2024

Explainable and sparse predictive models with applications in reproductive health and oncology

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

EXPLAINABLE AND SPARSE PREDICTIVE MODELS WITH APPLICATIONS IN REPRODUCTIVE HEALTH AND ONCOLOGY

by

ZAHRA ZAD

B.S., University of Tehran, 2013
M.S., Boston University, 2023

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2024

Approved by

First Reader

Ioannis Ch. Paschalidis, PhD
Distinguished Professor of Engineering
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Biomedical Engineering
Professor of Computing & Data Sciences

Second Reader

Lauren A. Wise, ScD
Professor of Epidemiology

Third Reader

Shruthi Mahalingaiah, MD
Adjunct Associate Professor of Obstetrics & Gynecology
Assistant Professor of Epidemiology
Assistant Professor of Pharmacology, Physiology & Biophysics

Fourth Reader

Pirooz Vakili, PhD
Research Associate Professor of Mechanical Engineering
Research Associate Professor of Systems Engineering

Love is such a powerful force.
It's there for everyone to embrace –
that kind of unconditional love for all of humankind.

That is the kind of love that impels people to go into
the community and try to change conditions for others,
to take risks for what they believe in. Coretta Scott King

Acknowledgments

I would like to express my deepest gratitude to the many individuals who have supported me throughout my PhD journey.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Yannis Paschalidis, for his invaluable guidance, unwavering support, and continuous encouragement throughout my PhD journey. His kindness, warmth, and consistently supportive and positive demeanor have been a constant source of inspiration. His extraordinary patience and thoughtful assignment of projects have afforded me countless invaluable learning opportunities. His profound knowledge, innovative ideas, and remarkable intelligence have been instrumental in shaping my research and academic growth.

I would like to extend my sincere appreciation to the members of my dissertation committee, Professor Lauren A. Wise, Professor Shruthi Mahalingaiah, and Professor Pirooz Vakili, for their invaluable feedback, constructive criticism, and steadfast support. Their guidance has greatly enhanced the quality of my research.

I am especially thankful to Professor Pirooz Vakili, who encouraged me to pursue this path five years ago. His belief in my potential gave me the inspiration and courage to embark on this journey. His support and mentorship have been invaluable throughout my PhD.

I am immensely grateful to Taiyao Wang, whose mentorship and generous supervision have been a cornerstone of my development as a researcher. His constant support and insightful feedback have significantly enriched my work.

My heartfelt thanks go to Yeping Jin for his generous assistance in the methodological aspects of my research. His expertise and willingness to help have been crucial in overcoming numerous challenges along the way.

I am also deeply thankful to my colleagues and friends, Luciana Mayumi Gutiyama, Julia Lima Fleck, Boran Hao, Shahab Sotoudian, Yang Hu, Victoria Jiang, Jennifer Jean Yland, and Majid Heidarifar, for their camaraderie, encouragement, and assistance. Their support has made this journey a collaborative and enriching experience.

Lastly, but by no means least, I would like to express my deepest appreciation to my husband, Nasser, whose unwavering support, understanding, and love have been my greatest source of strength and motivation. He recognized my potential and gave me the courage to pursue this path. His constant encouragement and support at every step have been invaluable, always pushing me forward. His belief in me has been a driving force behind my success and achievements. I am also profoundly grateful to my family for their endless love, encouragement, and support, which have been the foundation of all my accomplishments.

Thank you all for making this journey possible.

Zahra Zad
PhD
Division of Systems Engineering

EXPLAINABLE AND SPARSE PREDICTIVE MODELS WITH APPLICATIONS IN REPRODUCTIVE HEALTH AND ONCOLOGY

ZAHRA ZAD

Boston University, College of Engineering, 2024

Major Professor: Ioannis Ch. Paschalidis, PhD
Distinguished Professor of Engineering
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Biomedical Engineering
Professor of Computing & Data Sciences

ABSTRACT

This dissertation develops explainable and sparse predictive models applied to two main healthcare applications: reproductive health and oncology. Through the application of advanced machine learning techniques and survival analysis, we aim to enhance predictive accuracy and provide actionable insights in these critical areas. The thesis is structured into four distinct problems, each focusing on a particular research question.

