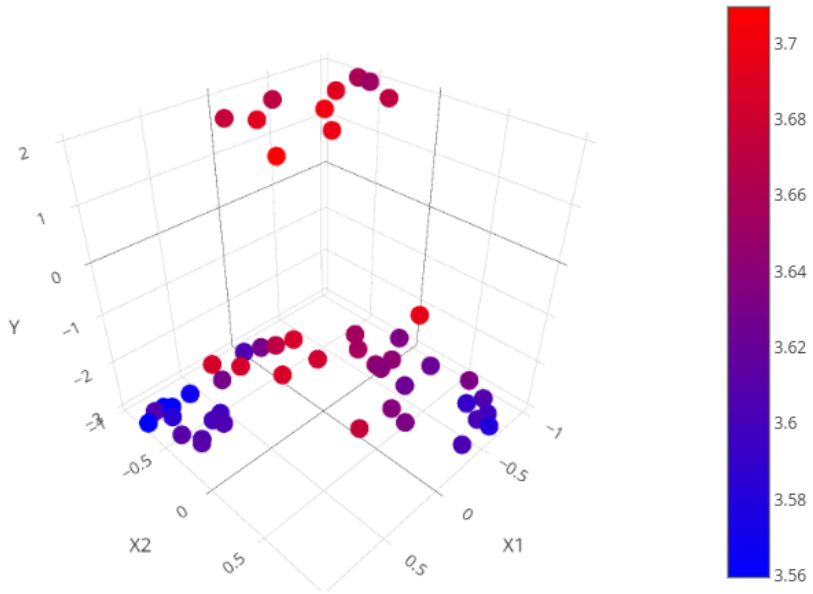

Figure 4.10: Δ values from an RF model in the regression data setting with a true underlying local relation between X and Y and $n = 25, p = 2$. Deeper blue points have lower relative Δ, while brighter red points have higher relative Δ. Note that high Δ values are no longer restricted to extreme values of X, a tendency that carries over from the $p = 1$ case. A fully interactive version of this plot can be found at https://plot.ly/ mtlawson/25/#/.

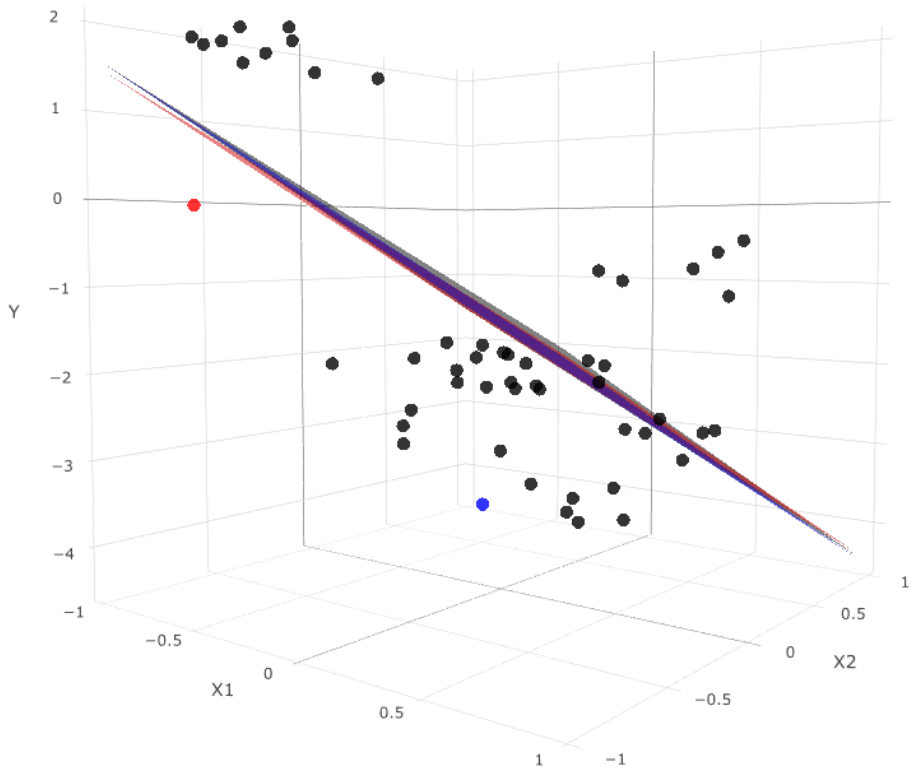

Figure 4.11: OLS prediction surfaces incorporating the observed data, minimal impact mismeasurement, and maximal impact mismeasurement, based on Δ values from an OLS model when the underlying data structure is nonlinear. The black points and black plane correspond to the observed data and the prediction surface from them, the blue point and plane correspond to the minimum-Δ observation after mismeasurement and the prediction surface after incorporating this point, and the red point and plane correspond to the maximum-Δ observation after mismeasurement and the prediction surface after incorporating this point. While all three prediction planes are close together, note that the red plane is more distant from the black than the blue plane.

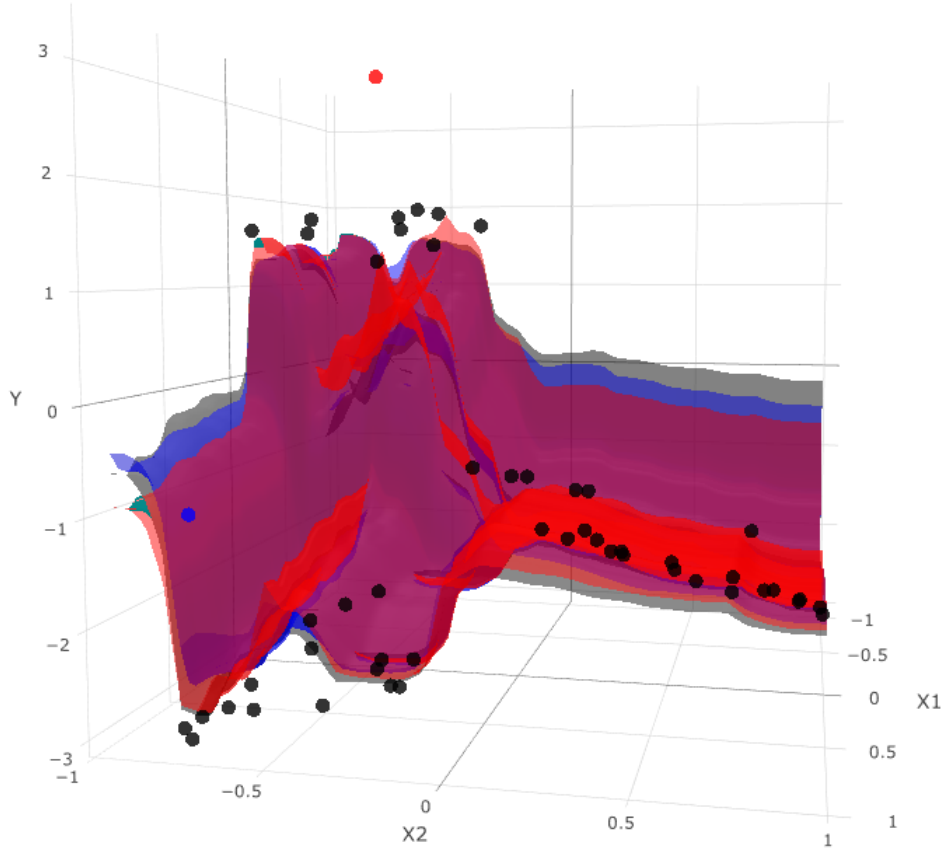

Figure 4.12: RF prediction surfaces incorporating the observed data, minimal impact mismeasurement, and maximal impact mismeasurement, based on Δ values from an RF model when the underlying data structure is nonlinear. The black points and black surface correspond to the observed data and the prediction surface from them, the blue point and surface correspond to the minimum-Δ observation after mismeasurement and the prediction surface after incorporating this point, and the red point and surface correspond to the maximum-Δ observation after mismeasurement and the prediction surface after incorporating this point. Note the differences in how the blue and red surfaces depart from the black. The blue surface departs from the black only to a small degree and only in the region where X_1 values are high and X_2 values are low. The red surface, meanwhile, has a large peak fairly close to the origin that juts above the black surface, visually represented by a lighter red.

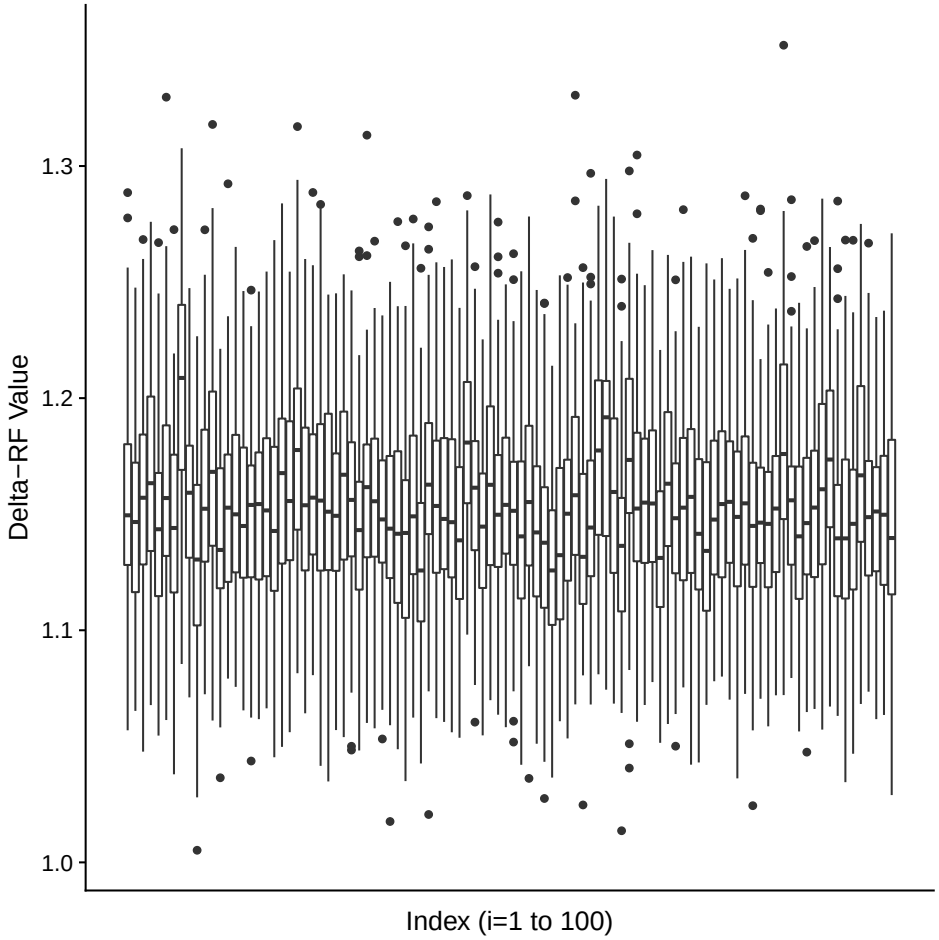

Figure 4.13: Side-by-side boxplots of Δ_{im} values computed via RF according to Algorithm 4 for $n = 100$ observations across $M = 100$ model runs in the regression data setting. Note the wide range for each observation, though some observations' main IQRs are nonoverlapping with others.

