Statistical Framework for Handling Variable Importance and Missing Data in Electronic Health Records

by
Nickolas Lewis
B.A., New York University, 2019
M.A., Brown University, 2023

A Dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Biostatistics at Brown University

Providence, Rhode Island

February 2026

This dissertation by Nickolas Lewis is accepted in its present form
by the Department of Biostatistics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Joseph W. Hogan, Advisor

Recommended to the Graduate Council

Date _____ _____

Arman Oganisian, Reader

Date _____ _____

Rami Kantor, Reader

Date _____ _____

Jon Steingrimsson, Reader

Approved by the Graduate Council

Date _____ _____

David Lindstrom, Dean of the Graduate School

Curriculum Vitae- Nickolas A. Lewis

Email: nickalewis1@gmail.com GitHub: NickolasLewis
Phone: (504) 432-4492 LinkedIn: nickolasalewis

EDUCATION

2019 – 2025 **Brown University** **Providence, RI**
Ph.D. in Biostatistics, December 2025
Thesis: Statistical Framework for Variable Importance and Missing Data in Electronic Health Records
Advisor: Joseph W. Hogan, ScD

Brown University **Providence, RI**
M.A. in Biostatistics, May 2023
M.A. in Applied Mathematics, May 2023

New York University **New York, NY**
B.A. in Mathematics with honors, May 2019

RESEARCH EXPERIENCE

Statistical framework for clinical decision support *Brown University | Sept 2020 – Present*

- Designed a principled statistical framework to predict retention status at future return visits for people living with HIV in Western Kenya who are enrolled in HIV care. The framework accurately captures the data generating process at each scheduled return visit
- Development of multinomial prediction framework through probabilistic chained classifiers to account for competing risk. Implementation of pattern-mixture model framework to account for missing data in the covariate space.
- Implementation of sampling strategy that accurately reflects visits and patients currently enrolled in care.

Quantifying unit-level variable importance *Brown University | Sept 2020 – Present*

- Developed methodology to quantify the impact of individual covariates on driving a high risk score for unit-level explanations in binary outcome models.
- Implementation and development of a fast, scalable sampling method that draws directly from the appropriate conditional distribution based on predictive mean matching from the missing data literature.
- Evaluated performance of sampling method by defining joint covariate distributions with sufficient nonlinear and interaction terms across different covariate types.

Variable Importance with Missing Data *Brown University | Sept 2020 – Present*

- Conducted a large-scale literature review of missing data methods for prediction (multiple imputation, regression imputation, pattern submodels) when missing covariate data is present at the model development and model deployment stage.
- Modified existing global level variable importance methods such as dropped variable importance (i.e. Leave COvariates Out (LOCO)) to account for the contribution of observed covariate data.

Survival Analysis on COVID-19 Reinfection *Brown University | July 2021 – July 2022*

- Graduate Biostatistician for statewide (Rhode Island) study to estimate the effectiveness of vaccination in preventing COVID-19 reinfection in comparison to natural immunity funded by the Rhode Island Department of Health.
- Analyzed longitudinal COVID-19 reinfection rates from across the state from over 95,000 residents using stratified Cox models to estimate vaccine efficacy in three populations of interest: Long term congregate care (LTCC) residents, LTCC employees and the general population, and Kaplan-Meier estimates to estimate effectiveness 30-, 60- and 90- days post secondary infection.
- First authored and submitted accepted manuscript for publication at JAMA Open Network that showed that completion of the primary vaccination series was associated with 49% protection from reinfection among LTCC residents, 47% protection among LTCC employees, and 62% protection in the general population during periods when wild type, Alpha, and Delta strains of SARS-CoV-2 were predominant.

PUBLICATIONS

Published

Lewis N, Chambers LC, Chu HT, et al. Effectiveness Associated With Vaccination After COVID-19 Recovery in Preventing Reinfection. JAMA Netw Open.

Newbolt, J., **Lewis, N.**, Bleu, M., Ramananarivo, S., Ristroph, L. Flow interactions lead to self-organized flight formations disrupted by self-amplifying waves. Nature Communications

Chambers LC, Chu HT, **Lewis N**, Kamat G, Fortnam T, De Vito R, Chan PA, Lasher L, McDonald J, Alexander-Scott N, Hogan JW. Effectiveness of Monoclonal Antibody Therapy for Preventing COVID-19 Hospitalization and Mortality in a Statewide Population. RI Med J 2022; In press.

Working

Lewis, N., Sang, E., Oganisian, A., Fraser, H., Kimaina, A., Kantor, R., Hogan, J., *Quantifying unit-level feature importance in machine learning in terms of conditional risk score distributions*.

Lewis, N., Sang, E., Oganisian, A., Fraser, H., Kimaina, A., Kantor, R., Hogan, J., *Statistical Framework for developing and implementing models for clinical decision support*.

PROGRAMMING

Advanced R, Python, LaTeX

Intermediate C++, PyTorch/Tensorflow

Familiar SQL, Stan, GitHub, MATLAB

TEACHING EXPERIENCE

2023 Jan – May **Brown University**
Applied Longitudinal Analysis Teaching Assistant
∘ Responsible for grading homework assignments, midterm and final examinations, and hosting weekly office hours.

2022 Jan – May **Brown University**
Bayesian Statistical Methods (PhD Level Course) Teaching Assistant
∘ Responsible for grading homework assignments, midterm and final examinations, and hosting weekly office hours.

AWARDS

2019 Initiative to Maximize Student Development NIH Grant
Brown University School of Public Health

2019 University Honor's Scholar
New York University

PRESENTATIONS

2025 **International Workshop on HIV and Hepatitis Observational Databases (IWHOD)**

March *Toledo, Spain*

∘ Poster Presentation (Accepted) *Quantifying unit-level feature importance in machine learning in terms of conditional risk score distributions.*.

2025 **Eastern North American Region (ENAR)**

March *New Orleans, LA*

∘ Poster Presentation (Accepted) *Quantifying unit-level feature importance in machine learning in terms of conditional risk score distributions.*.

2024 **International Workshop on HIV and Hepatitis Observational Databases (IWHOD)**

March *Vilamoura, Portugal*

∘ Poster Presentation (Accepted) *Statistical Framework for developing and implementing models for clinical decision support*.

2023 **Joint Statistical Meeting**

Aug *Toronto, ON, Canada*

∘ Oral Presentation of *Statistical Framework for developing and implementing models for clinical decision support*.

Acknowledgements

As I reflect on my academic and professional pursuits that have culminated in this dissertation, I can't help but think of the many people who have made achieving my goals possible and, at the same time, a truly positive experience. There are countless individuals that deserve some share of the credit for this achievement, and I can only hope to have mentioned all of them.

I am grateful for the tremendous amount of mentorship and guidance I have received from my dissertation advisor, **Joseph Hogan**. He always asked me thought-provoking questions that emphasized the depth and relevance of my research which taught me to truly think critically about public health problems. His training helped me develop into a rigorous and thoughtful statistician. His positive outlook, unwavering support, and genuine kindness transformed the dissertation process into a truly rewarding experience that has shaped my development as a genuine scholar.

I would like to thank my dissertation committee members, **Arman Oganisian**, **Jon Steingrimsson**, and **Rami Kantor**, for their time, insight, and support. Arman often served as a second advisor, and I am especially grateful for his thoughtful guidance and intellectual rigor. Jon has been an outstanding teacher and mentor who helped make the School of Public Health a welcoming and intellectually stimulating environment. Rami consistently emphasized the clinical relevance and real-world impact of my work, which was invaluable in shaping the focus of this dissertation.

I am grateful to the faculty, staff, and students of the Department of Biostatistics for fostering a supportive and enjoyable doctoral experience. In particular, I would like to thank my cohort-mates, **Patrick Gravelle** and **Anthony Sisti**, whose friendship made this journey especially meaningful. I would also like to thank **Robert Medeiros** for his consistent help with technological issues and being a kind presence around the school of public health.

This dissertation would not have been possible without the contributions of my collaborators at AMPATH, **Edwin Sang**, **Ann Mwangi** and **Allan Kimaina**, and at Brown, **Hamish Fraser** and **Allison Delong**. I am especially thankful to **Edwin Sang** for managing the AMRS

data collection, which formed the foundation of this work.

Finally I would like to extend the greatest thanks to my friends and family for always supporting and encouraging me. To my Mom, **Kristy**, and my Dad, **Alex**, whose sacrifices and unwavering commitment to my education made this achievement possible. Their belief in me has shaped my confidence and determination throughout my life. To my best friends **Landon Stein**, **Ali Rahman** and **Jackie Li**, and my friends **Dana Guan** and **Lizzie Lee**, who are some of the greatest people I've had the pleasure of encountering during my life. I am eternally grateful for their consistent help and support, and to have them in my life.

Contents

List of Tables

List of Figures

Introduction

Consistent medical care is essential for the health of people living with HIV (PLWH). Individuals living with HIV who receive regular medical care are more likely to have access to antiretroviral therapy (ART), less likely to develop Acquired Immune Deficiency Syndrome (AIDS), and have improved survival rates compared to HIV-positive individuals who do not receive regular medical care. Certain population-level health benchmarks can best be understood through the HIV care cascade. The HIV care cascade is a conceptual model describing key benchmarks for PLWH, and consists of (a) HIV diagnosis through testing, (b) linkage to care, (c) engagement and retention in care, (d) initiation of ART through retention, and (e) sustained suppression of viral load. The cascade provides a framework for monitoring and evaluating key care benchmarks for PLWH and monitoring the effectiveness of HIV healthcare systems while also providing a guideline to identify poor adherence metrics. The objective of this dissertation is to develop and implement data-driven tools to aid health-related decision making at patient, clinic and district levels, and evaluate the efficacy of using these methods.

The increasing availability of large-scale patient level cohort data is providing new opportunities for data-driven analyses of the care cascade. This dissertation investigates statistical issues of modeling retention in HIV care using electronic health record (EHR) data from the Academic Model Providing Access to Healthcare (AMPATH), a program delivering HIV care in Western Kenya. Clinical visit information from individuals who are engaged in medical care within AMPATH is stored in an electronic database known as the AMPATH Medical Record System (AMRS) which is generated by a browser-based front-end application deployed at the

point of care in over 300 facilities (hospital and clinics). Twice each year, raw data from the AMRS are formatted into a research-grade database that can be used to investigate various clinical and epidemiologic questions related to patient care. Information from AMRS has been used to monitor and evaluate comprehensive intervention and treatment programs in Western Kenya.

The objective of this dissertation is to develop a statistical framework for modeling retention in care that enables meaningful assessment of variable importance in clinical decision making while also explicitly accounting for missing data. Currently, AMPATH dedicates considerable resources to patient outreach following a missed visit - including follow up calls and at-home visits; however, the human resources used for outreach are not unlimited. Proactive outreach conducted prior to a scheduled return visit may increase the likelihood of continued engagement in care which is critical for patients to stay current with their ARV medications, motivating the development of models that predict a patient's risk of missing an upcoming scheduled clinic visit. These models can inform outreach strategies by identifying which patients should be prioritized for advanced outreach. The effectiveness of advanced outreach may be further enhanced if clinicians and outreach workers understand which patient-level factors drive "high risk" predictions and how these factors contribute to the model's assessment of missed-visit risk. Below we describe the three specific aims of this dissertation in greater detail,

1. **Aim 1: Develop a statistical framework for modeling the retention in care process for clinical decision support**

   Keeping patients engaged in HIV care is a significant goal for the clinics that comprise AMPATH. Regular clinical visits provide uninterrupted access to anti-retroviral drugs (ARVs), consistent monitoring of disease progression, and promote medication compliance. The Kenyan Ministry of Health (MOH) has defined several clinically meaningful benchmarks for engagement in care for programmatic reporting such as a default (missing by a day), an interruption in treatment (missing by more than 28 days) and a loss to follow-up (missing by more than 90 days). An interruption in treatment is of clinical significance

since ARVs are prescribed to last around 28 days; if a patient has not shown up to clinic within this time frame it is assumed they are no longer taking their life-saving medications. Loss to follow-up may indicate complete disengagement from care, unreported mortality, or an unreported transfer out of care. AMPATH dedicates considerable resources to patient outreach following a missed visit - including follow up calls and at-home visits; however, the human resources used for outreach are not unlimited. Developing strategies to target patients both before and after scheduled visits can improve the efficiency of outreach efforts and increase the likelihood of timely return to care. Thus the goal for the first chapter of this dissertation is to propose a statistical framework for modeling the visit return process that equips clinicians and outreach workers with a tool to determine which patients should be prioritized for advanced outreach prior to their next scheduled clinic visit.

The purpose of this model is to provide AMPATH outreach workers with a tool to identify patients classified as "high risk" of missing their next scheduled visit and enable targeted outreach prior to the scheduled return date. Several challenges arise, however, in building a principled framework for predicting retention in care. Missingness can occur in the outcome space and covariate space. Competing risks such as death that precludes a missed visit can prevent a patient's retention status at the next visit from being known. Defining an appropriate analysis sample from historical EHR data that accurately reflects the patients currently enrolled in AMPATH care is also nontrivial. To address these challenges we propose a flexible Bayesian multinomial framework to account for competing risks, along with a pattern-mixture modeling approach to address missingness in the covariate space. In addition, we present visualizations of global trends derived from the appropriate conditional distributions, which may assist clinicians in identifying patients to prioritize for advanced outreach. By enabling identification of patients at high risk for disengagement, this framework allows for outreach workers to provide targeted, proactive outreach that can both optimize the use of limited clinical resources, and potentially

reduce preventable interruptions in HIV treatment.

2. **Aim 2: quantify local variable importance for machine learning models in terms of conditional risk score distributions**

   The second aim of this dissertation builds on the pre-visit outreach strategy proposed in the first aim. Advanced outreach can be more effective if the outreach workers are aware of which variables the model uses to generate a high-risk score for a specific patient. Knowledge of these specific factors can provide outreach workers in clinics a way to develop strategies and personalize the outreach, and potentially prevent patients from missing their scheduled return visits. At the individual level, machine learning (ML) algorithms produce a risk score, but typically do not provide individual level insights about which covariates are driving the subject's high risk score. Local variable importance measures address this gap by explaining how specific covariates contribute to an individual's risk score, in contrast to global feature importance measures that quantify population-level trends in the data. Calculating these local importance measures generally requires computing a statistic (e.g. expected value, median) from an appropriate conditional distribution. Popular local variable importance measures like SHAPley Additive exPlanations (SHAP) provide local explanations by making simplifying assumptions such as covariate independence and then sampling from the empirical marginal distribution. Other methods assume the joint distribution of the data is multivariate Gaussian and then derive the appropriate conditionals.

   The second chapter of this dissertation seeks to design and implement a method for obtaining local variable importance as well as a method to sample from the appropriate conditional distributions. Both the unit-level variable importance method and sampling method may be applied to any machine learning model. Specifically, this aim introduces a conditional sampling method based on predictive mean matching, an imputation technique from the missing data literature that is scalable and robust to model mis-

specification. We simulate joint distributions where we induce sufficient interactions and nonlinearities in the covariate space to demonstrate the utility of this sampling method over the marginal and Gaussian approximation, and then apply the method to the stratified visit models built from AMPATH data from the first chapter. By enabling reliable, patient-specific interpretation of complex prediction models this work supports more informed clinical decision-making and targeted intervention strategies.

3. **Aim 3: Investigate the role of variable importance in prediction models when covariates have missing entries**

   The third aim of this dissertation builds on the model developed in the first chapter and focuses on global variable importance, in contrast to the local variable importance methods discussed in the second chapter, particularly in the presence of missing data. Missing data are ubiquitous, particularly in health-related settings where regular collection of key clinical measurements can be expensive. This can present several unique challenges when operationalizing EHR data to design and validate prediction models for the purpose of clinical decision making. We distinguish between the different sets of data in the prediction pipeline. There is the development data that is used to build the ML model and the deployment data which consists of individuals that the ML model will be applied to in order to forecast some quantity of risk of an event of interest. There are four general cases where missingness can occur in the development of a statistical model, specifically missingness in the covariate space.

   The focus of the third chapter of this dissertation is the situation where missingness in the covariate space occurs both at model development and model deployment. When missingness is present in both the development and deployment data the notion of variable importance can be obfuscated due to the need to handle missingness at the point of prediction. In this setting, imputation models are part of the prediction pipeline. As a result, when variable importance methods are applied at both the local and global level, the ob-

served covariate information in specific missingness patterns provides greater predictive utility. To address this, we use a popular model-agnostic variable importance technique known as Leave Out COvariates (LOCO) that measures the loss in predictive accuracy between a full model built using all of the covariate data and a reduced model developed on the covariate data with a specific covariate omitted. We modify LOCO to take into account missing data and then demonstrate discrepancies between the expected LOCO with full data available and LOCO with missingness in the covariate space. By explicitly accounting for missingness in the covariate space at both model development and deployment, this work produces variable importance estimates that more accurately reflect how ML models use covariate information to generate predictions in real-world clinical data, thereby improving the reliability of model-guided decisions about which factors elevate a patient's risk of missing a visit and how to prioritize interventions.

Chapter 1

Statistical Framework for developing and implementing models for clinical decision support

1.1 Introduction

Consistent medical care is essential for the health of people living with HIV (PLWH). The Joint United Nations Programme on HIV/AIDS (UNAIDS) recently announced the new global 95-95-95 (Ehrenkranz et al., 2021) targets for eradicating HIV worldwide, which calls for diagnosis of 95% of individuals who have HIV, initiating antiretroviral (ART) treatment for 95% of those who have been diagnosed, and achieving suppression of viral load (VL) for 95% of those who are on treatment. HIV-positive individuals who receive regular medical care are more likely to receive antiretroviral therapy, less likely to develop Acquired Immune Deficiency Syndrome (AIDS), and have improved survival rates compared to HIV-positive individuals who do not receive regular medical care. Despite tremendous progress toward achieving these objectives, challenges remain.

The work presented addresses the link between the second 95 and the third 95 of the UNAIDS goal, specifically as it relates to keeping PLWH retained in HIV care. Retention is a

necessary condition for maintaining patients on antiretroviral drugs (ARVs). In Kenya at the Academic Model Providing Access to Healthcare (AMPATH), viral load testing for adult clients is typically done six months after treatment initiation and annually thereafter. Even after a measured viral load indicating suppression, viral failure due to nonadherence or drug resistance can occur well before the next follow up one year later. Regular clinical visits promote regular access to ARVs, consistent monitoring of disease progression and medication compliance. This makes retention in care fundamental in reaching the last 95 goal, viral suppression.

Currently, AMPATH dedicates considerable resources to patient outreach following a missed visit - including follow up calls and at-home visits; however, the human resources used for outreach are not unlimited. This motivates the development of strategies that target disengagement in advance, specifically the development of prediction models that can identify patient's who are at "high risk" of missing their next scheduled visit. In recent years advances in machine learning (ML) models have opened up new avenues for prediction in health outcome settings and have gained considerable traction in predicting retention in HIV care (Ogbechie et al., 2023, Olatosi et al., 2021, Oliwa et al., 2021, Ramachandran et al., 2020, Ridgway et al., 2022). These models can guide outreach efforts in AMPATH clinics by identifying patients at "high risk" of missing their next visit. Identifying these high risk patients and tailoring pre-visit outreach is critical to keeping patients engaged in care so that they may stay current with their antiviral medications and so that clinicians can regularly monitor key biomarkers such as viral load and prevent further clinical progression and ensure it is identified and addressed if it occurs.

In this work we design a statistical framework for predicting retention in care for the purposes of targeting patients for advanced outreach ahead of their scheduled return visit. We develop prediction models around this framework that will be used to quantify individual patient risks of missing their next visit to inform who should be targeted for advanced outreach. We illustrate our approach through the analysis of data from 70,626 individuals enrolled in HIV care in AMPATH.

1.1.1 Previous Work

Modeling retention in HIV care through machine learning (ML) has received considerable attention. Existing work has focused on utilizing electronic health records (EHR) to build prediction models to (1) identify patients who have fallen out of care, and (2) target patients before they disengage from care. Patients can either fall out of care or do "silent transfers" (transferring out of care or to a different facility without notifying their current facility). Targeting patients before they disengage aligns with the goals of the work presented in this paper. A large spectrum of machine learning models were chosen to cover a possible predictive models over logistic regression, tree-based and boosting models (Random Forest, XGBoost), and natural language processing models. To date none of the papers appear to have deployed their models and reported on the efficacy on real time data. Ogbechie et al. (2023) integrated their ML models into the point of care, but have yet to report results.

At the same time there exists many definitions of retention in care. examples include attending at least 2 HIV care visits Ramachandran et al. (2020), Ridgway et al. (2021), having two CD4+ cell counts or viral load tests greater than 90 days apart within a 12-month period Olatosi et al. (2021), Semerdjian et al. (2018), having a single HIV care visit within a 6-month period Ramachandran et al. (2020), Oliwa et al. (2021), and having a single HIV care visit within a 30 day window from the scheduled visit Ogbechie et al. (2023). Lee et al. (2018) provided a multistate framework stratified by 200 day intervals to model the process of engagement in care though said model was built to study the transition framework within AMPATH. The Kenyan Ministry of Health (MOH) characterizes those who miss their visit by a day as experiencing a default in treatment, those who miss by 28 days as experiencing an interruption in treatment (IIT), and those who miss by 90 days or more as being loss to follow up (LTFU) (Kenya, 2022). IIT's are of clinical significance because antiviral medication prescriptions are frequently given in 28-day supplies; if a patient has not returned to clinic within this time frame it is assumed they are no longer taking their life-saving medications. A loss to follow up is usually indicative of a patient who has completely disengaged from care or, in

some cases, unreported mortality.

The rest of the chapter is organized as follows: Section 1.2 provides details about the AMPATH cohort and provides motivation behind the paper. Section 1.3 presents statistical challenges associated with prediction models built using EHR data. Section 1.4 provides details of the model specification and methods to address challenges raised in section Section 1.3. The application of our framework to the AMPATH data is presented in section Section 1.5 and Section 1.6 concluded the paper with a discussion of future directions and extensions.

1.2 Data Source & Goals

Clinical visit information from individuals who are engaged in medical care within AMPATH is stored in an electronic database known as the AMPATH Medical Record System (AMRS) which is generated by a browser-based front-end application deployed at the point of care in aasociated facilities (hospital and clinics) Tierney et al. (2007). Twice each year, raw data from the AMRS are formatted into a research-grade database that can be used to investigate various clinical and epidemiologic questions related to patient care.

Once a patient is linked to HIV care in AMPATH a set of immunological, clinical, and demographic markers are recorded at enrollment and subsequent visits. Important markers of HIV disease progression such as CD4 count is usually measured at the enrollment visit, and viral load is measured every six months or every year depending on the disease progression status of the patient Kenya (2022).

At the end of each visit a patient is then prescribed two return to care (RTC) dates; a date for which they should return to clinic for symptom monitoring and a date for which they should return to clinic to pick up their ARV prescription medicine. Often times these two can coincide. Follow up visits are usually scheduled every 1-28 weeks depending on the clinical condition of the patient with visits following initial enrollment being scheduled within shorter time frames (e.g. every 1-4 weeks). Patients in good status are eligible to be put in the differentiated care program where they are given longer RTC dates.

The overall goals of this analysis are to (1) design a principled framework for using patient level data derived from the AMRS to model the part of the HIV cascade in AMPATH that relates to retention in care and (2) generate accurate predictions of defaulting and experiencing an IIT and ultimately prioritize those who should be called by an outreach worker in advance of their visit. The end-goal is to use these prediction models to intervene on patients who are at high risk of missing their next scheduled visit with some pre-visit outreach in hopes of increasing their likelihood of returning to care.

1.3 Statistical Challenges for clinical prediction models for EHR Data

The task of modeling retention in care presents several key challenges. Although electronic health records (EHR) present rich opportunities for analyses, data is collected for record-keeping rather than research. As one of the largest providers of medical care in sub-Saharan Africa, AMPATH provides other medical services in addition to HIV/AIDS care. Consequently not all records within the AMRS are relevant for prediction and construction of the sampling data must be done with care and reflect the goals of the analysis. The analysis sample built from the historic EHR data must accurately reflect the current population of patient's enrolled into HIV care at AMPATH. It is tempting to use all available data to make predictions however this can cause issues if the sampled visits and patients are not aligned with the goals of the analysis.

The collection of covariate information can occur due to human error on the part of busy clinicians, a covariate falling out of use or the covariate only being measured on a fixed schedule. Missingness in both the outcome and covariate space presents another distinct challenge. When modeling retention as the outcome the retention time can be subject to right-censoring due to database closure or a competing risk that precludes a patient from making their next scheduled visit. Previous work has removed patients with incomplete outcome information

due to competing risk (Fentie et al., 2022, Nshimirimana et al., 2022). Information on key covariates is also subject to missingness since key clinical measurements such as CD4 count and viral load are expensive to measure and thus are sporadically and sparsely measured. The longitudinal nature of the AMRS implies missingness on certain key covariates is propagated throughout a patient's history, which potentially makes the missingness an informative predictor of clinical outcomes (Van Ness et al., 2023).

These challenges require us to address the operationalization of the cascade, data preparation, visit sampling and specification of the model. To address some of these challenges, we propose using probabilistic estimation models stratified by visit number and motivated by well-defined binary endpoints that meet the retention baselines defined by the Kenyan Ministry of Health mentioned in Section 1.1.1. We illustrate our approach through the analysis of data from 70,626 individuals enrolled in HIV care in AMPATH.

1.4 Framework for Modeling Retention in Care

1.4.1 Notation and Definitions

Clinical visit information from individuals who are engaged in medical care within AMPATH is stored in an electronic database known as the AMPATH Medical Record System (AMRS) which is generated by a browser-based front-end application deployed at the point of care in associated facilities (hospital and clinics) Tierney et al. (2007). For the purposes of model specification patients are indexed by i and visits by ℓ. For every patient i we observe L_i recorded visits indexed by $\ell = 0, \ldots, L_i$, where L_i is subject-specific. We define $L \coloneqq max_i L_i$ to be the maximum visit number observed within the data. Once a patient is linked to HIV care in AMPATH, a set of immunological, clinical, and demographic covariates are recorded at enrollment and at subsequent visits;denote these by $\{\boldsymbol{V}_\ell : \ell = 0, 1, 2, \ldots\}$, where $\boldsymbol{V}_\ell$ is a $J \times 1$ vector. Because certain markers of disease progression such as CD4 count and viral load are usually measured on a fixed schedule, we define a corresponding $J \times 1$ vector $\boldsymbol{M}_\ell$ of binary indicators for whether the each component of $\boldsymbol{V}_\ell$ has been observed. For completeness we define $\boldsymbol{M}_\ell$ for

every covariate in $\boldsymbol{V}_\ell$ but in practice only a subset are used in modeling.

At the end of visit ℓ, a patient is assigned a return to care (RTC) date for a regular return visit or a medication pickup; often times the two may coincide. The time between visit ℓ and the RTC date for visit $(\ell+1)$ is defined by $S_{\ell+1}$. The endpoint for our prediction model is a binary indicator $Y_{\ell+1}$ that takes value 1 if the patient fails to return before their RTC date, and 0 if they do return on time; i.e., $Y_{\ell+1}$ is a 'missed visit' indicator. We can formalize this as follows.

For a patient at visit ℓ, let $W_{\ell+1} > 0$ denote the waiting time until that patient shows up for their encounter at visit $\ell+1$, and let $\Delta \geq 0$ denote a 'grace period' interval. A censoring indicator $\delta_\ell = \{1, 0, -1\}$ is defined to denote whether a patient has survived up until visit ℓ, was administratively censoring (i.e. censoring due to database closure) or died before their next scheduled visit, whichever comes first. This motivates the definition of an intermediate outcome $D_{\ell+1}$ between return to care where $D_{\ell+1} = 1$ denotes survival to visit $\ell+1$ and $D_{\ell+1} = 0$ denotes death before visit $\ell+1$. Our binary missed-visit indicator, now annotated by Δ, is defined as $Y^{\Delta}_{\ell+1} = \mathbb{I}(W_{\ell+1} > S_{\ell+1} + \Delta)$. We use this formulation because specific values of Δ have contextual relevance for programmatic reporting: setting $\Delta = 0$ flags 'defaulters', or those who do not make their visit on or before the assigned date; $\Delta = 28$ flags those at risk for interruption in treatment, because antiviral medication prescriptions are frequently given in 28-day supplies; $\Delta = 90$ flags those who are at risk for loss to follow up or possibly mortality. This Δ-retention framework provides flexibility for modeling alternative definitions of retention and facilitates clinically meaningful interpretation of risk predictions, and can be extended from prediction settings to causal settings (Oganisian et al., 2024).

Figure 1.1 details four patient trajectories within the AMRS. Patient A and Patient C enroll into care and are consistently late to their scheduled return visits before finally experiencing a censoring event, death and loss to follow up, respectively. Patient B is consistently on time, and on one occasion early, for their scheduled return visits and is censored due to database closure. Patient D makes their last recorded scheduled return visit, but transfers out of care at this visit.

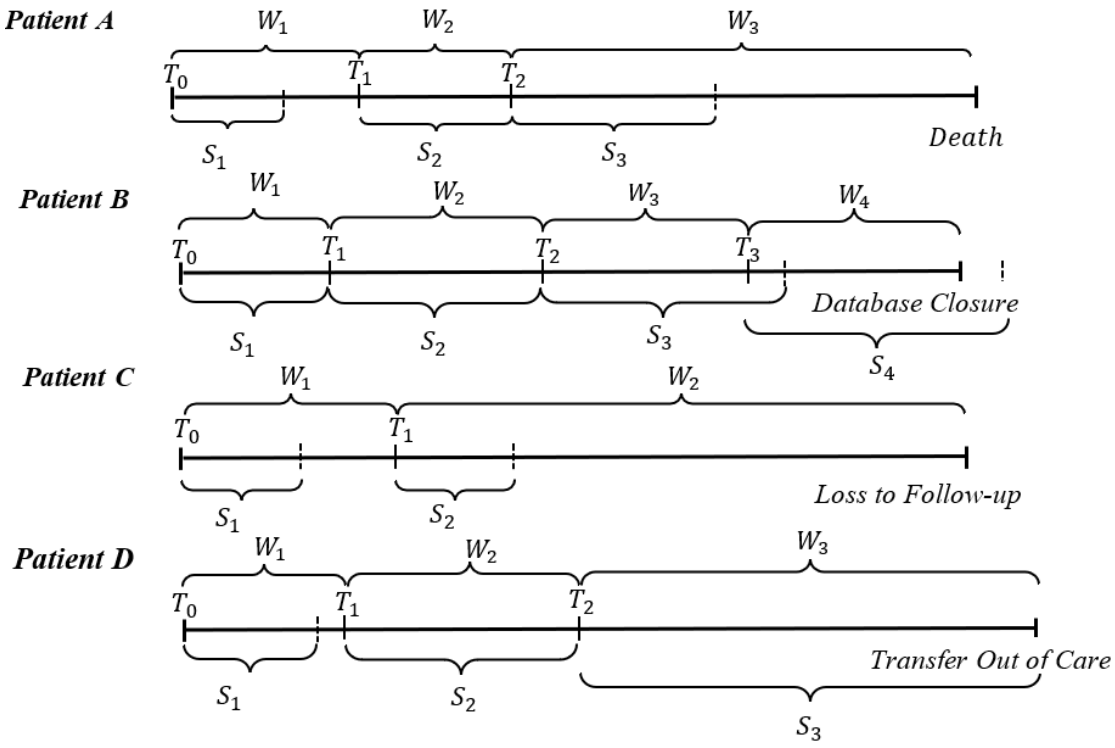

Figure 1.1: Examples of four specific patient trajectories in the AMRS. T_ℓ is defined as the time from enrollment to the current visit ℓ with $T_0 := 0$.

Covariate and retention information available at visit ℓ can be summarized as $\boldsymbol{X}_\ell = (\boldsymbol{V}_\ell, \boldsymbol{M}_\ell, \boldsymbol{W}_\ell, Y_\ell, S_{\ell+1})$. All clinical and retention information up to and including visit ℓ is denoted by $\overline{\boldsymbol{X}}_\ell = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_\ell)$ using overbar notation. In all notations, we denote the random variables with capital letters and observed values of the random variable with corresponding lowercase letters.

1.4.2 Modeling

Our goal is to estimate the probability

$$\pi(\overline{\boldsymbol{X}}_\ell) = Pr(Y^{\Delta}_{\ell+1} = 1 \,|\, \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1)$$

This is the strata of the population with covariate data $\overline{\boldsymbol{X}}_\ell$ who have been retained in care past visit ℓ (i.e. $\bar{\delta}_\ell = 1$). The outcome $Y^{\Delta}_{\ell+1}$ is only defined for patients who survive to visit $(\ell+1)$. Individuals who die or are administratively censored before their scheduled visit have an un-

defined value of $Y^{\Delta}_{\ell+1}$. In some cases, however, censoring is informative. Patients who fail to return within Δ days and are then censored, regardless of the censoring reason, provide information about the outcome, specifically that $Y^{\Delta}_{\ell+1} = 1$. In our analysis, administrative censoring is assumed non-informative conditional on $\overline{\boldsymbol{X}}_\ell$; thus, we exclude records where informative censoring could occur. To handle the competing risk of death we introduce a multinomial framework similar to Lee et al. (2018) through the use of probabilistic classifier chain to account for the competing risk of death before the next visit. Our expansion of the original probability then becomes,

$$\begin{aligned} Pr(Y^{\Delta}_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1) &= \sum_{d\in\{0,1\}} Pr(Y^{\Delta}_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1, D_{\ell+1} = d) Pr(D_{\ell+1} = d \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1) \\ &= Pr(Y^{\Delta}_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1, D_{\ell+1} = 1) Pr(D_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1) \end{aligned}$$

This simplification follows since the probability $\Pr(Y^{\Delta}_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, D_{\ell+1} = 0)$ is identically zero; i.e., a patient who dies before their scheduled visit cannot miss their scheduled visit.

Conditional on visit history $\overline{\boldsymbol{X}}_\ell$ and having survived past visit ℓ (i.e. $\bar{\delta}_\ell = 1$), we assume that both retention status and survival follow Bernoulli distributions $[Y^{\Delta}_{\ell+1} \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1, D_{\ell+1} = 1] \sim \text{Ber}\left\{\pi_1(\overline{\boldsymbol{X}}_\ell)\right\}$ and $[D_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1] \sim \text{Ber}\left\{\pi_2(\overline{\boldsymbol{X}}_\ell)\right\}$ respectively, where $\pi_1(\overline{\boldsymbol{X}}_\ell) = \Pr(Y^{\Delta}_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1, D_{\ell+1} = 1)$ and $\pi_2(\overline{\boldsymbol{X}}_\ell) = \Pr(D_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell, \bar{\delta}_\ell = 1)$. More specifically,

$$\begin{aligned} \pi_1(\overline{\boldsymbol{X}}_\ell; \boldsymbol{\theta}_{1\ell}) &= g^{\Delta}_{1,\ell+1}(\overline{\boldsymbol{X}}_\ell, \boldsymbol{\theta}_{1\ell}) \\ \pi_2(\overline{\boldsymbol{X}}_\ell; \boldsymbol{\theta}_{2\ell}) &= g^{\Delta}_{2,\ell+1}(\overline{\boldsymbol{X}}_\ell, \boldsymbol{\theta}_{2\ell}), \end{aligned}$$

where $\boldsymbol{\theta}_\ell$ is a finite-dimensional parameter with prior distribution $\boldsymbol{\theta}_\ell \sim p(\boldsymbol{\theta}_\ell) = p(\boldsymbol{\theta}_{1\ell}, \boldsymbol{\theta}_{2\ell})$, for $\ell = 1, \ldots, \max_i L_i$. The probabilistic chained classifier can be further generalized to categories greater than 3 (Sparapani et al., 2021, Martin et al., 2021) and has substantial flexibility as it can be applied to any machine learning model that can handle binary outcomes and the functional form $g(\cdot)$ may differ by model. Previous work has shown similar operating charac-

teristics to a multinomial logistic regression (Martin et al., 2021).