The first problem concerns the prediction of the probability of conception among couples actively trying to conceive. Using self-reported health data from a North American preconception cohort study, we analyzed factors such as sociodemographics, lifestyle, medical history, diet quality, and specific male partner characteristics. Machine learning algorithms were employed to predict the probability of conception demonstrating improved discrimination and potential clinical utility.

The second problem explores the application of machine learning algorithms to electronic health record (EHR) data for identifying predictor variables associated with polycystic ovarian syndrome (PCOS) diagnosis. Employing gradient boosted trees and feed-forward multilayer perceptron classifiers, we developed a scoring system that improved the model's performance, providing a valuable tool for early detection and intervention.

The third problem focuses on predicting the risk of miscarriage among female participants who conceived during the study period. Utilizing both static and survival analysis, including Cox proportional hazard models, we developed predictive models to assess miscarriage risk. The study revealed that most miscarriages were due to random genetic errors during early pregnancy, indicating that miscarriage is not easily predicted based on preconception sociodemographic and lifestyle characteristics.

Finally, the fourth problem focuses on the development of predictive models for managing Chronic Myeloid Leukemia (CML) patients. We developed models to predict whether patients will achieve deep molecular response (DMR) at later treatment stages and maintaining this status up to 60 months post-treatment initiation. These models offer insights into treatment effectiveness and patient management, aiming to support clinical decision-making and improve long-term patient outcomes.

By emphasizing the explainability of these models, this dissertation not only aims to provide accurate predictions but also to ensure that the results are interpretable and actionable for healthcare professionals. Overall, this thesis showcases the potential of predictive modeling to improve reproductive health and oncology-related outcomes. The development and validation of various models in these contexts underscore the value of machine learning algorithms in healthcare research, analysis of epidemiologic data, and prediction of critical health events. The findings have significant implications for enhancing patient care, informing clinical practices, and guiding healthcare policy decisions.

Keywords: predictive modeling, machine learning, survival analysis, artificial intelligence in healthcare research, reproductive health, Chronic Myeloid Leukemia, Polycystic Ovarian Syndrome, conception, miscarriage, electronic health record.

Contents

1 Introduction 1

1.1 Machine Learning in Healthcare 1

1.2 Survival Analysis in Healthcare 2

1.3 Motivation 3

1.4 Contributions of the Thesis 4

1.5 Overview of the Dissertation Structure 6

1.6 Bibliographic Notes 6

2 Methods 8

2.1 Machine Learning in Healthcare 9

2.1.1 Classification methods 9

2.1.2 Data pre-processing 10

2.1.3 Model training-testing 11

2.1.4 Performance metrics 13

2.1.5 Feature Selection Techniques to Drive Sparse models 15

2.1.6 Our MLP score 21

2.2 Survival Analysis in Healthcare 23

2.2.1 Survival Analysis Definition and Importance 23

2.2.2 Survival Function Definition and Formulation 23

2.2.3 Cox Proportional Hazard Models 24

2.2.4 Original Cox Models 26

2.2.5 Research Gap 28

2.2.6 Distributionally Robust Cox Models 29

3 Predictive Models of Pregnancy Based on Data from a Preconception Cohort Study 34

3.1 Introduction 34

3.2 Methods 35

3.2.1 Study population 35

3.2.2 Data collection 36

3.2.3 Outcomes 36

3.2.4 Predictive Models 39

3.2.5 Sensitivity analysis 42

3.3 Results 42

3.3.1 Numerical Experiments for Comparison of Original Cox Model and DRO Cox Models 53

3.4 Discussion 54

3.5 Conclusions 58

4 Predicting Polycystic Ovary Syndrome (PCOS) with Machine Learning Algorithms from Electronic Health Records 65

4.1 Introduction 65

4.2 Materials and Methods 66

4.2.1 Data acquisition 66

4.2.2 Ethical approval 67

4.2.3 Reference label definitions 67

4.2.4 Predictive models 71

4.2.5 Development of the MLP score 76

4.2.6 Recursive feature elimination (RFE) 77

4.2.7 Final predictive models 77

4.3 Results 78
4.3.1 Results of data acquisition and data pre-processing 78
4.4 Model Performance 78
4.5 Discussion 90

5 Predictive Models of Miscarriage Based on Data From a Preconception Cohort Study 100
5.1 Introduction 100
5.2 Materials and Methods 101
5.2.1 Study population 101
5.2.2 Data collection 102
5.2.3 Assessment of miscarriage 103
5.2.4 Statistical analysis 107
5.2.5 Sensitivity analysis 107
5.3 Results 108
5.3.1 Survival models 108
5.3.2 Static models 112
5.3.3 Numerical Experiments for Comparison of Original Cox Model and DRO Cox Models 114
5.4 Discussion 115
5.5 Conclusions 118