γ	% within 10 ranks
0.5	18%
1	20%
2	26%
3	18%
4	23%
6	21%
8	27%
12	35%
16	40%

Table 4.1: Percentage of observations with $|\text{rank}(\Delta_{i\ell}) - \text{rank}(\Delta_{i,\ell-1})| < 10$, by value of γ_ℓ.

Note also that modeling uncertainty appears to affect the value of Δ less as the maximal perturbation range $\Gamma(Y)$ grows. Using the setup of the paragraph above, instead suppose we hold model runs constant and vary the radius of the mismeasurement interval. That is, we specify $\ell = 1, \ldots, L$ values of γ, $\gamma_1, \ldots, \gamma_L$, then compute $\Delta_{i\ell}$ corresponding to each. Panel (a) of Figure 4.14 demonstrates the values of $\Delta_{i\ell}$ for each of the $n = 100$ observations across each value of γ_ℓ. Note that the average magnitude of Δ rises as γ rises, which is expected—as such, it is more informative to consider how Δ values change relatively. Panel (b) depicts this concept by plotting the rank of each observation's Δ_i among all $n = 100$ values of Δ_i at a given value of γ, across the different values of γ. As γ increases in magnitude, we see fewer crossing lines, and fewer lines that cross by large relative amounts, between subsequent values of γ. Table 4.1 summarizes this phenomenon numerically by reporting the number of observations i that, at a given value of γ, had their Δ rank less than 10 ranks distant from the rank at the next smallest value of γ. This fraction tends to increase as γ grows, suggesting that modeling uncertainty disrupts Δ values less when the measurement uncertainty is relatively large.

4.4.2 Precision Medicine Setting

We now consider the precision medicine setting. As outlined in Section 4.2, in this setting we observe the data $\{X_i, A_i, R_i\},\ i = 1, \ldots, n$, where $X_i \in \mathcal{X} \subset \mathbb{R}^p$ is a vector of covariates independent of treatment, $A_i \in \{-1, 1\}$ is the assigned treatment, and $R_i \in \mathcal{Y} \subset \mathbb{R}$ is a clinical

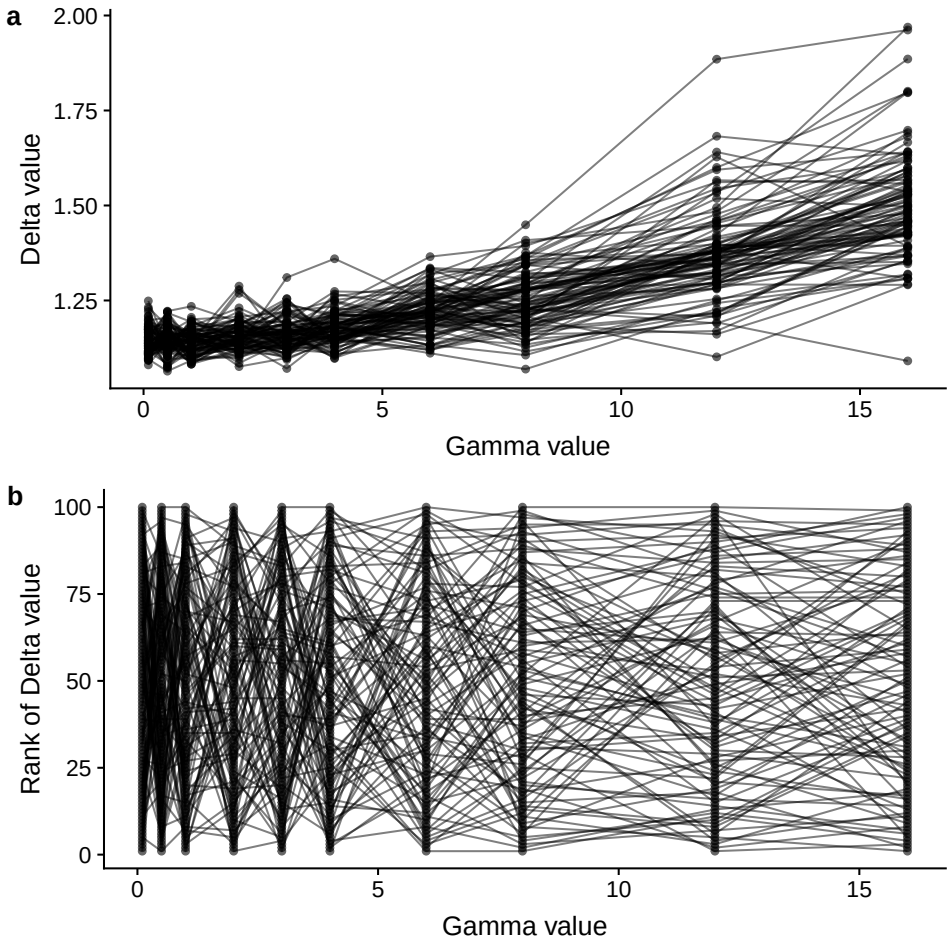

Figure 4.14: Values of (a) $\Delta_{i\ell}$ and (b) rank$(\Delta_{i\ell})$ by value of γ_ℓ for $n = 100$ observations and $L = 10$ values of γ. Note that the average magnitude of $\Delta_{i\ell}$ rises as γ_ℓ rises, visible in (a) but the amount of large relative change in $\Delta_{i\ell}$ drops. The dropoff in large crossing lines is more clearly visible in (b), where the scale is held constant.

reward. Following conventions in the field, we assume that higher values of the clinical reward indicate more favorable rewards; as in Zhou et al. (2017), we will allow rewards to be real-valued rather than strictly non-negative. The model P estimates an ITR $\pi : \mathcal{X} \to \{-1, 1\}$, a function that takes a value in the covariate space and recommends a value of treatment. We will index π by P for the remainder of this section to stress the dependence of the estimated ITR on the model P. In particular, P estimates an ITR that is optimal in the sense of maximizing the value V attainable within a class of policies, where $V(\pi) = \mathbb{E}_\pi[Y]$ is the expected reward obtained by following the policy π. The natural summary of model performance for this setting is given by $\Psi(P) = \hat{V}(\pi_P)$, where $\hat{V}$ is a consistent estimator of V; in particular, we use $\hat{V}(\pi_P) = \sum_{i=1}^{n} R_i I\{A_i = \pi_P(X_i)\} / \sum_{i=1}^{n} I\{A_i = \pi_P(X_i)\}$. As $\Psi(P) \in \mathbb{R}$, a natural choice of distance metric δ is squared univariate distance. We assume that potential mismeasurement occurs symmetrically at the level of the clinical reward up to a prespecified amount γ—that is, $\Gamma(R_i) = [R_i - \gamma, R_i + \gamma],\ i = 1, \ldots, n$.

In our numerical experiment, we set $n = 100$ and $p = 8$. We generated $X \sim U(-1, 1)$ and $A \sim$ Bernoulli($\frac{1}{2}$) independent of X. We mimicked data setup 1 from the simulations of Zhou et al. (2017), setting $\mu(X) = 1 + X_1 + X_2 + 2X_3 + 0.5X_4$ and $\nu(X) = 1.8(0.3 - X_1 - X_2)$ and then generating $R \sim \mathcal{N}(\mu(X) + \nu(X), 1)$. As discussed in that paper, this creates a moderate treatment effect size with a true linear decision boundary that depends on X_1 and X_2 alone. We use Residual Weighted Learning (RWL) as our model P, and we set $\gamma = 2$.