In AMPATH the frequency of observation times are heterogeneous within and between individuals, and key clinical information can be measured sporadically and at irregular times. To account for this we create separate models for visits $\ell = 1, 2, 3$ and a fourth model that combines all visits $\ell \geq 4$. We further make the assumption that a patient's status at visit $\ell + 1$, conditional on $\overline{\boldsymbol{X}}_\ell$, follows a third order Markov assumption where the most recent retention status depends only on the most recent clinical information and the retention history of the three most recent visits. In other words the covariate history up to visit ℓ simplifies to $\overline{\boldsymbol{X}}_\ell = (\boldsymbol{V}_\ell, \boldsymbol{M}_\ell, S_{\ell+1}, Y_\ell^\Delta, Y_{\ell-1}^\Delta, Y_{\ell-2}^\Delta)$. For earlier visit models, the conditioning set is truncated accordingly: no retention history is included for $\ell = 1$, only Y_1^Δ is included for $\ell = 2$, and Y_1^Δ, Y_2^Δ are included for $\ell = 3$. This visit-specific stratification is analogous to a stratified pattern submodel formulation (Fletcher Mercaldo and Blume, 2020). Missing a visit depends heavily on duration in care, with highest risk of missed visit occurring closest to enrollment. Previous work (Ridgway et al., 2022, Ogbechie et al., 2023) also suggest that a patient's clinical retention history is a strong predictor of future retention.

Missing Data in the outcome and covariate space

Missing data are ubiquitous in healthcare settings and can occur due to human error or a clinical measurement being expensive to measure. In the previous section we detailed methodology to account for the competing risk in the outcome space such as death that precludes a visit. We make the assumption that conditional on data $\overline{\boldsymbol{X}}_\ell$ administrative censoring is non-informative and thus drop records where this may occur. Missing data in the covariate space can be handled either through imputation (Little and Rubin, 2019) or inclusion of the missingness process in the model via a pattern mixture model approach with pattern submodels (Fletcher Mercaldo and Blume, 2020). We take a pattern mixture model approach in this work and incorporate missingness in the covariate space into the predictive framework, akin to Little and Wang (1996). We implement a smoother version of the pattern submodel approach from Fletcher Mercaldo and Blume (2020) where the missingness indicators $\boldsymbol{M}_\ell$ are instead

included as direct effects into the model rather than implementing a full stratification approach by missingness pattern. We justify this simplification due to key covariates such as WHO stage and education level containing small percentages of missing values which induces sparse missingness patterns in the data. Given the available data, we prefer this smoother approximation to the full pattern-mixture specification.

1.4.3 Sampling Strategy, Notations and Definitions

A common problem with prediction modeling with EHR data is the decision of choosing a sampling window. The sampling window is defined as the time period from where records and individuals are sampled for entry into the cohort and the data for which the model will be built from. Sampling can occur at two levels: (1) sampling patients who are actively enrolled in HIV care within a specified time frame and (2) sampling visits within that time frame that are relevant for prediction. It is tempting to use all available data to make predictions however this can cause issues if the sampled visits and patients are not aligned with the goals of the analysis. Changes in the standard of care can affect the utility of certain variables along with the quality of the data at hand. For example, certain standards of care may become outdated leading to some clinical measurements not being collected anymore, and not all visits are informative for HIV prediction. Thus patients and visits from earlier years may not reflect the current population. Older records reflect clinical decision rules that no longer apply, so a learner may implicitly pick up associations that are clinically obsolete. The goal of a prediction model is to forecast on new individuals who are currently enrolled in care at AMPATH and thus the sample used to develop the model should be an accurate reflection of patients, and visits, that occur today. Restricting to patients actively enrolled in care improves the model's relevance to the current population of patients and enhances generalizability, since future visits will follow similar clinical protocols. Similarly EHR datasets can be massive; finding an appropriate sampling window can reduce the computational burden on clinics.

We construct the sample by selecting a relevant sampling date and then (1) identifying

patients actively engaged in care beyond this date, defined as having at least one visit on or after the start of the chosen sampling date, and (2) retaining all visits occurring after the sampling date for the patients from (1). Figure 1.2 provides a visualization of patient trajectories and which patients are considered for cohort entry.

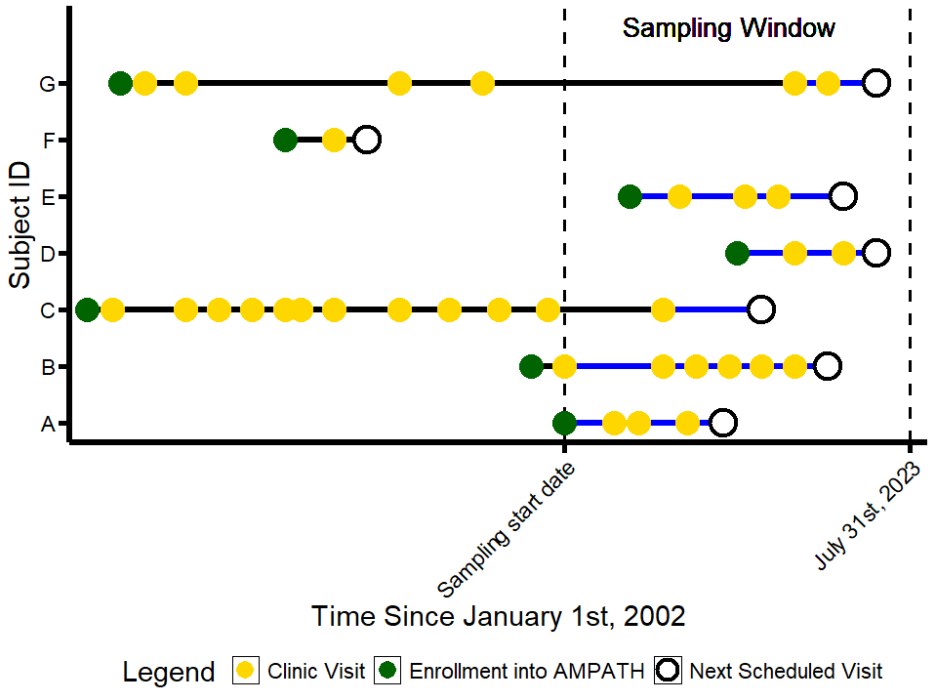

Figure 1.2: Examples of patient included and excluded from the analysis sample. Green denotes enrollment into AMPATH while yellow denotes when a clinic visit was made. The black circle denotes the next scheduled visit. The black lines that connect visits for each patient denote that the outcome at that visit is not used as an outcome for prediction. Blue lines indicate the time for which a patient is included in the model cohort.

Patients who have a visit within or at the start of the sampling window are entered into the cohort at the time of their first visit within the window. For example patients A, D and E all enroll on or after the start of the sampling window and are included in the analysis. Patients B, C and G enrolled into care prior to the start of the sampling window, but since they have remained in care and have had a visit within the sampling window they are included in the analysis sample. Patient F is censored before the start of the sampling window and is therefore not included into the analysis sample. The outcome space of our models is formed from the outcome of visits that occur after the first visit within the sampling window; not the

outcomes associated with the first visit that occurs within the sampling window. Information from visits prior to entry into the sample can still be used to make predictions; specifically retention information and clinical measurements from these previous visits enter as covariate information.

Given the stratified visit models from section 1.4.2 it is important to note that visits are indexed by the total number of visits a patient has had since enrollment into care at AMPATH, not just by the number of visits made within the sampling window. Thus only patient's A,D and E contribute to the risk set of the $\ell = 1$ model. Although Patient B's first return visit ($\ell = 1$) occurs at the start of the sampling window, thereby granting them entry into the cohort, they are not included in the risk set for the $\ell = 1$ visit model. Patient B enters into the risk set for the $\ell = 2$ visit model and remains in the risk set for later visit models until censoring occurs. Once a patient is effectively censored their contribution to the risk set ends. Patient D contributes to the risk set of the $\ell = 1, 2, 3$ models but will not for the $\ell \geq 4$.

1.4.4 Variable Importance

A common goal in prediction-settings is to understand the contributions of covariates as it relates to the overall predictive variance in the observed population (i.e. global variable importance) and for the individual risk score (i.e. local variable importance). The scope of this paper is limited to global variable importance, specifically the z-statistic in generalized linear models and split based importance for tree based methods such as Bayesian Additive Regression Trees (BART). We calculate these importance metrics for each stratified model from section 1.4.2 to identify key covariates that affect retention. Given the current model formulation these importance metrics can only be calculated for $\pi_1(\overline{\boldsymbol{X}}_\ell)$, but still may provide useful information.

We use these model specific measures of variable importance to flag important predictors and then provide visualizations of these key covariates to understand their effects and how they influence each model through the use of conditional effect plots. These conditional effect plots take the form of a function $h_j(X_{\ell j}) = E[s(\pi(\overline{\boldsymbol{X}}_\ell)) \,|\, X_{\ell j}]$ where $s(\cdot)$ is the log odds function

and $\pi(\overline{\boldsymbol{X}}_\ell) = \pi_1(\overline{\boldsymbol{X}}_\ell)\pi_2(\overline{\boldsymbol{X}}_\ell)$. The log odds transformation is performed for interpretability. Visualizations of the effect of each covariate can help to inform clinicians of how the model utilizes the covariate and how the average predicted risk varies on the support of the covariate $X_{\ell j}$.

Estimates $\hat{h}_j(x_j)$ are obtained for each covariate $X_{\ell j}$ by (i) clustering the predicted probabilities $\hat{\pi}(\overline{\boldsymbol{X}}_\ell)$ based upon the values in the support of $X_{\ell j}$ and (ii) taking the expected value of the predicted probabilities within each cluster. We can then plot the individual mean values along with their associated 95% credible interval (CI) which then gives us an idea of, on average, what we expect certain values on the distribution to do. A large benefit of the Bayesian framework is the natural quantification of uncertainty which is particularly important when dealing with variable importance metrics as they help us to understand average trends and their statistical significance. A key difference between other variable importance visualizations such as partial dependence plots (PDP) Goldstein et al. (2015) in the literature and this method is that we directly take into account the empirical conditional distribution $[\boldsymbol{X}_{\ell(-j)} \,|\, X_{\ell j}]$ rather than just assuming covariate independence and integrating over the empirical distribution $\hat{P}(\boldsymbol{X}_{\ell(-j)})$.

1.4.5 Performance Metrics & Model Evaluation

In prediction based settings the goal is to evaluate the model based on some measure of goodness of fit to ascertain its clinical utility. Thus, performance metrics should reflect how the model is used in practice. To evaluate the model we use area under the Receiving Operator Characteristic (AUROC), positive predictive value (PPV), sensitivity and specificity. A point estimate for each metric can be found by taking the median while a 95% credible interval can be found by taking the 2.5th and 97.5th percentiles of the draws.

In resource-constrained settings such as AMPATH, the outreach team is limited in their ability to target patients in advance of their scheduled visit. This means out of N patients who are scheduled return to care within a specific time frame only K can be feasibly targeted for intervention where $K < N$. We make the assumption that at most only 20% of the N patients

scheduled to return can feasibly be targeted and provide performance metrics for individual clinics.

The model is trained via a modified time-based k-fold cross validation procedure in order to accurately mimic the data generating process and how the model will be used in practice. The data are split into a training set consisting of patients who were scheduled to return to clinic between January 1st 2021 to December 31st 2022, and a test set of patients who were scheduled to return between January 2023 to June 2023. We then further split the test data by the month of the scheduled RTC date (e.g. Jan 2023, Feb 2023,..., May 2023, June 2023) to achieve 6 test folds. This model training strategy naturally helps to deal with patients who are censored due to database closure since they are forced into the test set and out of the training set. We train the model one time on the training data, and evaluate the model's performance on each of the test folds. This not only mimics how the model will be used in practice but also provides valuable insight into how the model performs from month to month. A visual explanation is provided in Figure 1.3.

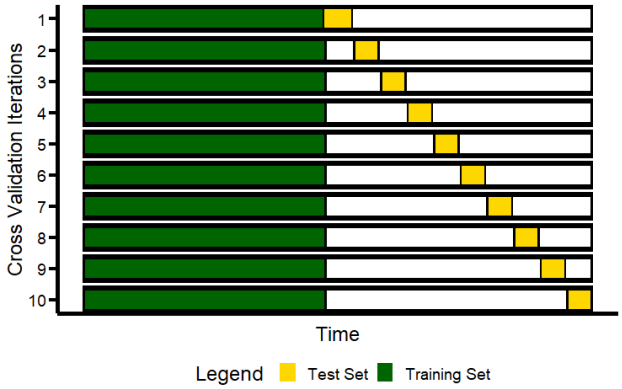

Figure 1.3: Time Series Cross Validation procedure. The test set is split into different time periods based upon when the visit was scheduled (i.e. month of RTC date). The green space represents the historical data used to fit the model and stays constant. The yellow squares are representative of future times to which the model will be applied, i.e. the month of the RTC date of future patients.

1.5 Application to AMPATH Data

1.5.1 Overview

We illustrate the application of the stratified visit modeling approach using data from AMPATH as described in Section 1.4. Our analysis uses data on 70,626 HIV infected adults aged 18 years or greater who were actively enrolled in care on or after January 1st, 2021. The analysis used data on 685,715 visits that have occurred since January 1st, 2021. Information from visits prior to January 1st, 2021 was used, however, the visits were not subject to prediction. This date was selected due to recency and that this date marks when AMPATH began to loosen COVID-19 related restrictions. The date of database closure was July 31st, 2023.

Table 1.1 describes the prevalence of the outcomes of interest among those included in the sample for each model. Patients within the "died" category are patients who died some time after visit ℓ, but before visit $(\ell + 1)$ could occur. Patients within the database closure category are patients who are censored after visit ℓ, but before visit $(\ell + 1)$ could occur. The rate of censoring due to death is of particular interest with deaths making up a relatively small percentages of each sample. In fact the rate of death drop substantially as patients remain in care, a sign of improved health outcomes for patients who remain engaged in HIV care.

Table 1.1: Prevalence of patient statuses for the $\Delta = 0$ delay models.

	Visit Total	Database Closure	On Time/Early	Missed	Dead
$\ell = 1$	10,778	236 (2.2)	6,064 (56.3)	4,303 (39.9)	175 (1.6)
$\ell = 2$	10,725	349 (3.3)	5,810 (54.2)	4,434 (41.3)	132 (1.2)
$\ell = 3$	10,210	431 (4.2)	5,374 (52.6)	4,320 (42.3)	85 (0.8)
$\ell \geq 4$	654,002	44,464 (6.8)	394,458 (60.3)	213,893 (32.7)	1,187 (0.2)

Retention statistics are also described visually in Figure 1.4. Note that delays in returning to clinic tend to decrease the longer a patient stays in care. While the overall proportion of missed visits by 1-7 days fluctuates, the overall proportion of delays longer than 28 days begins to decrease as a patient remains in care for longer. The rates of death and dropping out of care (e.g. LTFU status) also decreases as an overall proportion of those on their next visit.

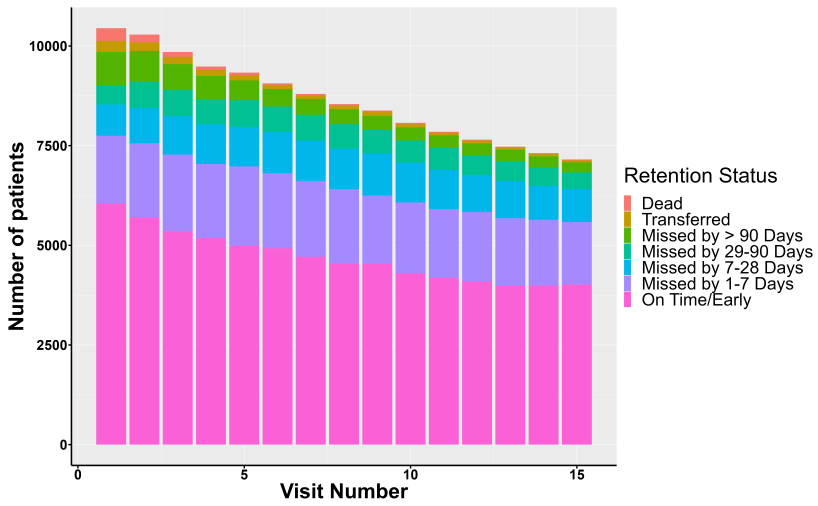

Figure 1.4: Event history barplot of patient outcomes by visit number for the first 15 visits after the start of the sampling window, January 1st, 2021.

Covariates in $\boldsymbol{X}_\ell$ include baseline characteristics such as gender, calendar year of enrollment (2002,...,2023), education level (missing, none, some primary level, secondary level, college level) and age at enrollment. Time varying covariates include scheduled return time, month of the scheduled appointment, body mass index (BMI), WHO stage (primary infection, clinically asymptomatic stage, symptomatic HIV infection, and AIDS), marital status (not married, married), visit type (drug pickup, differentiated care), clinic type (referral/regional hospital, county hospital, sub-county hospital, health center/dispensary, private hospital), clinic size and the most recent CD4 count in log cells/mm3. Covariates with missing entries (WHO stage, education level, BMI, and CD4 count) were coded with missingness indicators as described in Section 3. Table 1.2 describes the distribution of these covariates for members of the cohort, and is further stratified by patients who enter into the cohort after January 1st, 2021 and patients who enter into the cohort who were enrolled in HIV care at AMPATH before the beginning of the sampling window. From the information in Table 1.2 we justify the smooth pattern submodel approach due to the sparsity of key covariates such as WHO stage and education level, especially for earlier visit models which contain a larger proportion of the post-2021 enrollee cohort members.

Table 1.2: Baseline Characteristics for patients at the first visit on or after January 1st, 2021. The table reports "median (25th percentile, 75th percentile)" for continuous variables and percentage of true for binary variables or each level of categorical variables.

	Baseline Characteristics at entry into cohort		
Variable	Total (N = 70,626)	Post-2021 (N = 10,778)	Pre-2021 (N = 59,848)
Age (years)			
Median (Q1, Q3)	45 (35, 53)	36 (29, 45)	46 (37, 54)
Scheduled Appointment Time (Weeks)			
Median (Q1, Q3)	11.29 (4, 12)	2 (2, 3.43)	12 (4.71, 12)
Sex			
Female	44,041 (62.4)	6,434 (59.7)	33,607 (62.8)
Male	26,585 (37.6)	4,344 (40.3)	22,241 (37.2)
ARV Status			
Not On	1,496 (2.1)	1,207 (11.2)	289 (0.5)
On	69,130 (97.9)	9,571 (88.8)	59,559 (99.5)
Facility Type			
Regional/Referral Hospital	11,218 (15.9)	1253 (11.6)	9,965 (16.7)
County Hospital	17,734 (25.1)	2765 (25.7)	14,969 (25.0)
Sub-County Hospital	30,017 (42.5)	4032 (37.4)	25,985 (43.4)
Health Center/Dispensary	10,826 (15.3)	2604 (24.2)	8,222 (13.7)
Private	831 (1.2)	124 (1.2)	707 (1.2)
Body Mass Index ($\frac{kg}{m^2}$)			
Median (Q1, Q3)	22 (20, 25)	21 (19, 23)	22 (20, 26)
Missing	6,903 (9.8)	3,518 (32.6)	3,386 (5.6)
CD4 Count			
Median (Q1, Q3)	407 (252, 577)	244 (94, 437)	413 (261, 581)
Missing	24,228 (34.3)	8,711 (80.8)	15,517 (25.9)
WHO Stage*			
Missing	6,015 (8.5)	658 (6.1)	5,357 (9.0)
First Stage	35,112 (49.7)	6358 (50.2)	28,756 (48.0)
Second Stage	11,443 (16.2)	1838 (16.1)	9,605 (16.0)
Third Stage	14,975 (21.2)	1640 (20.9)	13,335 (22.3)
Fourth Stage	3,081 (4.4)	286 (4.3)	2,795 (4.7)
Education Level*			
Missing	12,411 (17.6)	233 (2.1)	12,178 (20.3)
None	3,659 (5.2)	873 (8.1)	2,786 (4.7)
Primary School	28,394 (40.2)	5025 (46.6)	23,369 (39.0)
Secondary School	19,158 (27.1)	3485 (32.3)	15,673 (26.2)
College	5,842 (9.8)	1162 (10.8)	5,842 (9.8)

1.5.2 Model Specification

We parameterize the models $g^{\Delta}_{1,\ell+1}(\boldsymbol{X}_\ell, \boldsymbol{\theta}_{1\ell})$ and $g^{\Delta}_{2,\ell+1}(\boldsymbol{X}_\ell, \boldsymbol{\theta}_{2\ell})$ using Bayesian Additive Regression Trees (BART) (Chipman et al., 2010) and main-effects logistic regression where continuous variables are modeled using regression splines. We assume independent priors $\boldsymbol{\theta}_\ell \sim p(\boldsymbol{\theta}_\ell) = p(\boldsymbol{\theta}_{1\ell}, \boldsymbol{\theta}_{2\ell}) = p(\boldsymbol{\theta}_{1\ell})p(\boldsymbol{\theta}_{2\ell})$, for $\ell = 1, \ldots, L$ and thus obtain independent posterior distributions. Under the assumption of independent priors this amounts to running two separate Bayesian inferences for each model. We implemented BART using a probit formula-

tion

$$\Phi^{-1}\{g(\boldsymbol{X};\boldsymbol{\theta})\} \quad = \quad \sum_{t=1}^{T} f(\boldsymbol{X};\tau_t,\mathscr{M}_t),$$

where $\Phi(\cdot)$ is the normal CDF, T is the number of trees, and $f(\boldsymbol{X};\tau_t,\mathscr{M}_t)$ is a distinct regression tree. The tree structure is parameterized by $\boldsymbol{\theta} = \{\tau_t,\mathscr{M}_t\}_{t=1}^{T}$ where τ_t is the tree structure and $\mathscr{M}_t = \{\mu_k\}_{k=1}^{b_t}$ is a set of b_t terminal nodes for τ_t. The priors are specified by the distributions $p(\mu_{b_t} \mid \tau_t)$ and $p(\tau_t)$ for $t = 1,\ldots,T$ and $b = 1\ldots,B$. The prior $p(\mu_{bt} \mid \tau_t)$ is normally distributed with location parameter equal to 0 and scale parameter equal to $3/(k\sqrt{t})$. The prior $p(\tau_t)$ on the trees is proportional to $\alpha(1+d)^{-\beta}$, where d is the tree depth. In this analysis, k is set to 2, and $(\alpha,\beta) = (0.95, 2)$. BART is fit using the BART package (Sparapani et al., 2021) in R Version 4.4.0. For logistic regression we run a metropolis hastings sampling procedure by specifying a $MVN(0, 5^2 I)$ prior on the coefficients $\boldsymbol{\theta}$. For each model we run 10000 MCMC iterations taking the first 5000 as burn-in and then taking every 5th sample thereafter as a posterior draw.

Covariates $X_{\ell j}$ with missing values enter into the model as $X_{\ell j}(1 - M_{\ell j})$ along with the corresponding missingness indicator $M_{\ell j}$. The covariates that contain missing values, and thus are accompanied by a missingness indicator, include BMI, CD4 count, WHO stage, marital status and education level. Given the longitudinal structure of the data we implement a last value carried forward approach where we assume a patient's value for a covariate will stay constant until the next update. This means that if a covariate $X_{\ell j}$ remains unmeasured starting at enrollment and remains unmeasured until visit ℓ' the missingness indicator $M_{\ell j} = 1$ for all visits $\ell < \ell'$ and $M_{\ell j} = 0$ for all visits $\ell \geq \ell'$.

1.5.3 Results

Performance metrics and model evaluation

Two trends are observable from the results reported in Table 1.3. Firstly that the predictive accuracy of the models increases with more visits. Predictive accuracy is higher for models

based on later visits, primarily because previous visit history is a strong predictor of future return to care. Moreover, the more complex BART model has similar accuracy to logistic regression across all settings. This is in agreement with other studies which also showed that logistic regression did as well as less interpretable models (Ridgway et al., 2021).

In Table 1.4, we report clinic-level performance metrics by applying the stratified model to test data from each clinic, regardless of visit number. Clinics were selected to represent a wide range of sizes and missed-visit prevalence. The calibration slope is included to provide additional insight into how well the model generalizes at the clinic level. Model performance varies across clinics, with AUC values ranging from 0.61 to 0.74. In particular, clinic 71 shows poor performance in both AUC and calibration slope. For the remaining clinics, performance is generally adequate, suggesting that this model formulation is useful at the clinic level for ranking patients and identifying those at risk of missing scheduled visits.

As shown in Table 1.3 and 1.4 the PPV exceeds the outcome prevalence for every visit-level model and across all reported clinics in Table 1.4. Under the practical constraint that a clinic can at most pro-actively contact K of N patients prior to their next scheduled visit, this indicates that prioritizing patients by their predicted risk of missing a visit results in identifying more true missed visits than would be achieved by randomly selecting K patients from the N scheduled to return within a given time frame.

Table 1.3: Model Performance Metrics for the $\Delta = 0$ retention models by visit number. The table reports the posterior median with the 95% credible interval (CI) for performance metrics along with the number of visits and the prevalence of a missed visit. Metrics are calculated using the 80^{th} percentile as the threshold for "high risk" classification. The prevalence of a missed visit is reported for context.

	$\Delta = 0$				
Visits Modeled	AUROC	Sensitivity	Specificity	PPV	Prevalence
($\ell = 1$)					
Logit	62.0 (60.0, 64.0)	28.1 (25.6, 30.7)	84.4 (83.0, 85.9)	50.6 (46.1, 55.2)	40.7
BART	62.0 (59.0, 64.0)	28.0 (24.7, 31.2)	84.3 (82.2, 86.2)	50.5 (44.3, 56.1)	40.7
($\ell = 2$)					
Logit	66.0 (64.0, 68.0)	32.2 (29.0, 34.8)	87.9 (85.8, 89.6)	63.8 (57.6, 68.8)	41.8
BART	66.0 (64.0, 69.0)	31.9 (28.8, 35.5)	87.8 (85.3, 89.5)	64.4 (58.3, 69.6)	41.8
($\ell = 3$)					
Logit	67.0 (65.0, 69.0)	31.0 (28.4, 34.1)	87.8 (86.0, 90.0)	64.4 (59.1, 70.7)	43.5
BART	68.0 (66.0, 70.0)	32.0 (29.3, 34.5)	88.6 (86.6, 90.5)	67.5 (61.9, 72.7)	43.5
($\ell \geq 4$)					
Logit	72.0 (71.0, 72.0)	35.8 (35.6, 35.9)	88.0 (87.9, 88.1)	60.4 (60.1, 60.6)	35.0
BART	73.0 (73.0, 73.0)	37.3 (37.2, 37.5)	88.8 (88.7, 88.9)	62.9 (62.6, 63.2)	35.0

Table 1.4: Model Performance Metrics stratified by clinic for the $\Delta = 0$ retention models. The table reports the posterior median with the 95% credible interval (CI) for performance metrics along with the number of visits and the prevalence of a missed visit. Metrics are calculated using the 80^{th} percentile as the threshold for "high risk" classification. The prevalence of a missed visit is reported for context. Results are reported from the logistic regression model.

		$\Delta = 0$			
Clinic	Total Visits	AUC	Cal. Slope	PPV	Prev.
Clinic 18	10,009	0.74 (0.73, 0.75)	1.13 (1.06, 1.19)	43.3 (42.3, 44.2)	18.3
Clinic 2	4464	0.73 (0.71, 0.74)	0.96 (0.88, 1.04)	74.9 (73.6, 76.2)	47.2
Clinic 4	3,098	0.70 (0.69, 0.72)	0.87 (0.78, 0.97)	73.7 (72.1, 75.2)	51.5
Clinic 74	2,113	0.68 (0.65, 0.70)	0.87 (0.73, 1.01)	43.1 (41.0, 45.2)	24.9
Clinic 200	1,739	0.68 (0.66, 0.71)	0.83 (0.70, 0.96)	77.5 (75.6, 79.5)	55.4
Clinic 189	1,147	0.69 (0.66, 0.72)	0.89 (0.73, 1.06)	83.4 (81.3, 85.6)	58.2
Clinic 372	764	0.68 (0.65, 0.72)	0.84 (0.65, 1.04)	63.8 (60.4, 67.2)	38.0
Clinic 71	512	0.61 (0.55, 0.67)	0.57 (0.28, 0.86)	36.3 (32.1, 40.4)	23.4
Clinic 112	293	0.74 (0.68, 0.79)	1.09 (0.77, 1.43)	65.5 (60.1, 71.0)	40.6
Clinic 56	63	0.70 (0.56, 0.84)	0.88 (0.15, 1.68)	41.7 (29.5, 53.8)	31.7

Sampling Window

The choice of sampling window was partially informed through this procedure. We developed models based on training data from 2010 to 2022, 2011 to 2022,..., 2016 to 2022 and up

to a model trained exclusively on data from 2022. Each model was then evaluated by forecasting outcomes in a common test set consisting of visits from January 2023 through June 2023, regardless of visit number, and predictive performance was assessed using the AUC. Results presented in Figure 1.5 show no discernible difference in the test AUC between these different models suggesting that later start dates for the sampling window do not meaningfully degrade predictive performance. This finding indicates that, when covariate distributions remain relatively stable over time, more recent training windows may be used to improve computational efficiency without sacrificing accuracy. An important direction for future research is to repeat this analysis in settings where key covariate distributions shift over time due to changes in measurement practices or clinical protocols.

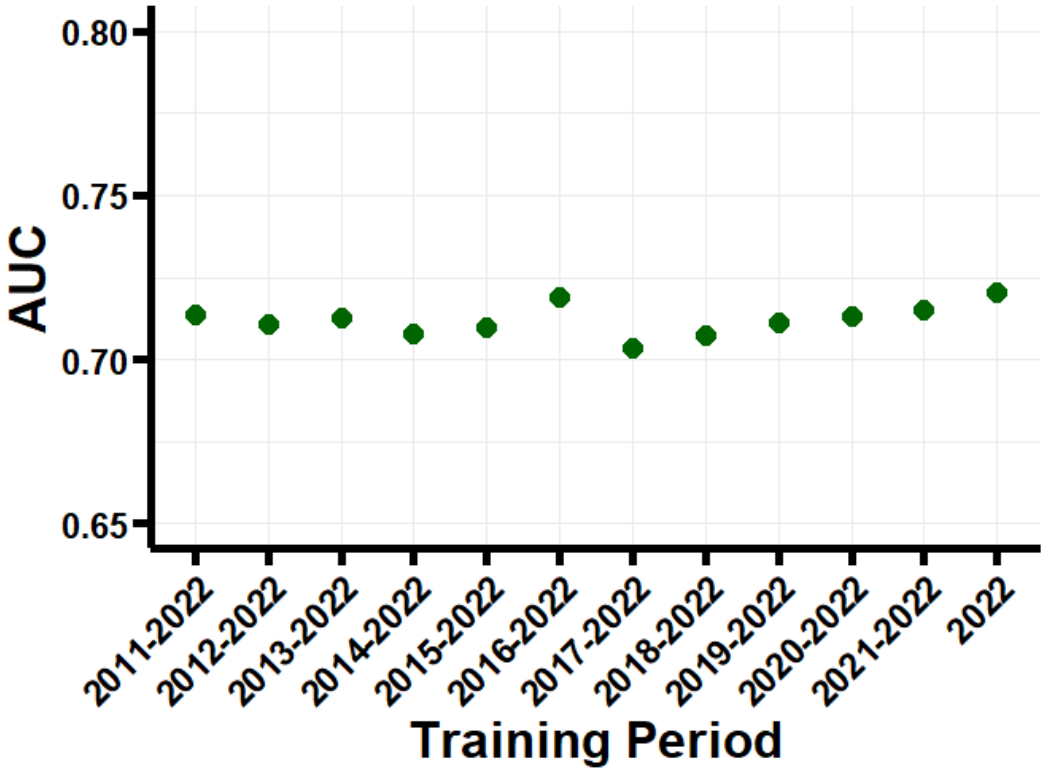

Figure 1.5: Test AUC for models trained from different time periods within the AMRS for the entire model regardless of visit number. The y-axis is restricted between 0.65 and 0.80 in order to visualize the results.

Covariate Importance

Global covariate importance measures from BART and logistic regression models are summarized in Figures 2.3 and 2.4, respectively. Variable importance for BART is based on the number of times that a specific covariate is split upon. In agreement with other studies (Ramachandran et al., 2020, Ogbechie et al., 2023) retention history is an important marker of whether someone will miss a visit or not. In fact splits using retention history covariates in later visit models begin to dominate the importance. Earlier visit models such as the first visit highlight clinical and demographic characteristics as being most important along with the timing of the scheduled visit. For logistic regression, we calculate global variable importance as the square root of the Wald statistic divided by its degrees of freedom; this leads to a scalar value for covariate effects represented by multiple terms, as is the case for spline representations of continuous variable effects. For visualization, covariate importance scores were normalized within each model by dividing by the sum of all scores, yielding relative contributions that sum to one. As with BART, prior visit history takes on a more prominent role for models designed to predict return to care at later visits; in fact these are the primary predictive variables for models of visits $\ell \geq 4$.