6 Predictive Models of Deep Molecular Response to Imatinib Treatment in Chronic Myeloid Leukemia patients 125
6.1 Introduction 125
6.2 Materials and methods 126
6.2.1 Data description 126
6.2.2 Predictive Models 133

6.3 Results 142

6.4 Discussion 142

7 Conclusions 150

7.1 Summary 150

7.2 Future work 151

References 153

Curriculum Vitae 171

List of Tables

3.1 Complete list of variables included in analysis. Part 1. 37

3.2 Complete list of variables included in analysis. Part 2. 38

3.3 Distribution of class and number of menstrual cycles of attempt time at study entry. 44

3.4 Distribution of each class by the number of menstrual cycles of attempt time at study entry. 44

3.5 Performance metrics for the parsimonious models, PRESTO 2013-2019. 45

3.6 Performance metrics for the full and sparse Models I-II, PRESTO 2013-2019. 47

3.7 Variables selected by the parsimonious Model I (predicting pregnancy in 12 cycles) using the L2LR algorithm, PRESTO 2013-2019, n=3,195 participants. 49

3.8 Variables selected by the parsimonious Model II (predicting pregnancy within 6 cycles) using the L2LR algorithm, PRESTO 2013-2019, n=3,476 participants. 50

3.14 Comparison of concordance indices for different ratios of outliers with $\epsilon = 0.05$ in a subset of the pregnancy dataset 54

3.9 Variables selected by the parsimonious Model III (fecundability), PRESTO 2013-2019, n=4,133 participants. 60

3.10 Performance metrics for the full and sparse Models IV-V, restricting to nulligravid women with no history of infertility, PRESTO 2013-2019. 61

3.11 Variables selected by the parsimonious Model IV (restricted to nulligravid women with no history of infertility) predicting pregnancy within 12 cycles, using the L2LR algorithm, PRESTO 2013-2019, n=1,571 participants. 62

3.12 Variables selected by the parsimonious Model V (restricted to nulligravid women with no history of infertility) predicting pregnancy within 6 cycles, using the L2LR algorithm, PRESTO 2013-2019, n=1,722 participants. 63

3.13 Variables selected by the parsimonious Model VI (restricted to nulligravid women with no history of infertility) predicting pregnancy within 6 cycles, using the L2LR algorithm, PRESTO 2013-2019, n=1,957 participants. 64

4.1 Exclusion ICD-9 Coding Associations. 68

4.2 Definitions for Predictive PCOS models generated from the Boston Medical Center Clinical Data Warehouse. 72

4.3 Summary of missingness in continuous variables. 73

4.4 Variables before statistical feature selection (SFS) for each model. Part 1. 74

4.5 Variables before statistical feature selection (SFS) for each model. Part 2. 75

4.6 Highly correlated variables and the retained variable. 76

4.7 Demographic characteristics of the study population and by model. . 82

4.8 Most significant variables for PCOS diagnosis prediction in Model I. . 83

4.9 Most significant variables for PCOS diagnosis prediction in Model II. 83

4.10 Most significant variables for PCOS diagnosis prediction in Model III. 83

4.11 Most significant variables for PCOS diagnosis prediction in Model IV. 85

4.12 Features for each model, associated with LR-L1 algorithm after SFS. 86

4.13 Features for each model, associated with LR-L1 algorithm after SFS. 87

4.14 Model performance over the test set, in the format of mean percentage (SD percentage) over 5 repetitions. 88

4.15 Performance of all models and all algorithms, before and after SFS. The means and standard deviations of AUC and weighted-F1 scores on the test set over the five repetitions are listed in the format of mean percentage (SD percentage). 89

5.1 Complete list of variables included in analysis to generate predictive models of miscarriage in PRESTO, 2013-2022. Part 1. 104

5.2 Complete list of variables included in analysis to generate predictive models of miscarriage in PRESTO, 2013-2022. Part 2. 105