We examine the influence of measurement error in this setting through a series of three graphs. Figure 4.15, which displays the ITR assignment by R, X_1, and X_2, demonstrates that our RWL model is estimating the true linear decision boundary well. Figure 4.16, which shows the values of Δ_i for each of the $n = 100$ observations by R, X_1, and X_2, demonstrates that RWL does not strictly seek either points near the decision boundary or points near the extremes of X, instead finding what appears to be a mixture of such points. And Figure 4.17, which demonstrates which points switch ITR assignment from the observed ITR to the minimally and maximally different ITR, in terms of estimated value $\hat{V}$, demonstrates why the point with the largest

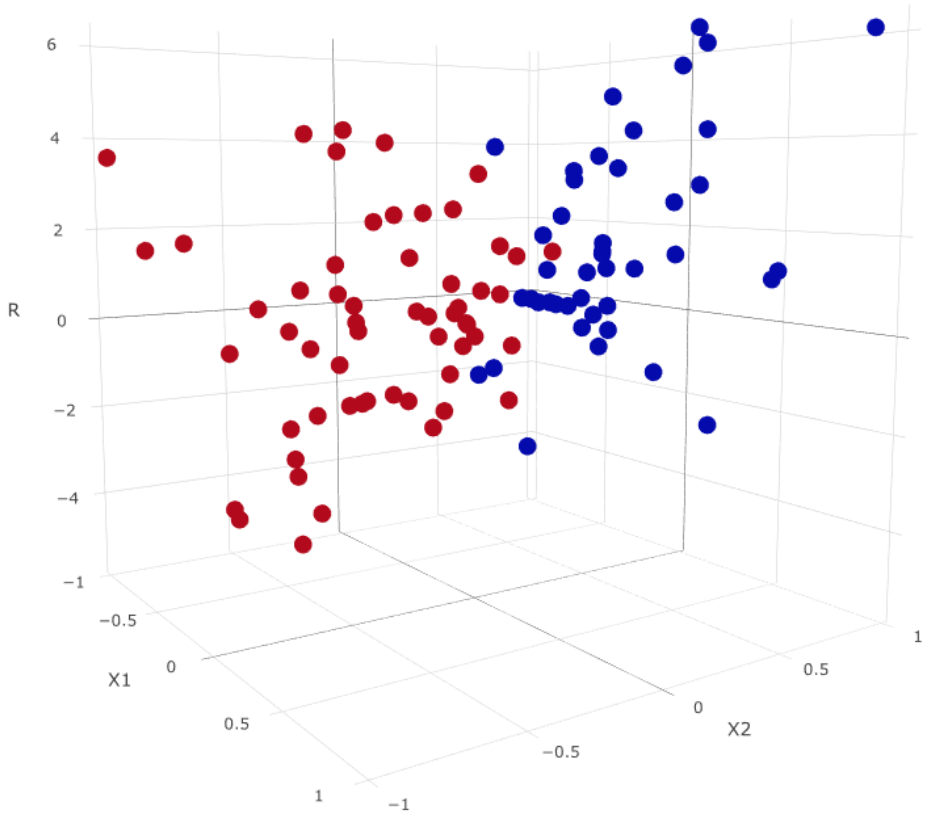

Figure 4.15: RWL ITR assignments from the precision medicine simulation described in Section 4.4.2, by values of X_1, X_2, and R. Note the decision boundary's approximate linearity in X_1 and X_2, which matches the true data-generating mechanism despite the presence of noise covariates.

Δ_i value has high measurement influence: when it is mismeasured, observations with large reward magnitudes switch ITR assignment.

Note that, in this experiment, the ITR with maximal perturbation has higher estimated value than the observed ($\hat{V}(\pi_P) = 2.26,\ \hat{V}(\pi_{(i_{\max})}) = 2.49$). This is not at all required, though—the ITR with maximal perturbation under measurement may achieve worse value instead. If gains and losses in value are not equally weighted—if achieving a worse value than the observed is more costly than failing to achieve a higher value than the observed, for instance—users of this method may wish to calibrate Δ to consider these cost weights.

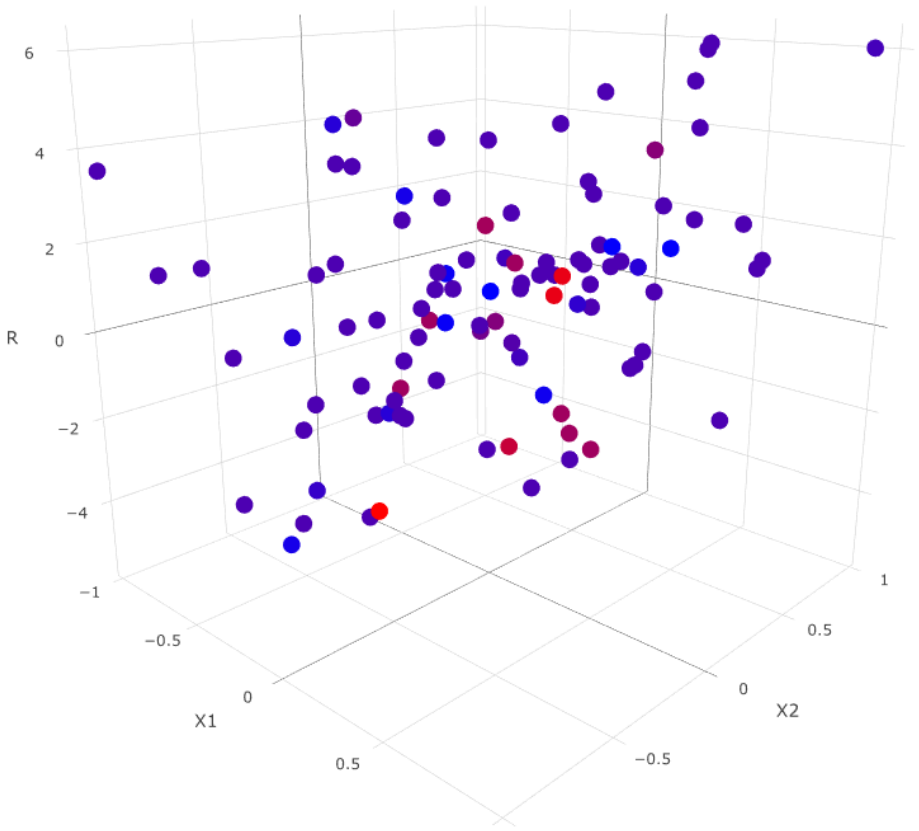

Figure 4.16: Values of Δ computed via RWL from the precision medicine simulation described in Section 4.4.2, by values of X_1, X_2, and R. Deeper shades of blue correspond to lower Δ values, while brighter shades of red correspond to higher Δ values. Note that high Δ values do appear near the decision boundary, and near the extremes of X, but are not confined to these locations deterministically.

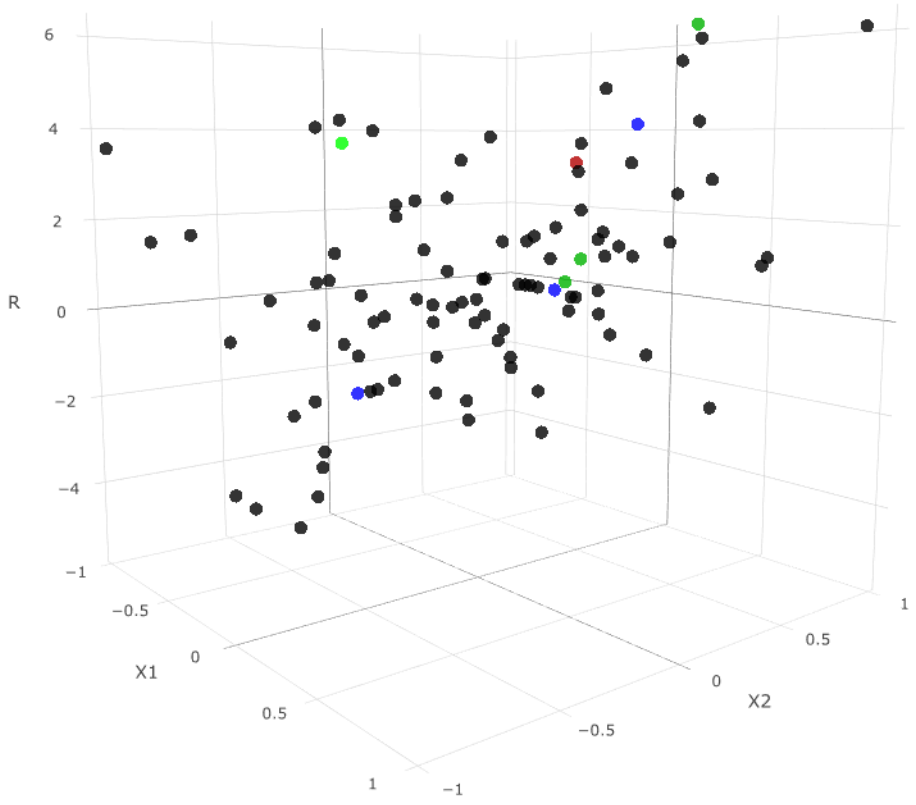

Figure 4.17: Depiction of ITR assignment switching between the observed, minimal impact, and maximal impact measurement errors. Black points have the same ITR assignment in all three ITRs. Red points switch assignment between the observed and maximal impact ITRs, blue points switch assignment between the observed and minimal impact ITRs, and green points switch assignment beween the observed and both the minimal and maximal impact ITRs. In this case, the minimal impact ITR departs only slightly from the observed because the blue points essentially balance each other out in terms of reward, leaving only the impact of the green points shared by the maximal impact ITR; meanwhile, the maximal impact ITR gains an additional high-reward point. Befitting this scenario, we observe $\hat{V}(\pi_P) = 2.26$, $\hat{V}(\pi_{P_{(i_{\min})}}) = 2.32$, and $\hat{V}(\pi_{P_{(i_{\max})}}) = 2.49$.

4.5 Environmental Applications

We illustrate the utility of our method by applying it to two datasets arising from the environmental sciences. The first, presented in Section 4.5.1, predicts the burn area of a forest fire based on a set of environmental and geological covariates, and therefore presents an application of our method to the regression setting. The second, presented in Section 4.5.2, evaluates the efficacy of an engineering intervention on the microbial content of a household compound's water source, and therefore presents an application of our method to the ITR estimation setting. In each case, aspects of the dataset complicate the construction of $\Gamma(Y)$—especially Section 4.5.2, in which the mechanism of the outcome's measurement must be taken into account.

4.5.1 Forest Fires

We illustrate the use of our method by applying it to a dataset arising from the environmental sciences. Cortez and Morais (2007) present an open-source dataset listing the hectares of forest burned in 517 wildfires in the Montesinho natural park in the Trás-os-Montes region of Portugal between January 2000 and December 2003. The dataset contains 12 design and meteorological covariates. Design and geological covariates comprise the month and day of the week the fire occurred and the X and Y coordinates within the Montesinho park area. Meteorological covariates comprise the temperature in Celsius, the relative humidity in %, the wind speed in km/h, the amount of rain in mm/m^2, and four components of the forest Fire Weather Index: the Fine Fuel Moisture Code (FFMC), the Duff Moisture Code (DMC), the Drought Code (DC), and the Initial Spread Index (ISI). Cortez and Morais (2007) propose support vector regression, single-layer neural networks, and random forest as candidates for P; we elect to use the random forest model. As the goal is prediction of the area burned by forest fires, we let $\Psi(P) = \hat{Y}$, and we choose δ to be the Euclidean distance between n-vectors.