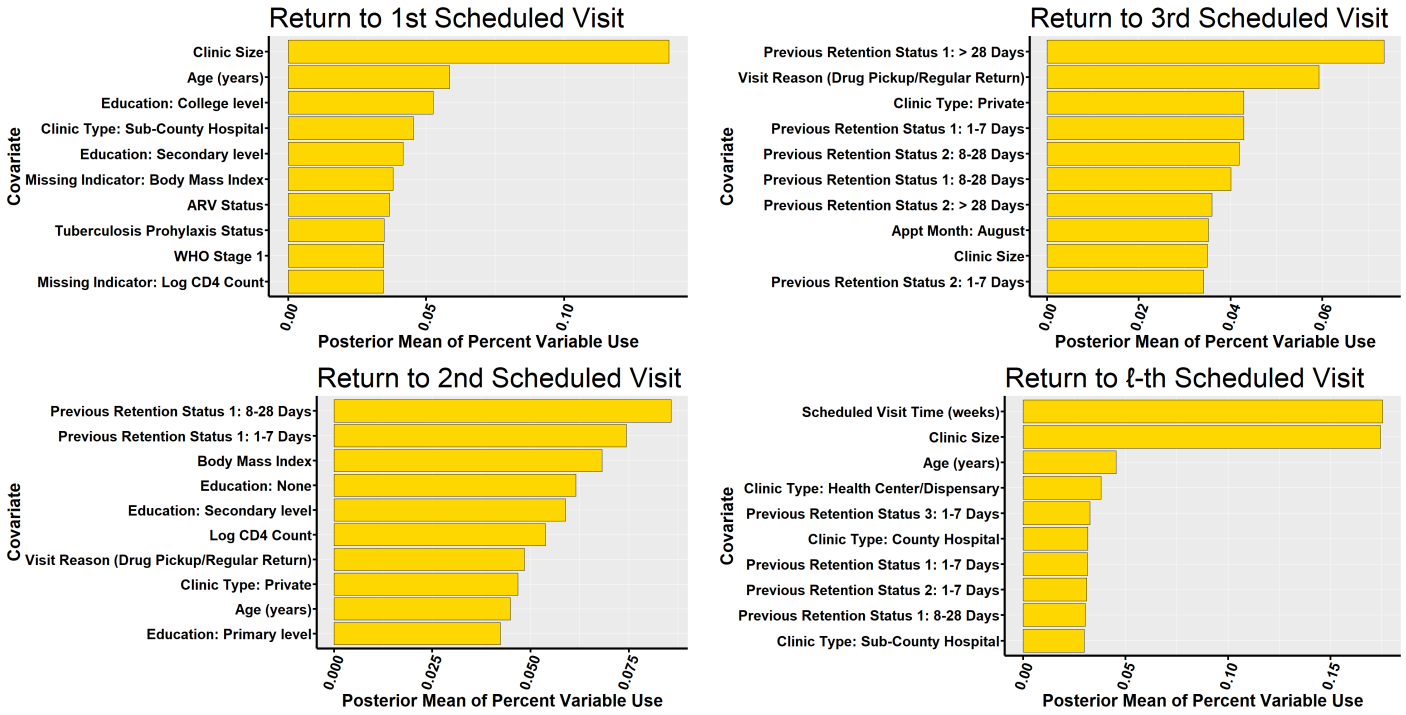

Figure 1.6: Global feature importance plots generated from Bayesian additive regression tree (BART) model fit to AMPATH data. Top 10 features from each model are shown.

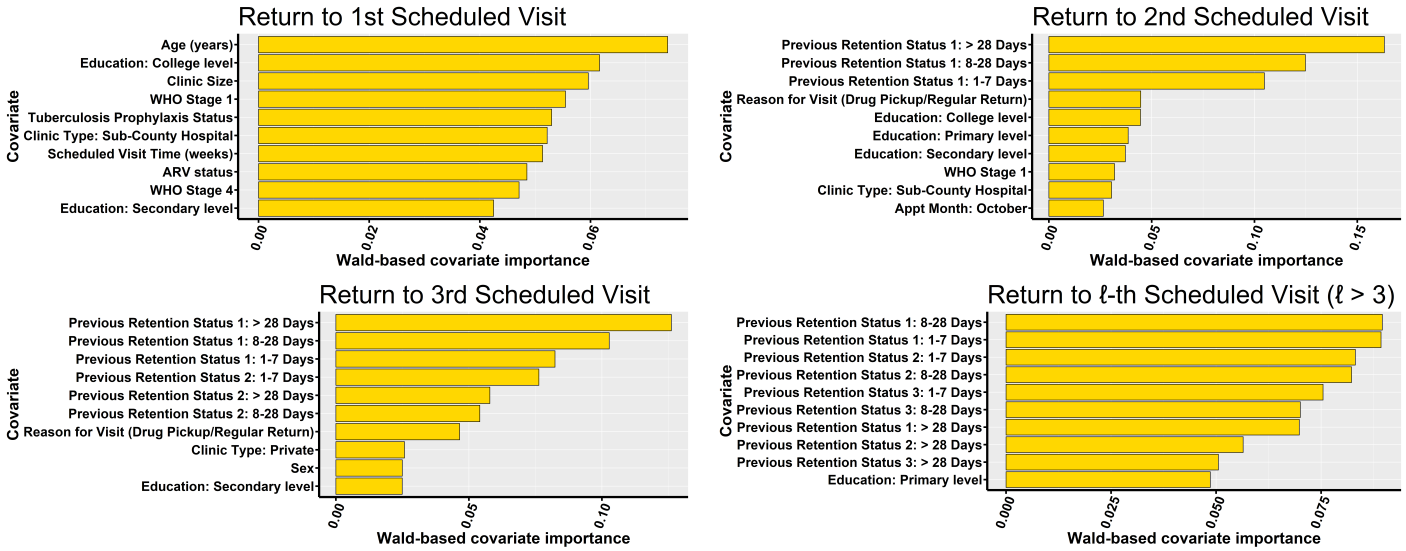

Figure 1.7: Global feature importance plots from logistic regression models fitted to AMPATH data. The top 10 features from each model are shown. Feature importance is quantified using the square root of the Wald statistic, and subsequently scaled by dividing each covariate's score by the sum of all scores.

Plots of $\hat{h}_j(X_j)$ from section 1.4.4 are shown in figure 1.8 for key covariates in the $\ell = 1$

model for BART and logistic regression, namely return visit date, age, education level and WHO stage. Age has a downward trend where the average risk for patients decreases with age implying younger patients are at higher risk for missing their next scheduled visit. For ages above 80, the credible intervals for the importance estimates become extremely wide due to the very small in-sample size (n=9). Compared with logistic regression, BART produces narrower intervals, likely reflecting shrinkage toward a piecewise-constant fit in sparsely populated regions of the covariate space, in which values above a split share the same effect. We similarly observe that scheduled return visits between 2–4 weeks from the current visit yield the lowest and most stable average risk, likely reflecting assignment from clinics with more consistent care practices. With respect to WHO stage (a measure of disease severity graded from 1 to 4) the highest average predicted risk is observed among patients in WHO stage 1. Risk then decreases for individuals in WHO stages 2 and 3, followed by an increase at WHO stage 4, a pattern that is not unexpected given advanced disease severity. This reflects the phenomenon that many patients who are in early stage of HIV may not be experiencing negative effects of the infection; as such, they can be prone to skipping visits. When patients get sicker, i.e. in later WHO stages, they are more likely to keep their scheduled visit. We find almost identical risk and shape for the plots of $h_j(x_j)$ for both BART and logistic regression with the most significant difference occurring in the size of the credible intervals which most likely can be attributed to the functional forms of each model.

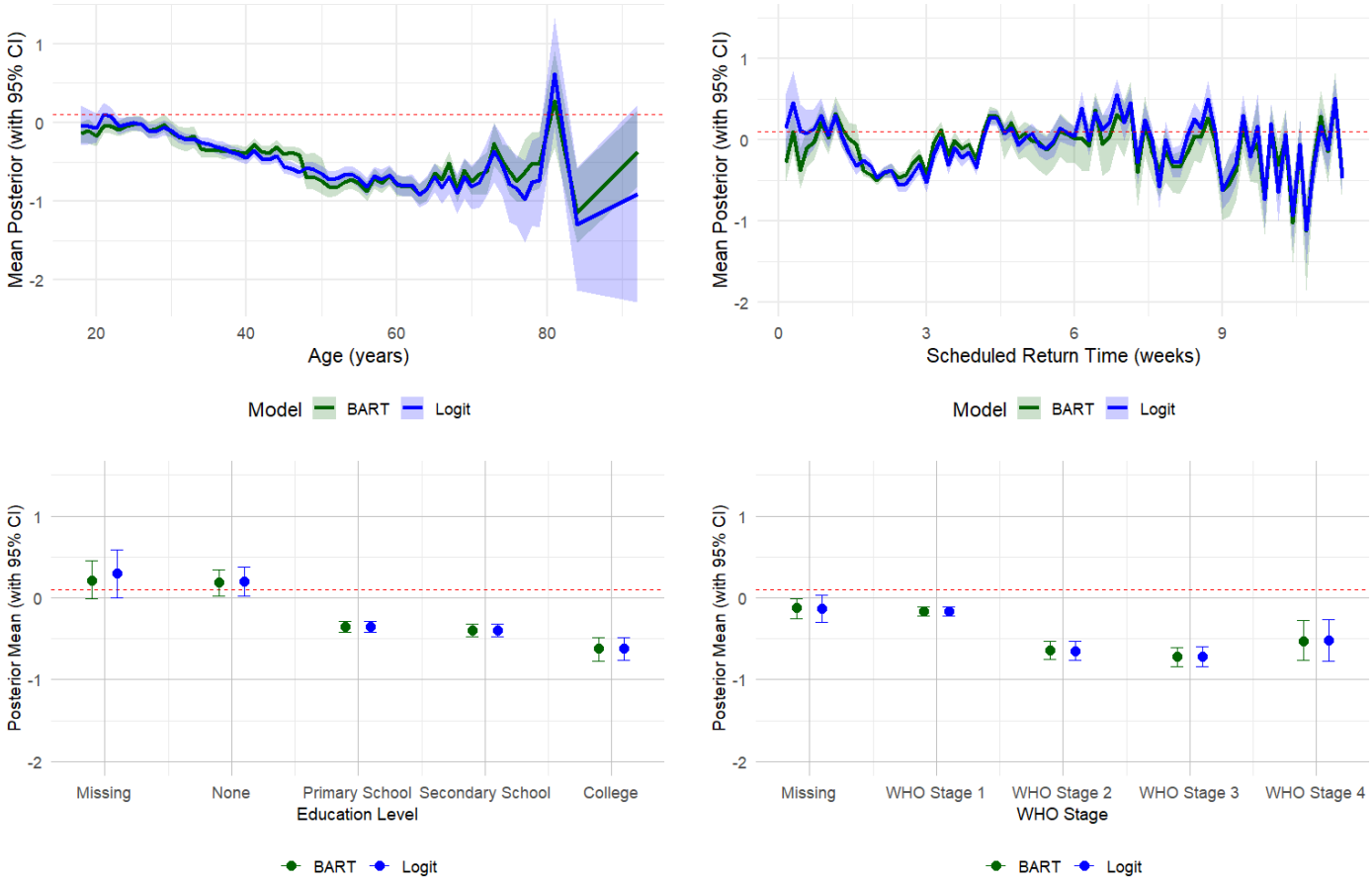

Figure 1.8: Visual plot of conditional effect of key covariates for the $\ell = 1$ visit model for Age (top left), scheduled return time (top right), education level (bottom left) and WHO stage (bottom right). The mean estimate and the associated 95% credible intervals are plotted for the discrete covariates. Green denotes the marginal effects from BART while blue denotes the marginal effects from Logistic regression. The red dashed line represents the log-odds of the 80th percentile of the predicted probabilities, i.e. the threshold for a "high risk" classification.

1.6 Discussion

This chapter presents a modeling framework for the HIV care cascade with a specific focus on retention in care. Retention is defined according to clinically relevant benchmarks established by the Kenyan Ministry of Health and is modeled using a series of stratified binary prediction models at each scheduled visit that explicitly account for competing risks such as death and missingness in the covariate space. The primary contribution of this work is the

development of a principled statistical framework that aligns both the prediction models and the associated sampling strategy with the underlying data-generating process at each return visit. A key motivation for this methodology is to provide AMPATH outreach workers with a tool to identify patients at high risk of disengagement, enabling targeted interventions ahead of scheduled visits. Outreach ahead of the scheduled return visit has the potential to increase the likelihood that the patients at high risk for disengagement will keep their appointments. Such interventions are critical to ensure patients remain current with their antiviral medications (i.e. ARVs), reduce the risk of clinical progression, and enable timely identification and management of any emerging health issues.

We find that the prediction models developed within this framework perform adequately in predicting disengagement events at the visit-level and clinic-level, with predictive accuracy improving as additional retention history becomes available. While some heterogeneity in performance exists, in particular larger clinics tend to achieve higher accuracy than smaller clinics, this observation highlights the need for further investigation into the factors contributing to lower performance in smaller clinics. Although BART can capture complex interactions without requiring explicit specification, this flexibility does not yield sufficient gains to offset the advantages of simpler, computationally faster models such as logistic regression in terms of predictive performance. This finding is consistent with prior work cautioning against the routine use of black-box models in high-stakes decision-making settings (Rudin, 2019). Notably, BART and logistic regression identified largely the same covariates as key contributors to missed-visit risk and revealed similar trends in the data.

Within the scope of this work we identify several limitations that suggest avenues for future development. Firstly, we focus on binary indicators of whether a delay occurs, without accounting for the reason for the delay or its duration. A patient with long periods in and out of care is markedly different from a patient who is consistently late by only a few days, but this distinction is not reflected in our binary framework. Modeling the actual event times could provide more nuanced insights into retention patterns. A second limitation of this work re-

lates to systemic changes in the standard of care. When certain clinical measurements fall out of routine use, covariates may become sparse or inconsistently recorded, potentially affecting predictive performance due to heterogeneity between development and external data (Luijken et al., 2019, Pajouheshnia et al., 2019, Luijken et al., 2020). This limitation underscores the importance of continuously re-fitting models and updating the sampling window as new data become available to ensure predictions accurately reflect current care practices. Finally, mis-classification of deaths and dropouts is a common issue in studies of loss to follow-up. A patient who has not attended for extended periods may have disengaged from care, but unreported deaths are also possible. Such mis-classification could affect the observed prevalence of missed visits, as hypothetical deaths may have occurred prior to the patient's next scheduled visit.

Finally, future work takes us into implementation of our strategy. While our model can accurately identify patients at risk of missing their next scheduled visit, it is also essential to evaluate the feasibility and effectiveness of the advanced outreach interventions to ensure that targeted calls and visits actually improve patient retention.

Chapter 2

Quantifying unit-level feature importance in machine learning in terms of conditional risk score distributions

2.1 Introduction

Advances in machine learning (ML) have opened up new avenues for prediction in health outcome settings. In recent years ML models have gained considerable traction in predicting retention in care in HIV settings (Ogbechie et al., 2023, Olatosi et al., 2021, Oliwa et al., 2021, Ramachandran et al., 2020, Ridgway et al., 2022). Retention is a necessary condition for maintaining people living with HIV (PLWH) on anti-retroviral therapy (ART) and successful suppression of HIV viral load, a key driver of HIV disease progression. Even after a measured viral load indicating suppression, viral failure due to nonadherence or drug resistance can occur well before the next follow up. Regular clinical visits promote consistent access to anti-retroviral drugs (ARVs) and monitoring of disease progression and medication compliance. Hence there is vested interest in identifying patients who are at risk for disengagement in care before their scheduled visit.

Implementation of ML models in clinical settings such as at the Academic Model Providing

Access to Healthcare (AMPATH) in Kenya offers a promising approach to address this challenge. We have developed a ML prediction tool that can be used to identify and provide outreach to patients who are at high risk of missing a visit and disengaging from care, with the goal of increasing attendance at scheduled appointments and maintaining engagement in care. While ML models can capture complex outcome-covariate relationships to generate accurate predictions, this usually comes at the cost of model interpretability. While there is no universal agreement on what makes a model 'interpretable' (Rudin, 2019, Lipton, 2018), understanding how specific covariates drive an individual prediction has well-recognized utility. For example, knowing the reasons that a patient has been identified as being at risk for missing their scheduled visit can then allow outreach workers to personalize the outreach encounter to the individual patient.

The component of explainability relating to variable importance can be divided into *global* and *local* variable importance, or explanations. Global explanations rank covariates based on their contributions to overall predictive variance in the model. Two common methods of global explanation are permutation based covariate importance (Fisher et al., 2019, Breiman, 2001) and partial dependence plots (Friedman, 2001). Local explanations identify the degree to which individual covariates influence a unit-level prediction from the model. Commonly-used measures include Shapley additive explanations (SHAP) (Lundberg and Lee, 2017, Shapley, 1953) and individual conditional expectation plots (ICE) (Goldstein et al., 2015). Most of these methods quantify marginal unit-level variable importance, measured unconditionally on other variables in the model. In this chapter we propose a new measure of unit-level covariate importance that conditions on the values of other covariates. Calculating the measure requires information about appropriate conditional distributions of the covariates, which rarely have a closed analytic form. Our approach uses predictive mean matching (Rubin, 1986, Little, 1988) for calculations of the conditional distributions required for quantifying unit-level variable importance.

The rest of this chapter is organized as follows. In Section 2.2, we review existing meth-

ods for quantifying local variable importance; in Section 2.3, we introduce and derive our measure of conditional variable importance, and use the example of exponential family distributions to provide intuition from a statistical point of view. In Section 2.4, we use simulated data to illustrate key differences between marginal and conditional measures of variable importance, and to demonstrate the ability of predictive mean matching to accurately capture unit level covariate importance for various covariate types (integer, skewed, categorical). Section 2.5 provides a detailed illustration of the method on real-world data from AMPATH. We develop and cross-validate a ML model using data from 685,715 visits on 70,626 patients from AMPATH; from that model, we calculate both marginal and conditional patient-level variable importance measures for a variety of patients who are flagged as being high risk for missed visit, and illustrate how the measures might be used to personalize pre-visit outreach. Section 2.6 concludes with a discussion and summary of directions for further work.

2.2 Measures of unit-level variable importance

2.2.1 Data Structure & Notation

We focus here on machine learning models for predicting a binary outcome. Let $\boldsymbol{X} = (X_1, \ldots, X_J)$ denote the $1 \times J$ vector of covariates, with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})$ its realized value for individual i, where $i = 1, \ldots, n$. We use the notation $\boldsymbol{x}_{i(-j)} = (x_{1i}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots, x_{iJ})$ to represent the list of covariates excluding x_{ij}. Conditional on $\boldsymbol{X}$, the binary response of interest Y follows a Bernoulli distribution $[Y \mid \boldsymbol{X}] \sim \text{Ber}\{\pi(\boldsymbol{X})\}$, where $\pi(\boldsymbol{X}) = \Pr(Y = 1 \mid \boldsymbol{X})$ is parameterized as

$$\pi(\boldsymbol{X}) \;=\; g(\boldsymbol{X}; \boldsymbol{\theta}),$$

and $\boldsymbol{\theta}$ is a finite-dimensional parameter with prior distribution $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$. The function $g : \mathscr{X} \to (0,1)$ is unknown. We further define a function $s : (0,1) \to \mathbb{R}$ that is a monotonic, 1-1 transformation of the predicted probability to the real line. This function can be thought

of a risk score. Common examples include the log-odds, the probit function and the identity function. For the remainder of this paper s will be referred to as the *risk score*, and generally written as $s(\boldsymbol{X})$ or $s(\boldsymbol{X};\boldsymbol{\theta})$, with the understanding that it is a transformation of $g(\boldsymbol{X};\boldsymbol{\theta})$. The work presented in this article considers Bayesian models but the methodology can be applied to any existing ML method. We elaborate in later sections the key features of the Bayesian approach.

2.2.2 Existing methods for unit-level covariate importance

In this section we review some existing unit-level importance measures, starting with SHAP (Lundberg and Lee, 2017). Let $\mathscr{J} = \{1,\ldots,J\}$ denote the set of covariate indices, and let $\mathscr{A} = 2^{\mathscr{J}}$ denote the power set of $\mathscr{J}$ – or the set of all possible subsets of $\mathscr{J}$. For an element $a \in \mathscr{A}$, define $\boldsymbol{X}_a$ and $\boldsymbol{X}_{(-a)}$ to be the collection of covariates that, respectively, are and are not indexed by elements of a.

The goal of SHAP is to explain why an individual's risk score deviates from the expected value of risk scores over the population, defined as $\varphi_0 = E[s(\boldsymbol{X};\boldsymbol{\theta})]$. The expected value can also be taken over a subset of the population (Ghalebikesabi et al., 2021), such as those who do not experience the event of interest ($Y = 0$). SHAP operationalizes unit-level explanations for each individual i by assuming that this difference can be decomposed into a linear, additive sum of values φ_{ij}, where each φ_{ij} is interpreted as the marginal contribution of the value of x_{ij} to an individual's deviation from the average risk score φ_0. The values of φ_{ij} are estimated using a method known as KernelSHAP (Lundberg and Lee, 2017), which solves for the φ_{ij} using the weighted least squares (WLS) loss function,

$$L_i(\boldsymbol{\varphi}) \;=\; \sum_{a \in \mathscr{A}} k(J,a)\left\{ v_i(a) - \left(\varphi_0 + \sum_{j \in \mathscr{A}} \varphi_{ij}\right)\right\}^2,$$

where $k(J,a) = (J-1)\Big/\left\{\binom{J}{|a|}|a|(J-|a|)\right\}$ denotes the weight associated with the subset of the covariates a, where $|a|$ is the cardinality of a and $v_i(a) = E[s(\boldsymbol{X}_{(-a)}, \boldsymbol{x}_{ia};\boldsymbol{\theta}) \mid \boldsymbol{X}_a = \boldsymbol{x}_{ia}]$ is

the expected value taken over the distribution $[\boldsymbol{X}_{(-a)} \mid \boldsymbol{X}_a = \boldsymbol{x}_{ia}]$. Now let $\boldsymbol{Z}$ be $2^J \times (J+1)$ matrix of binary indicators representing all possible combinations of inclusions/exclusions of the J features, $\boldsymbol{\nu}_i$ a $2^J \times 1$ vector containing all values $\nu_i(a)$, and $\boldsymbol{W}$ a $2^J \times 2^J$ diagonal weight matrix containing the corresponding weights $k(p,a)$ for each subset a, it can be shown that the WLS problem is optimized by $\boldsymbol{\varphi}_i = (\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{\nu}_i$ for individual i. The result is then a linear function $\varphi_0 + \sum_{j=1}^{J} \varphi_{ij}$ that approximates the original risk score of individual i.

In practice, there are specific challenges to calculating the SHAP values using this approach. Most notably, the calculation of the expected value $\nu_i(a)$ is difficult for each a due to the intractability of the conditional distributions $[\boldsymbol{X}_{(-a)} \mid \boldsymbol{X}_a = \boldsymbol{x}_{ia}]$ in real world data. Many implementations of SHAP simplify this calculation by assuming feature independence and performing Monte Carlo integration by sampling draws $\boldsymbol{x}^*_{(-a)}$ from the empirical distribution $\hat{P}(\boldsymbol{X}_{(-a)})$. This particular feature of kernelSHAP can be a limitation to interpreting the resulting variable importance measures in real world data (Frye et al., 2020, Aas et al., 2021, Olsen et al., 2024). The use of the marginal distribution $[\boldsymbol{X}_{(-a)}]$ over the appropriate conditional $[\boldsymbol{X}_{(-a)} \mid \boldsymbol{X}_a = \boldsymbol{x}_{ia}]$ can create covariate combinations that don't exist clinically and in general can increase the variance of the estimate. To address this, several methods have been proposed to approximate the needed conditional distributions. Aas et al. (2021) assumes joint normality of the covariate distribution $[\boldsymbol{X}] = [\boldsymbol{X}_a, \boldsymbol{X}_{(-a)}] \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are estimated using the training data. Using this method, samples $\boldsymbol{x}^*_{(-a)} \sim [\boldsymbol{X}_{(-a)} \mid \boldsymbol{X}_a = \boldsymbol{x}_{ia}]$ can be drawn from the implied conditional normal and plugged into the risk score to calculate $\nu_i(a)$. Olsen et al. (2024) proposes the use of an ML model to estimate $\nu_i(a)$ by regressing the risk score $s(\boldsymbol{X}_{(-a)}, \boldsymbol{x}_a; \boldsymbol{\theta})$ on the covariates X_a.

Other methods for calculating individual-level variable importance include individual conditional expectation (ICE) plots. ICE plots generate a curve of the chosen feature X_j against $s(X_j, \boldsymbol{x}_{i(-j)}; \boldsymbol{\theta})$ by permuting values of X_j from its marginal distribution while the fixing the valued for $\boldsymbol{X}_{(-j)}$ at the observed values $\boldsymbol{x}_{i(-j)}$ for an individual patient i. The average of all of the curves is the partial dependence plot (PDP) Goldstein et al. (2015). However, like SHAP, ICE

plots rely on sampling from the marginal distribution rather than the appropriate conditional distribution.

2.3 Methodology

2.3.1 Statistical Method for Patient Level Insights

Our proposed measure is designed to be applied in settings where the estimated risk score leads to classification of an individual into one of two classes. Specifically, we assume a decision rule $R : \mathbb{R} \to \{0,1\}$ that maps the risk score $s(\boldsymbol{X})$ to a binary classification $\widehat{Y} \in \{0,1\}$; i.e., for a specific covariate input $\boldsymbol{X} = \boldsymbol{x}$, $\widehat{Y} = R(s(\boldsymbol{x}))$. For ordered risk scores, a simple version of the rule is $R_c(s) = \mathbb{I}(s \geq c)$ for some threshold c. The rule R is assumed to be derived based on a pre-specified loss or utility function, and possibly other considerations; importantly, our method does not depend on how R is derived.

For convenience, we refer to those with $\widehat{Y} = 1$ as being classified as 'high risk'. The variable importance measure we describe here is designed to quantify, among those with $\widehat{Y} = 1$, the degree to which each covariate drives the 'high risk' classification. In the case of AMPATH, knowing which covariates drive an individuals's predicted risk of missing a visit can provide outreach workers in clinics a way to personalize outreach before the visit is missed, and increase the chance of preventing a missed visit.

Operationally, the measure is driven by comparing, for those with $\widehat{Y} = 1$, the observed value of the risk score $s(\boldsymbol{x}_i) = s(x_{ij}, \boldsymbol{x}_{i(-j)})$ to its expected value with respect to the conditional distribution $[X_j \mid \boldsymbol{X}_{(-j)} = \boldsymbol{x}_{i(-j)}, Y = 0]$. In words, we are comparing $s(x_{ij}, \boldsymbol{x}_{i(-j)})$ to its expected value among those with $Y = 0$, conditionally on the observed values $\boldsymbol{x}_{i(-j)}$ of all other covariates for that individual. The extent to which the observed risk score evaluated at x_{ij} differs from its expected conditional risk among those with $Y = 0$ quantifies the degree to which x_{ij} is driving the classification $\widehat{Y}_i = 1$. Formally, we define the importance of covariate X_j for individual i as

$$\phi(x_{ij}; \boldsymbol{\theta}) \;=\; s(\boldsymbol{x}_i; \boldsymbol{\theta}) - E\left\{s(X_j, \boldsymbol{x}_{i(-j)}; \boldsymbol{\theta}) \mid \boldsymbol{X}_{(-j)} = \boldsymbol{x}_{i(-j)}, Y = 0\right\}, \tag{2.1}$$

or more concisely, $\phi(x_{ij}) = s(\boldsymbol{x}_i) - E\{s(\boldsymbol{x}_i)|\boldsymbol{X}_{(-j)} = \boldsymbol{x}_{i(-j)}, Y = 0\}$. (We will sometimes write $\phi(x_{ij};\boldsymbol{\theta})$ when emphasis on the paramaterization of the prediction model is needed.)

The interpretation of $\phi(x)$ can be understood by considering some simple cases. Consider a single covariate X that arises from a two component mixture, with mixture components defined by the outcome Y and each mixture component coming from an exponential family density,

$$p_X(x) = \sum_{y=0}^{1} \lambda_y h(x) \exp\left\{\eta(\boldsymbol{\gamma}_y)T(x) - A(\boldsymbol{\gamma}_y)\right\},$$

where $\lambda_y = \Pr(Y = y)$, $\boldsymbol{\gamma}_y$ is the natural parameter, $T(x)$ is the sufficient statistic, $h(\cdot)$ is a function of x, and $A(\cdot)$ is a normalization function. Recall that $g(X;\boldsymbol{\theta}) = \Pr(Y = 1|X;\boldsymbol{\theta})$, and set the risk score $s(X;\boldsymbol{\theta})$ to be the log odds. With some algebra, it can be shown that the risk score takes the form

$$\begin{aligned} s(x;\boldsymbol{\theta}) &= \log\left\{\frac{g(x;\boldsymbol{\theta})}{1-g(x;\boldsymbol{\theta})}\right\} \\ &= \{\eta(\boldsymbol{\gamma}_1) - \eta(\boldsymbol{\gamma}_0)\}T(x) + \{A(\boldsymbol{\gamma}_1) - A(\boldsymbol{\gamma}_0)\} + \log(\lambda_1/\lambda_0), \end{aligned}$$

where $\boldsymbol{\theta} = (\lambda_0, \lambda_1, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1)$. In this setting, the implied prediction model for $[Y|X]$ is a logistic regression

$$\operatorname{logit}\Pr(Y = 1|X = x;\boldsymbol{\theta}) = \beta_0 + \beta_1 T(x),$$

where the coefficient of $T(x)$ is the difference $\eta(\boldsymbol{\gamma}_1) - \eta(\boldsymbol{\gamma}_0)$ in the canonical parameter and the intercept is $\beta_0 = \{A(\boldsymbol{\gamma}_1) - A(\boldsymbol{\gamma}_0)\} + \log(\lambda_1/\lambda_0)$.

In this illustrative example, $\phi(x;\boldsymbol{\theta})$ can be written as

$$\begin{aligned} \phi(x;\boldsymbol{\theta}) &= \{\eta(\boldsymbol{\gamma}_1) - \eta(\boldsymbol{\gamma}_0)\}\left[T(x) - E\{T(X)\,|\,Y = 0\}\right] \\ &= \beta_1\left[T(x) - E\{T(X)\,|\,Y = 0\}\right]. \end{aligned}$$

Hence the measure of individual-level covariate importance is the product of the difference between the *observed* covariate $T(x)$ and its expected value among the control cases, $E\{T(X) \mid Y = 0\}$ and the model coefficient β_1 (here, the difference in canonical parameters for the two mixture components). Hence in this unconditional case, unit-level importance is influenced by two factors: (i) the degree to which the *unit-specific* observed value x deviates from the distribution $[X \mid Y = 0]$; and (ii) the degree to which the conditional distributions of $[X \mid Y = 1]$ and $[X \mid Y = 0]$ differ from each other. This second factor reflects *global* variable importance, and is captured by the coefficient β_1 that represents the difference in canonical parameter values indexing the component distributions.

In a model with multiple predictors, the coefficient β_1 captures global importance conditional on other covariates in the model, and the expectation $E\{T(X) \mid Y = 0\}$ would be replaced with its conditional-distribution analogue. A key goal of our approach is to ensure that the individual-level comparison between x and $E(X \mid Y = 0)$ conditions on values of the other covariates for the same individual – which is not typically the case for other unit-level importance measures. To make this more concrete, Table 2.1 shows the functional form of $\phi(x; \boldsymbol{\theta})$ for cases where the mixture component distributions are well-known exponential family distributions.

Table 2.1: Functional form of covariate importance measure $\phi(x;\boldsymbol{\theta})$ for settings where the covariate X is a mixture of exponential family distributions over the outcome values $Y \in \{0,1\}$.

Covariate mixture distribution		
$[X \mid Y=1]$	$[X \mid Y=0]$	$\phi(x;\boldsymbol{\theta})$
$N(\mu_1, \sigma_1^2)$	$N(\mu_0, \sigma_0^2)$	$\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right)(x-\mu_0) + \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\left\{x^2 - (\mu_0^2 + \sigma_0^2)\right\}$
$N(\mu_1, \sigma^2)$	$N(\mu_0, \sigma^2)$	$\left(\frac{\mu_1 - \mu_0}{\sigma^2}\right)(x-\mu_0)$
$\text{Ber}(\mu_1)$	$\text{Ber}(\mu_0)$	$\{\text{logit}(\mu_1) - \text{logit}(\mu_0)\}(x-\mu_0)$
$\text{Mult}(1;\mu_{11},\ldots,\mu_{1K})$	$\text{Mult}(1;\mu_{01},\ldots,\mu_{0K})$	$\sum_{k=1}^{K} \log(\mu_{1k}/\mu_{0k})\,(x_k - \mu_{0k})$

2.3.2 Estimation of the $\phi(x;\boldsymbol{\theta})$

The value $\phi(x_{ij};\boldsymbol{\theta})$ for an individual i requires calculating both the risk score $s(\boldsymbol{x}_i;\boldsymbol{\theta})$ and the expected value of the risk score over the distribution $[X_j \mid \boldsymbol{X}_{(-j)} = \boldsymbol{x}_{i(-j)}, Y=0]$. The risk score follows immediately from the modeling procedure while the calculation of its conditional expectation is non-trivial due to potential correlation between covariates, combined with the fact that the joint distribution of the covariates is unknown in most practical settings. Assuming independence between covariates can lead to biased importance measures (Aas et al., 2021). In general, valid calculation of unit-level covariate importance requires knowledge about the joint distribution (Hooker et al., 2021) which, in practice, can be difficult to infer, and calculating unit-level importance measures can require substantial computational burden (Aas et al., 2021).

To address the need for incorporating covariate dependence in a way that is not computationally prohibitive, we proposed using predictive mean matching (Rubin, 1986, Little, 1988) to calculate the conditional expectations needed to estimate unit-level covariate importance. Predictive mean matching (PMM) originated in the missing data literature as a method to preserve the original covariate distribution and make imputations more robust to model mis-

specification. A key advantage of predictive mean matching is that it can be used to estimate a conditional mean without specifying a parametric model for $p_{\boldsymbol{X}}(\boldsymbol{x})$. It relies on a data reduction method, and approximates the needed conditional mean using samples drawn directly from the empirical distribution of the observed data, which preserves the properties of the original distribution.

Assuming we have a cross-validated prediction model $g(\boldsymbol{x};\boldsymbol{\theta})$, and a posterior sample $\{\boldsymbol{\theta}^{(b)} : b = 1,\ldots,B\}$, we use the following implementation of predictive mean matching to generate, for each individual i, a posterior sample of $E\left\{s(\boldsymbol{X};\boldsymbol{\theta}) \,|\, \boldsymbol{X}_{(-j)} = \boldsymbol{x}_{i(-j)}, Y = 0\right\}$, which is then used to obtain the posterior of the individual-level importance measure $\phi(x_{ij};\boldsymbol{\theta})$ using (2.1).

1. For a covariate X_j, fit the linear model $[X_j \,|\, \boldsymbol{X}_{(-j)}] \sim \mathcal{N}(\boldsymbol{X}_{(-j)}^{\mathrm{T}}\boldsymbol{\beta}, \sigma_j^2)$ to the subset of training data having $Y = 0$, and obtain a sample $\{\boldsymbol{\beta}^{(c)}, \sigma^{(c)} : c = 1,\ldots,C\}$ from the posterior $p(\boldsymbol{\beta}, \sigma \,|\, \text{data})$.

2. For each $i' \neq i$, calculate the predicted mean

$$\begin{aligned} \hat{x}_{i'j} &= E(x_{i'j} \,|\, \boldsymbol{x}_{i'(-j)}, \text{data}) \\ &= \int \boldsymbol{x}_{i'(-j)}^{\mathrm{T}}\boldsymbol{\beta}\; dP(\boldsymbol{\beta} \,|\, \text{data}). \end{aligned}$$

   Under the assumption of flat priors for $\boldsymbol{\beta}$ and σ, this is well approximated by $\boldsymbol{x}_{i'(-j)}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the OLS estimator.

3. For individual i, generate a sample $\tilde{x}_{ij}^{(c)} = \boldsymbol{x}_{i(-j)}^{\mathrm{T}}\boldsymbol{\beta}^{(c)}$, $c = 1,\ldots,C$ from the posterior predictive distribution of x_{ij}. In practice we find $C = 200$ to be a sufficient sample size.