5.3 Missing data among predictor variables in PRESTO. 106

5.4 Distribution of gestational age at miscarriage in PRESTO, 2013-2022. 109

5.5 Variables selected by the full survival model predicting miscarriage in PRESTO, 2013-2022. 110

5.6 Variables selected by sparse survival model to predict miscarriage in PRESTO, 2013-2022. 111

5.7 Variables selected by the sparse survival model predicting miscarriage after restricting to ≥8 gestational weeks in PRESTO, 2013-2022. . . . 112

5.8 Variables selected by the sparse survival model predicting miscarriage among primigravid participants in PRESTO, 2013-2022. 113

5.9 Variables selected by the sparse survival model predicting miscarriage after restricting to ≥8 gestational weeks among primigravid participants in PRESTO, 2013-2022. 113

5.10 Variables selected by sparse static model (L2LR) 120

5.11 Performance metrics for the static models predicting miscarriage in PRESTO, 2013-2022. 121

5.12 Variables selected by the sparse static model (logistic regression with an ℓ_2-norm regularization term) predicting miscarriage after restricting to ≥8 gestational weeks in PRESTO, 2013-2022. 122

5.13 Variables selected by the sparse static model (logistic regression with an ℓ_2-norm regularization term) predicting miscarriage among primigravid participants in PRESTO, 2013-2022. 123

5.14 Variables selected by the sparse static model (logistic regression with an ℓ_2-norm regularization term) predicting miscarriage after restricting to ≥8 gestational weeks among primigravid participants in PRESTO, 2013-2022. 124

5.15 Comparison of concordance indices for different ratios of outliers with $\epsilon = 0.005$ in a subset of the miscarriage dataset. 124

6.1 Description of features included in the Original Dataset. Part 1. 129

6.2 Description of features included in the Original Dataset. Part 2. 130

6.3 Description of features included in the new dataset (combination of Original Dataset and Additional Dataset) 130

6.4 The percentage of missing values and the count of non-missing values in the Original Dataset with 144 patients. 134

6.5 The percentage of missing values and the count of non-missing values in the new dataset (combination of Original Dataset and Additional Dataset) 135

6.6 Cohort statistics of the new dataset (combination of Original Dataset and Additional Dataset). 137

6.7 Cohort statistics of Original Dataset. Part 1. 140

6.8 Cohort statistics of Original Dataset. Part 2. 141

6.9 The most important predictors selected by feature selection (SFS and RFE) in parsimonious format of all models. We list the LR coefficients of each variable (**Coef**), the correlation of the variable with the outcome (**Y_corr**), the mean of the variable (**Y1_mean**) in the patients achieving DMR at the time corresponding to each model, and the mean of the variable (**Y0_mean**) in the remaining (non-DMR) patients. . 143

6.10 All models' evaluation metrics on the test set. 144

6.11 Probabilities calculated by our models alongside different categories of patients from Original Dataset. 146

List of Figures

2·1 MLP architecture. (Lenail, 2022) 22

3·1 Area under the precision-recall curves (AUPRC) for Models I, II, IV, and V. Note: Model I predicts pregnancy in <12 menstrual cycles; Model II predicts pregnancy in <7 menstrual cycles; Model IV predicts pregnancy in <12 menstrual cycles among nulligravid women with no history of infertility; Model V predicts pregnancy in <7 menstrual cycles among nulligravid women with no history of infertility. 46

3·2 Visualization of the Model I coefficients and their 95% confidence intervals. Note: Figures 3·2, 3·3 and 3·4 present the model coefficients with error bands equivalent to 95% confidence intervals. The variables are ordered according to their mean coefficient. All plots are associated with the parsimonious L2LR version of models, consistent with the rest of the Manuscript. Model I predicts pregnancy in $<$ 12 menstrual cycles. Model II predicts pregnancy in $>$ 7 menstrual cycles. Model III predicts the probability of pregnancy within each menstrual cycle for up to 12 cycles of follow-up. 51

3·3 Visualization of the Model II coefficients and their 95% confidence intervals. 52

3·4 Visualization of the Model III coefficients and their 95% confidence intervals. 52

4·1 Flow of patients from the BMC CDW into the dataset used by the study. 81

4·2 Feature importance graphs based on logistic regression coefficients (± 95% confidence interval), associated with parsimonious models utilizing the MLP score (LR-L2-MLP score). The absolute value of the logistic regression coefficients shows how much the variable affects the predicted probability of the outcome. A positive/negative coefficient implies that the larger the absolute value of the variable, the higher/lower the chance of having a PCOS diagnosis as defined by the model outcome. 84