As in the original analysis, we log-transform the outcome of interest to account for its right skew, which gives rise to some consideration on how to construct $\Gamma(Y)$. We propose poten-

tial mismeasurement at the scale of the original burn area in hectares. This mismeasurement is symmetrical but bounded below at the value of zero. That is, if we let $\tilde{Y}$ denote the pre-log-transformed outcome, we set $\Gamma(\tilde{Y}_i) = [(\tilde{Y}_i - \gamma)_+, \tilde{Y}_i + \gamma]$, where $(\cdot)_+ = \max(0, \cdot)$ is the positive part function. We obtain $\Gamma(Y)$ by log-transforming the points of $\Gamma(\tilde{Y})$.

Table 4.2 gives the fifteen highest measurement influence forest fires among the 517 observed, including the value of Δ, the raw area burned in hectares, the (X,Y) coordinates, the month, the FFMC, the ISI, the temperature in Celsius, the relative humidity in %, and the wind speed. Most covariates appear in general ranges consistent with their overall distributions, with some exceptions. The appearance of a December fire in the fifteen most measurement influential fires is noteworthy, as only 9 fires, or 1.7%, occurred in December, and especially given that this December fire was quite large, lying in the 84th percentile of all fires by size. Only 5.2% of fires had an FFMC of 84.1 or lower, so the appearance of two in the fifteen highest by Δ is also peculiar. Although several fires that burned 0 hectares, the minimum and mode in these data, were highly influential if mismeasured, notably, none of the very largest fires were—the largest fire in this top fifteen burned 11.19 hectares, placing it in only the 84th percentile of fires by area burned.

4.5.2 Water Source Microbial Content

[Author's note: These data are still undergoing cleaning, and we are still awaiting the full study committee's approval to use this dataset. A full analysis will follow once the data are available. In my view, though, this section's consideration of a mechanistic motivation for the construction of $\Gamma(Y)$ still warrants its inclusion at the current state.]

Studies of microbe concentration provide an intriguing application of the proposed method due to the mechanistic nature of the measurement error involved. In the MapSan study, household compounds in the vicinity of Maputo, Mozambique, were visited at baseline and 12 months after baseline. After the baseline visit, each household compound went through a standardized de-worming procedure, then approximately half of household compounds were assigned to an

Δ	Area	(X,Y)	Month	FFMC	ISI	Temp	RH	Wind Speed
1.627	0.43	(1,4)	Sep	91	7	21.7	38	2.2
1.593	0	(5,4)	Sep	92.1	9.6	10.1	75	3.6
1.591	0	(6,5)	Feb	84.1	2.2	5.3	68	1.8
1.584	11.19	(8,6)	Dec	84	5.3	5.1	61	8
1.582	0	(7,5)	Aug	93.7	8.4	26.4	33	3.6
1.578	0	(3,5)	Sep	93.5	8.1	17.2	43	3.1
1.578	4.53	(1,4)	Aug	90.2	8.9	20.3	39	4.9
1.577	0	(8,6)	Aug	92.3	8.5	24.1	27	3.1
1.575	0	(5,5)	March	90.9	8	11.6	48	5.4
1.575	2.69	(8,6)	Aug	85.6	6.6	17.4	50	4
1.569	0	(8,6)	Aug	91.1	5.8	23.4	22	2.7
1.566	0.17	(6,5)	Aug	94.3	22.7	19.4	55	4
1.560	1.75	(4,5)	Sep	91.1	12.5	15.9	38	5.4
1.559	0	(3,4)	Sep	91.8	9.2	18.9	35	2.7
1.556	7.21	(1,4)	Sep	92.8	7.5	16.8	28	4

Table 4.2: The Δ values, area burned (ha), X and Y coordinates within the Montesinho park area (both ordinal from 1 to 9), month, FFMC index, ISI index, temperature in degrees Celsius, relative humidity in %, and wind speed in km/h of the fifteen most measurement influential forest fires in the data of Cortez and Morais (2007). Δ values were obtained using the RF model with $\Psi(P) = \hat{Y}$ and $\Gamma(Y)$ lying symmetrically about the raw burned area $\tilde{Y}$, clipped below at zero, and then log-transformed before analysis as the outcome variable was.

engineering intervention intended to improve the cleanliness of the water supply. At the baseline and 12-month visits, water from the household compound's main water source was sampled, frozen, and transported to Chapel Hill, NC. There, the water samples were tested for the presence and abundance of certain microbial targets, such as *E. coli*, via quantitative polymerase chain reaction (qPCR). The goal of analysis is to assess the efficacy of the intervention in reducing the microbial content in the household compound's primary water source. We view this as essentially an ITR estimation problem, as an ITR can estimate those patients expected to benefit from receiving the intervention.

In qPCR, a target segment of DNA that corresponds to the microbe of interest is specified, then a given sample is run through many cycles of the PCR loop, with each cycle increasing the abundance of the target DNA in the sample. After a given cycle, if the target DNA is detected at a certain level of abundance, known as the *detection limit*, or greater, then the relative abundance of the DNA is calculated using the number of cycles until detection normalized by a control curve

constructed from a known quantity of DNA (Filion, 2012). After a given cycle, or even a full set of cycles, the abundance of the target DNA may still lie below the detection limit, indicating that the abundance is not distinguishable from zero; in practice, in these cases the abundance is often artificially set to either zero or the midpoint between zero and the detection limit. In practice, detection limits can be quite high relative to zero, meaning the amount of potential mismeasurement can be non-negligible (Filion, 2012).

The mechanism of measurement in qPCR suggests a particular form for $\Gamma(Y)$. Let λ_i denote the limit of detection for sample i. Let $\tilde{Y}_i$ denote sample i's partially-observed "true" microbial abundance, and let Y_i denote sample i's "processed" abundance output, where any values below the detection limit λ_i are considered not reliably measurable and are set to $\lambda_i/2$. Let Y_λ denote the set of processed outcomes lying below their detection limits, and let Y_λ^c denote the set of processed outcomes lying above it. If $Y_i \in Y_\lambda$, $\tilde{Y}_i$ could reasonably lie anywhere in $[0, \lambda_i]$. If $Y_i \in Y_\lambda^c$, it may still be reasonable to assume some level of potential mismeasurement $\gamma < \lambda_{\min}/2$, where $\lambda_{\min} = \min(\lambda_1, \ldots, \lambda_n)$. As such, we can define

$$\Gamma(Y_i) = \begin{cases} [0, \lambda_i], & Y_i = \lambda_i/2 \\ [Y_i - \gamma, Y_i + \gamma], & Y_i > \lambda_i. \end{cases} \tag{4.1}$$

To account for skew, microbial qPCR abundances are sometimes log-transformed before analysis; in this case, the limits of $\Gamma(Y)$ can be long-transformed along with Y.

4.6 Discussion

In this paper, we present a statistic quantifying the influence that a given observation's potential mismeasurement has on the performance of the model at large. To our knowledge, this combination of measurement error and observation influence is novel, and as such may represent a fruitful avenue for interesting future research. The measurement influence statistic is presented in a model-agnostic form that is adaptable to a variety of data settings, as well as to different

modeling goals within a given data setting. We explore the measurement influence statistic's performance in a series of numerical experiments, in which we clarify certain ways in which the statistic depends on its underlying model. And we show the influence statistic's real-world applicability by analyzing an environmental dataset to identify which potentially costly measurements are most important to have measured correctly.

We would like to highlight salient aspects of this approach that we believe to be strengths. First, this approach is straightforward to conceptualize. The focus of this method began with, and was centered on, utility in the real world—the ability to simply explain this method's rationale for selecting an observation to spend actual cost on remeasuring should be immediately apparent as a boon. Second, this approach is flexible and allows easy incorporation of domain knowledge. We demonstrate this fact clearly in both environmental applications in Section 4.5: knowledge of the underlying data measurement mechanism is reflected in the construction of $\Gamma(Y)$. Third, this approach concurs with well-studied properties of some models. In particular, the tendency of values of Δ computed via OLS to seek the extremes of covariate space, as seen in Section 4.4.1, mirrors the approach of Zhou et al. (2002), which notes that sampling from the extremes can increase efficiency in the estimation of OLS parameters.

This approach is not without caveats. First, although the measurement influence statistic is model-agnostic, it is not free of impact from the chosen model, as seen quite plainly in Section 4.4.1. While we choose to view this as a positive factor, as it clarifies aspects of the modeling approach chosen, some may view it otherwise. Second, computing our measurement influence statistic may incur a hefty computational burden, depending on the underlying model. At worst, it may multiply the computational cost by a factor of nM—albeit in a highly parallelizable way. Finally, it is possible that the proposed approach may be misled in some data settings by the fact that it only perturbs one observation at a time. Extensions of this approach, in which k-tuples of points are perturbed at once, may be a natural approach to tackle this limitation—though we will note these extensions incur a further combinatorial burden of computation.

In *Animal Farm*, George Orwell wrote that "All animals are equal, but some animals are more equal than others." Removing the totalitarian timbre of that quote, we contend that the same holds true for errors in measurement: while measurement errors may occur in similar ways, for the same reasons, and with equal probabilities across a sample, not all potential measurement errors have the same importance when it comes to the end goal of an analysis. The proposed method provides a way of identifying the highest-impact errors in measurement. In a world where measurement errors may be costly and difficult to avoid, this represents a powerful new tool available to investigators.