4. For each c, calculate the $n-1$ absolute differences $d_{ii'j}^{(c)} = \left|\hat{x}_{i'j} - \tilde{x}_{ij}^{(c)}\right|$. Randomly select one value of $d_{ii'j}^{(c)}$ from the five smallest values in the sample, and let x_{cj}^{*} denote the value of $x_{i'j}$ corresponding to the randomly-selected difference measure. Repeating this process across each posterior predictive value $\tilde{x}_{ij}^{(c)}$ yields a 'donor set' $\{x_{cj}^{*} : c = 1,\ldots,C\}$ that can be viewed as approximating a sample from the distribution $[X_j \,|\, \boldsymbol{X}_{(-j)} = \boldsymbol{x}_{i(-j)}, Y = 0]$.

5. Use the donor set to calculate the approximation

$$E[s(\boldsymbol{X};\boldsymbol{\theta}^{(b)})\,|\,\boldsymbol{X}_{(-j)}=\boldsymbol{x}_{i(-j)},Y=0] \;\approx\; \frac{1}{C}\sum_{c=1}^{C} s(x^{*}_{cj},\boldsymbol{x}_{i(-j)};\boldsymbol{\theta}^{(b)}).$$

6. For posterior draw $\boldsymbol{\theta}^{(b)}$, the individual-specific importance measure is calculated as

$$\phi_{ij}(\boldsymbol{\theta}^{(b)}) \;=\; s(\boldsymbol{x}_i;\boldsymbol{\theta}^{(b)}) - \frac{1}{C}\sum_{c=1}^{C} s(x^{*}_{cj},\boldsymbol{x}_{i(-j)};\boldsymbol{\theta}^{(b)}),$$

which yields a sample $\{\phi_{ij}(\boldsymbol{\theta}^{(b)}) : b = 1,\ldots,B\}$ from its posterior distribution.

Using a multivariate version of step 1 above, this process can also extend to quantifying variable importance for a group of covariates rather than just one. For a subset $a \in \mathscr{A} = 2^{[J]}$ of covariate indices – where $|a| = k_a$, step 1 is to fit a multivariate regression $[\boldsymbol{X}_a \,|\, \boldsymbol{X}_{(-a)}, Y = 0] \sim \text{MVN}(\boldsymbol{X}_a\boldsymbol{B}, \boldsymbol{\Sigma})$, where $\boldsymbol{B}$ is a coefficient matrix of dimension $k_a \times (J - k_a)$ and the variance matrix $\boldsymbol{\Sigma}$ has dimension $k_a \times k_a$ – and to obtain the posterior distribution $p(\boldsymbol{B}\,|\,\text{data})$. Steps 2 through 5, appropriately modified, are then used to calculate an individual-level importance value for the covariates in set a. A flat prior is chosen for $p(\boldsymbol{\beta}, \sigma)$ in step 1, but a proper prior may be specified as well when the sample size is insufficient, or the end-user requires regularization.

The end result provides the user with a distribution of importance values $\phi_{ij}(\boldsymbol{\theta})$ for any covariate X_j or for groups of covariates $\boldsymbol{X}_a$ for $j \in a$. Subsequently, with each distribution, we can then rank features in terms of $\phi_{ij}(\boldsymbol{\theta})$ for each patient, providing clinicians with an understanding of which features have the greatest impact on an individual patient's risk score. Provided with the importance distributions for each feature, one can also rank features per individual

in terms of their impact for greater stability using the expected posterior rank, calculated as,

$$\begin{aligned} \text{rank}(X_{ij}) &= \sum_{j' \neq j} \Pr\left\{\phi_{ij}(\boldsymbol{\theta}) > \phi_{ij'}(\boldsymbol{\theta})\right\} \\ \Pr\left\{\phi_{ij}(\boldsymbol{\theta}) > \phi_{ij'}(\boldsymbol{\theta})\right\} &\approx \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{\phi_{ij}(\boldsymbol{\theta}^{(b)}) > \phi_{ij'}(\boldsymbol{\theta}^{(b)})\right\} \end{aligned}$$

Thus the end-user can be provided both with a distribution for the importance of features and a ranked list of features which can be provided to a clinician.

2.4 Simulation Study

We evaluate our proposed sampling methodology through a simulation study with varying feature type, prevalence and correlation structure. The aim of the simulation study is to assess the performance of predictive mean matching for quantifying individual-level conditional variable importance. Because we simulate from known – but complex – prediction models, we are able to calculate the true value of both marginal and conditional variable importance. The second objective is to compare our sampling procedure to existing measures of conditional variable importance (Aas et al., 2021), and to standard measures that rely on feature independence assumptions in calculating variable importance (i.e., marginal variable importance).

2.4.1 Simulated data models

Our model uses 3 features, $\boldsymbol{X} = (X_1, X_2, X_3)$. The joint distribution $[Y, \boldsymbol{X}]$ is specified as $[\boldsymbol{X} \mid Y][Y]$, with the marginal distribution of $\boldsymbol{X}$ following the mixture distribution

$$p(X_1, X_2, X_3) = \sum_{y \in \{0,1\}} p(X_1 \mid X_2, X_3, Y = y)\, p(X_2 \mid X_3, Y = y)\, p(X_3 \mid Y = y)\, \Pr(Y = y) \quad (2.2)$$

This formulation requires that we specify each of the 6 conditional distributions under the summation sign above, together with a value for $\psi = \Pr(Y = 1)$. For certain specifications of the distributions in (2.2), $[Y \mid \boldsymbol{X}]$ corresponds to a logistic regression (see Table 2.1).

For a given iteration of the simulation, a sample of individual-level data were generated sequentially using the conditional distributions; i.e., (i) $Y^* \sim \text{Ber}(\psi)$; (ii) $X_3^* \sim P(X_3 \,|\, Y^*)$; (iii) $X_2^* \sim P(X_2 \,|\, X_3^*, Y^*)$; (iv) $X_1^* \sim P(X_1 \,|\, X_2^*, X_3^*, Y^*)$. Next, data $(Y^*, \boldsymbol{X}^*)$ derived from steps (i-iv) were used to fit a model $Y \,|\, \boldsymbol{X} \sim \text{Bern}(\pi(\boldsymbol{X}))$, where $\pi(\boldsymbol{X}) = g(\boldsymbol{X}, \boldsymbol{\theta})$ was paramaterized using BART and logistic regression. Therefore, for each distinct specification of $[Y, \boldsymbol{X}]$, we have two parameterizations for $g(\boldsymbol{X}, \boldsymbol{\theta})$; for each combination.

To investigate performance across a wide variety of covariate types and association structures, our simulations were based on four model specifications of the conditional distribution $[\boldsymbol{X} \,|\, Y]$. For each scenario, the goal is to estimate unit-level variable importance for X_1, which requires calculating

$$\eta_0(\boldsymbol{x}_{(-1)}) \;=\; E\left[s(X_1, \boldsymbol{x}_{(-1)}; \boldsymbol{\theta}) \,|\, \boldsymbol{X}_{(-1)} = \boldsymbol{x}_{(-1)}, Y = 0\right],$$

where $s : (0,1) \to \mathbb{R}$ is the function (e.g., inverse logit) that transforms the $g(\boldsymbol{X}; \boldsymbol{\theta})$ into a risk score.

Model 1: Continuous covariate, multivariate normal model. We assume normal distributions $[\boldsymbol{X} \mid Y = y] \sim \text{MVN}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}_0 = (0,0,0)^{\text{T}}$, and $\boldsymbol{\mu}_1 = (1,1,1)^{\text{T}}$, and

$$\boldsymbol{\Sigma} \;=\; \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

with $\rho \in \{0, 0.25, 0.50, 0.75\}$.

Model 2: Binary covariate, Bernoulli model. Here we assume $[X_1 \,|\, X_2, X_3, Y] \sim \text{Ber}(\pi(X_2, X_3, y))$, with

$$\begin{aligned} \text{logit}(\pi(X_2, X_3, Y)) \;&=\; \alpha_0 + \alpha_1 X_2 + \alpha_2 X_3 + \alpha_3 X_2 X_3^2 \\ &\quad + Y\left(\beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_2 X_3^2\right), \end{aligned}$$

and that $[(X_2, X_3) \,|\, Y = y] \sim \text{MVN}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}_0 = (0,0)^{\text{T}}$, $\boldsymbol{\mu}_1 = (1,1)^{\text{T}}$, and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}.$$

We set $\alpha_1 = 0.70$, $\alpha_2 = 1$, $\alpha_3 = 0.50$, $\beta_1 = -0.10$, $\beta_2 = -0.80$, $\beta_3 = 0.30$. To modify the mean of X_1, we vary (α_0, β_0) over the set $\{(0.50, -1), (1, -2), (1.5, -3)\}$.

Model 3: Multinomial covariate. We assume $[X_2, X_3 \,|\, Y = y] \sim \text{MVN}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ using the same specification as in Model 2. For X_1, we use a 3-category multinomial specification $[X_1 \mid X_2, X_3, Y] \sim \text{Mult}(\boldsymbol{\pi}(X_1, X_2, Y))$, where

$$\begin{aligned} \log\left\{\frac{\pi_m(X_2, X_3)}{\pi_3(X_2, X_3)}\right\} &= \alpha_{m0} + \alpha_{m1} X_3 + \alpha_{m2} X_3^2 + \alpha_{m3} X_2 + \alpha_{m4} \sin(X_2 X_3) \\ &\quad + Y\left\{\beta_{m0} + \beta_{m1} X_3 + \beta_{m2} X_3^2 + \beta_{m3} X_2 + \beta_{14} \sin(X_2 X_3)\right\}, \qquad m = 1, 2. \end{aligned}$$

For category $m = 1$ we set $\beta_{10} = 2$, $\beta_{11} = -0.60$, $\beta_{12} = -0.10$, $\beta_{13} = 0.10$, $\beta_{14} = -0.20$, $\beta_{15} - 1.20$, $\beta_{16} = -0.70$, $\beta_{17} = -0.20$, $\beta_{18} = -0.20$, $\beta_{17} = 0.40$. For category $m = 2$ we set $\beta_{20} = 0.10$, $\beta_{21} = -0.10$, $\beta_{22} = 0.05$, $\beta_{23} = 0.50$, $\beta_{24} = 0.50$, $\beta_{25} = 0.90$, $\beta_{26} = 0.40$, $\beta_{27} = -0.25$, $\beta_{28} = 0.40$, $\beta_{29} = -0.90$.

Model 4: Lognormal covariate. We assume that $[X_1 \,|\, X_2, X_3, Y]$ is distributed as a log normal distribution, $[X_2 \,|\, X_3, Y]$ is Bernoulli, and $[X_3 \,|\, Y]$ is normal, according to the following

parameterizations:

$$
\begin{aligned}
[X_3 \,|\, Y = y] &\sim N(\mu_y, 1), \\
\mu_y &= y
\end{aligned}
$$

$$
\begin{aligned}
[X_2 \,|\, X_3, Y] &\sim \text{Ber}(\pi(X_3, Y)), \\
\text{logit}(\pi(X_3, Y)) &= \beta_0 + \beta_1 X_3 + \beta_2 log(1 + |X_3|) + \beta_3 Y + \beta_4 Y X_3 + \beta_5 Y log(1 + |X_3|)
\end{aligned}
$$

$$
\begin{aligned}
[X_1 \,|\, X_2, X_3, Y] &\sim \text{log-Normal}(\nu(X_2, X_3, Y), 1), \\
\log(\nu(X_2, X_3, Y)) &= \gamma_0 + \gamma_1 X_2 + \gamma_2 X_3 + \gamma_3 sin(X_2 X_3^2) + \gamma_4 Y + \gamma_5 Y X_2 + \gamma_6 Y X_3 + \gamma_7 Y sin(X_2 X_3^2)
\end{aligned}
$$

We set coefficient values as follows: For the Bernoulli distribution governing X_2, $\beta_0 = -1$, $\beta_1 = 1$, $\beta_2 = -0.10$, $\beta_3 = 2$, $\beta_4 = 0.10$ and $\beta_5 = -0.40$; for the log normal model of X_3, $\gamma_0 = -1$, $\gamma_1 = 0.50$, $\gamma_2 = 0.20$, $\gamma_3 = 0.20$, $\gamma_4 = -0.50$, $\gamma_5 = 0.20$, $\gamma_6 = -0.10$, and $\gamma_7 = 0.20$. The parameter values were chosen to ensure that the data-generating mechanisms were stable and well-behaved across simulation runs. Coefficients were chosen to induce sufficiently strong but plausible associations so that the performance of the proposed method could be meaningfully evaluated.

2.4.2 Prediction models

For each of the data generating distributions described in section 2.4.1, we used Bayesian Additive Regression Trees (BART) (Chipman et al., 2010) and logistic regression to parameterize $g(\boldsymbol{X}, \boldsymbol{\theta}) = \pi(\boldsymbol{X}) = \Pr(Y = 1 \,|\, \boldsymbol{X})$. The logistic regression assumes $g(\boldsymbol{X}; \boldsymbol{\theta})$ is additive in the logit scale, with effects of continuous covariates specified using natural cubic splines with four degrees of freedom and categorical covariates with C levels coded with $C - 1$ indicator variables.

We implemented BART using a probit formulation

$$\Phi^{-1}\{g(\boldsymbol{X};\boldsymbol{\theta})\} \quad = \quad \sum_{t=1}^{T} f(\boldsymbol{X};\tau_t,\mathscr{M}_t),$$

where $\Phi(\cdot)$ is the normal CDF, T is the number of trees, and $f(\boldsymbol{X};\tau_t,\mathscr{M}_t)$ is a distinct regression tree. The tree structure is parameterized by $\boldsymbol{\theta} = \{\tau_t,\mathscr{M}_t\}_{t=1}^{T}$ where τ_t is the tree structure and $\mathscr{M}_t = \{\mu_k\}_{k=1}^{b_t}$ is a set of b_t terminal nodes for τ_t. The priors are specified by the distributions $p(\mu_{b_t} \mid \tau_t)$ and $p(\tau_t)$ for $t = 1,\ldots,T$ and $b = 1\ldots,B$. The prior $p(\mu_{bt} \mid \tau_t)$ is normally distributed with location parameter equal to 0 and scale parameter equal to $3/(k\sqrt{t})$. The prior $p(\tau_t)$ on the trees is proportional to $\alpha(1+d)^{-\beta}$, where d is the tree depth. In this analysis, k is set to 2, and $(\alpha,\beta) = (0.95,2)$. BART is fit using the BART package (Sparapani et al., 2021) in R Version 4.4.0. We run 10000 MCMC iterations taking the first 5000 as burn-in and then taking every 5th sample thereafter as a posterior draw to obtain a total of 1000 posterior draws.

2.4.3 Calculating variable importance

Recall that the key quantity required for calculating unit-level variable importance of X_1 is

$$\eta_0(\boldsymbol{x}_{-1}) \quad = \quad E[s(X_1,\boldsymbol{x}_{(-1)};\boldsymbol{\theta}) \mid \boldsymbol{X}_{(-1)} = \boldsymbol{x}_{(-1)}, Y = 0].$$

For each distinct data model described in Section 2.4.1 we first simulate a test set of size $N_{Te} = 5000$ from $[Y,\boldsymbol{X}]$ that is independent of the simulation iterations. The size of the test set is chosen to be sufficiently large in order sufficiently represent all plausible points in the distribution. At each iteration $t = 1,\ldots,T$ we then carried out the following steps of the simulation: (i) simulate a sample of size $N = 2000$ from $[Y,\boldsymbol{X}]$ which constitutes the training data; (ii) fit the logistic and BART models described above to the training data from step (i); (iii) calculate individual-specific estimates $\hat{\eta}_0^{(t)}(\boldsymbol{x}_{-1,i})$ of $\eta_0(\boldsymbol{x}_{-1,i})$ for each individual $i = 1,\ldots,N_{Te}$ in the test set for the fitted logistic and BART models using three different methods;

(iv) calculate the sample-specific bias associated with each method. These steps were repeated $T = 250$ times.

The iteration-specific estimates $\hat{\eta}_0^{(t)}(\boldsymbol{x}_{-1,i})$ were calculated using (i) the marginal distribution $[X_1 \mid Y = 0]$, which is the basis for commonly-used implementations of the SHAP algorithm; (ii) the method proposed by Aas (Aas et al., 2021) based on assuming $[X_1 \mid \boldsymbol{X}_{-1} = \boldsymbol{x}_{-1i}, Y = 0]$ follows a conditional normal distribution with linear dependence on $\boldsymbol{x}_{-1i}$; and (iii) our proposed method based on predictive mean matching. These distribution were estimated at each iteration t of the simulation using the iteration-specific simulated data. We drew $C = 200$ samples from each method.

The bias of each estimate is obtained by comparing $\hat{\eta}_0(\boldsymbol{x}_{-1i})$ to the true value of $\eta_0(\boldsymbol{x}_{-1i})$ which can be calculated by using the analytic form of the conditional distribution $[X_1 \mid \boldsymbol{X}_{-1} = \boldsymbol{x}_{-1i}, Y = 0]$. We calculate the average value of $\hat{\eta}_0(\boldsymbol{x}_{-1i}) = (1/T)\sum_{t=1}^{T} \hat{\eta}_0^{(t)}(\boldsymbol{x}_{-1i})$ for all individuals in the test set, and calculate the bias $(1/N_{Te})\sum_{i=1}^{N_{Te}} \left\{\hat{\eta}_0^{(t)}(\boldsymbol{x}_{-1i}) - \eta_0^{(t)}(\boldsymbol{x}_{-1i})\right\}$ for each iteration. We report the mean bias and 95% Monte Carlo intervals of the bias for each method over all of the simulations. Individual level biases $\{\hat{\eta}_0(\boldsymbol{x}_{-1i}) - \eta_0(\boldsymbol{x}_{-1i})\}$ were also calculated for each subject in the test set.

2.4.4 Simulation: Results

To illustrate the nature of the covariate effects, Figure 2.1 shows the partial effect plots for model 4 from the simulation. Tables 2.2 and 2.3 present results of the simulation study, showing the bias in estimating $\eta_0(\boldsymbol{x}_{-1i})$ across a range of data generating mechanisms. In Table 2.2, the marginal probability $\psi = P(Y = 1)$ is set to $\psi = .50$; in Table 2.3, $\psi = .30$. The first section of each table shows bias when the covariates follow a multivariate normal distribution. When covariates are uncorrelated ($\rho = 0$), all three methods of quantifying unit-level variable importance (PMM, Gaussian conditional, and marginal) are essentially unbiased. When the covariates are correlated ($\rho > 0$), the Gaussian conditional method remains unbiased while, not surprisingly, the marginal method shows bias that increases in magnitude with ρ. A similar

pattern emerges for cases where X_1 is binary. For each of the binary cases, X_1 is correlated with (X_2, X_3), leading to bias when using marginal VI methods. The Gaussian conditional method begins to break down as dependence increases; meanwhile, the PMM method exhibits little or no bias. Similar patterns hold for the cases where X_1 is conditionally multinomial and log-normal, with the Gaussian conditional model performing equally well to the PMM in the log-normal case.

While the simulations do not cover all covariate dependence structures, they highlight the ability of our predictive mean matching based method to produce results of similar accuracy to the analytic distribution regardless of prevalence and across a variety of covariate types. PMM shows low or no bias across the full range of scenarios considered, performing equally well or better than variable importance methods based on marginal distributions and Gaussian approximations. One benefit that predictive mean matching has over other methods is the preservation of the original data distribution and is relatively simple to calculate analytically, relying on simple matrix multiplication that can be easily stored for future use.

To further illustrate the importance of accounting for feature correlation, Figure 2.2 shows unit-level variable importance measures for SHAP calculated for two distinct covariate profiles using one dataset simulated from model 1 (multivariate normal). The unit-level covariate values, for hypothetical individuals a and b, are $\boldsymbol{X}_a = (3, 0.5, 0.25)$ and $\boldsymbol{X}_b = (-0.25, 1.5, 2.5)$. These values are chosen such that $\pi(\boldsymbol{X}_a) \approx \pi(\boldsymbol{X}_b)$. While $\pi(\boldsymbol{X}_a)$ and $\pi(\boldsymbol{X}_b)$ vary as a function of the pairwise correlation ρ, the equality holds. Specifically, for $\rho = 0, .3, .5$, we have $s(\pi(\boldsymbol{X}_a)) = s(\pi(\boldsymbol{X}_b)) = 2.25, 1.41, 1.12$, respectively, where $s(\cdot)$ is the logit transformation. The SHAP values are calculated using the conditional distribution and marginal distribution of the covariates from Model 1.

It should be noted that when using the log-odds as the risk score to be explained that importance values are independent of the prevalence. In Figure 2.2, we see that the unit-level VI is equivalent for both the marginal and conditional calculation methods. The variable with the highest importance is X_1 for individual a and X_3 for individual b. As ρ increases to .3

and .5, the ordering of variable importance is preserved; however the magnitude and even direction of the variable importance measures differ between the marginal and conditional methods. The discrepancies are particularly notable for covariate X_3 of individual a, where marginal and conditional importance have opposite sign; and covariate X_1 of individual b, where the conditional importance measure amplifies the importance as ρ goes from 0 to .5.

2.5 Application

We illustrate the application of our method on data from AMPATH. Our analysis makes use of clinical level data on 70,626 adults with HIV aged 18 years or greater who were actively enrolled in care on or after January 1st, 2021. We define being 'actively enrolled' as having at least one AMPATH clinic visit on or after January 1st, 2021. The analysis used data on 685,715 visits that have occurred since January 1st, 2021. Covariate data collected prior to January 1, 2021 was used, when available, for patients who were included in the analysis sample. The cutoff date for inclusion into the sample was selected because it marks the period where AMPATH began to loosen COVID-19 related restrictions for in-person clinical encounters. The date of database closure was July 31st, 2023.

2.5.1 Data structure, notation, endpoint definition

Clinical visit information from individuals who are engaged in medical care within AMPATH is stored in an electronic database known as the AMPATH Medical Record System (AMRS) which is generated by a browser-based front-end application deployed at the point of care in associated facilities (hospital and clinics) Tierney et al. (2007). For the purposes of model specification, recall that patients are indexed by i and visits by ℓ. Once a patient is linked to HIV care in AMPATH, a set of immunological, clinical, and demographic covariates are recorded at enrollment and at subsequent visits; denote these by $\{\boldsymbol{V}_\ell : \ell = 0, 1, 2, \ldots\}$, where $\boldsymbol{V}_\ell$ is a $J \times 1$ vector. Because certain markers of disease progression such as CD4 count and viral load are usually measured on a fixed schedule, we define a corresponding $J \times 1$ vector $\boldsymbol{M}_\ell$ of binary in-

dicators for whether the each component of $\boldsymbol{V}_\ell$ has been observed. For completeness we define $\boldsymbol{M}_\ell$ for every covariate in $\boldsymbol{V}_\ell$ but in practice only a subset are used in modeling. Specifically, missingness indicators are coded for CD4 count, WHO stage, education level, marital status and body mass index (BMI).

At the end of visit ℓ, a patient is assigned a return to care (RTC) date for a regular return visit for symptom monitoring or a medication pickup for their ARV medicine, denoted by $S_{\ell+1}$. Often times these two can coincide. The endpoint for our prediction model is a binary indicator $Y_{\ell+1}$ that takes value 1 if the patient fails to return before their RTC date $S_{\ell+1}$, and 0 if they do return on time; i.e., $Y_{\ell+1}$ is a 'missed visit' indicator.

We can formalize this as follows. For a patient at visit ℓ, let $W_{\ell+1} > 0$ denote the waiting time until that patient shows up for their encounter at visit $\ell+1$, and let $\Delta \geq 0$ denote a 'grace period' interval. Our binary missed-visit indicator, now annotated by Δ, is defined as $Y^{\Delta}_{\ell+1} = \mathbb{I}(W_{\ell+1} > S_{\ell+1} + \Delta)$. We use this formulation because specific values of Δ have contextual relevance for programmatic reporting: setting $\Delta = 0$ flags 'defaulters', or those who do not make their visit on or before the assigned date; $\Delta = 28$ flags those at risk for interruption in treatment, because antiviral medication prescriptions are frequently given in 28-day supplies; $\Delta = 90$ flags those who are at risk for loss to follow up or possibly mortality (Kenya, 2022). Covariate and retention history available at visit ℓ can be summarized as $\boldsymbol{X}_\ell = (\boldsymbol{V}_\ell, \boldsymbol{M}_\ell, \boldsymbol{W}_\ell, Y^{\Delta}_\ell, S_{\ell+1})$. All clinical and retention information up to and including visit ℓ is denoted by $\overline{\boldsymbol{X}}_\ell = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_\ell)$ using overbar notation. We summarize observed covariate information in Table 3.2 for the cohort and stratify by those who enrolled before the date of the sampling window and those who enrolled after the date of the sampling window.

2.5.2 Model specification

We specify our model according to the form described in Section 2.2.1. Specifically we assume that missed-visit status $Y^{\Delta}_{\ell+1}$ follows a Bernoulli distribution $[Y^{\Delta}_{\ell+1} \mid \overline{\boldsymbol{X}}_\ell] \sim \text{Ber}\left\{\pi^{\Delta}(\overline{\boldsymbol{X}}_\ell)\right\}$, where $\pi^{\Delta}(\overline{\boldsymbol{X}}_\ell) = \Pr(Y^{\Delta}_{\ell+1} = 1 \mid \overline{\boldsymbol{X}}_\ell)$ is parameterized as $\pi^{\Delta}(\overline{\boldsymbol{X}}_\ell) = g^{\Delta}_{\ell+1}(\overline{\boldsymbol{X}}_\ell, \boldsymbol{\theta}_\ell)$, and $\boldsymbol{\theta}_\ell$ is a finite-

dimensional parameter with prior distribution $\boldsymbol{\theta}_\ell \sim p(\boldsymbol{\theta}_\ell)$ for $\ell = 1, \ldots, \max_i L_i$. For this analysis we focus on modeling probability of defaulting ($\Delta = 0$), although this method can be applied for any other value of Δ.

To parameterize $g^{\Delta}_{\ell+1}(\overline{\boldsymbol{X}}_\ell, \boldsymbol{\theta}_\ell)$, we use both Bayesian Additive Regression Trees (BART) and main-effects logistic regression where continuous variables are modeled using regression splines. We implement a smoother version of the pattern submodel approach from Fletcher Mercaldo and Blume (2020) where the missingness indicators $\boldsymbol{M}_\ell$ are instead included as direct effects into the model rather than implementing a full stratification approach by missingness pattern. We justify this simplification due to key covariates such as WHO stage and education level containing small percentages of missing values which induces sparse missingness patterns in the data. Covariates with missing values then enter into the outcome model, and the predictive mean matching models from section 2.3.2, as $X_{\ell j}(1 - M_{\ell j})$ along with the corresponding indicator $M_{\ell j}$. For logistic regression, the knots for the spline terms are chosen from the empirical observed distribution of $X_{\ell j}$. This avoids the need for imputation and enables missingness itself to be predictive of the outcome. If a covariate $X_{\ell j}$ has missing values we fit the linear regressions for the PMM sampling method to the subset of the observed values of $X_{\ell j}$ (i.e. $M_{\ell j} = 0$). Given the longitudinal structure of the data we implement a last value carried forward approach where we assume a patient's value for a covariate will stay constant until the next update. This means that if a covariate $X_{\ell j}$ remains unmeasured starting at enrollment until visit ℓ' the missingness indicator $M_{\ell j} = 1$ for all visits $\ell < \ell'$ and $M_{\ell j} = 0$ for all visits $\ell \geq \ell'$.

We make the assumption that a patient's status at visit $\ell + 1$, conditional on $\overline{\boldsymbol{X}}_\ell$, follows a third order Markov assumption where the most recent retention status depends only on the most recent clinical information and the retention history of the three most recent visits. We create separate models for $\ell = 1, 2, 3$ and a fourth model that combines all visits $\ell \geq 4$.

In the AMPATH data, the waiting time $W_{\ell+1}$ can be administratively censored by database closure date for cases where that date precedes a scheduled visit. We assume administrative

censoring is non-informative and exclude records where this can occur. Censoring by death before a visit is also possible and represents a competing risk in the estimation of retention. This is handled through a probabilistic chained classifier approach where we introduce an auxiliary binary outcome variable $D_{\ell+1}$ defined by survival long enough to have visit $\ell+1$ ($D_{\ell+1}=1$) or death before visit $\ell+1$ ($D_{\ell+1}=0$). Our expansion of the original probability then becomes,

$$\begin{aligned}\Pr(Y^{\Delta}_{\ell+1}=1\mid\overline{\boldsymbol{X}}_{\ell}) &= \sum_{d\in\{0,1\}}\Pr(Y^{\Delta}_{\ell+1}=1\mid\overline{\boldsymbol{X}}_{\ell},D_{\ell+1}=d)\Pr(D_{\ell+1}=d\mid\overline{\boldsymbol{X}}_{\ell})\\ &= \Pr(Y^{\Delta}_{\ell+1}=1\mid\overline{\boldsymbol{X}}_{\ell},D_{\ell+1}=1)\Pr(D_{\ell+1}=1\mid\overline{\boldsymbol{X}}_{\ell})\end{aligned}$$

This simplification occurs because the probability $\Pr(Y^{\Delta}_{\ell+1}=1\mid\overline{\boldsymbol{X}}_{\ell},D_{\ell+1}=0)$ is identically zero; i.e., a patient who dies before their scheduled visit cannot miss their scheduled visit. We describe the prevalence of the outcomes (i.e. early/on time visit, missed visit, death before a visit) for each visit model in Table 3.1.

2.5.3 Performance metrics and model evaluation

We implement a modified time-based K-fold cross validation procedure to train the model in order to accurately mimic the data generating process and how the model could be used in practice. The data are split into a training set consisting of patients who were scheduled to return to clinic between January 2021 and December 2023, and a test set of patients who were scheduled to return between January 2023 and June 2023. The test data are further split by the month of the RTC date (e.g. January 2023,..., May 2023, June 2023) to achieve 6 test folds. We train the model one time on the training data, and evaluate the model's performance on each of the test folds. We evaluate model performance using area under the receiver-operator characteristic curve (AUROC), as well as positive predictive value (PPV), sensitivity and specificity. For these latter measures, we use a cutoff corresponding to the 80^{th} percentile of predicted probabilities, motivated by considerations related to the capacity for

pre-visit patient outreach.

2.5.4 Model performance and global variable importance measures

Performance metrics for the cross-validated prediction models are given in Table 2.4. Predictive accuracy is higher for models based on later visits, primarily because previous visit history is a strong predictor of future return to care. Moreover, the more complex BART model has similar accuracy to logistic regression across all settings. This is in agreement with other studies which also showed that logistic regression did as well as less interpretable models (Ridgway et al., 2021).

Global covariate importance measures from BART and logistic regression models are summarized in Figures 2.3 and 2.4, respectively. Variable importance for BART is based on the number of times that a specific covariate is split upon, and for logistic regression we calculate a scalar value for the global variable importance as the square root of the Wald statistic divided by its degrees of freedom as previously done in the first chapter. We see similar trends, specifically that prior visit history takes on a more prominent role for models designed to predict return to care at later visits; in fact these are the primary predictive variables for models of visits $\ell \geq 4$. Earlier visit models such as the first visit highlight clinical and demographic characteristics as being most important along with the timing of the scheduled visit.

2.5.5 Conditional and marginal unit-level variable importance

We calculated individual-level variable importance values $\phi_{ij}(\boldsymbol{\theta})$ from the $\ell = 1, 2, 3$ retention models for six patients based on the marginal empirical distribution and using our conditional method based on the predictive mean matching described in Section 2.3.2. Because we are using Bayesian inference, our representation of variable importance is done in terms of a posterior distribution rather than fixed values; this enables inference about variable importance by comparing the posterior distribution to a reference null value (zero for log odds scale).

Figure 2.5 is designed to illustrate how individual-level variable importance profiles can

be heterogeneous across patients who have similar predicted risk scores. It shows conditional individual-level variable importance for four patients, labeled A, B, C, D, whose probability of missed visit is between .64 and .67. Patients A and C are registered at the same county hospital clinic and are scheduled for a return visit in January 2023; patients B and D are registered at the same health center / dispensary and are scheduled to return in April 2023.

For patients A and B, the model is predicting probability of missing the first follow-up visit after enrollment ($\ell = 1$). The panels in the top row of Figure 2.5 show that the most influential predictors of risk are, for patient A, timing of the next scheduled visit, clinic size, and being WHO stage 1 (a measure of HIV disease severity ranked from 1 to 4 in terms of disease progression). For patient B, appointment month, age, and scheduled visit time contribute to higher risk score, while clinic size and being WHO stage 2 keep the score from being even higher. The WHO stage indicator is particularly interesting. For patient A, being WHO stage 1 contributes to a *higher* missed visit risk, while for patient B, being WHO stage 2 contributes to a lower risk. This reflects the phenomenon that many patients who are in early stage of HIV may not be experiencing negative effects of the infection; as such, they can be prone to skipping visits, and when they get sicker they are more likely to make their visit on time.

We now turn to patients C and D, whose predictions correspond to probability of missed visit at the third follow up time ($\ell = 3$), and whose variable importance measures are shown in the bottom two panels of Figure 2.5. For patient C, the explanation for having increased risk is multi-factorial: reason for previous visit (drug pickup), a delay of 1-7 days in keeping the previous appointment, and clinic size (large clinic) all contribute to being at higher risk. At the same time, female sex and having shown up early on a previous visit, which have negative variable importance, appear to keep the risk score from being even higher. For patient D, the key drivers of risk are having delays of 1-7 days on two of the most recent 3 scheduled visits. In practical terms, variable importance plots such as these can help outreach workers customize their pre-visit messaging to patients who are flagged as being high risk for a missed visit.

Figure 2.6 focuses on comparisons between marginal and conditional variable importance.

It shows results for two patients, E and F, whose predicted probability of a missed visit is around .8. For patient E, while many of the measures are similar based on method, the importance of clinic size goes in opposite direction for conditional (negative impact) and marginal (positive impact) measures – to the extent that the respective 95% credible intervals (CI) both exclude zero. For patient F, we see that the marginal method is likely overstating the impact of a long delay in keeping a previous appointment. The conditional method places the importance of that variable to be roughly equal to reason for previous visit (drug pickup) and being at a county hospital. This is supported by Figure 2.7 which presents heatmaps of the inverse correlation coefficients for the covariates used in the stratified return to visit models. In practical terms, this could lead an outreach worker to focus on whether the instructions for future appointments are equally clear at drug pickups – which can be informal – compared to regular clinic visits.

2.6 Discussion

In this chapter we proposed a method to quantify feature importance for statistical and machine learning models in a binary setting by comparing the individual risk score to the conditional expectation of the risk score under a reference distribution – in this case, for the population where $Y = 0$. The key contribution is a predictive mean matching algorithm that enables individual-level variable importance measures to take into account the values of other, possibly correlated, variables for that same individual. At present, standard implementations of methods for individual-level variable importance, such as SHAP and ICE, do not condition on other variables. Recent developments include using a normal approximation to the joint covariate distribution, but this can be inappropriate when pairwise associations are nonlinear or covariates are not continuous. Our method is model-agnostic in the sense that it can be used in conjunction with any rule or algorithm that is designed to predict probability of a binary outcome. A key motivation for our methodology is the need to customize pre-visit outreach for patients in HIV care in order to improve the likelihood of patients at high risk for

disengagement to keep their scheduled appointments. This in turn is critical for patients to stay current with their antiviral medications, monitor key biomarkers such as viral load, and prevent further clinical progression and ensure it is identified and addressed if it occurs.