4·3 Example of receiver operator characteristic (ROC) curves associated with parsimonious logistic regression models utilizing the MLP score (LR-L2-MLP score). 91

4·4 Example of receiver operator characteristic (ROC) curves associated with parsimonious logistic regression models utilizing the MLP score (LR-L2-MLP score). 92

6·1 The exclusion criteria of total CML patients, the number of patients in each model, and the number/percentage of patients who achieve DMR (responders, represented as [R]) and do not achieve DMR (non-responders, represented as [NR]) at the time of each model's outcome. 127

6·2 The baseline characteristics of patients.
[1] The respondent and non-respondent definition is based on Model_long_12 where respondents are the patients who reach DMR at 24 months after the start of the treatment and maintain this status until 60 months after treatment initiation.
[2] MR_0_3 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the first three-month interval after start of IM.
[3] MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from 3 to 6 months after the start of IM. 128

6·3 An overview of the steps in model development. 132

6·4 The feature importance in L2LR parsimonious models based on Logistic Regression coefficients +- 95% CI. The median of the patient's recorded BCR-ABL1 ratios, white blood cell count at diagnosis, and the achievement of Major Molecular Response (MMR, defined as $\geq$MR3.0) are among the most important predictive variables. MR_0_3 indicates achieving MMR in the 3-month interval from time 0 to 3 months after the start of IM. 145

6·5 ROC curves associated with the L2LR parsimonious version of each model. The Area Under the ROC Curve (AUC), is used to evaluate prediction performance. A perfect predictor has an AUC of 1 and a predictor which makes random guesses has an AUC of 0.5. 145

6·6 Test set performance of the parsimonious models using LR, SVM, RF, MLP, and LightGBM. 146

6·7 The feature importance in L2LR parsimonious models based on Logistic Regression coefficients +- 95% CI. MR_3_6 indicates achieving Major Molecular Response (MMR, defined as $\geq$MR3.0) in the 3-month interval from time 3 to 6 months after the start of IM. 147

6·8 ROC curves associated with the parsimonious version of each model. The Area Under the ROC Curve (AUC) is used to evaluate prediction performance. A perfect predictor has an AUC of 1 and a predictor which makes random guesses has an AUC of 0. 148

6·9 The feature importance in parsimonious models based on Random Forest feature importance. MR_0_3 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from time 0 to 3 months after the start of IM. MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from time 3 to 6 months after the start of IM. 149

List of Abbreviations

AUC	Area Under the Curve
BMC	Boston Medical Center
BMI	Body mass index
CML	Chronic myeloid leukemia
DMR	Deep molecular response
E2	Estradiol
EHR	Electronic health records
ELN	European Leukemia Net
FSH	Follicle stimulating hormone
GBM	Gradient Boosting Machine
GBT	Gradient Boosted trees
HA	Hyperandrogenism
ICD	International Classification of Diseases
IM	Irregular menses
INCA	Brazilian National Cancer Institute
LH	Luteinizing hormone
LR	Logistic regression
MLP	Feed Forward Multilayer Perceptron Neural Network
MR	Molecular response
PCOM	Polycystic ovarian morphology
PCOS	Polycystic ovary syndrome
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristics
SFS	Statistical feature selection
SHBG	Sex hormone binding globulin
SVM	Support vector machines
TFR	Treatment-free remission

Chapter 1

Introduction

In healthcare, predictive models are increasingly utilized to enhance patient outcomes, reduce costs, and optimize resource utilization. By analyzing patterns and trends in patient data, predictive models assist healthcare professionals in making precise and timely diagnoses, identifying patients at risk for specific health issues, and providing early interventions and preventive care. In the long run, this can lead to better patient outcomes and decreased healthcare expenses.