CHAPTER 5: DISCUSSION AND FUTURE RESEARCH

In this chapter, we discuss several directions for future research. We begin with the subgroup determination method of Chapter 2. We briefly mentioned one potential direction for future research in Section 2.6.1.3. Namely, as Figures 2.1, 2.2, and 2.3 demonstrate, subgroup recovery sensitivity and specificity, particularly sensitivity regarding the muted group and specificity regarding the intervention and control groups, may be improved by an ITR estimation method that is less sensitive to differences between treatment and control. In the case of an indirect estimation method like RLT, this may take the form of ϵ-insensitivity to differences in predicted values. That is, where the RLT-based ITR is currently assigned by

$$\hat{\pi}(X_i) = \begin{cases} -1, & \hat{Q}^{-1}(X_i) - \hat{Q}^{1}(X_i) > 0 \\ 0, & \hat{Q}^{-1}(X_i) - \hat{Q}^{1}(X_i) = 0 \\ 1, & \hat{Q}^{-1}(X_i) - \hat{Q}^{1}(X_i) < 0, \end{cases} \tag{5.1}$$

it would instead be assigned by

$$\hat{\pi}(X_i) = \begin{cases} -1, & \hat{Q}^{-1}(X_i) - \hat{Q}^{1}(X_i) > \epsilon \\ 0, & |\hat{Q}^{-1}(X_i) - \hat{Q}^{1}(X_i)| \leq \epsilon \\ 1, & \hat{Q}^{-1}(X_i) - \hat{Q}^{1}(X_i) < -\epsilon, \end{cases} \tag{5.2}$$

where $\hat{Q}^{-1}$ and $\hat{Q}^{1}$ are defined in Section 2.6.1. The details of finding an automated procedure to choose the most desirable ϵ in (5.2)—and indeed the correct metric to evaluate the desirability of a given choice of ϵ—remain for future research. Yuan and Wegkamp (2010) provide a method for

"rejecting" the choice between intervention and control in an OWL framework by penalizing misclassification more heavily than failing to classify. While this procedure creates a third group, the algorithmic justification of this group is not identical to that of the muted group. The "rejection" group of Yuan and Wegkamp (2010) consists of those patients who are likely to be misclassified if the algorithm is forced to choose; the muted group described in Chapter 2 consists of those patients who experience no difference in predicted outcome between intervention and control (or, if the algorithm is amended as in (5.2), a sufficiently small difference). It is reasonable to presume that patients likely to be misclassified by OWL are also likely to experience no (or very small) RLT-predicted differences between intervention and control. This is, however, merely a presumption, and it must be verified.

Another natural extension of the work presented in Chapter 2 is to more complex intervention options than binary. For indirect estimation methods such as RLT, the K-intervention setting does not induce a particular mathematical burden in ITR estimation, with the potential exception of needing to adaptively scale ϵ in (5.2) by intervention pair. The concept of the muted group, however, must be extended to "muted for a given intervention set." A patient may, for instance, experience identical predicted outcomes under interventions A and B but identical, higher predicted outcomes under interventions C and D. Such a case clearly recommends one set of interventions over another set, but does not distinguish between the interventions within either set. In the same way, extending indirect estimation methods to the ordinal treatment or dose-finding setting may not pose especially great mathematical challenges, but the challenges in interpretation are non-trivial. For the direct estimation method of OWL, extensions to the K-treatment (Fu et al., 2016), ordinal (Chen et al., 2018), and dose-finding (Chen et al., 2016) settings that do not consider a muted group are feasible given existing methods. None of these methods are currently adapted to include a "rejection" region, and the ordinal treatment and dose-finding settings in particular would seem to present mathematical challenges for doing so. Even still, the potential for disconnect between the muted group and the "rejection" group remains here.

There are similarly several extension of interest for the differential equation model presented in Chapter 3. First, one potential factor that hinders the flexibility of our model is its strict additivity. Adding first-order interactions between the functional forms in (3.5) would relax the strictness of our model's additivity while maintaining computational feasibility. In particular, we can estimate the interaction term $f_{jk}\left(Z_j(t), Z_k(t)\right)$ with a tensor product of the basis functions in phi, then enforce the proper interaction hierarchy using overlapping group penalties in LASSO, as outlined in Bien et al. (2013). An analogous procedure will apply for interaction terms g_{jk} in the $dW(t)$ piece of (3.5). This extension of the model should not disrupt any of the theory presented in Chapter 3, and its impact on computational complexity should be minor in terms of ease to implement, as LASSO is a method that is suited for $n \ll p$. The primary cost is likely to be computation time, as each individual run of LASSO with overlapping group penalties is slower than the standard group LASSO.

A further extension of the model in Chapter 3 is to relax the assumption that our error term takes a Brownian motion form—that is, to replace $dW(t)$ with $dL(t)$, where $L(\cdot)$ is a more general Lévy process. This substitution would complicate the presented approach in several ways. First, the calculus that leads from (3.13) to (3.16) would no longer apply, causing the estimating equation to take a potentially less tractable form. Second, the rates in Theorem 3.1, which depend on the properties of Brownian motion, would no longer apply, likely leading to slower rates of convergence.

Another extension of the work presented in Chapter 3 is an alternative method for solving for $\hat{\beta}$ in (3.16). Namely, rather than solve for $\hat{\beta}$ through our bilinear estimation scheme, we could proceed in a manner similar to that of forward stagewise regression, starting from 0 and taking many small steps in directions governed by the correlations between the covariates and the current residual vector (Efron et al., 2004). As Efron et al. (2004) describes, forward stagewise regression is not preferred in linear models, where the Least Angular Regression (LARS) algorithm provides a more efficient solution for asymptotically equivalent covariate paths. In the setting of (3.16), however, where the tensor structure of the data violates the linear independence assump-

tions of LARS, an adaptation of forward stagewise regression that accounts for the tensor nature of the data may still prove feasible, and furthermore this algorithm may not prove inefficient compared to fitting LASSO many times, as our current estimation approach requires.

The research of Chapter 4, to our knowledge, represents a novel union of two concepts, measurement error and influence statistics. As such, there are many possible avenues to extend the work presented in this chapter. A natural concern with the present approach of Algorithms 4 and 5 is that perturbing a single observation at a time may fail to reveal the underlying trends related to measurement error, as a single observation may not have a substantial enough impact on the overall summary of model performance. The clearest way to address these concerns is to perturb not a single observation, but a q-tuple of observations simultaneously, where $1 < q \ll n$. It should be immediately noted that, without improvements in the underlying computational algorithm, this extension of the method would incur a combinatorial increase in the computational burden of the algorithm.

In the current approach, we specify a grid of values $\Gamma_k(Y),\ k = 1, \ldots, K$, then select $\Delta_i = \max_{k \in 1,\ldots,K} \Delta_{ik}$. Implicitly, this approach ignores all perturbations within the range of measurement error $\Gamma(Y)$ except the perturbation with maximum impact. While there are arguments for why this approach is logical—focusing on the worst case allows us to take a minmax approach to the risk associated with measurement error—it may hinder the ability to make probabilistic statements. Approaches that incorporate multiple values of $\Gamma_k(Y)$, or Bayesian formulations of this problem, may provide an avenue toward doing so.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 3

A.1 Proofs

In this section, we prove Theorems 3.1 and 3.2 from Section 3.3 in the main paper. Subsection A.1.1 gives the full explanation of commonly used notation in this section. Subsection A.1.2 presents the proof of Theorem 3.1. Subsection A.1.3 presents several additional technical conditions required for the proof of Theorem 3.2, then presents the proof.

A.1.1 Notation

We begin by introducing notation that will be used throughout this section. We first define a frequently referenced index set, $S_\sigma^0 = (0, S_\sigma)$. That is, S_σ^0 denotes the set of true regulators of $\sigma(\cdot)$ in union with the intercept β_0.

We next note that we will frequently use sets to index several quantities, including various variations of β. For instance, $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_2})^T$ is an M_2-vector, while $\beta_{S_\sigma^0}$ is an $s_\sigma M_2$-vector, where $s_\sigma = |S_\sigma|$, corresponding to $\bigcup_{j \in S_\sigma} \beta_j$ arranged in the proper order. We will apply set subscripts to several other vector quantities in the same manner in the course of the proof.

In Section 3.2.4, we introduced $\hat{U}_i$, an $(M_2 p + 1) \times (M_2 p + 1)$ matrix corresponding to the ith time point. In that section, we noted that $\hat{U}_i$ had a first row and column consisting of a single 1 and many 0s, and the rest of $\hat{U}_i$ had a block structure, with each of the p^2 $M_2 \times M_2$ blocks corresponding to one $(j, k) \in \{1, \ldots, p\} \times \{1, \ldots, p\}$. That is, we can write

$$\hat{U}_i = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & \hat{U}_{i(11)} & \hat{U}_{i(12)} & \ldots & \hat{U}_{i(1p)} \\ 0 & \hat{U}_{i(21)} & \hat{U}_{i(22)} & \ldots & \hat{U}_{i(2p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \hat{U}_{i(p1)} & \hat{U}_{i(p2)} & \ldots & \hat{U}_{i(pp)} \end{pmatrix}. \tag{A.1}$$

We will use parenthetical indices, as shown above, to reference these blocks. Set notation applies here as well: $\hat{U}_{i(S_\sigma^0 S_\sigma^0)}$ is an $(s_\sigma M_2 + 1) \times (s_\sigma M_2 + 1)$ matrix, for instance. We will use the centered dot $(\cdot)$ to denote the entire row or column set: for instance, $\hat{U}_{i(j\cdot)}$ is the $M_2 \times (pM_2 + 1)$ matrix equal to the jth "block row" of $\hat{U}_i$, and $\hat{U}_{i(\cdot\cdot)}$ is another way to write $\hat{U}_i$.

A.1.2 Proof of Theorem 3.1

The error in covariates nature of our problem requires additional care—before we can establish more desirable theoretical results, we must first ensure that the smoothed processes $\hat{Z}_j, j = 1, \ldots, p$ are sufficiently close to the true processes $Z_j^*, j = 1, \ldots, p$.