Through simulations, we showed that our sampling method for calculating the conditional expectation of the risk score over the distribution $[X_j \mid \boldsymbol{X}_{(-j)}, Y = 0]$ can accurately capture the true effect regardless of prevalence, correlation and data type, and mimic knowing the actual conditional distribution. In the simulation studies, bias estimates for both BART and logistic regression were broadly similar across methods, demonstrating the applicability of the proposed approach across a range of modeling techniques. Bias was most pronounced when samples from the marginal distribution were used to approximate conditional expectations, underscoring the importance of properly accounting for correlation among covariates. For continuous covariates, bias levels were comparable between predictive mean matching (PMM) and the Gaussian conditional approximation; however, for binary and categorical covariates, PMM yielded substantially lower bias, highlighting its advantage over commonly used sampling strategies. With our approach we obtain samples from the appropriate conditional and therefore can calculate other summary statistics of interest such as the median. In our data application, BART's ability to capture complex interactions without explicit specification did not outweigh the advantages of a simpler, computationally faster model such as logistic regression in terms of prediction. This finding echoes prior work cautioning against the use of black-box models in high-stakes decision-making settings Rudin (2019).

There are several limitations of our approach in this work. First is that predictive mean matching does not induce a coherent joint distribution over the covariates, in contrast to the Gaussian approximation. While our simulation studies suggest that this limitation does not materially affect performance, it remains a conceptual drawback. Second, because PMM relies on matching observed values, estimates of conditional expectations may be unstable or inaccurate for rare covariate combinations or in regions of the covariate space with limited empirical support. Addressing these challenges represents an important direction for future research.

Finally, the proposed methodology is developed for binary prediction settings. Extending this framework to multinomial outcomes, where multiple mixture components may be present, or to continuous outcomes, where expectations over regions of the outcome support may be of interest, is a natural avenue for future work.

In conclusion, by developing both a local variable importance measure and a conditional sampling approach that appropriately accounts for dependence among covariates, this work enables more informative, patient-specific risk assessment and supports targeted pre-visit outreach for patients in HIV care at AMPATH, with the potential to improve appointment adherence and reduce preventable disengagement from care.

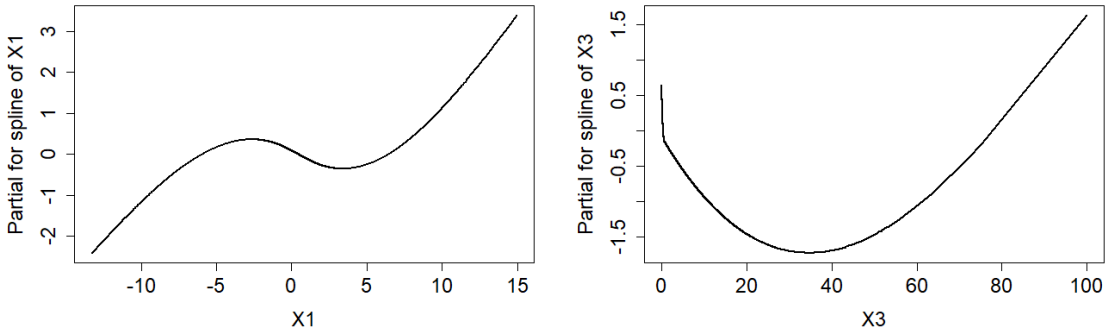

Figure 2.1: Marginal effect plots for covariates X_1 and X_3 from model 4 of the simulation study, where $[X_1 \,|\, Y]$ has a conditional normal distribution and $[X_3 \,|\, X_2, X_3, Y]$ is conditionally lognormal.

Table 2.2: Bias and 95% Monte-Carlo intervals for the bias of the expected value $\eta_0(\boldsymbol{x}_{i(-1)}) = E[s(g(\boldsymbol{X};\boldsymbol{\theta})) \mid \boldsymbol{X}_{(-1)} = \boldsymbol{x}_{i(-1)}, Y = 0]$ for different covariate types. The prevalence is set to $\psi = P(Y = 1) = 0.50$.

	Bias		
Covariate Type	PMM	Gaussian Conditional	Marginal
Normal ($\rho = 0$)			
BART	–0.02 (–0.23, 0.16)	0.00 (–0.23, 0.23)	0.01 (–0.20, 0.22)
Logit	–0.01 (–0.22, 0.16)	0.01 (–0.21, 0.21)	0.00 (–0.19, 0.21)
Normal ($\rho = 0.25$)			
BART	–0.03 (–0.19, 0.14)	–0.02 (–0.19, 0.14)	0.28 (0.09, 0.45)
Logit	–0.01 (–0.13, 0.16)	–0.01 (–0.16, 0.15)	0.26 (0.08, 0.41)
Normal ($\rho = 0.50$)			
BART	–0.02 (–0.18, 0.13)	–0.02 (–0.17, 0.13)	0.35 (0.12, 0.52)
Logit	–0.01 (-0.17, 0.12)	–0.01 (–0.16, 0.13)	0.33 (0.12, 0.51)
Normal ($\rho = 0.75$)			
BART	0.02 (–0.10, 0.15)	0.02 (–0.10, 0.15)	0.38 (0.17, 0.61)
Logit	–0.01 (–0.13, 0.11)	–0.01 (–0.13, 0.11)	0.33 (0.13, 0.57)
Binary Case 1			
BART	0.00 (–0.02, 0.03)	0.01 (–0.02, 0.06)	0.09 (–0.15, 0.26)
Logit	0.02 (–0.04, 0.08)	0.03 (–0.03, 0.10)	0.11 (-0.11, 0.30)
Binary Case 2			
BART	0.01 (–0.05, 0.06)	0.06 (–0.01, 0.13)	0.44 (0.20, 0.61)
Logit	0.02 (–0.07, 0.10)	0.07 (–0.02, 0.15)	0.45 (0.21, 0.64)
Binary Case 3			
BART	0.01 (–0.08, 0.09)	0.13 (0.02, 0.24)	0.78 (0.54, 1.01)
Logit	0.00 (–0.13, 0.10)	0.13 (0.02, 0.24)	0.77 (0.55, 1.02)
Multinomial			
BART	0.03 (–0.10, 0.18)	0.09 (–0.02, 0.22)	0.35 (0.23, 0.47)
Logit	0.05 (–0.09, 0.19)	0.11 (–0.02, 0.25)	0.37 (0.23, 0.50)
Log Normal			
BART	0.05 (0.00, 0.13)	0.05 (–0.06, 0.19)	–0.26 (–0.40, –0.13)
Logit	0.04 (–0.05, 0.17)	0.04 (–0.16, 0.23)	–0.27 (–0.45, –0.11)

Table 2.3: Bias and 95% Monte-Carlo intervals for the bias of the expected value $\eta_0(\boldsymbol{x}_{i(-1)}) = E[s(g(\boldsymbol{X};\boldsymbol{\theta})) \mid \boldsymbol{X}_{(-1)} = \boldsymbol{x}_{i(-1)}, Y = 0]$ for different covariate types. The prevalence is set to $\psi = P(Y = 1) = 0.30$.

	Bias		
Covariate Type	PMM	Gaussian	Marginal
Normal ($\rho = 0$)			
BART	0.06 (–0.27, 0.40)	0.04 (–0.17, 0.30)	0.05 (–0.15, 0.28)
Logit	–0.01 (–0.31, 0.33)	–0.01 (–0.22, 0.22)	–0.01 (–0.22, 0.21)
Normal ($\rho = 0.25$)			
BART	0.08 (–0.09, 0.25)	0.05 (–0.10, 0.21)	0.33 (0.17, 0.54)
Logit	0.02 (–0.12, 0.17)	0.01 (–0.11, 0.16)	0.28 (0.11, 0.46)
Normal ($\rho = 0.50$)			
BART	0.04 (–0.09, 0.21)	0.03 (–0.11, 0.20)	0.37 (0.18, 0.59)
Logit	0.00 (–0.13, 0.15)	0.00 (–0.14, 0.15)	0.33 (0.15, 0.53)
Normal ($\rho = 0.75$)			
BART	0.04 (–0.08, 0.16)	0.02 (–0.10, 0.15)	0.38 (0.17, 0.61)
Logit	0.01 (-0.11, 0.13)	0.00 (-0.12, 0.12)	0.35 (0.14, 0.60)
Binary Case 1			
BART	0.00 (–0.01, 0.03)	0.02 (–0.01, 0.06)	0.12 (–0.02, 0.29)
Logit	0.05 (0.00, 0.12)	0.06 (0.00, 0.12)	0.16 (0.02, 0.36)
Binary Case 2			
BART	0.01 (–0.04, 0.06)	0.06 (0.00, 0.13)	0.45 (0.25, 0.69)
Logit	0.05 (–0.02, 0.14)	0.10 (0.02, 0.19)	0.50 (0.26, 0.79)
Binary Case 3			
BART	0.00 (–0.08, 0.09)	0.13 (0.01, 0.31)	0.77 (0.51, 1.08)
Logit	0.03 (–0.07, 0.15)	0.16 (0.04, 0.32)	0.80 (0.56, 1.08)
Multinomial			
BART	0.02 (–0.08, 0.12)	0.08 (–0.01, 0.19)	0.34 (0.21, 0.49)
Logit	0.06 (–0.06, 0.19)	0.12 (0.01, 0.25)	0.38 (0.26, 0.53)
Log Normal			
BART	0.04 (–0.01, 0.10)	0.05 (–0.07, 0.17)	–0.24 (–0.42, –0.12)
Logit	0.07 (–0.01, 0.21)	0.08 (–0.07, 0.25)	–0.20 (–0.44, –0.04)

Table 2.6: Prevalence of retention states by $\Delta = 0$ day delay for the four outcome models.

Visits Modeled	Visit Total	Database Closure	On Time/Early	Missed	Dead
$\ell = 1$	10,778	236 (2.2)	6,064 (56.3)	4,303 (39.9)	175 (1.6)
$\ell = 2$	10,725	349 (3.3)	5,810 (54.2)	4,434 (41.3)	132 (1.2)
$\ell = 3$	10,210	431 (4.2)	5,374 (52.6)	4,320 (42.3)	85 (0.8)
$\ell \geq 4$	654,002	44,464 (6.8)	394,458 (60.3)	213,893 (32.7)	1,187 (0.2)

Table 2.4: Performance metrics for cross-validated predictive models fit to longitudinal visit data from AMPATH. Table entries represent posterior mean and 95% posterior predictive interval for each measure.

	$\Delta = 0$				
Visits Modeled	AUROC	Sensitivity	Specificity	PPV	Prevalence
$(\ell = 1)$					
Logit	62.0 (60.0, 64.0)	28.1 (25.6, 30.7)	84.4 (83.0, 85.9)	50.6 (46.1, 55.2)	40.7
BART	62.0 (59.0, 64.0)	28.0 (24.7, 31.2)	84.3 (82.2, 86.2)	50.5 (44.3, 56.1)	40.7
$(\ell = 2)$					
Logit	66.0 (64.0, 68.0)	32.2 (29.0, 34.8)	87.9 (85.8, 89.6)	63.8 (57.6, 68.8)	41.8
BART	66.0 (64.0, 69.0)	31.9 (28.8, 35.5)	87.8 (85.3, 89.5)	64.4 (58.3, 69.6)	41.8
$(\ell = 3)$					
Logit	67.0 (65.0, 69.0)	31.0 (28.4, 34.1)	87.8 (86.0, 90.0)	64.4 (59.1, 70.7)	43.5
BART	68.0 (66.0, 70.0)	32.0 (29.3, 34.5)	88.6 (86.6, 90.5)	67.5 (61.9, 72.7)	43.5
$(\ell \geq 4)$					
Logit	72.0 (71.0, 72.0)	35.8 (35.6, 35.9)	88.0 (87.9, 88.1)	60.4 (60.1, 60.6)	35.0
BART	73.0 (73.0, 73.0)	37.3 (37.2, 37.5)	88.8 (88.7, 88.9)	62.9 (62.6, 63.2)	35.0

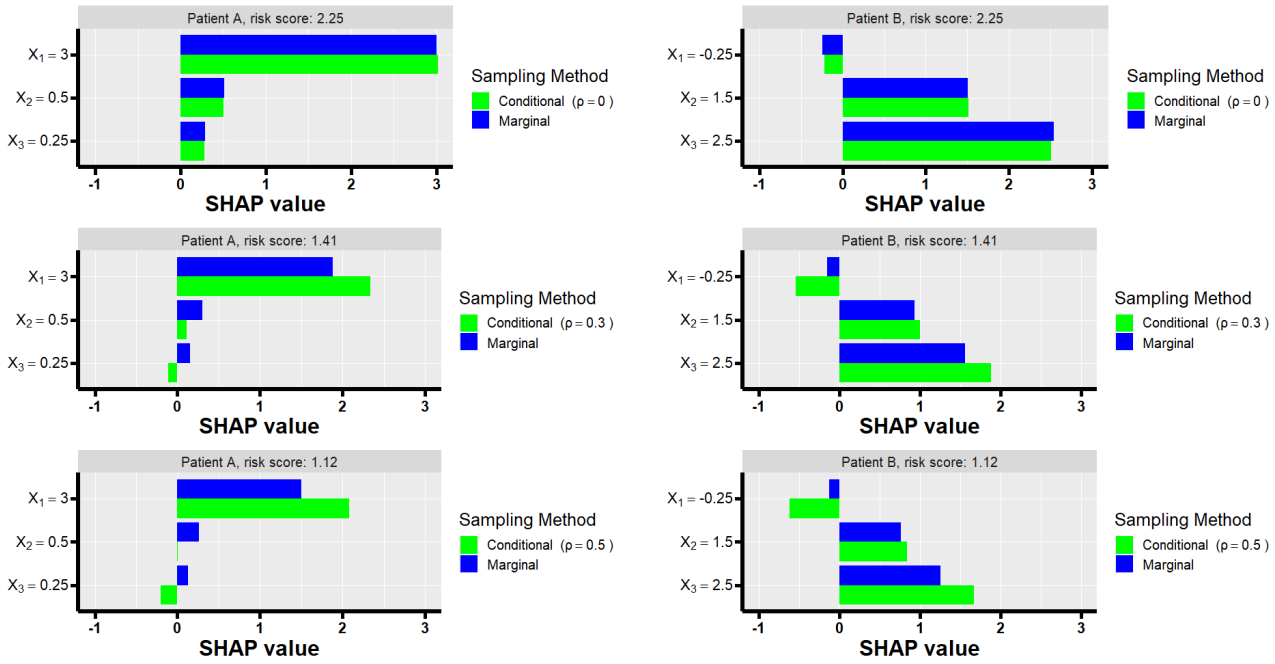

Figure 2.2: Barplots of SHAP values for two individuals from simulated data where the data generating distribution is a trivariate normal where the correlation coefficient between each covariate takes values $\rho \in \{0, 0.25, 0.50\}$. Blue indicates that draws from the six marginal distributions were used to calculate the SHAP values. Green indicates that draws from the six conditional distributions were used to calculate the SHAP values.

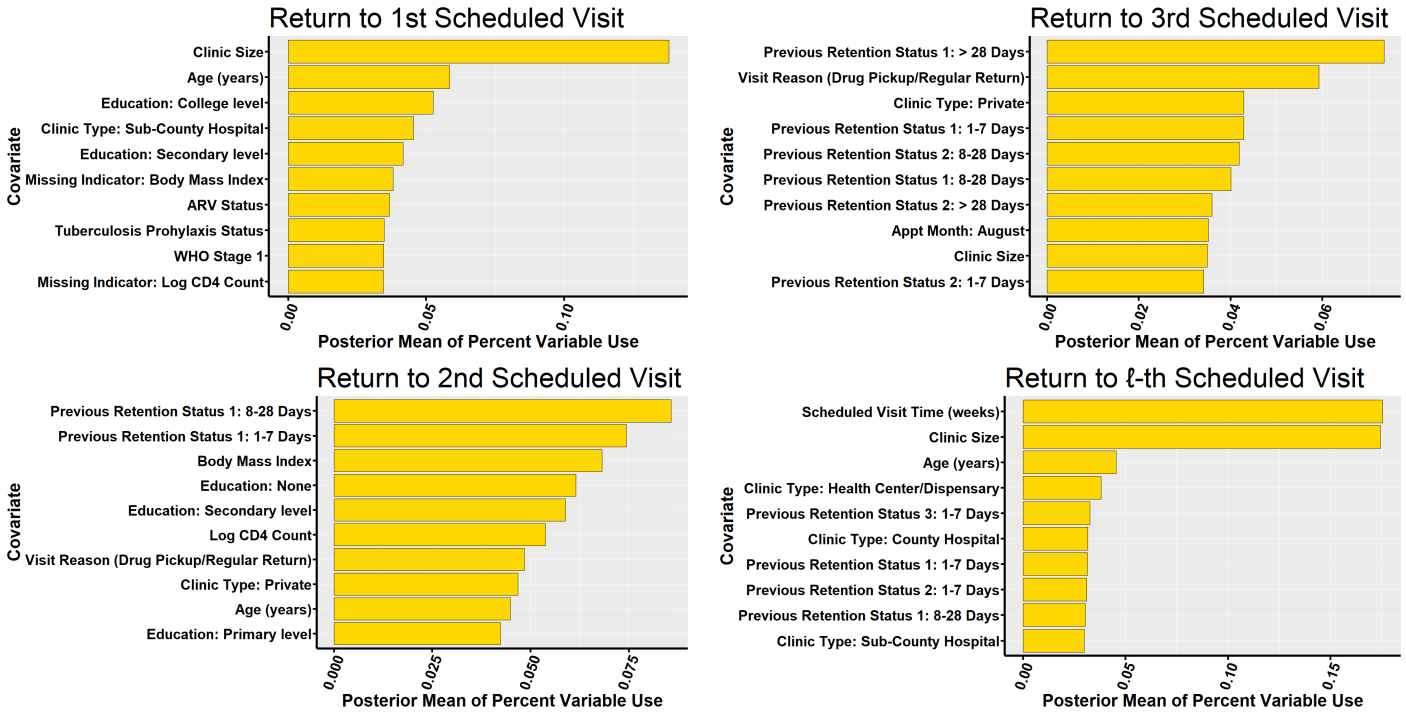

Figure 2.3: Global feature importance plots generated from Bayesian additive regression tree (BART) model fit to AMPATH data. Top 10 features from each model are shown.

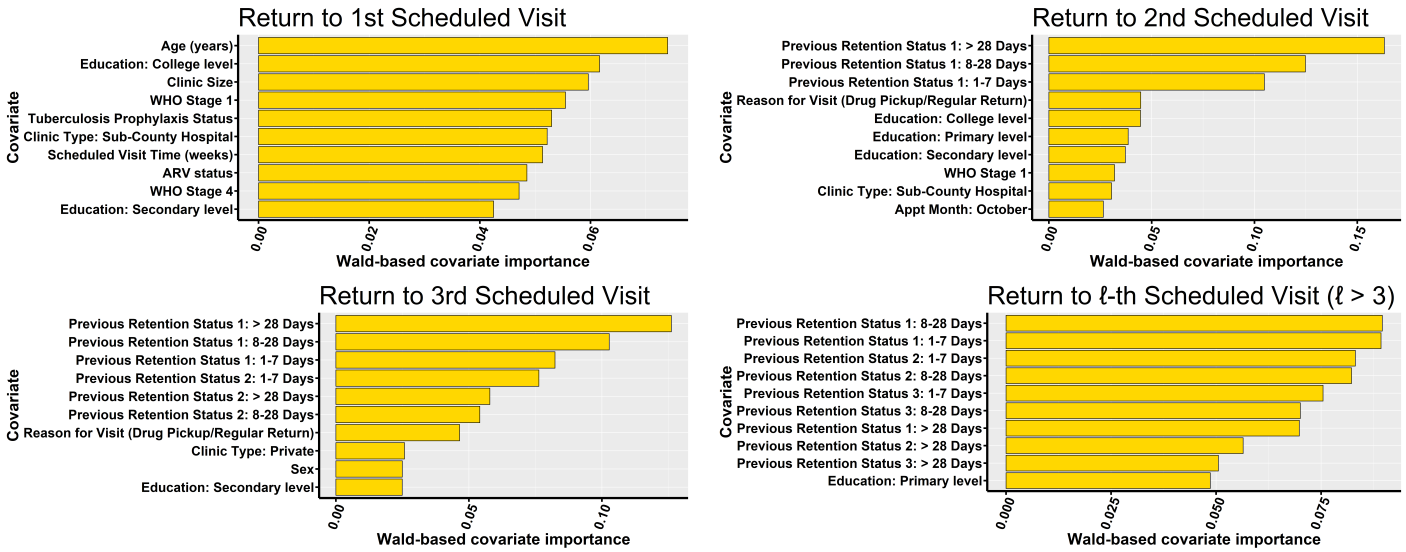

Figure 2.4: Global feature importance plots from logistic regression models fitted to AMPATH data. The top 10 features from each model are shown. Feature importance is quantified using the square root of the Wald statistic, and subsequently scaled by dividing each covariate's score by the sum of all scores.

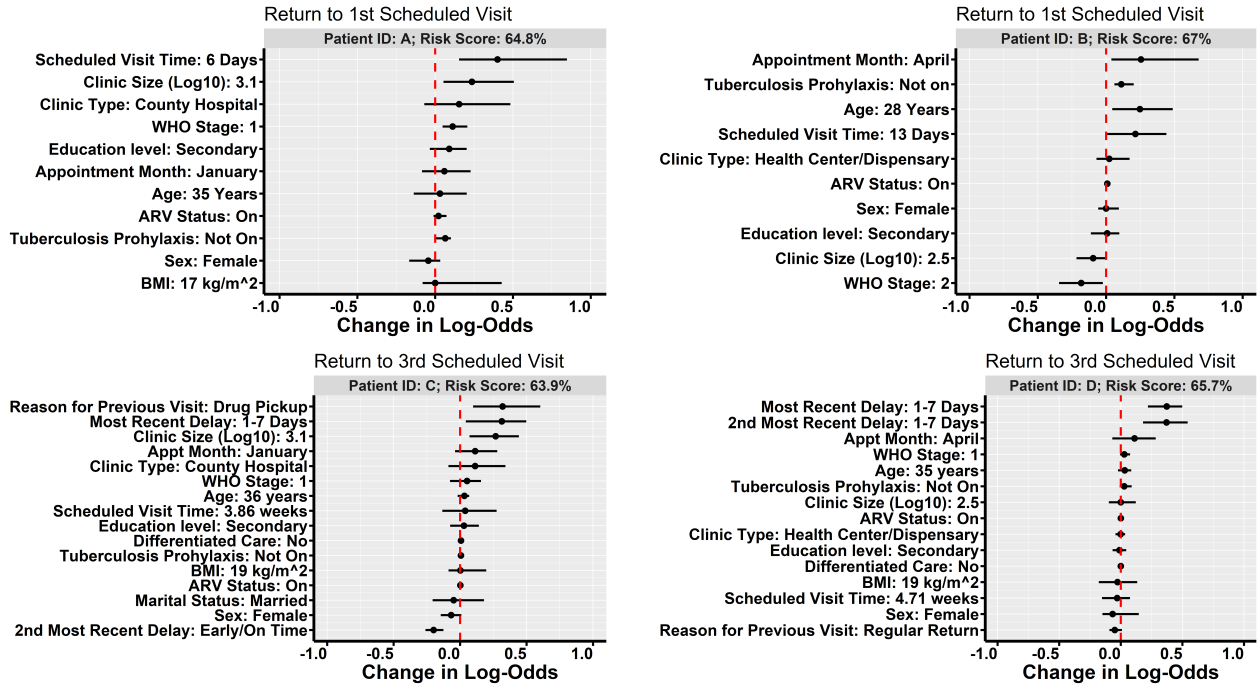

Figure 2.5: Conditional patient-level variable importance for four patients with similar predicted probability of missed visit. Plots show posterior median and 95% credible intervals for individual-level variable importance measure $\phi_{ij}(\boldsymbol{\theta})$, where i indexes patient and j indexes covariate.

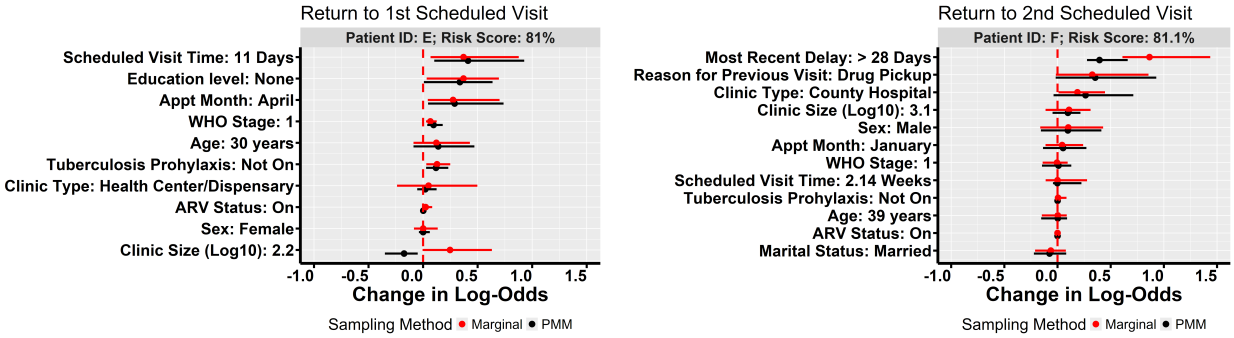

Figure 2.6: Marginal (red) and conditional (black) patient-level variable importance for two patients with similar predicted probability of missed visit. Plots show posterior median and 95% credible intervals for individual-level variable importance measure $\phi_{ij}(\boldsymbol{\theta})$, where i indexes patient and j indexes covariate.

Table 2.5: Baseline Characteristics for patients at entry into the cohort. The table reports "median (25th percentile, 75th percentile)" for continuous variables and percentage of true for binary variables or each level of categorical variables. Q1 refers to the 25th percentile and Q3 refers to the 75th percentile.

	Baseline Characteristics at entry into cohort		
Variable	Total (N = 70,626)	Post-2021 (N = 10,778)	Pre-2021 (N = 59,848)
Age (years)			
Median (Q1, Q3)	45 (35, 53)	36 (29, 45)	46 (37, 54)
Scheduled Appointment Time (Weeks)			
Median (Q1, Q3)	11.29 (4, 12)	2 (2, 3.43)	12 (4.71, 12)
Sex			
Female	44,041 (62.4)	6,434 (59.7)	33,607 (62.8)
Male	26,585 (37.6)	4,344 (40.3)	22,241 (37.2)
ARV Status			
Not On	1,496 (2.1)	1,207 (11.2)	289 (0.5)
On	69,130 (97.9)	9,571 (88.8)	59,559 (99.5)
Facility Type			
Regional/Referral Hospital	11,218 (15.9)	1253 (11.6)	9,965 (16.7)
County Hospital	17,734 (25.1)	2765 (25.7)	14,969 (25.0)
Sub-County Hospital	30,017 (42.5)	4032 (37.4)	25,985 (43.4)
Health Center/Dispensary	10,826 (15.3)	2604 (24.2)	8,222 (13.7)
Private	831 (1.2)	124 (1.2)	707 (1.2)
Body Mass Index ($\frac{kg}{m^2}$)			
Median (Q1, Q3)	22 (20, 25)	21 (19, 23)	22 (20, 26)
Missing	6,903 (9.8)	3,518 (32.6)	3,386 (5.6)
CD4 Count			
Median (Q1, Q3)	407 (252, 577)	244 (94, 437)	413 (261, 581)
Missing	24,228 (34.3)	8,711 (80.8)	15,517 (25.9)
WHO Stage*			
Missing	6,015 (8.5)	658 (6.1)	5,357 (9.0)
First Stage	35,112 (49.7)	6358 (50.2)	28,756 (48.0)
Second Stage	11,443 (16.2)	1838 (16.1)	9,605 (16.0)
Third Stage	14,975 (21.2)	1640 (20.9)	13,335 (22.3)
Fourth Stage	3,081 (4.4)	286 (4.3)	2,795 (4.7)
Education Level*			
Missing	12,411 (17.6)	233 (2.1)	12,178 (20.3)
None	3,659 (5.2)	873 (8.1)	2,786 (4.7)
Primary School	28,394 (40.2)	5025 (46.6)	23,369 (39.0)
Secondary School	19,158 (27.1)	3485 (32.3)	15,673 (26.2)
College	5,842 (9.8)	1162 (10.8)	5,842 (9.8)

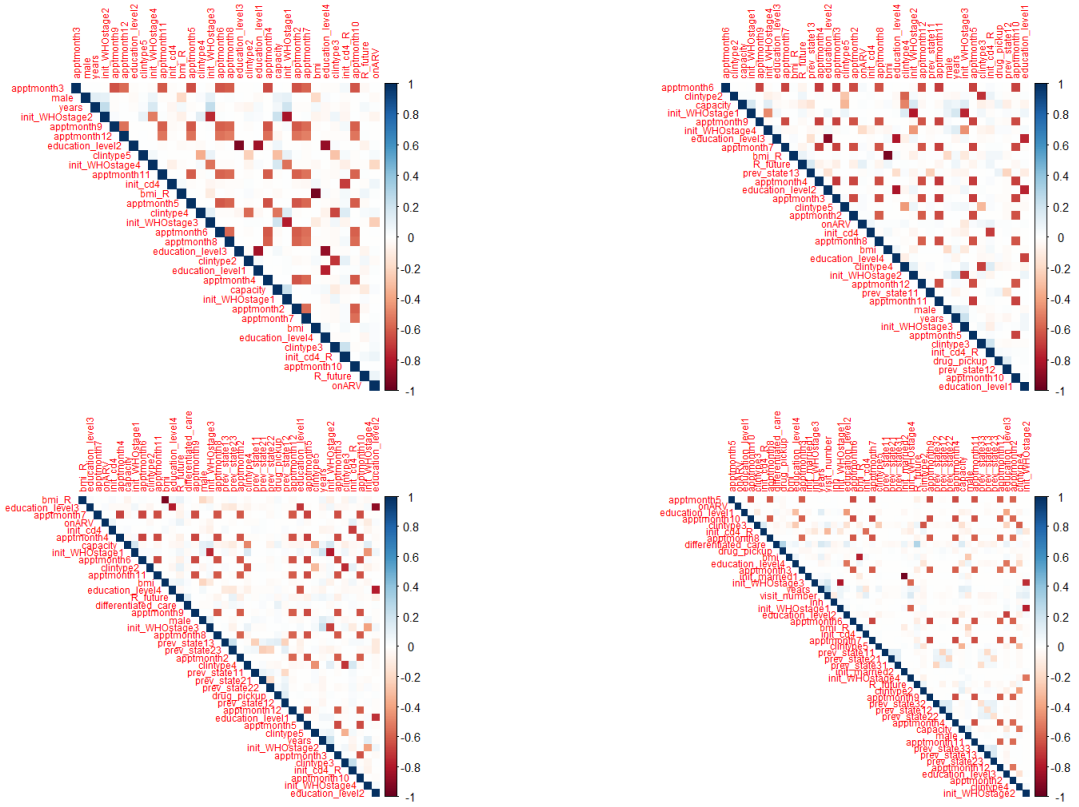

Figure 2.7: Heatmaps of the inverse correlations for the predictors in the return to visit models $\ell = 1, 2, 3$ and $\ell \geq 4$.

Chapter 3

Variable importance quantification for prediction models with missing covariates

3.1 Introduction

Advances in machine learning (ML) have opened up new avenues for prediction in health outcome settings. In many applications in the healthcare setting, researchers seek to leverage electronic health records (EHR) to build prediction models for diagnosis, prevention, and to aid clinical decision making. In addition to accurate risk prediction, knowing how the variables in the model influence the outcome is an important goal for model interpretability. Influence of individual variables is quantified by *variable importance* measures that, in a generic sense, are designed to represent the degree to which each variable contributes to predictive variation or overall variation in the outcome variable. In a clinical setting knowing which covariates contribute the most to predictive variation can inform clinicians how the model relies on certain variables to make decisions and reveal to which degree predictions are driven by actionable clinical signals and structural features of care delivery. The methods used to calculate variable importance can have formal representation in terms of loss or variance, in which case they can be assigned a model-related scale; or they can be more empirical. For example, a popular variable importance measures for random forests is the number of times a variable is used for

the first split in one of the individual trees comprising the ensemble.

The choice of variable importance measure is often determined by the modeling technique being employed. For generalized linear models such as linear regression and logistic regression, the Z statistic (Williamson et al., 2023) provides a quantitative way to measure variable importance for each covariate. As indicated above, tree based methods such as random forest, XGBoost and Bayesian additive regression trees often use some function of the number of splits for a variable. Additionally, because the variable importance measures native to different modeling techniques generally have a different interpretation, comparisons across techniques can be difficult.

Missing covariate data can complicate the definition and calculation of variable importance. To understand how, it is useful to first observe that covariate missingness can occur in the development (training) stage, the deployment (validation, prediction, forecasting) stage, or both, yielding four distinct possibilities. We follow the principle that the prediction model should be developed on a training set whose distribution is aligned with the validation set; i.e., development should be done in a way that mimics the circumstances of deployment. Or more generally, variable importance measures, which are calculated at the development stage, should be relevant to the setting where they will be deployed. For example, if age is missing for a large fraction of cases used to develop the model, variable importance measures for the prediction model should have a way to account for this.

The default case is where covariates are complete in both development and deployment. Usually we assume that the training and test sets come from the same joint distribution, and from there, standard variable importance measures can be applied and properly interpreted in context. If missingness occurs in the development stage, but data are expected to be complete at deployment, then a key first step in the development stage may be to learn the complete-data distribution corresponding to the deployment setting, and possibly impute completed datasets for training. Work from Wood et al. (2015) has found that multiply imputing (Little and Rubin, 2019) K complete datasets and subsequently fitting a model to each of the imputed datasets

works best in terms of producing good predictive performance. However the imputations are derived from other variables in the model, which may impact how to formulate and interpret variable importance.

The setting addressed in this chapter is where covariate missingness is expected at both the development and deployment stages, and where the joint distribution of covariates and missing indicators is similar across the stages. When missingness is present in both the development and deployment data the notion of variable importance can be obfuscated due to the need to handle missingness at the point of prediction. In this setting, imputation models are part of the prediction pipeline. A naive approach is to treat imputed values of missing covariates as 'not missing', and calculate variable importance based on the imputed data. Covariates that are missing at the deployment stage typically will be imputed using the values of other, observed, covariates. At both development and deployment stages, because imputed values of missing covariates are functions of the observed covariates, proper allocation of variable importance is not obvious.