1.1 Machine Learning in Healthcare

Machine learning (ML) significantly benefits healthcare by enhancing diagnostic accuracy, improving treatment plans, and boosting patient outcomes. By analyzing extensive medical data, ML algorithms can identify patterns that may be missed by human practitioners, leading to earlier and more accurate diagnoses and personalized treatments (Jiang et al., 2017) (Esteva et al., 2017). ML excels in predictive analytics, helping healthcare providers anticipate disease outbreaks, predict patient deterioration, and manage chronic diseases, which allows for timely interventions that save lives and reduce costs (Shickel et al., 2017) (Choi et al., 2016) (Amini et al., 2021) (Amini et al., 2023) (Amini et al., 2024). Personalized medicine leverages ML's capability to analyze genetic, lifestyle, and patient-specific data, resulting in more effective and personalized treatments with fewer side effects (Obermeyer and Emanuel, 2016) (Topol, 2019). In addition, machine learning techniques like federated learning enable collaborative training of models across multiple healthcare institutions without sharing patient data, thus enhancing privacy and security while improving diagnostic accuracy and personalized treatment plans (Talaei and Izadi, 2024a) (Talaei and Izadi, 2024c) (Talaei and Izadi, 2024b). Additionally, ML streamlines healthcare operations by optimizing scheduling, managing supply chains, and enhancing resource allocation, thus reducing waste and increasing efficiency (Rudin and Radin, 2019) (Rajkomar et al., 2018) (Talaei et al., 2024). In drug discovery and development, ML identifies potential drug candidates, predicts their efficacy, and optimizes clinical trial designs, significantly reducing the time and cost required to bring new drugs to market (Chen et al., 2018) (Vamathevan et al., 2019) (Hashemi et al., 2023). ML-powered wearable devices and health apps facilitate continuous real-time monitoring of patients' vital signs, enabling early detection of anomalies and proactive interventions, thereby preventing hospitalizations and improving the quality of life for patients with chronic conditions (Ravì et al., 2016). Overall, ML's capabilities demonstrate its broad and significant impact on advancing healthcare practices and outcomes.

1.2 Survival Analysis in Healthcare

Survival analysis is crucial for understanding time-to-event data in healthcare, where the timing of events such as disease recurrence or patient death is essential. Traditional regression models often fall short in these scenarios, but survival analysis excels by effectively handling censored data, where the event has not occurred for some subjects within the study period. This ensures that all available data is utilized less bias (Klein et al., 2003) (Collett, 2023).

A key strength of survival analysis is its ability to estimate hazard functions, offering insights into the risk of events occurring at different times. This is particularly valuable for identifying high-risk periods and tailoring medical interventions accordingly (Cox, 1972) (Kalbfleisch and Prentice, 2011). Additionally, survival analysis can incorporate time-dependent covariates, allowing for more dynamic and accurate modeling of scenarios where influencing factors change over time, such as patient conditions and treatment regimens (Therneau, 1997). The Cox regression model effectively conditions time out of the model, making it more parsimonious than other regression models that estimate rate ratios, such as Poisson regression. It allows the baseline hazard to vary over time, provided that the proportional hazards assumption (i.e., constant hazard ratio) holds throughout the study period. A significant advantage of using survival analysis is its ability to account for varying lengths of follow-up, precisely capture the timing of events, and censor participants with unobservable person-time after the date of last contact. This ensures that rates are less likely to be underestimated (Cox, 1972) (Klein et al., 2003).

Beyond understanding past events, survival analysis enhances predictive modeling, helping forecast future occurrences and improving preventive measures and resource allocation in healthcare. Its applications extend beyond healthcare to fields like engineering, economics, and social sciences, demonstrating its versatility and broad relevance (Bradburn et al., 2003) (Lee and Wang, 2003) (Kleinbaum and Klein, 1996).

1.3 Motivation

Reproductive health and oncology were chosen as focal points due to their profound impact on individual and public health. Infertility and miscarriage are prevalent issues affecting millions of couples, while cancers like chronic myeloid leukemia pose significant treatment challenges. Improving predictive capabilities in these areas can lead to substantial improvements in patient care and outcomes.

Despite the advancements in predictive modeling, challenges such as data quality, integration of diverse data sources, and the need for explainable models remain. This dissertation addresses these challenges by employing robust machine learning algorithms and advanced statistical methods to create models that are both accurate and interpretable.

1.4 Contributions of the Thesis

This dissertation aims to develop and validate predictive models for various health outcomes, with a particular focus on reproductive health and oncology. Through the application of advanced machine learning techniques and survival analysis, we seek to enhance early diagnosis, improve treatment plans, and ultimately contribute to better patient outcomes.