Minor, and standard, assumptions of smoothness of the estimators $\hat{Z}_j$ must be met. In particular, if we use the local polynomial estimator for $\hat{Z}_j$,

$$\hat{Z}_j(t; h) = \sum_{i=1}^{n} Y_{ji} W_{ni}(t; h), \tag{A.2}$$

where W_{ni} is defined as in Section 1.6 of Tsybakov (2009), then we must make three additional assumptions. For the sake of brevity, those assumptions are omitted here; we direct the interested reader to that book. These assumptions allow us to use Lemma 1.3 of Tsybakov (2009), presented here for clarity:

Lemma A.1. *Under the given assumptions, for all* $n > n_0$, $h > 1/(2n)$, *and* $t \in [0, 1]$, *the weights* W_{ni} *in* (A.2) *satisfy:*

1. $\sup_{i,t} |W_{ni}(t; h)| \leq C_3/(nh)$;
2. $\sum_{i=1}^{n} |W_{ni}(t; h)| \leq C_3$.

We are now ready to prove Theorem 3.1 of the main paper.

Proof.

$$\left\|\hat{Z}_j - Z_j^*\right\|^2 = \int_0^1 [\hat{Z}_j(s;h) - Z_j^*(s;h)]^2 ds = \int_0^1 \left\{\frac{1}{n}\sum_{i=1}^n Y_{ji} W_{ni}(s;h) - Z_j^*(s)\right\}^2 ds \quad \text{(A.3)}$$
$$= \int_0^1 \left\{\sum_{i=1}^n [Z_j^*(t_i) + \epsilon_{ji}] W_{ni}(s;h) - Z_j^*(s)\right\}^2 ds$$
$$\leq 2\int_0^1 \left\{\sum_{i=1}^n [Z_j^*(t_i) - Z_j^*(s)] W_{ni}(s;h)\right\}^2 ds + 2\int_0^1 \left\{\sum_{i=1}^n \epsilon_{ji} W_{ni}(s;h)\right\}^2 ds,$$

where the last inequality follows from the fact that the weight sum to one and the fact that $(a + b)^2 \leq 2a^2 + 2b^2$. Thus we can characterize our inequality as

$$\left\|\hat{Z}_j - Z_j^*\right\|^2 \leq 2\int_0^1 \text{bias}^2(s) ds + 2\int_0^1 v^2(\epsilon_j/\sigma, s, h) ds, \quad \text{(A.4)}$$

where bias$(\cdot)$ is clearly defined in the final line of (A.3) and

$$v(a, s, h) = \sigma \frac{1}{n}\sum_{i=1}^n a_i W_{ni}(s;h), \quad \epsilon_j = (\epsilon_{j1}, \ldots, \epsilon_{jn})^T, \quad \text{(A.5)}$$

with σ arising from Assumption 3.1 in the main paper.

We scrutinize (A.4). First, from Assumption 3.2, we note that for any $t \in [0, 1]$ and $h > 1/n$,

$$|\text{bias}(t)| \leq \sum_{i=1}^n |Z_j^*(t_i) - Z_j^*(t)||W_{ni}(t;h)| \quad \text{(A.6)}$$
$$\leq \sum_{i=1}^n L_1 |t_i - t|^{\tau_1} |W_{ni}(t)|$$
$$\leq \sum_{i=1}^n L_1 h^{\tau_1} |W_{ni}(t)|$$
$$\leq L_1 h^{\tau_1} C_3 \equiv q_1 h^{\tau_1},$$

where C_3 comes from Lemma A.1. Thus we can bound the first term in (A.4). Next, we can use Theorem 5.6 from Boucheron et al. (2013) to bound $\int_0^1 v^2(\epsilon_j/\sigma, s, h) ds$ by $n^{\nu-1}h^{-1}$ for some

positive $\nu < 1$ with at least probability $1 - 2\exp\{-n^{\nu}/(2\sigma^2 C_3)\}$ in the exact same manner as done in Chen et al. (2017). For a full explication of this portion of the proof, including a proof that v is Lipschitz, we direct the reader to Section A of the supplementary materials of that paper.

Hence we can say

$$\left\|\hat{Z}_j - Z_j^*\right\|^2 \leq 2q_1^2 h^{2\tau_1} + 2n^{\nu-1}h^{-1}, \tag{A.7}$$

with probability no smaller than $1 - 2\exp\{-n^{\nu}/(2\sigma^2 C_3)\}$. When we minimize the right-hand side of (A.7) with respect to h, we find that the minimizer h_n satisfies $2\tau_1 q_1^2 h_n^{2\tau_1+1} = n^{\nu-1}$. With this bandwidth, the error bound is therefore

$$\left\|\hat{Z}_j - Z_j^*\right\|^2 \leq C_2 n^{\frac{2\tau_1}{2\tau_1+1}(\nu-1)},$$

where C_2 is a global constant. □

Note that this particular proof uses the local polynomial model (A.2). This is not necessary; a similar proof will hold for other well-behaved and sufficiently smooth estimators $\hat{Z}_j$.

A.1.3 Proof of Theorem 3.2

The estimation scheme presented in 3 rests on two facets: finding an overall consistent estimator from (3.21), then updating it with a standardized LASSO-like penalty, as in Simon and Tibshirani (2012), to correctly induce variable selection. As discussed in Chen et al. (2017), the fact that the regressors $\hat{U}_i, i = 1, \ldots, n$ are estimated poses additional challenges.

Define the true coefficients β^* by

$$\mathbb{E}[V^*(t_i)] = \beta^{*T} U_i^* \beta^*,$$

where U_i^* is analogous to $\hat{U}_i$ using the true function values $\psi(Z^*(t_i))$. Here we build upon the work of Chen et al. (2017) and extend the result of variable selection consistency for group LASSO regression with errors in variables to the estimation of process volatility. In order for

this consistency to hold, we need one additional assumption and four conditions, presented here. Of note, several of these conditions rely on an existing consistent estimate $\tilde{\beta}$; this mimics the proposed estimation scheme, which begins with a consistent estimator and improves it to induce variable selection.

Assumption A.1. *Assume that $\beta \in \mathcal{B}$, where $\mathcal{B}$ is compact.*

Condition A.1. *Suppose that, for any almost surely consistent estimator $\tilde{\beta}$, the following holds almost surely:*

$$
\begin{aligned}
&0 < \frac{1}{2} D_{\min} \leq \Lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_{i(S_\sigma^0 \cdot)} \tilde{\beta} \tilde{\beta}^T \hat{U}_{i(\cdot S_\sigma^0)} \right), \\
&\Lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_{i(S_\sigma^0 \cdot)} \tilde{\beta} \tilde{\beta}^T \hat{U}_{i(\cdot S_\sigma^0)} \right) \leq 2 D_{\max}, \\
&0 < \frac{1}{2} D_{\min} \leq \Lambda_{\min} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_{i(j \cdot)} \tilde{\beta} \tilde{\beta}^T \hat{U}_{i(\cdot j)} \right), j \notin S_\sigma^0,
\end{aligned}
$$

where $D_{\min}$ and $D_{\max}$ are introduced in Assumption 3.4.

Condition A.2. *Suppose that, for any almost surely consistent estimator $\tilde{\beta}$, we have almost surely that*

$$
\max_{k \notin S^0} \left\| \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_{i(j \cdot)} \tilde{\beta} \tilde{\beta}^T \hat{U}_{i(\cdot S_\sigma^0)} \right) \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_{i(S_\sigma^0 \cdot)} \tilde{\beta} \tilde{\beta}^T \hat{U}_{i(\cdot S_\sigma^0)} \right)^{-1} \right\|_2 \leq 2\kappa, \tag{A.8}
$$

where κ is introduced in Assumption 3.5.

Condition A.3. *For $j = 1, \ldots, p$, let $\Gamma = \max_{j=1,\ldots,p} \left\| \hat{Z}_j - Z_j^* \right\|$. Assume that, for any almost surely consistent estimator $\tilde{\beta}$, we have almost surely that*

$$
\left\| \frac{1}{n} \sum_{i=1}^{n} \hat{U}_{i(j \cdot)} \tilde{\beta} \tilde{\beta}^T \hat{U}_{i(\cdot S_\sigma^0)} \beta_{S_\sigma^0}^* - \frac{1}{n} \sum_{i=1}^{n} \tilde{\beta}^T \hat{U}_{i(S_\sigma^0 \cdot)} (V_i + \tilde{\beta}^T \hat{U}_i \tilde{\beta}) \right\|_2 \leq \eta, \tag{A.9}
$$

where η depends only on n, M_2, Γ, $\| \beta_{S_\sigma}^* \|$, and global constants.

Condition A.4. *The following inequalities hold:*

$$\frac{2\sqrt{s+1}}{D_{\min}}\eta + \lambda_n \frac{2\sqrt{2sD_{\max}}}{D_{\min}} \leq \frac{2}{3}\beta_{\min},$$
$$\frac{2\kappa\sqrt{s+1}+1}{\lambda_n}\eta + 2\kappa\sqrt{2sD_{\max}} \leq \sqrt{D_{\min}/2},$$

where $\beta_{\min} = \min_{j \in S_\sigma^0} \|\beta_j^*\|_2$, *and* $\kappa, \eta, D_{\min}$, *and* $D_{\max}$ *are introduced in Assumptions 3.4-3.6.*

We are now ready to give the proof of Theorem 3.2.

Proof. The proof is divided into two main steps. First, we verify the consistency (but not support recovery consistency) of the Ridge-like estimator proposed in (3.21). Second, we verify that, given the starting point of any consistent estimator $\tilde{\beta}$ of β^*, a one-step Lasso-like improvement of $\tilde{\beta}$ has support recovery consistency.