In recent years a multitude of strategies have been developed to deal with missing data in prediction based settings (Jones, 1996, Kapelner and Bleich, 2015, Hoogland et al., 2020a, Fletcher Mercaldo and Blume, 2020, Sisk et al., 2023), specifically when missingness is present at the deployment stage. While modeling strategies have been developed to account for this, the question of how to calculate and interpret variable importance for partially observed covariates has received relatively little attention (Hu et al., 2021). Several papers calculate variable importance when covariates are missing (Ramachandran et al., 2020, Wallace et al., 2023, Ogbechie et al., 2023) but the calculation is not the main focus of the work and it is unclear if the data the models will be deployed to will also contain missingness values. Other work (Vo et al., 2024) modifies local variable importance methods such as Shapley additive explanations (Lundberg and Lee, 2017) when missing covariates are imputed using a constant. Based on the existing literature, there is relatively little standardization of methodology.

The purpose of this chapter is to provide a principled framework for defining and calculat-

ing variable importance in the presence of missing covariates, and to illustrate through simulation and real-data application how the framework applies for generalized linear models. This chapter is organized as follows. Section 3.2 provides an overview of notation and background on the presence of missing data at the development stage and at the deployment stage. Section 3.3 provides an overview of the variable importance measure that is the focus of this article. Section 3.4 identifies and discusses a variety of techniques to handle missing data in machine learning settings where forecasting is the primary goal of interest. Section 3.5 adjusts the variable importance metric from section 3 to account for missing data. Section 3.6 describes simulations to compare the variable importance and predictive performance of the techniques described from Section 3.4 while Section 3.7 provides a detailed illustration of the method on real-world data from the Academic Model for Providing Access to Healthcare (AMPATH) for predicting retention in HIV care. Section 3.8 concludes with a discussion. Motivated by our case study, we focus on a binary outcome. However, all methods considered can be straightforwardly applied to a continuous outcome.

3.2 Data structure, notation, and prediction objectives

In clinical settings the purpose of a prediction model is to take patient and clinical information (covariates/predictors) as inputs and, output an estimated risk that a patient currently has or will develop a condition of interest. These prediction models can then be implemented at the point of care to guide clinical decision making as well as facilitate interventions to reduce future patient risks. Most methodology to estimate a prediction model assumes that there will be complete data for model development and complete data at model deployment, but in health-related settings with electronic health records (EHR) missing data is ubiquitous, meaning the prediction model must be designed in such a way as to account for the missing data mechanism (MDM).

Let $Y \in \{0,1\}$ denote the outcome of interest and let $\boldsymbol{X} = (X_1, \ldots, X_p)$ denote the covariates indexed by $j = 1, 2, \ldots, p$, where X_j denotes the random variable and x_{ij} is the observed value

of X_j for an individual indexed by $i = 1, \ldots, N$. We write $\boldsymbol{X}_{(-j)} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$ to denote the vector containing all covariates except X_j. When there is missingness in the covariate space we define an accompanying set of indicators $\boldsymbol{M} = (M_1, \ldots, M_p)$, where $M_j = 0$ if X_j is observed and $M_j = 1$ if X_j is missing. For this paper, we assume all responses Y are observed and that missingness only occurs in the covariates.

The vector $\boldsymbol{M}$ allows up to 2^p possible unique patterns of missingness in the data; in practice only a few are observed. Using $\{0,1\}^p$ to denote the full collection of unique configurations of $\boldsymbol{M}$, define $\mathscr{S}_p \subset \{0,1\}^p$ to be the set of the *observed* missingness patterns, with each unique pattern indexed by $s \in \mathscr{S}_p$. Define $\boldsymbol{X}_{\text{obs}(s)}$ as the set of covariates that is observed in pattern s and $\boldsymbol{X}_{\text{mis}(s)}$ those that are missing, so that for any missingness pattern s, we have $\boldsymbol{X} = (\boldsymbol{X}_{\text{obs}(s)}, \boldsymbol{X}_{\text{mis}(s)})$.

When there are no missing covariates, the general goal is to derive a rule $D : \mathbb{R}^p \to \{0,1\}$ that maps an observed set of covariates $\boldsymbol{x} \in \mathbb{R}^p$ to a binary classification. This is typically accomplished in two steps. First, a model or algorithm is used to learn a function $f(\boldsymbol{x}) = \Pr(Y = 1 \,|\, \boldsymbol{X} = \boldsymbol{x})$; the function can be parameterized in terms of a finite-dimensional parameter $\boldsymbol{\theta}$ such that $f(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta})$. For example, logistic regression with additive covariate effects corresponds to $f(\boldsymbol{x}; \boldsymbol{\theta}) = 1 - \{1 + \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})\}^{-1}$. Next, the rule D is typically selected to minimize a pre-specified loss function for classification; for example, $D_c(\boldsymbol{x}) = \mathbb{I}\{f(\boldsymbol{x}) > c\}$, for a value of c that minimizes misclassification loss, is a common choice.

When covariates are missing at both the development and deployment stages, the model for $\Pr(Y = 1 \,|\, \boldsymbol{X})$ must also account for missing data. Our approach is to use a pattern-mixture formulation that makes the missingness indicators an explicit component of the model input; i.e., to model $\Pr(Y = 1 \,|\, \boldsymbol{X}, \boldsymbol{M})$. The most general version of this model relies only on the observed covariates for each missing data pattern, leading to pattern-specific models

$$f_s(\boldsymbol{x}_{\text{obs}(s)}, s; \boldsymbol{\theta}_s) \;\; = \;\; \Pr(Y = 1 \,|\, \boldsymbol{X} = \boldsymbol{x}_{\text{obs}(s)}, S = s), \quad s \in \mathscr{S}_p.$$

This formulation allows prediction to depend *only* on covariate information that is observed for

a particular individual; it also allows prediction to depend on the missing data pattern. Hence the effect of a specific covariate X_j could depend on missing data pattern s. In subsequent sections, we show how the prediction model can be implemented using imputation, and how that approach relates to the pattern-mixture approach described here.

3.3 Measures of Variable Importance

Variable importance can be divided into *global* and *unit-level* variable importance, or explanations. Global explanations rank covariates based on their contributions to overall predictive variance while unit-level explanations identify the degree to which individual covariates influence an individual prediction from the model. Our focus is on global variable importance.

Among global variable importance measures, we restrict attention to those that are model-agnostic. A variable importance measure is said to be *model-agnostic* if it can be calculated in a way that does not depend on the technique being used to estimate $f(\boldsymbol{x})$. For example, a measure based on likelihood functions is not model-agnostic because it is not well defined for tree-based methods; similarly, measures based on number of splits for tree-based methods are not model-agnostic because they are generally not applicable to regression-based prediction functions. Common model-agnostic methods for global explanation include permutation based covariate importance (Fisher et al., 2019, Breiman, 2001), partial dependence plots (Friedman, 2001), and dropped covariate importance (Lei et al., 2018, Williamson et al., 2021, 2023), the last of which we describe in Section 3.3.1 and use as our variable importance measure of choice.

Finally, although it might seem intuitively obvious, variable importance measures are interpretable only within the family of models used to estimate the prediction function $f(\boldsymbol{x})$, even if they are model agnostic. For example, the importance of variable X_j is only interpretable if the model $f(\boldsymbol{x})$ and the model $f_{(-j)}(\boldsymbol{x}_{(-j)})$ come from the same family of models, such as logistic regression or random forest. It is not meaningful to quantify variable importance when $f(\cdot)$ is a logistic regression and $f_{(-j)}(\cdot)$ is a random forest.

3.3.1 Dropped Variable Importance

The measure of variable importance used in this article represents a normalized difference in predictive loss between a prediction function $f(\boldsymbol{x})$ and $f_{(-j)}(\boldsymbol{x}_{(-j)})$, where $f_{(-j)} : \mathbb{R}^{p-1} \rightarrow \mathbb{R}$ is a reduced form of $f(\cdot)$, and includes only the covariates $\boldsymbol{X}_{(-j)}$ as predictors. Following (Williamson et al., 2021, 2023), we define variable importance using

$$\psi(j) \quad = \quad \frac{E_{Y,\boldsymbol{X}}\left\{L(Y, f(\boldsymbol{X}))\right\} - E_{Y,\boldsymbol{X}}\left\{L(Y, f_{-j}(\boldsymbol{X}_{-j}))\right\}}{E_{Y,\boldsymbol{X}}\left\{L(Y, \mu_0)\right\}},$$

where $L(y, f)$ is a well-defined loss function and $\mu_0 = E(Y)$ is the null model with no covariate information. As indicated above, $f(\cdot)$ and $f_{(-j)}(\cdot)$ are assumed to belong to a common family $\mathscr{F}$ of models or algorithms. The expectation is taken over the distribution $P(Y,\boldsymbol{X})$; in practice the expectation is taken over the empirical data distribution $\hat{P}(Y,\boldsymbol{X})$. This is sometimes referred to as the leave-one-covariate-out (LOCO) approach (Lei et al., 2018, Williamson et al., 2021, 2023) and can be summarized as the change in risk prediction when either a covariate X_j, or more generally a group of covariates $\boldsymbol{X}_a$, are removed from the model. We detail some examples below using well-known generalized linear models.

Example: Linear Models

Consider the case where the loss function is squared error loss $L(y, f) = (y - f)^2$, the prediction function is specified as $f(\boldsymbol{X}; \boldsymbol{\theta}) = E(Y \,|\boldsymbol{X}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta}$, and the prediction function excluding X_j is $f(\boldsymbol{x}_{(-j)}; \boldsymbol{\theta}') = E(Y \,|\boldsymbol{X}_{(-j)}) = \boldsymbol{x}_{(-j)}^{\mathrm{T}}\boldsymbol{\theta}'$. Specifically the form of the reduced model is the optimal predictor for squared error loss when only information $\boldsymbol{X}_{(-j)}$ is available. In this case, the parameter $\boldsymbol{\theta}'$ indexing the reduced model is a $1 \times (p-1)$ vector, but importantly the coefficients of $\boldsymbol{\theta}$ from the first model will generally take different values than the corresponding elements

of $\boldsymbol{\theta}'$ in the reduced model. For this case, variable importance for X_j is

$$\begin{aligned}\psi(j) &= \frac{E_{Y,\boldsymbol{X}}\left\{(Y-f(\boldsymbol{X}))^2\right\}-E_{Y,\boldsymbol{X}}\left\{(Y-f_{-j}(\boldsymbol{X}_{-j}))^2\right\}}{E_{Y,\boldsymbol{X}}\left\{(Y-E(Y))^2\right\}} \\ &= \frac{E_{Y,\boldsymbol{X}}\left\{f(\boldsymbol{X})-f_{-j}(\boldsymbol{X}_{-j}))^2\right\}}{\text{var}(Y)} \\ &= \frac{\theta_j^2 E\left[\left\{X_j-E(X_j\,|\boldsymbol{X}_{-j})\right\}^2\right]}{\text{var}(Y)}.\end{aligned}$$

The full proof is detailed in the appendix. The coefficient θ_j measures the strength of the relationship between X_j and the outcome Y while the term $E[(X_j - E[X_j\,|\boldsymbol{X}_{-j}])^2]$ captures variation in X_j that is not explained by other variables in the model. Hence variable importance is a combination of association between Y and X_j (i.e. through θ_j) and variation in X_j after conditioning on $\boldsymbol{X}_{-j}$ (i.e. through $E[(X_j-E[X_j\,|\boldsymbol{X}_{-j}])^2]$).

Example: Binary Outcome

Now suppose Y is binary and that the prediction function follows a logistic regression $f(\boldsymbol{x};\boldsymbol{\theta}) = E(Y\,|\boldsymbol{X}) = \frac{\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})}$. Using the log-loss function $L(y,f) = y\log f + (1-y)\log(1-f)$ yields

$$\begin{aligned}\psi(j) &= \frac{E_{\boldsymbol{X}}\left(E_{Y\,|\boldsymbol{X}}\left[Y\log\left\{\frac{f(\boldsymbol{X})}{f_{(-j)}(\boldsymbol{X}_{(-j)})}\right\}+(1-Y)\log\left\{\frac{1-f(\boldsymbol{X})}{1-f_{(-j)}(\boldsymbol{X}_{(-j)})}\right\}\right]\right)}{E_{Y,\boldsymbol{X}}\left\{Ylog(E[Y])+(1-Y)log(1-E[Y])\right\}} \\ &= \frac{E_{\boldsymbol{X}}\left[f(\boldsymbol{X})\log\left\{\frac{f(\boldsymbol{X})}{f_{(-j)}(\boldsymbol{X}_{(-j)})}\right\}+(1-f(\boldsymbol{X}))\log\left\{\frac{1-f(\boldsymbol{X})}{1-f_{(-j)}(\boldsymbol{X}_{(-j)})}\right\}\right]}{\pi_0\log(\pi_0)+(1-\pi_0)\log(1-\pi_0)},\end{aligned}$$

where $\pi_0 = E(Y)$. The numerator is the Kullback-Leibler divergence between $f(\boldsymbol{X})$ and $f_{(-j)}(\boldsymbol{X}_{(-j)})$ and thus represents the information loss between the distributions when X_j is not present.

3.3.2 Definition of the reduced model

The dropped variable importance described requires the estimation of an outcome model $f(\boldsymbol{x})$ fit to the full set of covariates $\boldsymbol{X}$, and a reduced model $f_{(-j)}(\boldsymbol{x}_{(-j)})$ where X_j is removed.

The definition of a reduced model requires careful definition.

Consider the linear model $f(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2$ where the distribution of $\boldsymbol{X}$ is a standard bivariate normal with correlation $\rho > 0$. Simply discarding the covariate X_2 and re-fitting the model will provide a new function $f^{\text{nested}}_{(-2)}(\boldsymbol{x}_{(-2)}) = \theta_0' + \theta_1' x_1$. This is as a nested model because the functional form of the reduced model is contained within the space of the original full model. However, a different type of reduced model can be obtained by integrating X_2 out of the original model; i.e.,

$$\begin{aligned} f^{\text{integrated}}_{(-2)}(\boldsymbol{X}_{(-2)}) &= \int (\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_1 X_2)\, dF(x_2 \,|\, x_1) \\ &= \theta_0 + (\theta_1 + \rho\,\theta_2) x_1 + \rho\,\theta_3 x_1^2. \end{aligned}$$

This example serves to show that there are two types of reduced models, the integrated reduced model which is defined as $f_{(-j)}(\boldsymbol{X}_{(-j)}) = E[f(\boldsymbol{X}) \,|\, \boldsymbol{X}_{(-j)}]$ and the nested version which amounts to simply dropping X_j. The key component of variable importance is in the fact that the measure quantifies the performance if we never had access to X_j to begin with; not necessarily how the functional form of the model predicts if X_j were integrated out.

The dropped variable approach compares *nested* rather than *integrated* models. The nested approach to a reduced model does not require any additional assumptions about the joint covariate distribution $P(\boldsymbol{X})$; hence with generalized linear models it can be more straightforward to apply, and because our focus is on generalized linear models, we will quantify variable importance using nested models. It is important to note, however, that with machine learning methods such as random forest and BART, the concept of nested models is not well defined. Because the prediction function for these methods can capture interactions and nonlinearities, dropped variable importance is more likely to be using the integrated version of the reduced model, even without having to know the joint covariate distribution $P(\boldsymbol{X})$. It's possible that integrated reduced models for GLM can be approximated with nested models, for example, using regression splines or generalized additive models. We view this as an important topic for

further investigation.

3.4 Methods for handling missing data for prediction

In this section we provide a brief overview of methodology designed to handle missing data in prediction-based settings. Methods can roughly be divided into pattern submodel approaches (Sisk et al., 2023, Fletcher Mercaldo and Blume, 2020) and imputation-based approaches (Wood et al., 2015, Hoogland et al., 2020a, Sisk et al., 2023). In general the two are not necessarily mutually exclusive; one may implement imputation techniques into a pattern framework. As indicated earlier, our focus is on the case where missingness occurs at both development and deployment, sometimes referred to as 'pragmatic model performance', where missingness is to be expected when forecasting of future individuals (Wood et al., 2015).

In the methods descriptions below, recall that we are using a GLM framework to illustrate our approach. We assume that for a covariate vector $\boldsymbol{X}$, the prediction function is designed to estimate the mean $E(Y|\boldsymbol{X})$; in a GLM, $E(Y|\boldsymbol{X}=\boldsymbol{x}) = g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})$, where $g : \mathbb{R} \to \mathbb{R}$ is a monotone link function (e.g., identity link, logit link). Hence the prediction function is given by $f(\boldsymbol{x};\boldsymbol{\theta}) = g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\theta})$.

3.4.1 Pattern Submodels

The pattern submodel (PS) (Fletcher Mercaldo and Blume, 2020) handles missingness by partitioning the data according to the observable missingness patterns $s \in \mathscr{S}_p$ and then fitting a model within each pattern using only the covariates $\boldsymbol{X}_{\mathrm{obs}(s)}$ that are observed within that pattern. Specifically we fit a sequence of fully stratified models of the form

$$\begin{aligned} f_s(\boldsymbol{X}_{\mathrm{obs}(s)};\boldsymbol{\theta}_s) &= E(Y|\boldsymbol{X}_{\mathrm{obs}(s)}, S=s) \\ &= E(Y|\boldsymbol{X}_{\mathrm{obs}(s)}, \boldsymbol{M}=_s) \\ &= g(\boldsymbol{X}^{\mathrm{T}}_{\mathrm{obs}(s)}\boldsymbol{\theta}_s), \end{aligned}$$

for $s \in \mathscr{S}_p$, and where $_s$ denotes the configuration of $\boldsymbol{M}$ that corresponds to missing data pattern s.

To illustrate, consider a generalized linear model with outcome Y and three covariates, $\boldsymbol{X} = (X_1, X_2, X_3)$, where (X_1, X_2) may contain missing values. This implies four missing data patterns, with $(M_1, M_2) \in \{(0,0),(1,0),(0,1),(1,1)\}$ or, correspondingly, $\mathscr{S} = \{1,2,3,4\}$. A fully stratified pattern submodel approach yields four submodels of the form,

$$\begin{aligned} f_1(\boldsymbol{X}_{\text{obs}(1)}; \boldsymbol{\theta}_1) &= g(\theta_{01} + \theta_{11}X_1 + \theta_{21}X_2 + \theta_{31}X_3) \\ f_2(\boldsymbol{X}_{\text{obs}(2)}; \boldsymbol{\theta}_2) &= g(\theta_{02} + \theta_{22}X_2 + \theta_{32}X_3) \\ f_3(\boldsymbol{X}_{\text{obs}(3)}; \boldsymbol{\theta}_3) &= g(\theta_{03} + \theta_{13}X_1 + \theta_{33}X_3) \\ f_4(\boldsymbol{X}_{\text{obs}(4)}; \boldsymbol{\theta}_4) &= g(\theta_{04} + \theta_{34}X_3) \end{aligned} \tag{3.1}$$

Hence, for example, prediction for individuals with complete covariate information uses all 3 covariates, while predictions for individuals missing (X_1, X_2) uses only X_3. The model can also be represented by fully interacting the missing data indicators with the linear predictors. For example, if $g(\cdot)$ is the identity function, we could write

$$\begin{aligned} f(\boldsymbol{X}_{\text{obs}}, \boldsymbol{M}; \boldsymbol{\theta}) \;=\; & (1-M_1)(1-M_2)(\theta_{01} + \theta_{11}X_1 + \theta_{21}X_2 + \theta_{31}X_3) \\ & + M_1(1-M_2)(\theta_{03} + \theta_{13}X_1 + \theta_{33}X_3) \\ & + (1-M_1)M_2(\theta_{03} + \theta_{13}X_1 + \theta_{33}X_3) \\ & + M_1M_2(\theta_{04} + \theta_{34}X_3) \end{aligned} \tag{3.2}$$

Assuming the submodels are correctly specified, this example illustrates that the pattern submodel is the most general representation of the prediction function in that it (i) makes use of only the observed covariates for prediction and (ii) does not require any assumptions about the missing data mechanism. The representation (3.2) also helps draw connections to the 'missing indicator' method (Jones, 1996), where elements of $\boldsymbol{M}$ are treated as covariates in

the prediction function – as is implicitly the case in (3.2). In practice, there will be far more than 4 distinct missing data patterns, and simplifications are needed. As we show below, certain formulations of the missing data indicator method can be viewed as versions of pattern submodels where some of the patterns are collapsed or combined.

This approach extends to tree-based models (e.g. Random Forest (Breiman, 2001), XGBoost (Friedman, 2001), BART (Chipman et al., 2010)) by enforcing consistent hyperparameters and priors across pattern-specific submodels, and can be visualized in a similar manner as a large sum of trees where the covariate space is now defined by the interaction between $\boldsymbol{X}_{\text{obs}(s)}$ and the corresponding pattern s.

An important distinction between our formulation in (3.1) and (3.2) and that given by Fletcher Mercaldo and Blume (2020) is that covariates are not imputed; each submodel is fit conditional on both the observed covariates and the missingness pattern. This is the main underlying assumption behind the pattern submodel approach. Thus, estimates take the form $E[Y \mid \boldsymbol{X}_{\text{obs}(s)}, \boldsymbol{M}]$ rather than $E[Y \mid \boldsymbol{X}_{\text{obs}(s)}]$. Related methods seek to fit models of the form $E[Y \mid \boldsymbol{X}_{\text{obs}(s)}]$ (Hoogland et al., 2020a, Saar-Tsechansky and Provost, 2007, Wood et al., 2015),ignoring the missingness pattern and multiply imputing the missing data in order to fit models to the entire data.

The main advantages of the pattern submodel in its most general form is avoiding the need for imputation at either the development or deployment stages, and not having to make assumptions about the missing data mechanism. Predictions solely depend on the observed covariates $\boldsymbol{X}_{\text{obs}(s)}$. However, in most applied settings, the full stratification approach can be impractical. The number of required submodels can grow exponentially with the number of missing covariates, which can become computationally expensive even for a small number of covariates. The subdivision of the data by pattern can induce sparsity and thus lead to lack of identifiability or overfitting for rare patterns. This motivates collapsing over patterns, which can be represented using missing indicators. For example, consider a reduced form of (3.1)

given by

$$f(\boldsymbol{X}_{\text{obs}}, \boldsymbol{M}; \theta) = \quad g(\alpha_0 + \alpha_1 M_1 + \alpha_2 M_2 + \beta_3 X_1(1-M_1) + \beta_2 X_2(1-M_2) + \beta_3 X_3), \tag{3.3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. This is a collapsed version of (3.1), with

$$\begin{aligned} f_1(\boldsymbol{X}_{\text{obs}(1)}; \boldsymbol{\theta}_1) &= g(\alpha_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3) \\ f_2(\boldsymbol{X}_{\text{obs}(2)}; \boldsymbol{\theta}_2) &= g(\alpha_0 + \alpha_1 + \beta_2 X_2 + \beta_3 X_3) \\ f_3(\boldsymbol{X}_{\text{obs}(3)}; \boldsymbol{\theta}_3) &= g(\alpha_0 + \alpha_2 + \beta_1 X_1 + \beta_3 X_3) \\ f_4(\boldsymbol{X}_{\text{obs}(4)}; \boldsymbol{\theta}_4) &= g(\alpha_0 + \alpha_1 + \alpha_2 + \beta_3 X_3) \end{aligned} \tag{3.4}$$

As seen here, the inclusion of missingness indicators as covariates can be used to represent a smoother, pooled alternative to the full stratification across the missingness pattern. In (3.3), the additive effect of missingness pattern is captured by the pattern-specific combinations of α parameters, and the covariate effects, represented by β parameters, are collapsed and become shared across patterns. This smoother alternative may be justified if some missingness patterns are sparse. For the remainder of this paper the method used to generate equation (3.1) will be referred to as the stratified pattern submodel approach and equation (3.3) will be referred to as the smooth, or collapsed, pattern submodel approach.

3.4.2 Imputation-based methods

The goal of imputation-based methods is to fill in the missing values in the $\boldsymbol{X}$ matrix, generating a set of 'complete' datasets $\{\boldsymbol{X}^{(k)} : k = 1, \ldots, K\}$. This can be accomplished in a number of ways, which we describe below. In the prediction setting, both model formulation and variable importance need to be defined with respect to the particular context for model development and deployment. Specifically, it is important to define whether the objective is to generate a prediction model $f(\boldsymbol{x})$ that assumes availability of complete data on $\boldsymbol{X}$, or a model $f(\boldsymbol{x}, \mathbf{m})$ where the missing data patterns or indicators are informing the predictions; we discuss this in

more detail below.

There are three imputation techniques that are common in predictive modeling: constant-value imputation, regression imputation (RI), and multiple imputation (MI). Constant-value imputation fills in missing values of the covariates with a constant such as the sample mean. There are several well-documented drawbacks to this approach, and we will not consider it further.

Regression imputation, also known as conditional mean imputation (CMI), refers to the single-value imputation based on a set of regression models that use predicted values from of the conditional mean $E(\boldsymbol{X}_{\text{mis}}|\boldsymbol{X}_{\text{obs}})$. Returning to the the three covariate case from Section 3.4.1, regression imputation would be implemented by fitting regression models for $E(X_1|X_2,X_3)$, $E(X_2|X_1,X_3)$, $E(X_1|X_3)$ and $E(X_2|X_3)$. This covers the different possibilities for imputing a missing value of X_1 or X_2 from available data for an individual subject; e.g., for an individual i, if X_1 is missing and (X_2,X_3) are observed, the imputation is $\widehat{X}_{1i} = \widehat{E}(X_1|X_2 = x_{2i}, X_3 = x_{3i})$. In general, imputation models for RI do not need to be regression models; they can be fit using a any prediction algorithm or model. Throughout this work, however, RI will be implemented using regression models.

By contrast with single-value imputation, *multiple imputation* (Little and Rubin, 2019) replaces missing covariate data with values drawn from one or more imputation models based on the observed relationships between the covariates and outcome, typically under a missing at random (MAR) assumption. For prediction models, the imputation model may or may not condition on the dependent variable Y. While this is sensible for settings where the training data are incomplete but data at deployment are expected to be complete, it is less appropriate when covariates are missing at the deployment stage. In that case, the first step needed for generating an individual-level prediction is to impute the missing covariates; of course, this cannot be accomplished without the dependent variable Y. We therefore focus on multiple imputation using only the observed covariate data.

Two widely-used approaches to MI are joint modeling, which requires the user to specify

a model for the joint distribution $P(\boldsymbol{X} \mid \boldsymbol{\phi}) = P(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}} \mid \boldsymbol{\phi})$, full conditional specifications (FCS), which require the user to specify a series of conditional distributions for imputatin. The model-based approach usually assumes covariates are missing at random (MAR). In that case, imputations amount to taking draws from the posterior predictive distribution $P(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{X}_{\text{obs}})$. The FCS approach (Molenberghs et al., 2014) specifies a set of univariate conditional distributions $P(X_j \mid \boldsymbol{X}_{-j}, \boldsymbol{\eta})$ for covariates X_j that are subject to missingness. Imputations via FCS are carried out in an iterative manner that requires a set of starting values for components of $\boldsymbol{X}_{-j}$ that are missing. Compared to the JM approach, the FCS approach provides more flexibility for imputing different types of variables, but it may suffer from theoretical limitations because a proper joint distribution based on the different conditional models may not exist (Van Buuren et al., 2006). The FCS approach is implemented in the MICE (multiple imputation through chained equations) package in R (Van Buuren and Groothuis-Oudshoorn, 2011) and will be the imputation technique used for MI for the remainder of this work.

With either method, each set of imputations results in a completed dataset. MI therefore generates K plausible values for each missing value resulting in a set of K complete datasets $\boldsymbol{X}^{(k)}$, $k = 1, \ldots, K$. Each completed dataset is then used to estimate an imputation-specific prediction function $f(\boldsymbol{X}^{(k)})$ or $f(\boldsymbol{X}^{(k)}, \boldsymbol{M})$. The latter version retains the missing data indicators to distinguish cases where predictions are based on imputed values, and is sometimes referred to as 'imputation aware'. The K prediction models are then averaged using Rubin's rules to obtain an estimate of the risk score $f(\boldsymbol{X}_{\text{obs}})$ or $f(\boldsymbol{X}_{\text{obs}}, \boldsymbol{M})$ that is conditional only on observed data. (Wood et al., 2015).

Imputation methods have advantages and disadvantages for prediction. Regression imputation is a deterministic process and thus does not take into account uncertainty with the imputed values. Outcome models that implement RI also inherit similar limitations to the stratified pattern submodel approach since the end-user is required to specify as many imputation models as there are missingness patterns in the data. Model-based MI requires sampling from the posterior predictive distribution $P(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{X}_{\text{obs}}, \boldsymbol{\phi})$, which can be computationally ex-

pensive. Suggested modifications either fix $\boldsymbol{\phi}$ at some value $\boldsymbol{\phi}^*$, usually the last set of parameters at the end of the MI procedure, and sample only from the conditional predictive distribution $P(\boldsymbol{X}_{\text{mis}}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{\phi}^*)$ (Hoogland et al., 2020b) or re-fit the entire imputation procedure from scratch (Sisk et al., 2023, Janssen et al., 2009).

3.4.3 Relationship between pattern submodels and imputation

There is a relationship between 'imputation-aware' models, which use both imputation and missing data indicator, and the pattern submodel (Fletcher Mercaldo and Blume, 2020). For simplicity we will consider a two-variable case $\boldsymbol{X} = (X_1, X_2)$, where missing entries only occur in X_2, to illustrate. In this case, the pattern submodel specification is

$$\begin{aligned} f_1(\boldsymbol{X}_{\text{obs}(1)}; \boldsymbol{\theta}_1) &= g(\theta_{01} + \theta_{11}X_1 + \theta_{21}X_2) \\ f_2(\boldsymbol{X}_{\text{obs}(2)}; \boldsymbol{\theta}_2) &= g(\theta_{02} + \theta_{12}X_1) \end{aligned}$$

This full stratification can be written as a single model by interacting the terms in each submodel with the missingness indicator M_2. The resulting model is then

$$\begin{aligned} f(\boldsymbol{X}_{\text{obs}}, M_2; \boldsymbol{\theta}, \boldsymbol{\lambda}) &= g(\theta_{01} + \theta_{11}X_1 + \theta_{21}X_2 + (\theta_{02} - \theta_{01})M_2 + (\theta_{12} - \theta_{11})M_2X_1 - \theta_{21}M_2X_2) \\ &= g(\theta_{01} + \theta_{11}X_1 + \theta_{21}(1 - M_2)X_2 + \lambda_0 M_2 + \lambda_1 M_2 X_1) \end{aligned}$$

where $\lambda_0 = \theta_{02} - \theta_{01}$ and $\lambda_1 = \theta_{12} - \theta_{11}$.

Now consider a prediction function specified as

$$f^*(X_1, X_2; \boldsymbol{\beta}) = g(\beta_0 + \beta_1 X_1 + \beta_2 X_2),$$

where the missing entries of X_2 are generated using the imputation model $\hat{X}_2 = \alpha_0 + \alpha_1 X_1$.

This implies that the value of X_2 in the completed dataset is

$$X_2^* = (1-M_2)X_2 + M_2(\alpha_0 + \alpha_1 X_1),$$

which implies

$$\begin{aligned} f^*(X_1, X_2, M_2; \boldsymbol{\beta}, \boldsymbol{\gamma}) &= g(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + M_2(\beta_2\alpha_0 + \beta_2\alpha_1 X_1 - \beta_2 X_2)) \\ &= g(\beta_0 + \beta_1 X_1 + \beta_2(1-M_2)X_2 + \gamma_0 M_2 + \gamma_1 M_2 X_1), \end{aligned}$$

where $\gamma_0 = \beta_2\alpha_0$ and $\gamma_1 = \beta_2\alpha_1$. This shows that for a GLM with an additive linear predictor, combined with an imputation based on a linear regression, will reduce to a pattern submodel. Here, the functional forms of f and f^* are the same, though the coefficients for the submodels may take different values, owing to the use of imputation.

For variable importance considerations, the key observation is that *with imputation, prediction for cases with missing X_2 only uses information from X_1*. The pattern submodel structure is maintained even if a more general function $X_2^* = h(X_1)$ is used to carry out the imputation; the key difference is that the submodel corresponding to $M_2 = 1$ may not be linear or additive in X_1. These relationships are not expected to hold for cases where the linear predictor for the complete-data model has interactions or is otherwise not additive.

3.5 Implications for Variable Importance with Missing Data

In this section we describe how the LOCO measure of variable importance is calculated for pattern submodels and for imputation methods, with focus on whether and how variable importance reflects both contributions to predictive variance and degree of missingess. Specifically, key concept in how variable importance is defined in this work is the impact on predictive performance. Thus when determining how properly to calculate variable importance, it is important for the measure to properly reflect whether missingness or incomplete information on a variable is incorporated at the point of prediction. When missing data is expected at both

development and deployment the imputation strategy then becomes an integral part of the prediction pipeline and needs to be taken into account when calculating variable importance.

Stratification by missingness pattern for pattern submodels make it explicit how each covariate contributes to the submodels where they are observed, and more generally how each covariate contributes to the overall predictive variance. Similarly, variable importance calculated from imputed data must reflect that imputed values of a particular variable are actually functions of other observed variables in the model, which means the generated predictions actually rely directly on these observed values. In this section we show how to apply and calculate LOCO (dropped variable importance) for pattern submodels and imputation procedures. We also demonstrate that LOCO calculated from pattern submodels should be viewed as the 'correct' representation of variable importance because it directly accounts for incomplete information. The fully stratified pattern submodel is the most general formulation of a prediction model that uses only $\boldsymbol{X}_{\text{obs}}$.

3.5.1 Pattern Submodels

The stratification of the data by missingness pattern makes the calculation of dropped variable importance relatively straightforward for pattern submodels. For concreteness, we return to the three covariate example from the previous section, and illustrate the calculation of variable importance for X_1. Removing X_1 from the full specification given in (3.1) leads to the following set of reduced models,

$$\begin{aligned} f_1(\boldsymbol{X}_{\text{obs}(1)\backslash\{1\}};\widetilde{\boldsymbol{\theta}}_1) &= g(\widetilde{\theta}_{01} + \widetilde{\theta}_{21}X_2 + \widetilde{\theta}_{31}X_3) \\ f_2(\boldsymbol{X}_{\text{obs}(2)};\boldsymbol{\theta}_2) &= g(\theta_{02} + \theta_{22}X_2 + \theta_{32}X_3) \\ f_3(\boldsymbol{X}_{\text{obs}(3)\backslash\{1\}};\widetilde{\boldsymbol{\theta}}_3) &= g(\widetilde{\theta}_{03} + \widetilde{\theta}_{33}X_3) \\ f_4(\boldsymbol{X}_{\text{obs}(4)};\boldsymbol{\theta}_4) &= g(\theta_{04} + \theta_{34}X_3) \end{aligned} \tag{3.5}$$

where $\widehat{\theta}$ represents coefficients corresponding to the reduced models, and $\boldsymbol{X}_{\text{obs}(1)\backslash\{1\}}$ and $\boldsymbol{X}_{\text{obs}(3)\backslash\{1\}}$ refer to the fact that X_1 has been removed from patterns $s = 1$ and $s = 3$. This specification

does not collapse submodels 1 and 3; it only removes X_1 from the models where $M_1 = 0$ to isolate the effect of X_1 among the populations where it is observed. The models where $M_1 = 1$ do not change. The general version of variable importance is then calculated as a weighted average over the patterns as follows,

$$\begin{aligned}
\psi(j) &= E_{Y,\boldsymbol{X}}[L(Y, f(\boldsymbol{X})) - L(Y, f_{(-j)}(\boldsymbol{X}_{(-j)}))] \\
&= \sum_{s \in \mathscr{S}_p} E_{Y,\boldsymbol{X}}[L(Y, f(\boldsymbol{X})) - L(Y, f_{(-j)}(\boldsymbol{X}_{(-j)})) \,|\, S = s] \Pr(S = s) \\
&= \sum_{s \in \mathscr{S}_p} \psi_s(j) \Pr(S = s),
\end{aligned}$$

where $\psi_s(j)$ is the LOCO measure for X_j in stratum s. For a specific variable X_j, because X_j is not present in the patterns s where $M_j = 1$, the LOCO variable importance measure $\psi_s(j)$ within that pattern is identically zero and the estimator for global variable importance reduced to a weighted sum of $\psi_s(j)$ for the patterns s having $M_j = 0$. In our example, where $j = 1$, this implies $\psi_2(1) = \psi_4(1) = 0$. Hence the importance of a covariate reflects both its contribution to the pattern for where it is observed and the proportion of observations for which it is observed, and the variable importance can be calculated only over the strata where X_j is observed,

$$\psi(j) = \sum_{s : M_j = 0} \psi_s(j) \Pr(S = s)$$

It is important to distinguish here between the effect of X_j and the effect of (X_j, M_j). The above estimator quantifies the effect of X_j conditional on $M_j = 0$, but to get an estimate of the effect of both requires one to collapse the models over both (X_j, M_j) jointly. This is beyond the scope of this chapter and an avenue for future research.

3.5.2 Imputation-based Approaches

Imputation-based approaches require the end-user to produce two sets of models; firstly are the imputation models built from the observed data, and secondly is the outcome model. Again for concreteness and to illustrate key points, we returning to the two-covariate case from section 3.4.3. Recall that the full-data model and corresponding imputation model are given by

$$\begin{aligned} f^*(X_1, X_2; \boldsymbol{\beta}) &= g(\beta_0 + \beta_1 X_1 + \beta_2 X_2) \\ \hat{X}_2 &= \alpha_0 + \alpha_1 X_1. \end{aligned}$$

Denote the linear predictor, or the argument of $g(\cdot)$, as η. Applying the full-data model f^* to the imputed data yields

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 \left[(1 - M_2) X_2 + M_2 \hat{X}_2 \right]$$

We distinguish here between *conditional* and *unconditional* variable importance, where conditioning refers to whether or not one conditions on the imputation $\widehat{X}_2$. *Conditional variable importance* is calculated by treating $\hat{X}_2$ as the actual value of X_2; i.e., it does not distinguish between imputed and observed values of X_2. As we show below, this can have the effect of (mis)attributing variable importance of X_1 to X_2, or more generally attributing importance of variables used in the imputation model to the variable being imputed. *Unconditional variable importance* does not condition on the imputed values. It is calculated by removing, in this example, X_1 from both the prediction model *and* the imputation model.

Let us first illustrate unconditional imputation. Because there are no other variables in the model, removing X_1 from the imputation model implies that X_2 would be imputed as the mean

of its observed values, say μ. The reduced version of the linear predictor is

$$\eta_{\text{uncond.}} = \widetilde{\beta}_0 + \widetilde{\beta}_2[(1-M_2)X_2 + M_2\mu],$$

which is not a function of X_1. By contrast, the reduced version of the linear predictor for conditional imputation is

$$\eta_{\text{cond.}} = \widetilde{\beta}_0 + \widetilde{\beta}_2[(1-M_2)X_2 + M_2(\alpha_0 + \alpha_1 X_1)],$$

which clearly is a function of X_1.

When data are missing at both the development and deployment stages, the outcome and imputation model(s) should be considered as a unit in prediction when missingness is present at both model development and model deployment, and unconditional variable importance is the most appropriate approach. In practical terms, this means that in calculating variable importance, a covariate must be dropped from both the imputation and outcome models. Given that the imputed values are functions of this covariate we (1) re-fit the imputation models by dropping X_j, and (2) use the imputed dataset(s) derived from (1) to fit a reduced outcome model excluding X_j. This strategy only applies when the imputation model is used at the point of prediction. If missingness is not present at the point of prediction for new data, then it more appropriate to use conditional variable importance; i.e., it is fine to calculate the dropped importance without re-fitting the imputation models.

The calculation of LOCO for an outcome model with RI follows similarly since it results from one imputed dataset. For multiple imputation we calculate LOCO for each imputed dataset and combine the imputation specific LOCO's via Rubin's rules. In order to obtain reliable estimates we fit a large amount of outcome models, $K = 30$.

3.6 Simulation Study

Our simulation study is designed to compare covariate importance and predictive performance derived from four different techniques to handle missing data for prediction: (1) a fully stratified pattern submodel approach, (2) a smooth (collapsed) pattern submodel where the missingness indicators are modeled as direct effects and pattern specific covariate effects are collapsed, (3) a set of full outcome models estimated from multiply imputed data, and (4) a full outcome model estimated from regression imputation. Models (3-4) are fitted by interacting M with X and thus coefficients are estimated for the terms M, $X(1-M)$ and $\hat{X}M$. The full model with complete data is used as a reference to compare what the variable importance would look in the absence of missing data. This simulation section only considers a binary prediction problem and thus the loss function used to calculate the dropped variable importance is the log-loss defined in Section 3.3.1.

3.6.1 Data Generating Process

We present simulations for a generalized linear model under an MAR assumption for the missing data. These results, in principle, would extend to the setting of more complex models. Allow Y to be a binary outcome and $\boldsymbol{X} = (X_1, X_2, X_3)$ covariate data generated from a trivariate normal distribution MVN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) such that,

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

where the correlation coefficient ρ is set to $\rho = 0.50$ to induce dependencies between covariates. We set $g(\eta) = \exp\{\eta/(1+\eta)\}$ and specify the full-data model as

$$\Pr(Y = 1 \,|\, \boldsymbol{X}) \quad = \quad g(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3).$$

For illustrative purposes we set $\beta_1 = \beta_2 = \beta_3 = 1$ to allow the full-data relative importance of each covariate have equal value. This enables us to understand the impact of covariate missingness when calculating variable importance from various observed-data prediction models.

Missing values are induced in X_1 and X_2 using the following models,

$$\begin{aligned} \Pr(M_1 = 1 \,|\, \boldsymbol{X}) &= g(\xi_0 + \xi_1 X_1 + \xi_2 X_2 + \xi_3 X_3) \\ \Pr(M_2 = 1 \,|\, \boldsymbol{X}) &= g(\nu_0 + \nu_1 X_1 + \nu_2 X_2 + \nu_3 X_3), \end{aligned}$$

where (ξ_0, ν_0) are chosen so that the marginal probabilities $\Pr(M_1 = 1)$ and $\Pr(M_2 = 1)$ are approximately 30% and 70%, respectively. We set $\xi_3 = \nu_3 = 1$ and $\xi_1 = \eta_2 = \nu_1 = \nu_2 = 0$ so that the missingness mechanism is MAR.

3.6.2 Imputation Strategies

We calculate variable importance using both regression imputation and multiple imputation, and calculate both conditional and unconditional variable importance for each. Regression imputation of missing values are are fit using linear regression, with observed covariates used as predictors. For multiple imputation we use the MICE package (Van Buuren and Groothuis-Oudshoorn, 2011). We assume the conditional distributions $P(X_1 \,|\, X_2, X_3)$ and $P(X_2 \,|\, X_1, X_3)$ follow linear regressions with normal errors, consistent with the data generating model for the covariates.

Multiple imputation using chained equations (MICE) is implemented as follows:

1. Initialize each incomplete variable in the data matrix $\boldsymbol{X}$ by sampling from its observed marginal distribution, yielding an initial completed data matrix.

2. For each incomplete variable $X_{j'} \in \boldsymbol{X}_{mis}$, specify an imputation model

$$g_{j'}^{-1}\left\{E(X_{j'} \,|\, \boldsymbol{X}_{(-j')}, \boldsymbol{\eta}_{j'})\right\},$$

where $g_{j'}(\cdot)$ denotes a link function appropriate to the data type of $X_{j'}$ (e.g., binary, categorical, count). Each model is fit using only observed values of $X_{j'}$ (i.e., $M_{j'} = 0$). Independent priors are assumed across imputation models,

$$p(\boldsymbol{\phi}) = \prod_{j':X_{j'} \in \boldsymbol{X}_{mis}} p(\boldsymbol{\eta}_{j'}),$$

where $\boldsymbol{\phi} = \{\boldsymbol{\eta}_{j'} : X_{j'} \in \boldsymbol{X}_{mis}\}$.

3. Draw $\boldsymbol{\eta}^*_{j'}$ from its posterior distribution and, conditional on this draw, sample imputed values $x^*_{j'}$ from the posterior predictive distribution for individuals with $M_{j'} = 1$. Replace the previous imputed values of $X_{j'}$ with $x^*_{j'}$.

4. Repeat step (3) sequentially for each $X_{j'} \in \boldsymbol{X}_{mis}$ to complete one MICE iteration. This cycle is repeated for B iterations, with the completed data matrix from iteration b used as the starting point for iteration $b + 1$.

5. Repeat steps (1)–(4) independently K times to obtain K completed datasets $\{\boldsymbol{X}^{(k)}\}_{k=1}^{K}$, retaining the posterior draws $\boldsymbol{\phi}^{(k)}$ for each imputation model.

6. For each completed dataset $[Y, \boldsymbol{X}^{(k)}]$, fit the prediction model with risk function $f(\boldsymbol{x})$ (or $f(\boldsymbol{x}, \boldsymbol{m})$, when missingness indicators are included).

7. To generate predictions for a new individual in the validation set with incomplete covariates, initialize missing values using the mean or mode from the development data. Conditional on the fixed posterior draws $\boldsymbol{\phi}^{(k)}$, sample x^*_j from the posterior predictive distributions and repeat the imputation cycle described in step (4) (Hoogland et al., 2020b).

8. Compute individual-specific risk estimates by averaging predictions across imputations,

$$f(\boldsymbol{X}_{\text{obs}}) = \frac{1}{K} \sum_{k=1}^{K} f(\boldsymbol{X}^{(k)}).$$

3.6.3 Modeling Setup

The inclusion of the missingness indicators into the pattern mixture model framework to estimate $f(\boldsymbol{x}, \boldsymbol{m})$ can lead to an intractable amount of pattern submodels that need to be estimated and thus smoothening assumptions need to be made. In practice, however, with imputation techniques one usually seeks to estimate the quantity $f(\boldsymbol{x})$; i.e., the prediction function in terms of the covariates only. However, a question that arises is how to enter missingness indicators $\boldsymbol{M}$ into the model framework to reduce the number of potential parameters but still contain the pattern structure within the data.

As discussed previously when patterns are sparse the estimation of the kernels $f(\boldsymbol{x}, \boldsymbol{m})$ can become intractable and simplifying assumptions can be made, such as adding the missingness indicators into the model as direct effects. Thus we also seek to understand whether more parsimonious model specification can maintain predictive performance. For the imputation-based models, we use three specifications: (1) no missingness indicators, (2) the missingness indicators only enter as direct effects, and (3) the missingness indicators enter as direct effects along with interactions with the associated covariate. Specifically,

$$
\begin{aligned}
f_1(\boldsymbol{X};\boldsymbol{\theta}) &= g(\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3) \\
f_2(\boldsymbol{X},\boldsymbol{M};\boldsymbol{\alpha},\boldsymbol{\theta}) &= g(\alpha_0 + \alpha_1 M_1 + \alpha_2 M_2 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3) \\
f_3(\boldsymbol{X},\boldsymbol{M};\boldsymbol{\alpha},\boldsymbol{\theta}) &= g\left\{(\alpha_0 + \alpha_1 M_1 + \alpha_2 M_2) + (\theta_1(1-M_1) + \gamma_1 M_1)X_1 + (\theta_2(1-M_1) + \gamma_2 M_2)X_2 + \theta_3 X_3\right\}
\end{aligned}
$$

We code these two models for regression imputation and multiple imputation to best see how to borrow strength from patterns. In the simulation study, we focus on both f_1 and f_3 while in the application section we strictly focus on f_1.

3.6.4 Target and Performance measures

Our goal is to compare estimates of variable importance across the different approaches to modeling and handling missing covariates. To normalize the comparisons across methods,

we standardize relative importance of the covariates (X_1, X_2) in relation to X_3; specifically we use the relative importance quantities $\psi(1)/\psi(3)$ and $\psi(2)/\psi(3)$. We also calculate various predictive performance measures for each model using simulated validation samples.

3.6.5 Simulation Procedure

The full simulation procedure was carried out using the following steps:

1. $N = 2500$ samples of data $[Y, \boldsymbol{X}]$ are generated by first sampling $\boldsymbol{X}$ from the specified multivariate normal and subsequently sampling Y from the specified model;

2. missingness indicators $\boldsymbol{M}$ are generated according to the described MAR procedure and are then used to induce missingness in the covariates $\boldsymbol{X}$;

3. data $[Y, \boldsymbol{X}]$ and $[\boldsymbol{M}]$ are combined into a single dataset $[Y, \boldsymbol{X}, \boldsymbol{M}]$, and subsequently split into a training set of size $N_{\text{Tr}} = 2000$ and a test set of size $N_{\text{Te}} = 500$;

4. imputations are generated for models where it is required, and then an outcome model is fit for each of the four methods;

5. for $j = 1, 2, 3$ we fit a reduced outcome model by dropping X_j from the full outcome model and re-fitting. For models which require imputation we calculate variable importance using conditional and unconditional imputation;

6. calculate the performance metrics of interest for each method.

These steps are repeated $T = 250$ times and the median estimates of the relative variable importance and performance metrics, and associated 95% Monte Carlo intervals, are reported. For steps requiring multiple imputation we impute $K = 30$ completed datasets. Performance metrics are calculated using the set of K outcome models from multiple imputation and combined using Rubin's rules. Simulations and analyses were performed in R version 4.2.2 (Team et al., 2021). The pROC library (Robin et al., 2011) was used to calculate the AUROC while the

MICE package (Van Buuren and Groothuis-Oudshoorn, 2011) was used for multiple imputation as described above.

3.6.6 Results

Figure 3.1 shows the comparison of variable importance across the range of models that use prediction function f_3. As a reference point, for the full model (no missing covariates), the relative VI for X_1 and X_2, relative to X_3, are all distributed around 1. The stratified pattern submodel, which is the most general specification, shows lower variable importance for both X_2 and X_3, reflecting their proportion of missingness. The collapsed pattern submodel, which uses fewer stratification categories, shows similar results.

Turning to the imputation methods, we see that when the covariate is removed from the outcome model but not from imputation model (i.e., conditional variable importance), that we tend to overestimate the importance of X_1 and X_2 relative to X_3. This is particularly pronounced when using regression imputation, which is a single-imputation strategy and does not average over the distribution of the missing data as is done with multiple imputation. Single regression imputation is deterministic and thus collapses the conditional variance of the imputed variable, which changes the downstream model coefficients and measured importance more when the variable in question is removed from the imputation model. These patterns are repeated in Figure 3.2 where the f_1 model specification with no missingness indicators is used for outcome models built from imputed data, and in Figure 3.3, where the relative variable importance of X_1 and X_2 is larger ($\sqrt{2}$ instead of 1).

In terms of model performance, each of the models had similar AUROC, sensitivity and specificity. This is likely attributable to the fact that each model is comprised of linear prediction models for both the outcome and the imputations.

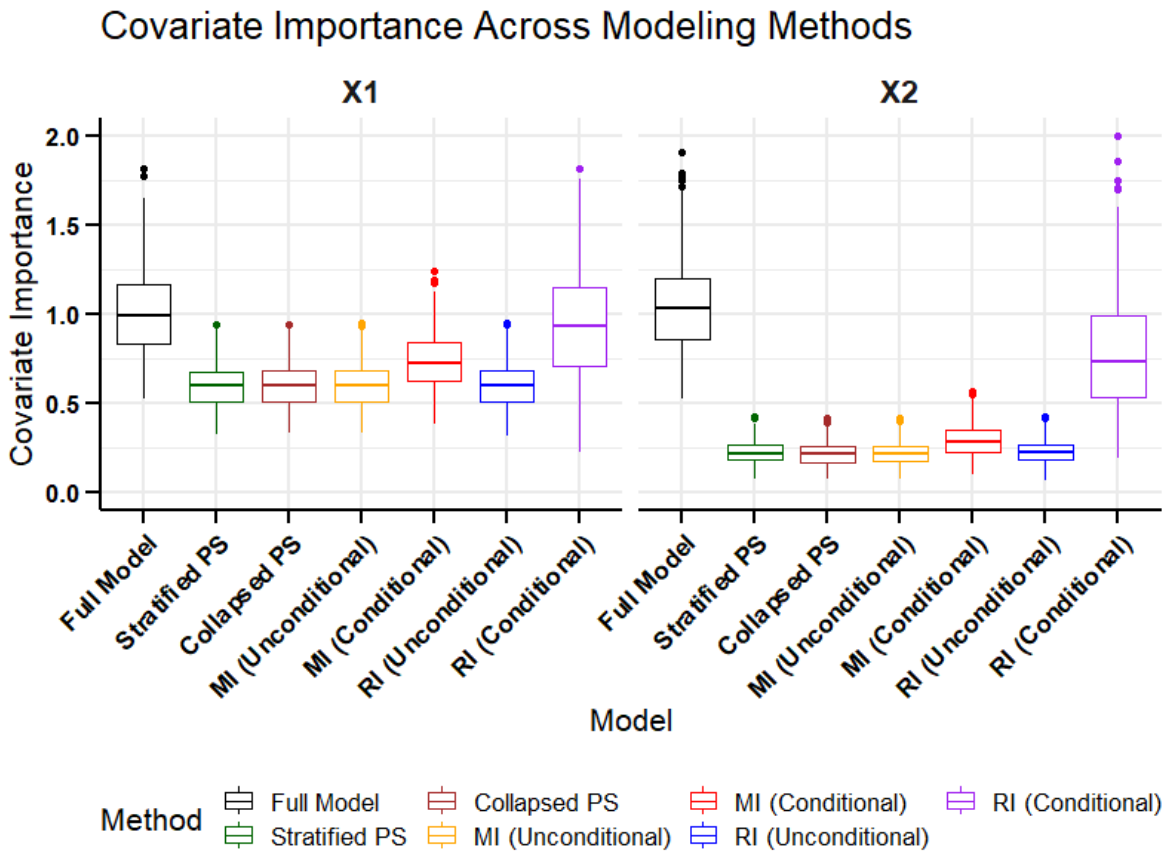

Figure 3.1: Boxplots for the LOCO importance relative ratios $\psi(j)/\psi(3)$ for $j = 1,2$ for the model where $(\beta_1, \beta_2, \beta_3) = (1,1,1)$. Models are fit using logistic regression for the linear case specified in Scenario 1. MI stands for multiple imputation, RI stands regression imputation and PS stands for pattern submodels. The full model refers to the model fit to the simulated data where there is no induced missingness. Models built from imputed data are fit with missingness indicators using the f_3 model specification.

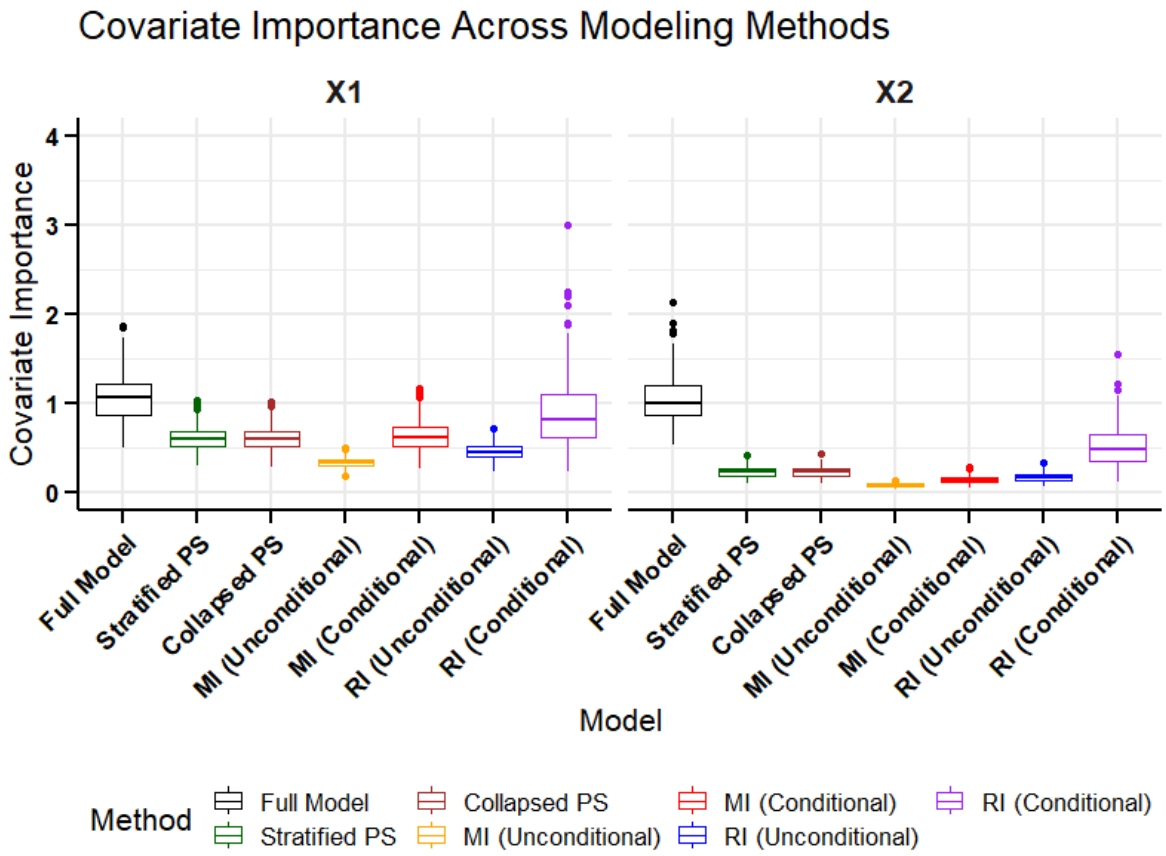

Figure 3.2: Boxplots for the LOCO importance relative ratios $\psi(j)/\psi(3)$ for $j = 1,2$ for the model where $(\beta_1, \beta_2, \beta_3) = (1,1,1)$. MI stands for multiple imputation, RI stands regression imputation and PS stands for pattern submodels. The full model refers to the model fit to the simulated data where there is no induced missingness. Models built from imputed data are fit using f_1 specification with no missingness indicators.

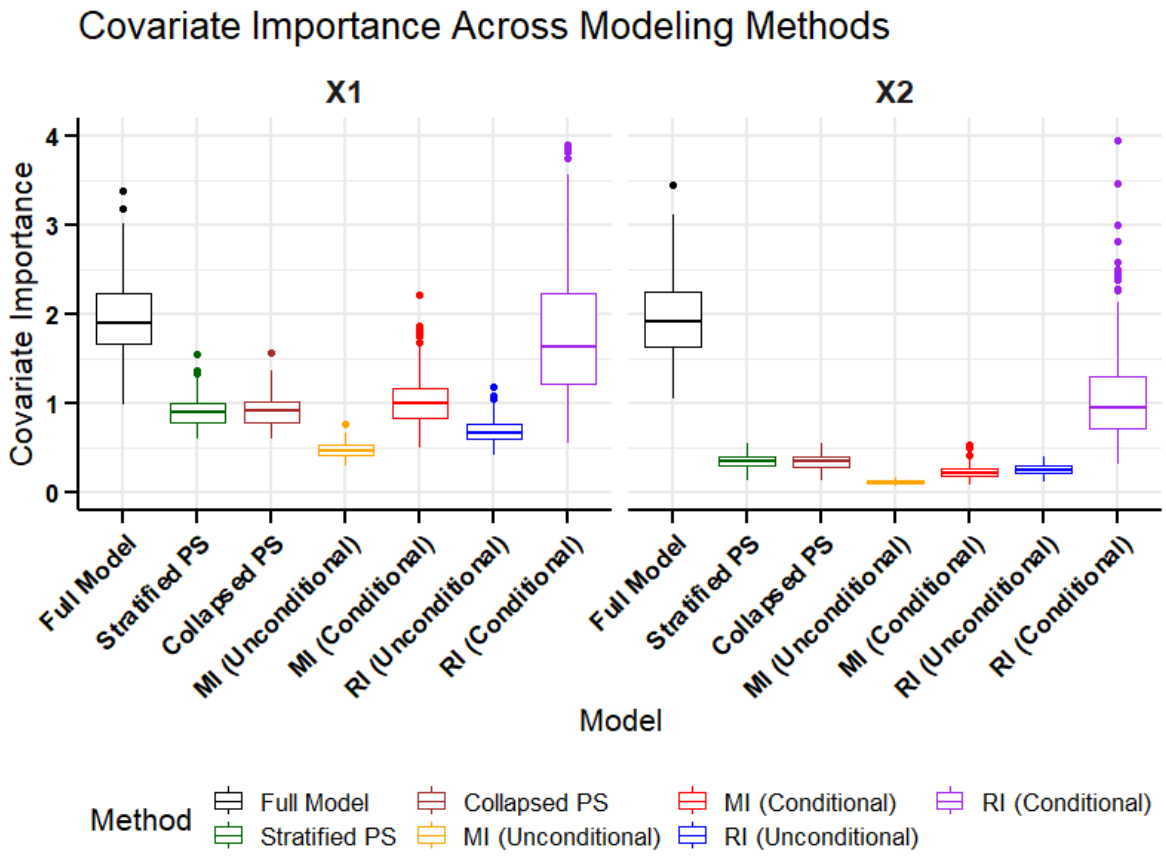

Figure 3.3: Boxplots for the LOCO importance relative ratios $\psi(j)/\psi(3)$ for $j = 1,2$ for the model where $(\beta_1, \beta_2, \beta_3) = (\sqrt{2}, \sqrt{2}, 1)$. MI stands for multiple imputation, RI stands regression imputation and PS stands for pattern submodels. The full model refers to the model fit to the simulated data where there is no induced missingness. Models built from imputed data are fit using f_1 specification with no missingness indicators.

3.7 Data Application

3.7.1 AMPATH data

We illustrate the application of our method on data from AMPATH for a cohort of 10,778 patients who have enrolled into care since January 1st, 2021. Clinical visit information from these individuals was derived from an electronic database known as the AMPATH Medical Record System (AMRS) which is generated by a browser-based front-end application deployed at the point of care in associated facilities (hospital and clinics) (Tierney et al., 2007). At the end of visit, a patient is assigned a return to care (RTC) date for a regular return visit or

a medication pickup. Often times these two can coincide. The endpoint for our prediction model is a binary indicator Y that takes value 1 if the patient fails to return before their RTC date, and 0 if they do return on time. The focus of the application section is on summarizing covariate importance for a model that predicts return to care after the enrollment visit. The binary outcome for Table 3.1 summarizes the distribution of possible states after the enrollment visit. We assume administrative censoring is non-informative and exclude records where this can occur. Among the 10,542 patients in the analysis sample, after excluding those censored due to database closure, 57% returned on time or early, 41% missed their next scheduled visit, and 2% died before their next visit could occur.

We split the data temporally into a training set of size 8,642 consisting of patient records from January 1st, 2021 to December 31st, 2022, and a test set of size 1,900 consisting of patients who were scheduled to return to care between January 1st, 2023 and July 1st, 2023. Covariates information used to develop the model include age, gender, body mass index (BMI), ARV status, Tuberculosis Prophylaxis status, CD4 count, clinic type, clinic size on the log base 10 scale, and the month of the scheduled appointment. Covariates containing missing records included BMI, CD4 count, WHO stage and education level and were coded with a corresponding missingness indicator leading to a configuration of 14 different missing data patterns, the prevalence of each pattern by outcome is described in table 3.3. Table 3.2 summarizes the observed distributions of the covariates for the population and stratified by outcome level. The month of the scheduled appointment was coded as a categorical variable, but was excluded from the table due to the large number of categories.

Table 3.1: Patient statuses for $\Delta = 0$ delays

	Visit Total	Database Closure	On Time/Early	Missed	Dead
$\ell = 1$	10,778	236 (2.2)	6,064 (56.3)	4,303 (39.9)	175 (1.6)

Table 3.2: Baseline Characteristics for patients at the time of entry into the cohort. The table reports "median (25th percentile, 75th percentile)" for continuous variables and percentage of true for binary variables or each level of categorical variables.

		Return to 1st Scheduled Clinical Visit After Enrollment		
Variable	Total (N = 10,778)	Early/On Time (N = 6,064)	Missed (N = 4,303)	Died (N=175)
Age (years)				
Median (Q1, Q3)	36 (29, 45)	37 (30, 47)	35 (28, 43)	40 (34, 49)
Scheduled Appointment Time (Weeks)				
Median (Q1, Q3)	2 (2, 3.29)	2 (2, 2.86)	2 (2, 3.86)	2 (2,2)
Sex				
Female	6,434 (59.7)	3,580 (59.0)	2,628 (61.1)	84 (48.0)
Male	4,344 (40.3)	2,484 (41.0)	1,675 (38.9)	91 (52.0)
ARV Status				
Not On	1,207 (11.2)	709 (11.7)	401 (9.3)	74 (42.3)
On	9,571 (88.8)	5,355 (88.3)	3,902 (90.7)	101 (57.7)
Facility Type				
Regional/Referral Hospital	1253 (11.6)	714 (11.8)	475 (11.0)	47 (26.9)
County Hospital	2765 (25.7)	1630 (26.9)	1034 (24.0)	42 (24.0)
Sub-County Hospital	4032 (37.4)	2421 (39.9)	1460 (33.9)	69 (39.4)
Health Center/Dispensary	2604 (24.2)	1244 (20.5)	1270 (29.5)	16 (9.1)
Private	124 (1.2)	55 (0.9)	64 (1.5)	1 (0.6)
Body Mass Index ($\frac{kg}{m^2}$)				
Median (Q1, Q3)	21 (19, 23)	21 (18, 23)	21 (19, 23)	19 (17, 21)
Missing	3,518 (32.6)	1,691 (27.9)	1,665 (38.9)	98 (56.0)
CD4 Count				
Median (Q1, Q3)	244 (94, 437)	226 (91, 407)	289 (125, 488)	39 (19, 90)
Missing	8,711 (80.8)	4,726 (77.9)	3,643 (84.7)	127 (78.5)
WHO Stage*				
Missing	658 (6.1)	334 (5.5)	318 (7.4)	2 (1.1)
First Stage	6358 (50.2)	3,448 (56.9)	2736 (63.6)	23 (13.1)
Second Stage	1838 (16.1)	1155 (19.0)	605 (14.2)	34 (19.4)
Third Stage	1640 (20.9)	991 (16.3)	540 (12.5)	73 (41.7)
Fourth Stage	286 (4.3)	136 (2.2)	104 (2.4)	43 (24.6)
Education Level*				
Missing	233 (2.1)	93 (1.5)	132 (3.1)	5 (2.9)
None	873 (8.1)	390 (6.4)	441 (10.2)	17 (9.7)
Primary School	5025 (46.6)	2858 (47.1)	1991 (46.3)	82 (46.9)
Secondary School	3485 (32.3)	2006 (33.1)	1339 (31.1)	57 (32.6)
College	1162 (10.8)	717 (11.8)	400 (9.3)	14 (8.0)

Table 3.3: Observable patterns of missing covariates in the AMPATH data at the enrollment visit.

(BMI, CD4, WHO Stage, Educ. Level)	Total (N = 10,778)	Database Closure (N= 236)	Early/On Time (N = 6,064)	Missed (N = 4,303)	Died (N=175)
Pattern 1 (Obs, Obs, Obs, Obs)	1,589 (14.7)	14 (5.9)	1,072 (17.7)	481 (11.2)	22 (12.6)
Pattern 2 (NA, Obs, Obs, Obs)	393 (3.6)	5 (2.1)	220 (3.6)	144 (3.3)	24 (13.7)
Pattern 3 (Obs, NA, Obs, Obs)	5,244 (48.7)	152 (64.4)	3,081 (50.8)	1,957 (45.5)	54 (30.9)
Pattern 4 (NA, NA, Obs, Obs)	2,678 (24.8)	58 (24.6)	1,269 (20.9)	1,283 (29.8)	68 (38.9)
Pattern 5 (Obs, Obs, NA, Obs)	46 (0.4)	1 (0.4)	26 (0.4)	19 (0.4)	0 (0)
Pattern 6 (NA, Obs, NA, Obs)	26 (0.2)	0 (0)	15 (0.2)	10 (0.2)	1 (0.6)
Pattern 7 (Obs, NA, NA, Obs)	291 (2.7)	3 (1.3)	150 (2.5)	138 (3.2)	0 (0)
Pattern 8 (NA, NA, NA, Obs)	278 (2.6)	0 (0)	138 (2.3)	139 (3.2)	1 (0.6)
Pattern 9 (Obs, Obs, Obs, NA)	9 (0.1)	1 (0.4)	5 (0.1)	3 (0.1)	0 (0)
Pattern 10 (NA, Obs, Obs, NA)	4 (0)	0 (0)	0 (0)	3 (0.1)	1 (0.6)
Pattern 11 (Obs, NA, Obs, NA)	77 (0.7)	1 (0.4)	39 (0.6)	36 (0.8)	1 (0.6)
Pattern 12 (NA, NA, Obs, NA)	126 (1.2)	1 (0.4)	44 (0.7)	78 (1.8)	3 (1.7)
Pattern 13 (Obs, Obs, NA, NA)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pattern 14 (NA, Obs, NA, NA)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pattern 15 (Obs, NA, NA, NA)	4 (0)	0 (0)	0 (0)	4 (0.1)	0 (0)
Pattern 16 (NA, NA, NA, NA)	13 (0.1)	0 (0)	5 (0.1)	8 (0.2)	0 (0)

3.7.2 Model Specification

We assume that missed-visit status Y follows a Bernoulli distribution $[Y \mid \boldsymbol{X}] \sim \text{Ber}\{\pi(\boldsymbol{X})\}$, where $\pi(\boldsymbol{X}) = \Pr(Y = 1|\boldsymbol{X})$ is parameterized as $\pi(\boldsymbol{X}) = g(\boldsymbol{X})$. To parameterize $g(\boldsymbol{X})$, we use main-effects logistic regression. Missingness in the covariate space is handled using the fully stratified pattern submodel, the collapsed pattern submodel, regression imputation and multiple imputation approaches. Due to the sparsity of the observable missingness patterns from (3.3) we focus on the largest four patterns induced by the combinations of CD4 count and BMI. Thus the stratified pattern submodel is fit to those four patterns with the missingness indicators of WHO stage and education level added as direct effects in the resulting four regression models.

The collapsed pattern submodel includes the four missingness indicators for the covariates with missing values as direct effects. The covariates are then represented as $X_j(1-M_j)$, which is modeled using cubic regression splines for BMI and CD4 count, with knots selected from the empirical distribution of $X_j(1-M_j)$. For categorical variables such as WHO stage and education level, missingness is treated as an additional category. Outcome models fitted using regression imputation or multiple imputation follow the f_1 specification from Section 3.6.3 and

are therefore fit without missingness indicators. In these models, covariates with missing data enter as X_j without distinguishing between observed and imputed values.