We use standard M-estimation theory to show that $\tilde{\beta}$ from (3.21) is consistent for β^*. Define $M_n(\beta)$ as

$$M_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}\left(V_i - \beta^T \hat{U}_i \beta\right)^2 + \lambda_n \left[\frac{1}{n}\sum_{i=1}^{n}\left(\beta_S^T \hat{U}_{i(SS)} \beta_S\right)^2\right], \tag{A.10}$$

and define its expectation $M_0(\beta)$ as

$$M_0(\beta) \equiv \mathbb{E}\left(V - \beta^T \hat{U} \beta\right)^2 + \lambda_n \left[\mathbb{E}\left(\beta_S^T \hat{U}_{(SS)} \beta_S\right)^2\right]. \tag{A.11}$$

As M_n is a sum of quadratic forms, only mild assumptions are necessary to ensure that it has a global minimizer β_0. Then, as we have assumed $\beta \in \mathcal{B}$, where $\mathcal{B}$ is compact, it suffices to show that

$$\sup_{\beta \in \mathcal{B}} |M_n(\beta) - M_0(\beta)| \underset{p}{\to} 0. \tag{A.12}$$

The verification of (A.12) is clear when we decompose M_n:

$$M_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} V_i^2 - 2\beta^T \frac{1}{n}\sum_{i=1}^{n}\left(V_i\hat{U}_i\right)\beta + \left(\beta^T \frac{1}{n}\sum_{i=1}^{n}\hat{U}_i\beta\right)^2 + \lambda_n\left[\frac{1}{n}\sum_{i=1}^{n}\left(\beta_S^T\hat{U}_{i(SS)}\beta_S\right)^2\right]. \tag{A.13}$$

Note that $M_0(\beta)$ has an analogous decomposition. The first term in (A.13) clearly converges to its analogue from $M_0(\beta)$ by the law of large numbers. The convergence of each other piece is guaranteed by Theorem 3.1. Hence we have shown consistency for the Ridge-like estimate $\tilde{\beta}_{\text{ridge}}$ in (3.21).

Now suppose we have any estimator $\tilde{\beta}$ that is consistent for β^*. As discussed in Section 3.3 in the main paper, we can examine the properties of a one-step improvement estimator $\hat{\beta} = \tilde{\beta} + \Delta$ by analyzing $\tilde{M}_n(\hat{\beta})$ from (3.29).

We will prove support recovery consistency using the primal-dual witness method given in Wainwright (2009). The estimator $\hat{\beta}$ minimizes (3.24) if it satisfies the Karush-Kuhn-Tucker (KKT) condition, which for this problem is given by

$$-\frac{1}{n}\sum_{i=1}^{n}\left[2\left(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - 2\tilde{\beta}^T\hat{U}_i\hat{\beta}\right)\hat{U}_{i(j\cdot)}\tilde{\beta}\right] + \lambda_n\hat{r}_j = 0, \qquad j = 1,\ldots,p, \tag{A.14}$$

where

$$\hat{r}_j^T\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot j)}\right)^{-1}\hat{r}_j < 1, \qquad \hat{\beta}_j = 0 \tag{A.15}$$

$$\hat{r}_j = \left[n^{-1}\hat{\Delta}_j^T\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot j)}\hat{\Delta}_j\right]^{-1/2}\left[n^{-1}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot j)}\hat{\Delta}_j\right], \qquad \hat{\beta}_j \neq 0.$$

We construct an oracle primal-dual pair $\left(\hat{\Delta}, \hat{r}\right)$ as follows:

1. Set $\hat{\beta}_j = 0$ for $j \notin S_\sigma^0$.

2. Let

$$\hat{\beta}_{S_\sigma^0} = \underset{\beta_{S_\sigma^0} \in \mathbb{R}^{s_\sigma M_2+1}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\left[V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - \beta_{S_\sigma^0}^T\hat{U}_{i(S_\sigma^0 S_\sigma^0)}\beta_{S_\sigma^0}\right]^2 + \lambda_n\left[\frac{1}{n}\sum_{i=1}^{n}\left(\beta_{S_\sigma}^T\hat{U}_{i(S_\sigma S_\sigma)}\beta_{S_\sigma}\right)^2\right]^{1/2}.$$

3. Define $\hat{r}_{S_\sigma^0} = (0, \hat{r}_{S_\sigma}^T)^T$ as in (A.15).

4. Solve $\hat{r}_j$ from the subgradient condition in (A.14) for $k \notin S_\sigma^0$.

We need to verify the following statements:

$$\max_{j \in S_\sigma^0} \|\hat{\beta}_j - \beta_j^*\|_2 \leq \frac{2}{3}\beta_{\min} \tag{A.16}$$

$$\max_{j \notin S_\sigma^0} \hat{r}_j^T\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot j)}\right)^{-1}\hat{r}_j < 1. \tag{A.17}$$

(A.16) ensures support recovery consistency, while (A.17) ensures strict dual feasibility.

To establish (A.16), we begin by stating a convenient formulation of the subgradient condition for (A.14):

$$-\frac{1}{n}\sum_{i=1}^{n}\left[2\left(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - 2\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\hat{\beta}_{S_\sigma^0}\right)\hat{U}_{i(S_\sigma^0\cdot)}\tilde{\beta}\right] + \lambda_n\hat{r}_j = 0. \tag{A.18}$$

After absorbing the constant 2 into λ_n and adding and subtracting $\frac{2}{n}\sum_{i=1}^{n}\hat{U}_{i(S_\sigma^0\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\beta_{S_\sigma^0}^*$, then rearranging terms, we have

$$\hat{\beta}_{S_\sigma^0} - \beta_{S_\sigma^0}^* = \left(\frac{2}{n}\sum_{i=1}^{n}\hat{U}_{i(S_\sigma^0\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\right)^{-1}(R_{S_\sigma^0} + \lambda_n\hat{r}_{S_\sigma^0}), \tag{A.19}$$

where

$$R_j = \frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\beta_{S_\sigma^0}^* - \frac{1}{n}\sum_{i=1}^{n}\tilde{\beta}^T\hat{U}_{i(S_\sigma^0\cdot)}(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta}). \tag{A.20}$$

We scrutinize (A.19). Condition A.3 states that $\|R_j\|_2 \leq \eta$ for each $j \in S_\sigma^0$. Thus Condition A.3 implies

$$\|R_{S_\sigma^0}\|_2 \leq \eta\sqrt{s_\sigma + 1}. \tag{A.21}$$

Additionally, Condition A.1 implies

$$\Lambda_{\max}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(S_\sigma^0\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\right)^{-1}\right\} \leq \frac{2}{D_{\min}}. \tag{A.22}$$

From the KKT conditions in (A.15) and the almost sure consistency of $\tilde{\beta}$, for $j \in S_\sigma$, we have

$$\|\hat{r}_j\|_2^2 \leq \left[\hat{r}_j^T\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot j)}\right)^{-1}\hat{r}_j\right]\Lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(S_\sigma^0\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\right) \leq 2D_{\max}, \tag{A.23}$$

and we know $\|\hat{r}_0\| = 0$ by construction. Hence we can say

$$\|\hat{r}_{S_\sigma^0}\|_2 = \|\hat{r}_{S_\sigma^0}\|_2 \leq \sqrt{2sD_{\max}}. \tag{A.24}$$

Finally, it follows from (A.19) and (A.21), (A.22), and (A.24) that for each $j \in S_\sigma^0$,

$$\|\hat{\beta}_j - \beta_j^*\|_2 \leq \|\hat{\beta}_{S_\sigma^0} - \beta_{S_\sigma^0}^*\|_2 \leq \frac{2\eta\sqrt{s+1}}{D_{\min}} + \lambda_n\frac{2\sqrt{2sD_{\max}}}{D_{\min}}.$$

Then clearly by Condition A.4 we have verified (A.16).

To prove strict dual feasibility, we begin with the subgradient condition for $j \notin S_\sigma^0$,

$$-\frac{1}{n}\sum_{i=1}^{n}\left[2\left(V_i + \tilde{\beta}^T\hat{U}_i\tilde{\beta} - 2\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\hat{\beta}_{S_\sigma^0}\right)\hat{U}_{i(j\cdot)}\tilde{\beta}\right] + \lambda_n\hat{r}_j = 0. \tag{A.25}$$

After adding and subtracting $\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\beta^*_{S_\sigma^0}$, rearranging terms, and plugging in (A.19) and (A.20), we have

$$\lambda_n\hat{r}_j = \frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(S_\sigma^0\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot S_\sigma^0)}\right)^{-1}(R_{S_\sigma^0}+\lambda_n\hat{r}_{S_\sigma^0}) - R_j. \tag{A.26}$$

Using (A.21) and (A.24) together with Condition A.2, we see from (A.26) that

$$\|\hat{r}_j\|_2 \leq \frac{2\kappa\sqrt{s_\sigma}+1}{\lambda_n}\eta + 2\kappa\sqrt{2sD_{\max}}, \qquad j \notin S_\sigma^0. \tag{A.27}$$

Therefore, applying Condition A.3, we observe that

$$\hat{r}_j^T\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_{i(j\cdot)}\tilde{\beta}\tilde{\beta}^T\hat{U}_{i(\cdot j)}\right)^{-1}\hat{r}_j \leq \frac{2\|\hat{r}_j\|_2^2}{D_{\min}}, \qquad j \notin S_\sigma^0. \tag{A.28}$$

Applying Condition A.4 leads directly to the desired result, (A.17).

Thus we have shown that the oracle primal-dual pair recovers the support of β^* and solves the optimization problem in (3.24). If the optimal solution to (3.24) is unique, then the oracle estimator is the unique estimator. If not, then the null set of any optimal solution should contain S_σ^{0c}, meaning any optimal solution should satisfy the construction of the oracle estimator, as explored in Roth and Fischer (2008). That is, any optimal solution to the optimization problem should recover the correct support S_σ^0.

Furthermore, we note that in Algorithm 3, we may use a finite series of one-step improvements rather than a single one-step improvement. The proof of support recovery consistency trivially holds for such an estimation scheme, as the one-step improvement of a consistent estimator $\tilde{\beta}$ will produce another consistent estimator $\tilde{\beta}'$.

Finally, we note that in Algorithm 3, we scale $\hat{\beta}$ by $\hat{\gamma}$, where $\hat{\gamma}$ is defined in step 3(b). Due to the almost sure convergence of $\tilde{\beta}$ and the argument in the preceding paragraph, we have that $\hat{\gamma}$ converges to 1 almost surely. Furthermore, multiplying $\hat{\beta}$ by a scalar does not change its esti-

mated support. Thus $\hat{\beta}$ from Algorithm 3 is almost surely consistent and almost surely recovers the correct support. □

BIBLIOGRAPHY

Association, A. D. (2016). 11. children and adolescents. *Diabetes Care*, 39(Supplement 1):S86–S93.

Association, A. D. et al. (2018). 15. diabetes advocacy: Standards of medical care in diabetes-2018. *Diabetes care*, 41(Suppl 1):S152.

Baum, A., Scarpa, J., Bruzelius, E., Tamler, R., Basu, S., and Faghmous, J. (2017). Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the look ahead trial. *The Lancet Diabetes & Endocrinology*, 5(10):808–815.

Beck, R. W., Connor, C. G., Mullen, D. M., Wesley, D. M., and Bergenstal, R. M. (2017). The fallacy of average: how using hba1c alone to assess glycemic control can be misleading. *Diabetes Care*, 40(8):994–999.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

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

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.

Burton, H., Sagoo, G. S., Pharoah, P., and Zimmern, R. L. (2012). Time to revisit geoffrey rose: strategies for prevention in the genomic era? *Italian Journal of Public Health*, 9(4).

Carlson, M. G. and Campbell, P. J. (1993). Intensive insulin therapy and weight gain in iddm. *Diabetes*, 42(12):1700–1707.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464.

Channon, S. J., Huws-Thomas, M. V., Rollnick, S., Hood, K., Cannings-John, R. L., Rogers, C., and Gregory, J. W. (2007). A multicenter randomized controlled trial of motivational interviewing in teenagers with diabetes. *Diabetes care*, 30(6):1390–1395.

Chen, G., Zeng, D., and Kosorok, M. R. (2016). Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association*, 111(516):1509–1521.

Chen, J., Fu, H., He, X., Kosorok, M. R., and Liu, Y. (2018). Estimating individualized treatment rules for ordinal treatments. *Biometrics*.

Chen, S., Shojaie, A., and Witten, D. M. (2017). Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112(520):1697–1707.

Cortez, P. and Morais, A. d. J. R. (2007). A data mining approach to predict forest fires using meteorological data. *APPIA Proceedings*.

Cryer, P. E., Davis, S. N., and Shamoon, H. (2003). Hypoglycemia in diabetes. *Diabetes care*, 26(6):1902–1912.

Danne, T., Nimri, R., Battelino, T., Bergenstal, R. M., Close, K. L., DeVries, J. H., Garg, S., Heinemann, L., Hirsch, I., Amiel, S. A., et al. (2017). International consensus on use of continuous glucose monitoring. *Diabetes Care*, 40(12):1631–1640.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Filion, M. (2012). *Quantitative real-time PCR in applied microbiology*. Horizon Scientific Press.

Fu, H., Zhou, J., and Faries, D. E. (2016). Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statistics in medicine*, 35(19):3285–3302.

Group, D. R. et al. (1988). Weight gain associated with intensive therapy in the diabetes control and complications trial. *Diabetes Care*, 11(7):567–573.

Gueorguieva, R., Mallinckrodt, C., and Krystal, J. H. (2011). Trajectories of depression severity in clinical trials of duloxetine: insights into antidepressant and placebo responses. *Archives of general psychiatry*, 68(12):1227–1237.

Hampson, S. E., Skinner, T. C., Hart, J., Storey, L., Gage, H., Foxcroft, D., Kimber, A., Cradock, S., and McEVILLY, E. A. (2000). Behavioral interventions for adolescents with type 1 diabetes: how effective are they? *Diabetes Care*, 23(9):1416–1422.

Henderson, J. and Michailidis, G. (2014). Network reconstruction using nonparametric additive ode models. *PloS one*, 9(4):e94003.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124.

Khoury, M. J., Gwinn, M. L., Glasgow, R. E., and Kramer, B. S. (2012). A population approach to precision medicine. *American journal of preventive medicine*, 42(6):639–645.

Kichler, J. C., Seid, M., Crandell, J., Maahs, D. M., Bishop, F. K., Driscoll, K. A., Standiford, D., Hunter, C. M., and Mayer-Davis, E. (2018). The flexible lifestyle empowering change (flex) intervention for self-management in adolescents with type 1 diabetes: Trial design and baseline characteristics. *Contemporary clinical trials*, 66:64–73.

Kilpatrick, E. S., Rigby, A. S., and Atkin, S. L. (2008). Hba1c variability and the risk of microvascular complications in type 1 diabetes: data from the dcct. *Diabetes care*.

Lehmann, E. and Deutsch, T. (1992). A physiological model of glucose-insulin interaction in type 1 diabetes mellitus. *Journal of biomedical engineering*, 14(3):235–242.

Lu, T., Liang, H., Li, H., and Wu, H. (2011). High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, 106(496):1242–1258.

Maahs, D. M., Daniels, S. R., De Ferranti, S. D., Dichek, H. L., Flynn, J., Goldstein, B. I., Kelly, A. S., Nadeau, K. J., Martyn-Nemeth, P., Osganian, S. K., et al. (2014). Cardiovascular disease risk factors in youth with diabetes mellitus: a scientific statement from the american heart association. *Circulation*, pages CIR–0000000000000094.

Maahs, D. M., Mayer-Davis, E., Bishop, F. K., Wang, L., Mangan, M., and McMurray, R. G. (2012). Outpatient assessment of determinants of glucose excursions in adolescents with type 1 diabetes: proof of concept. *Diabetes technology & therapeutics*, 14(8):658–664.

Mayer-Davis, E. J., Kahkoska, A. R., Jefferies, C., Dabelea, D., Balde, N., Gong, C. X., Aschner, P., and Craig, M. E. (2018a). Ispad clinical practice consensus guidelines 2018: Definition, epidemiology, and classification of diabetes in children and adolescents. *Pediatric diabetes*, 19:7–19.

Mayer-Davis, E. J., Maahs, D. M., Seid, M., Crandell, J., Bishop, F. K., Driscoll, K. A., Hunter, C. M., Kichler, J. C., Standiford, D., Thomas, J. M., et al. (2018b). Efficacy of the flexible lifestyles empowering change intervention on metabolic and psychosocial outcomes in adolescents with type 1 diabetes (flex): a randomised controlled trial. *The Lancet Child & Adolescent Health*, 2(9):635–646.

Monnier, L., Colette, C., and Owens, D. R. (2008). Glycemic variability: the third component of the dysglycemia in diabetes. is it important? how to measure it? *Journal of diabetes science and technology*, 2(6):1094–1100.

Nathan, D. M., Group, D. R., et al. (2014). The diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: overview. *Diabetes care*, 37(1):9–16.

Radloff, L. S. (1977). The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1(3):385–401.

Roth, V. and Fischer, B. (2008). The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th international conference on Machine learning*, pages 848–855. ACM.

Saisho, Y. (2014). Glycemic variability and oxidative stress: a link between diabetes and cardiovascular disease? *International Journal of Molecular Sciences*, 15(10):18381–18406.

Shepard, J. A., Vajda, K., Nyer, M., Clarke, W., and Gonder-Frederick, L. (2014). Understanding the construct of fear of hypoglycemia in pediatric type 1 diabetes. *Journal of pediatric psychology*, 39(10):1115–1125.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Simon, N. and Tibshirani, R. (2012). Standardization and the group lasso penalty. *Statistica Sinica*, 22(3):983.

Song, J., Carey, M., Zhu, H., Miao, H., Ramírez, J. C., and Wu, H. (2018). Identifying the dynamic gene regulatory network during latent hiv-1 reactivation using high-dimensional ordinary differential equations. *International Journal of Computational Biology and Drug Design*, 11(1-2):135–153.

Spencer, S. L., Berryman, M. J., Garcia, J. A., and Abbott, D. (2004). An ordinary differential equation model for the multistep transformation to cancer. *Journal of Theoretical Biology*, 231(4):515–524.

Stekhoven, D. J. and Bühlmann, P. (2011). Missforestnon-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

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

Trusheim, M. R., Berndt, E. R., and Douglas, F. L. (2007). Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nature reviews Drug discovery*, 6(4):287.

Tsybakov, A. B. (2009). Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats.

Van Der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

VanderWeele, T. J. and Knol, M. J. (2011). Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. *Annals of internal medicine*, 154(10):680–683.

Varni, J. W., Seid, M., and Kurtin, P. S. (2001). Pedsql™ 4.0: Reliability and validity of the pediatric quality of life inventory™ version 4.0 generic core scales in healthy and patient populations. *Medical care*, pages 800–812.

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome biology*, 15(2):R37.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.

Wang, Y.-C., Stewart, S. M., Mackenzie, M., Nakonezny, P. A., Edwards, D., and White, P. C. (2010). A randomized controlled trial comparing motivational interviewing in education to structured diabetes education in teens with type 1 diabetes. *Diabetes care*, 33(8):1741–1743.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Wright, L. A.-C. and Hirsch, I. B. (2017). Metrics beyond hemoglobin a1c in diabetes management: time in range, hypoglycemia, and other parameters. *Diabetes technology & therapeutics*, 19(S2):S–16.

Wu, H., Lu, T., Xue, H., and Liang, H. (2014a). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716.

Wu, S., Liu, Z.-P., Qiu, X., and Wu, H. (2014b). Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PloS one*, 9(5):e95276.

Wysocki, T., Buckloh, L. M., Antal, H., Lochrie, A., and Taylor, A. (2012). Validation of a self-report version of the diabetes self-management profile. *Pediatric diabetes*, 13(5):438–443.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Yuan, M. and Wegkamp, M. (2010). Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11(Jan):111–130.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhou, H., Weaver, M. A., Qin, J., Longnecker, M., and Wang, M. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187.

Zhu, R., Zeng, D., and Kosorok, M. R. (2015). Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784.