In the AMPATH data, censoring by death before a visit represents a competing risk in the estimation of retention and is handled through a probabilistic chained classifier approach where we introduce an auxiliary binary outcome variable D defined by survival long enough to have the next visit ($D = 1$) or death before the visit ($D = 0$). Our expansion of the original probability then becomes

$$\begin{aligned} \Pr(Y = 1 \mid \boldsymbol{X}) &= \sum_{d \in \{0,1\}} \Pr(Y = 1 \mid \boldsymbol{X}, D = d) \Pr(D = d \mid \boldsymbol{X}) \\ &= \Pr(Y = 1 \mid \boldsymbol{X}, D = 1) \Pr(D = 1 \mid \boldsymbol{X}). \end{aligned}$$

This simplification occurs because the probability $\Pr(Y = 1 \mid \boldsymbol{X}, D = 0)$ is identically zero; i.e., a patient who dies before their scheduled visit cannot miss their scheduled visit. Due to sparsity the functional form of $\Pr(D = 1 \mid \boldsymbol{X})$ is fit using a main effects logistic regression where continuous covariates are instead fit linearly instead of with regression splines. The functional form of $\Pr(Y = 1 \mid \boldsymbol{X}, D = 1)$ is fit using a main effects logistic regression where continuous covariates are specified as regression splines.

3.7.3 Imputation Model Specification

For regression imputation we specify four imputation models in total; one for BMI, CD4 count, WHO stage and education level. BMI and CD4 count are specified as linear models while WHO stage and education level are specified as multinomial logistic regressions. Due to the sparsity of missingness patterns shown in Table 3.3, we collapse patterns and include the missingness indicators as direct effects in the models. Covariates with missing values are entered as $X_j(1 - M_j)$, with continuous covariates represented using cubic regression splines. Continuous covariates are imputed using their conditional mean, while categorical covariates are imputed deterministically as the category with the highest conditional probability, i.e. $\hat{X}_j = \text{argmax}_{c \in \{1,2,\ldots,C\}} Pr(X_j = c \mid \boldsymbol{X}_{-j})$.

For multiple imputation a total of four imputation models are developed, one each for BMI, CD4 count, WHO stage and education level. The imputation models for BMI and CD4 count are specified using a linear model while WHO stage and education level are specified using a multinomial logistic regression. Continuous covariates are represented with cubic regression splines. We generate $K = 30$ completed datasets and fit the outcome model to each imputed dataset. Predictions for a new individual are obtained following the procedure outlined in Section 3.6.5.

3.7.4 Results

Figure 3.4 shows the relative variable importance estimates for the outcome model using regression imputation and the outcome model using multiple imputation to handle missing data. The ranking of covariates varies depending on whether a conditional or unconditional strategy is employed. When the covariate is removed from the outcome model but not the imputation model (conditional strategy) the relative importance of WHO stage changes substantially, with more extreme differences observed for regression imputation compared to multiple imputation. When missingness is present at both model development and deployment, failing to remove a covariate from both the imputation and outcome models can lead to underestimation of its relative importance. This may misrepresent how predictions are generated, and if variable importance metrics are used to guide clinical decision-making, it could result in critical measurements being collected less frequently or not at all.

For the pattern submodel, we report both the global variable importance and the stratified model variable importances. The global variable importance for the stratified pattern submodel is calculated using the estimand described in Section 3.5.1. Figure 3.5 compares the global variable importance across the stratified pattern submodels, the collapsed pattern submodels, the outcome model with regression imputation, and the outcome model with multiple imputation. All four approaches identify age as the most important predictor. In general, the stratified pattern submodels place greater emphasis on the month of the scheduled appoint-

ment, likely an artifact of the stratification structure. CD4 count and BMI are consistently ranked near the bottom in terms of relative importance, consistent with our observation that variables with high percentages of missing values contribute little to overall predictive variation. Overall, the variable importance estimates from the imputation outcome models and the collapsed pattern submodel are remarkably similar.

While CD4 count is consistently ranked near the bottom in the global variable importance, Figure 3.6 shows that within fully observed strata, CD4 count is the most important predictor. This illustrates that even covariates that are sparsely measured in the overall dataset can have a substantial predictive effect in the strata where they are observed. Notably, in the subgroup where both BMI and CD4 count are missing, clinic-level characteristics emerge as increasingly important contributors to the predictive variation.

Table 3.4 presents predictive performance metrics for the pattern submodel approaches and imputation-based approaches. Overall, the methods perform similarly, with the exception of the fully stratified pattern submodel for calibration intercept and slope, likely due to limited sample sizes within each pattern. This effect is visually apparent in Figure 3.7, where the stratified pattern submodel substantially underestimates risk at higher predicted probabilities. Other metrics, such as AUC and PPV, show minimal differences between the pattern submodel and imputation approaches. This result is expected: imputation does not introduce new information into the model, as imputations are derived from the observed data. Consequently, it is generally not surprising that imputation does not improve predictive performance compared with a pattern submodel-based approach.

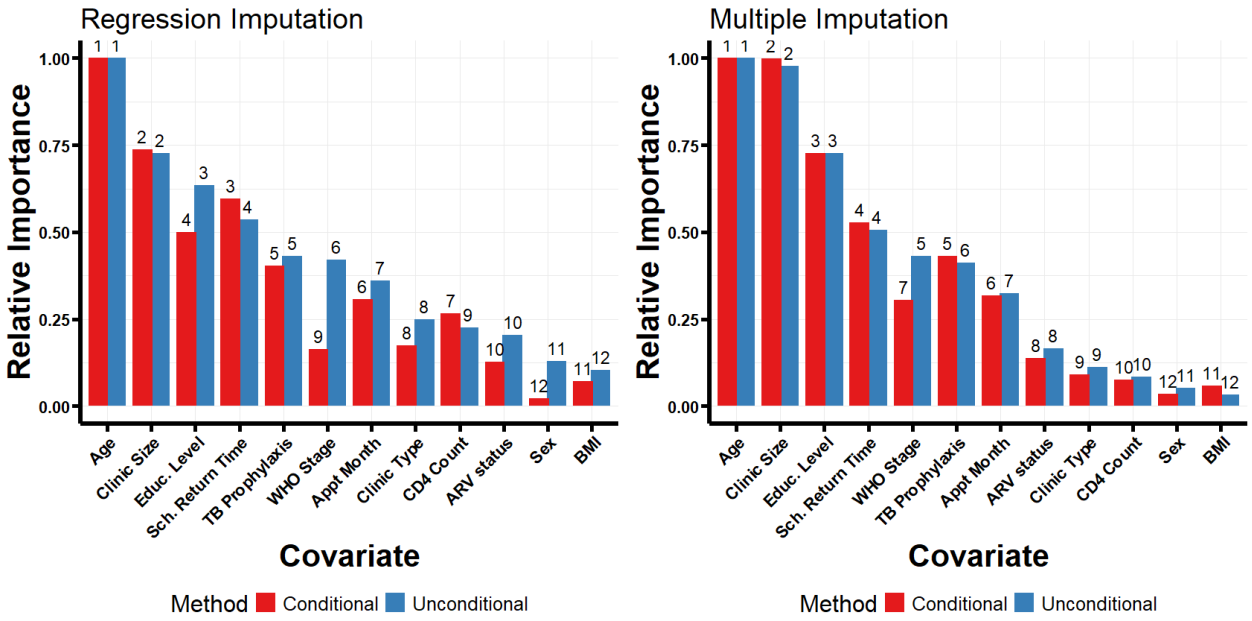

Figure 3.4: Relative global LOCO estimates for the variable importance $\psi(j)$ using regression imputation (left) and multiple imputation (right). Red denotes the conditional method of calculating variable importance with imputed values where X_j is removed from both the imputation models and outcome model. Blue denotes the unconditional method of calculating variable importance with imputed values where X_j is removed from the outcome model, but not the imputation model. Covariates are ranked in descending order from largest, denoted by 1, to smallest, denoted by 12.

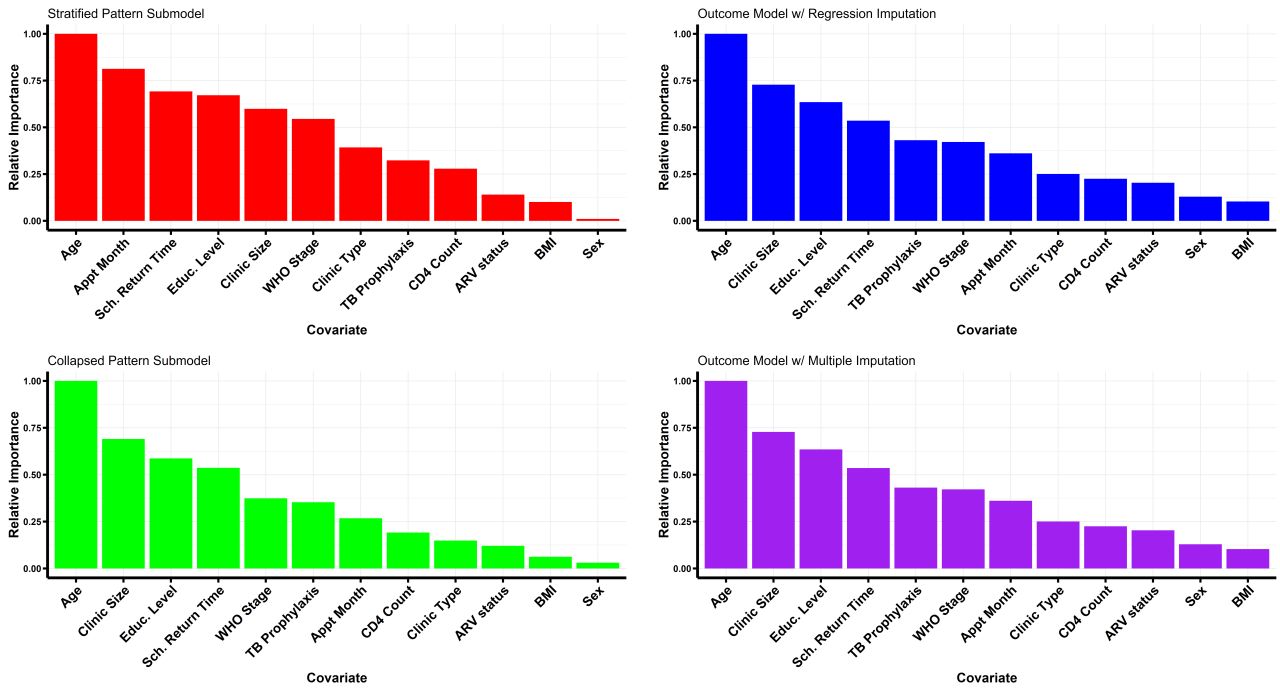

Figure 3.5: Relative global LOCO estimates for the variable importance $\psi(j)$ using stratified pattern submodels (top left), regression imputation (top right) using the unconditional strategy, the collapsed pattern submodel (bottom left) and multiple imputation (bottom right) using the unconditional strategy. Presented are the relative ratios $\psi(j)/\psi(k)$ where the reference $\psi(k)$ is the largest LOCO estimate, Age.

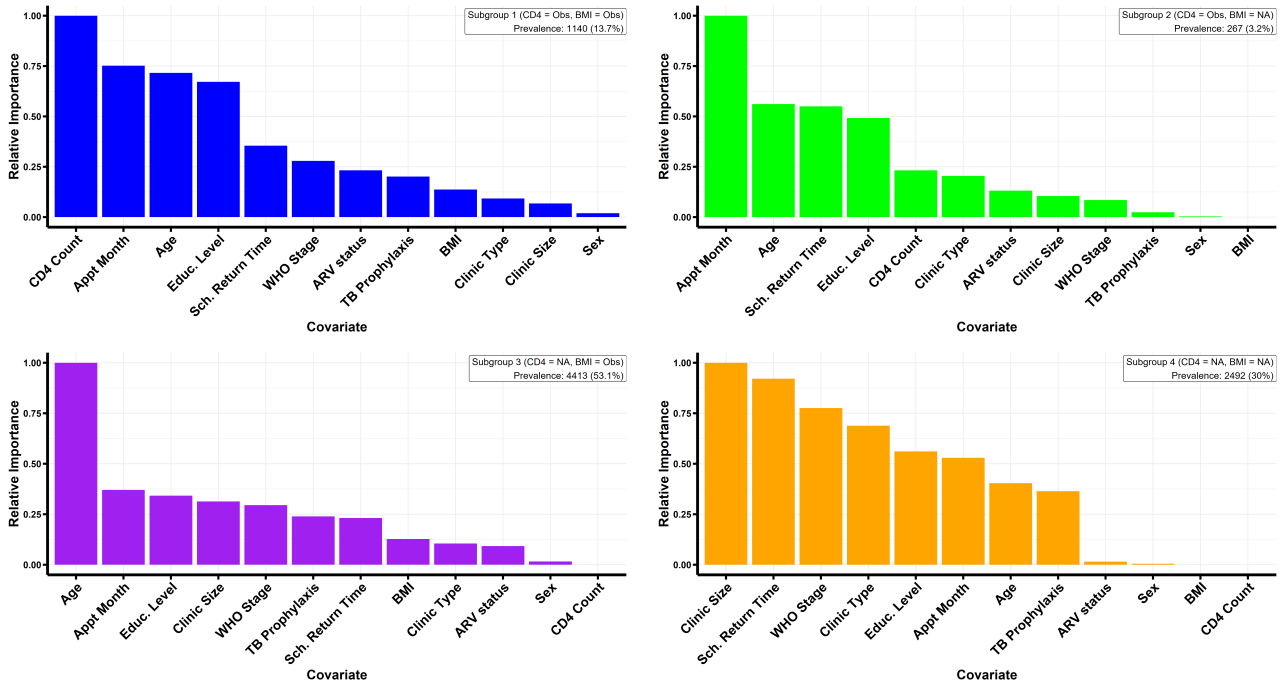

Figure 3.6: Relative variable importance estimates using LOCO within each of the defined subgroups using the stratified pattern submodels. Plotted are the relative importances where in each subgroup the $\psi(j)$ estimates are divided by the largest LOCO within each subgroup, $\max_j \psi(j)$.

Table 3.4: Model Performance Metrics for different missingness modeling strategies. Metrics such as positive predictive value are calculated using the 80^{th} percentile as the threshold for "high risk" classification. The table reports the estimate with the 95% confidence interval (CI). The prevalence is reported for context.

	$\Delta = 0$				
Missingness Strategy	AUROC	CITL	Cal. Slope	PPV	Prevalence
Stratified Pattern Submodels	62.9 (60.4, 65.5)	–0.19 (–0.29, –0.09)	0.57 (0.44, 0.70)	50.6 (46.1, 55.2)	40.7
Collapsed Pattern Submodel	62.7 (60.1, 65.2)	–0.17 (–0.27, –0.07)	0.74 (0.58, 0.90)	50.5 (44.3, 56.1)	40.7
Regression Imputation	62.9 (60.3, 65.5)	–0.16 (–0.26, –0.07)	0.74 (0.59, 0.91)	50.6 (46.1, 55.2)	40.7
Multiple Imputation	63.0 (60.4, 65.6)	–0.15 (–0.25, –0.05)	0.77 (0.60, 0.93)	50.6 (46.1, 55.2)	40.7

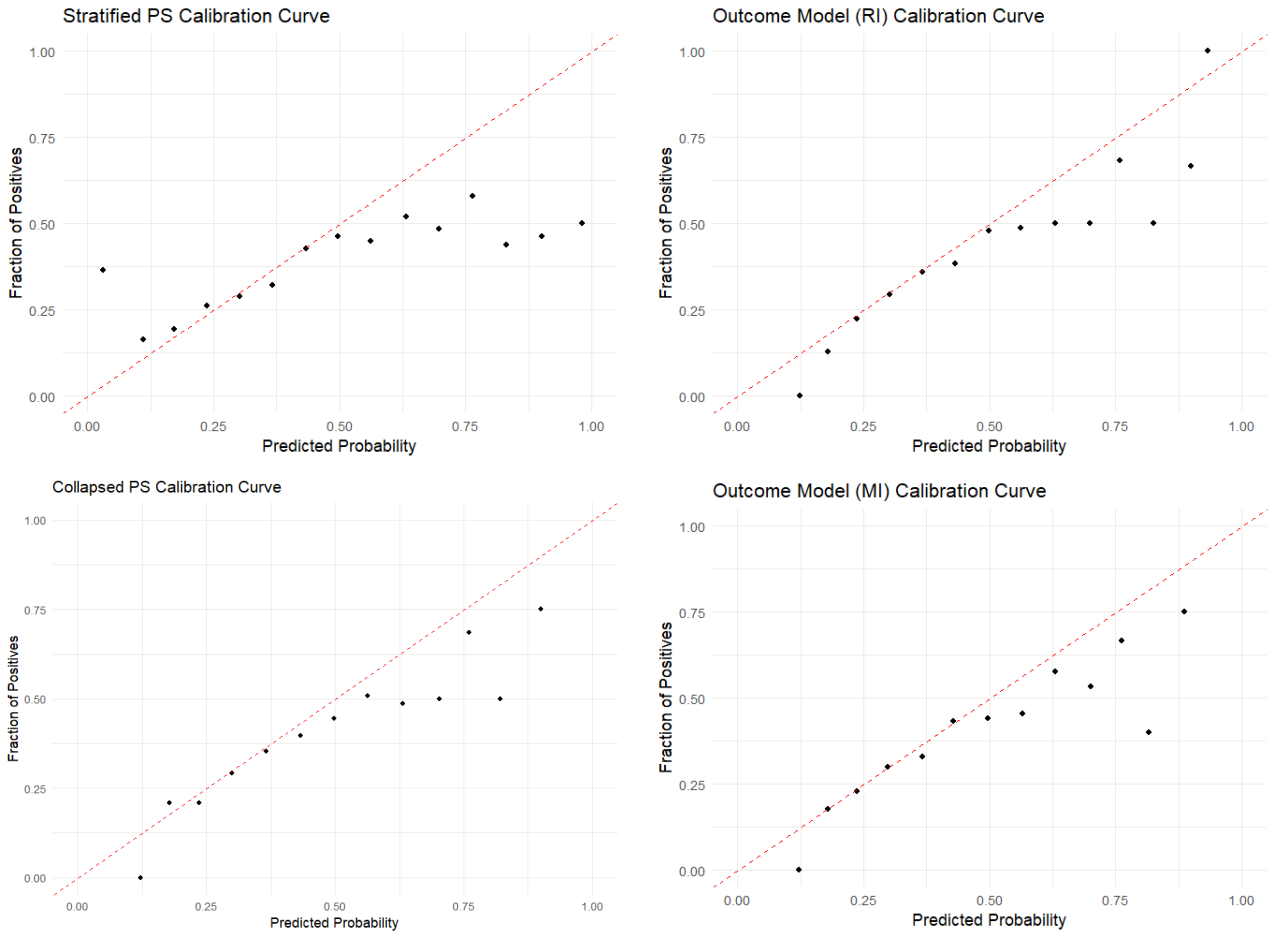

Figure 3.7: Calibration curves using stratified pattern submodels (top left), outcome model with regression imputation (top right), the collapsed pattern submodel (bottom left) and the outcome model with multiple imputation (bottom right). The dashed red line is a 45-degree diagonal line which represents perfectly calibrated prediction probabilities.

3.8 Discussion

We have conducted an extensive study regarding the role of variable importance in the presence of missing data when missingness is expected at both model development and model deployment. We have shown how the overall outcome model relies on certain key covariates and in the missingness patterns where they are observed. We demonstrate this approach through popular model agnostic variable importance methods such as dropped variable importance. These measures are defined as a contrast between the predictiveness of the best possible prediction function based on all available features versus all features but those under

consideration.

When dealing with missing data at both model development and model deployment in prediction settings we recommend the using either the smooth pattern submodel approach, or if imputation is required, an outcome model with multiple imputation. The smooth pattern submodel does not require imputation at the prediction stage, is straightforward to implement and has comparable predictive performance to other more computationally expensive methods. While regression imputation is deterministic and straightforward the technique still requires one to fit as many imputation models as there are missingness patterns and thus inherits similar weaknesses to the full pattern submodel stratification. Multiple imputation is already a computationally exhaustive process, and requires a large amount of imputations to reliably estimate the risk score.

We encounter several limitations within the scope of this chapter. Firstly missingness itself can be informative about the outcome Y. For example, a lab test not being ordered could indicate that the doctor thought the patient was low risk for an event, and a user not filling in a field may correlate with some behavioral pattern. Future work could explicitly investigate the contribution of missingness patterns to the outcome, rather than focusing solely on the observed components of the covariates. Second, the missing data mechanism (MDM) in the simulation study was assumed to be missing at random (MAR); additional investigation is needed to determine whether similar results hold under missing not at random (MNAR) mechanisms. With regard to the imputation models it is common for the end-user to include auxiliary variables into the imputation models which do not appear in the outcome model (Sisk et al., 2023). Similarly the quality of the imputation models was not investigated in this work; future work may seek to determine whether ML-based imputation models produce better predictions (Hu et al., 2021). Finally, the missing data mechanism may evolve across stages of model development and deployment. The use of a prediction tool may influence clinical behavior, encouraging more systematic data collection (Groenwold, 2020). Awareness that certain variables carry predictive value could alter decisions about whether and how of-

ten they are measured, potentially changing their predictive utility and affecting the model's ability to generalize to deployment data. These dynamics represent important directions for future research.

Appendix

Calculation of the conditional effect plots

To calculate the conditional effect plots from chapter 1 we implement the following scheme. The support of a covariate X is defined as the set of unique values that are observed for X in the data.

1. Given a fitted model $\hat{\pi}(\boldsymbol{X})$ choose a selected covariate X_j from $\boldsymbol{X}$.

2. group the predicted probabilities from (1) based upon the values in the support of X_j. For example, if X_j were categorical with C categories one would have C sets of probabilities $\mathscr{C}_k = \{i : X_j = c_k\}$

3. Calculate the average risk score per value of x in the support of X_j,

$$h_j(x) \quad = \quad \frac{1}{\#\{i : X_j = x\}} \sum_{i:X_j=x} s(\hat{\pi}(\boldsymbol{x}_{i(-j)}, x))$$

4. Repeat step (3) for each posterior value.

For continuous covariates where there may be sparsity for some values in the support one can then implement a quantile based approach where the average risk score is calculated per bin.

Derivation of LOCO importance for linear model

Linear Model with squared error loss

Consider the case where the loss function is squared error loss $L(y,f) = (y-f)^2$, the prediction function is specified as the optimal predictor $f(\boldsymbol{X};\boldsymbol{\theta}) = E(Y|\boldsymbol{X})$. Notice at this point no functional form has been specified for $f(\boldsymbol{X})$. For this case, variable importance for X_j is

$$\begin{aligned}
\psi(j) &= E_{Y,\boldsymbol{X}}\left\{(Y-f(\boldsymbol{X}))^2\right\} - E_{Y,\boldsymbol{X}}\left\{(Y-f_{-j}(\boldsymbol{X}_{-j}))^2\right\} \\
&= E_{Y,\boldsymbol{X}}\left\{(Y-f(\boldsymbol{X}))^2 - (Y-f_{-j}(\boldsymbol{X}_{-j}))^2\right\} \\
&= E_{Y,\boldsymbol{X}}\left\{(Y^2-2Yf(\boldsymbol{X})+f(\boldsymbol{X})^2) - (Y^2-2Yf_{(-j)}(\boldsymbol{X}_{(-j)})+f_{(-j)}(\boldsymbol{X}_{(-j)})^2)\right\} \\
&= E_{Y,\boldsymbol{X}}\left\{(f(\boldsymbol{X})^2 - f_{(-j)}(\boldsymbol{X}_{(-j)})^2)\right\} - 2E_{\boldsymbol{X}}\left\{E_{Y|\boldsymbol{X}}\left\{\left\{Y(f_{(-j)}(\boldsymbol{X}_{(-j)}) - f(\boldsymbol{X}))\right\}\right\}\right. \\
&= E_{\boldsymbol{X}}\left\{f(\boldsymbol{X})^2 - f_{(-j)}(\boldsymbol{X}_{(-j)})^2\right\} - 2E_{\boldsymbol{X}}\left\{f(\boldsymbol{X})(f_{(-j)}(\boldsymbol{X}_{(-j)}) - f(\boldsymbol{X}))\right\} \\
&= -E_{\boldsymbol{X}}\left\{f(\boldsymbol{X})^2 + 2f(\boldsymbol{X})(f_{(-j)}(\boldsymbol{X}_{(-j)}) - f_{(-j)}(\boldsymbol{X}_{(-j)})^2\right\} \\
&= E_{\boldsymbol{X}}\left\{\left(f(\boldsymbol{X}) - f_{(-j)}(\boldsymbol{X}_{(-j)})\right)^2\right\}
\end{aligned}$$

The optimal predictors for the loss functions $E_{Y,\boldsymbol{X}}\left\{(Y-f(\boldsymbol{X}))^2\right\}$ and $E_{Y,\boldsymbol{X}}\left\{(Y-f_{(-j)}(\boldsymbol{X}_{(-j)}))^2\right\}$ are $f(\boldsymbol{X}) = E[Y|\boldsymbol{X}]$ and $f_{(-j)}(\boldsymbol{X}_{(-j)}) = E[Y|\boldsymbol{X}_{(-j)}]$, respectively. When we specify $f(\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\theta}$, using the tower property we then deduce that $f_{(-j)}(\boldsymbol{X}_{(-j)}) = E[Y|\boldsymbol{X}_{(-j)}] = E[E[Y|\boldsymbol{X}]|\boldsymbol{X}_{(-j)}] = \theta_0 + \theta_j E[X_j|\boldsymbol{X}_{(-j)}] + \sum_{k\neq j}\theta_k X_k$. Respectively we can see that this form will reduce to,

$$\begin{aligned}
E_{\boldsymbol{X}}\left\{\left(f(\boldsymbol{X}) - f_{(-j)}(\boldsymbol{X}_{(-j)})\right)^2\right\} &= E_{\boldsymbol{X}}\left\{\left[\left(\theta_0 + \sum_{k=1}^{J}\theta_k X_k\right) - \left(\theta_0 + \theta_j E[X_j|\boldsymbol{X}_{(-j)}] + \sum_{k\neq j}\theta_k X_k\right)\right]^2\right. \\
&= E_{\boldsymbol{X}}\left\{\left[\left(\theta_j X_j - \theta_j E[X_j|\boldsymbol{X}_{(-j)}]\right]^2\right\} \\
&= \theta_j^2 E_{\boldsymbol{X}}\left\{\left(X_j - E[X_j|\boldsymbol{X}_{(-j)}]\right)^2\right\}
\end{aligned}$$

An thus the form of the LOCO estimator the linear additive case reduces to

$$\psi(j) \quad = \quad \frac{\theta_j^2 E\left[\left\{X_j - E(X_j \,|\boldsymbol{X}_{-j})\right\}^2\right]}{\text{var}(Y)}$$

Bibliography

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502.

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

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees.

Ehrenkranz, P., Rosen, S., Boulle, A., Eaton, J. W., Ford, N., Fox, M. P., Grimsrud, A., Rice, B. D., Sikazwe, I., and Holmes, C. B. (2021). The revolving door of hiv care: Revising the service delivery cascade to achieve the unaids 95-95-95 goals. *PLoS medicine*, 18(5):e1003651.

Fentie, D. T., Kassa, G. M., Tiruneh, S. A., and Muche, A. A. (2022). Development and validation of a risk prediction model for lost to follow-up among adults on active antiretroviral therapy in ethiopia: a retrospective follow-up study. *BMC Infectious Diseases*, 22(1):727.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Fletcher Mercaldo, S. and Blume, J. D. (2020). Missing data and prediction: the pattern submodel. *Biostatistics*, 21(2):236–252.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., and Feige, I. (2020). Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*.

Ghalebikesabi, S., Ter-Minassian, L., DiazOrdaz, K., and Holmes, C. C. (2021). On locality of local explanation models. *Advances in neural information processing systems*, 34:18395–18407.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65.

Groenwold, R. H. (2020). Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and prognostic research*, 4(1):8.

Hoogland, J., van Barreveld, M., Debray, T. P., Reitsma, J. B., Verstraelen, T. E., Dijkgraaf, M. G., and Zwinderman, A. H. (2020a). Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in medicine*, 39(25):3591–3607.

Hoogland, J., van Barreveld, M., Debray, T. P., Reitsma, J. B., Verstraelen, T. E., Dijkgraaf, M. G., and Zwinderman, A. H. (2020b). Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in medicine*, 39(25):3591–3607.

Hooker, G., Mentch, L., and Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31:1–16.

Hu, L., Joyce Lin, J.-Y., and Ji, J. (2021). Variable selection with missing data in both covariates and outcomes: Imputation and machine learning. *Statistical Methods in Medical Research*, 30(12):2651–2671.

Janssen, K. J., Vergouwe, Y., Donders, A. R. T., Harrell Jr, F. E., Chen, Q., Grobbee, D. E., and Moons, K. G. (2009). Dealing with missing predictor values when applying clinical prediction models. *Clinical chemistry*, 55(5):994–1001.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230.

Kapelner, A. and Bleich, J. (2015). Prediction with missing data via bayesian additive regression trees. *Canadian Journal of Statistics*, 43(2):224–239.

Kenya (2022). Kenya HIV prevention and treatment guidelines 2022.

Lee, H., Hogan, J. W., Genberg, B. L., Wu, X. K., Musick, B. S., Mwangi, A., and Braitstein, P. (2018). A state transition framework for patient-level modeling of engagement and retention in hiv care using longitudinal cohort data. *Statistics in medicine*, 37(2):302–319.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3):287–296.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons.

Little, R. J. and Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, pages 98–111.

Luijken, K., Groenwold, R. H., Van Calster, B., Steyerberg, E. W., and van Smeden, M. (2019). Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. *Statistics in medicine*, 38(18):3444–3459.

Luijken, K., Wynants, L., Van Smeden, M., Van Calster, B., Steyerberg, E. W., Groenwold, R. H., Timmerman, D., Bourne, T., and Ukaegbu, C. (2020). Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *Journal of clinical epidemiology*, 119:7–18.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Martin, G. P., Sperrin, M., Snell, K. I., Buchan, I., and Riley, R. D. (2021). Clinical prediction models to predict the risk of multiple binary outcomes: a comparison of approaches. *Statistics in medicine*, 40(2):498–517.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2014). *Handbook of missing data methodology*. CRC Press.

Nshimirimana, C., Ndayizeye, A., Smekens, T., and Vuylsteke, B. (2022). Loss to follow-up of patients in hiv care in burundi: A retrospective cohort study. *Tropical Medicine & International Health*, 27(6):574–582.

Oganisian, A., Hogan, J., Sang, E., DeLong, A., Mosong, B., Fraser, H., and Mwangi, A. (2024). Bayesian counterfactual prediction models for hiv care retention with incomplete outcome and covariate information. *arXiv preprint arXiv:2410.22481*.

Ogbechie, M.-D., Fischer Walker, C., Lee, M.-T., Abba Gana, A., Oduola, A., Idemudia, A., Edor, M., Harris, E. L., Stephens, J., Gao, X., et al. (2023). Predicting treatment interruption among people living with hiv in nigeria: Machine learning approach. *JMIR AI*, 2:e44432.

Olatosi, B., Sun, X., Chen, S., Zhang, J., Liang, C., Weissman, S., and Li, X. (2021). Application of machine-learning techniques in classification of hiv medical care status for people living with hiv in south carolina. *Aids*, 35:S19–S28.

Oliwa, T., Furner, B., Schmitt, J., Schneider, J., and Ridgway, J. P. (2021). Development of a predictive model for retention in hiv care using natural language processing of clinical notes. *Journal of the American Medical Informatics Association*, 28(1):104–112.

Olsen, L. H. B., Glad, I. K., Jullum, M., and Aas, K. (2024). A comparative study of methods for estimating model-agnostic Shapley value explanations. *Data Mining and Knowledge Discovery*, pages 1–48.

Pajouheshnia, R., Van Smeden, M., Peelen, L., and Groenwold, R. (2019). How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *Journal of clinical epidemiology*, 105:136–141.

Ramachandran, A., Kumar, A., Koenig, H., De Unanue, A., Sung, C., Walsh, J., Schneider, J., Ghani, R., and Ridgway, J. P. (2020). Predictive analytics for retention in care in an urban hiv clinic. *Scientific reports*, 10(1):6421.

Ridgway, J. P., Ajith, A., Friedman, E. E., Mugavero, M. J., Kitahata, M. M., Crane, H. M., Moore, R. D., Webel, A., Cachay, E. R., Christopoulos, K. A., et al. (2022). Multicenter development and validation of a model for predicting retention in care among people with hiv. *AIDS and Behavior*, 26(10):3279–3288.

Ridgway, J. P., Lee, A., Devlin, S., Kerman, J., and Mayampurath, A. (2021). Machine learning and clinical informatics for improving hiv care continuum outcomes. *Current HIV/AIDS Reports*, 18:229–236.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):77.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1):87–94.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. nat mach intell 1: 206–215. *DOI: https://doi.org/10.1038/s42256-019-0048-x*.

Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classification models.

Semerdjian, J., Lykopoulos, K., Maas, A., Harrell, M., Priest, J., Eitz-Ferrer, P., Zolopa, A., and Wyand, C. (2018). Supervised machine learning to predict hiv outcomes using electronic health record and insurance claims data. *AIDS*.

Shapley, L. S. (1953). A value for n-person games. *Classics in Game Theory*.

Sisk, R., Sperrin, M., Peek, N., van Smeden, M., and Martin, G. P. (2023). Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: a simulation study. *Statistical Methods in Medical Research*, page 09622802231165001.

Sparapani, R., Spanbauer, C., and McCulloch, R. (2021). Nonparametric machine learning and efficient computation with bayesian additive regression trees: the bart r package. *Journal of Statistical Software*, 97:1–66.

Team, R. C. et al. (2021). R: A language and environment for statistical computing. *R foundation for statistical computing, Vienna, Austria*.

Tierney, W. M., Rotich, J. K., Hannan, T. J., Siika, A. M., Biondich, P. G., Mamlin, B. W., Nyandiko, W. M., Kimaiyo, S., Wools-Kaloustian, K., Sidle, J. E., et al. (2007). The ampath medical record system: creating, implementing, and sustaining an electronic medical record system to support hiv/aids care in western kenya. *Studies in health technology and informatics*, 129(Pt 1):372–376.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

Van Ness, M., Bosschieter, T. M., Halpin-Gregorio, R., and Udell, M. (2023). The missing indicator method: From low to high dimensions. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5004–5015.

Vo, T. L., Nguyen, T., Lopez-Ramos, L. M., Hammer, H. L., Riegler, M. A., and Halvorsen, P. (2024). Explainability of machine learning models under missing data. *arXiv preprint arXiv:2407.00411*.

Wallace, M. L., Mentch, L., Wheeler, B. J., Tapia, A. L., Richards, M., Zhou, S., Yi, L., Redline, S., and Buysse, D. J. (2023). Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC medical research methodology*, 23(1):144.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658.

Wood, A. M., Royston, P., and White, I. R. (2015). The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biometrical Journal*, 57(4):614–632.