In Chapter 3, we applied machine learning algorithms to develop predictive models of pregnancy using three distinct, clinically relevant definitions: infertility, subfertility, and fecundability. Infertility affects 10 to 15% of couples in North America and up to 12% of reproductive-aged women and 9.4% of men aged 25-44 years in the US use fertility treatments, costing more than $5 billion annually (Chandra et al., 2013) (Macaluso et al., 2010). By developing predictive tools for couples attempting to conceive, we aim to provide essential information for clinical practice and minimize expenses. Accurate predictive models can help women who are anxious about their fertility status make informed decisions about postponing pregnancy or addressing other modifiable factors. Our models were based on comprehensive datasets and included features such as demographic, lifestyle, and environmental factors, providing a holistic approach to predicting pregnancy outcomes.

In Chapter 4, we aimed to determine predictor variables associated with polycystic ovarian syndrome (PCOS) diagnosis by applying machine learning algorithms to electronic health record (EHR) data. PCOS is the most common cause of anovulatory infertility in women of reproductive age, with more than 90% of anovulatory women seeking infertility treatment having PCOS (Azziz et al., 2009). Along with infertility, PCOS also increases the risk of endometrial hyperplasia and endometrial cancer and has been linked to the development of metabolic syndrome, diabetes, cerebrovascular disease, and hypertension compared to women without the condition (Barry et al., 2014) (Lim et al., 2019) (Anagnostis et al., 2018) (Wekker et al., 2020). Despite the serious health implications, PCOS often goes undiagnosed due to varying symptom severity on presentation, leading to delayed treatment and potentially severe clinical consequences (Barry et al., 2014). Predictive models have the potential to aid in the earlier diagnosis of PCOS and can be used to guide early detection and interventions for PCOS.

In Chapter 5, we employed various machine learning methods and Cox proportional hazard models to predict miscarriage based on self-reported preconception data. Approximately 20% of recognized pregnancies end in miscarriage, defined as pregnancy loss before 20 weeks of gestation (Rossen et al., 2018). While earlier studies have created predictive models for pregnancy loss utilizing early pregnancy characteristics, such as laboratory values and ultrasound measurements (Huang et al., 2022) (DeVilbiss et al., 2020) (Li et al., 2022), our study took a different approach by using prospectively collected data on lifestyle, environmental, and medical factors during the preconception period to develop predictive models for miscarriage. Our models identified significant predictors such as female age, history of miscarriage, and male partner age, demonstrating the importance of these factors in predicting miscarriage.

In Chapter 6, we developed predictive models to achieve deep molecular response (DMR) in chronic myeloid leukemia (CML) patients treated with imatinib. CML comprises approximately 15% of all leukemia cases, and it is estimated that one person in every 526 in the U.S. will suffer from CML during their lifetime (American Cancer Society, 2022). With the discovery of BCR-ABL1, CML became one of the major success stories in cancer history (Mughal et al., 2016). The current management approach for CML focuses on achieving a stable DMR and treatment-free remission (TFR) through an individualized therapy plan based on efficacy, tolerance, toxicity, and cost (Schiffer, 2019) (Bonifacio et al., 2019). Predictive models have the potential to improve the management of CML by identifying patients more likely to achieve DMR and those who do not, thus informing physician decisions to recommend TKI discontinuation or earlier indication of hematopoietic stem cell transplantation. Our models leveraged comprehensive clinical data and BCR-ABL1/ABL1IS quantification to predict the likelihood of DMR achievement, providing a robust tool for optimizing CML treatment strategies.

1.5 Overview of the Dissertation Structure

This dissertation is structured as follows: Chapter 1 introduces the scope and significance of the study. Chapter 2 discusses the methodologies used, including survival analysis and machine learning techniques. Chapters 3 to 6 present case studies on predictive models for pregnancy, miscarriage, PCOS, and CML. Finally, Chapter 7 includes the conclusion, summarizing the key findings and discussing future work.

1.6 Bibliographic Notes

Large parts of the thesis appear in published or working research papers: (Zad et al., 2022; Zad et al., 2024; Yland et al., 2022; Yland et al., 2024).

Notational conventions: All vectors are column vectors. For economy of space, we write $\mathbf{x} = (x_1, \ldots, x_{\dim(\mathbf{x})})$ to denote the column vector $\mathbf{x}$, where $\dim(\mathbf{x})$ is the dimension of $\mathbf{x}$. In case that we have y to represent the actual label, $\hat{y}$ represents the predicted value of y. Unless otherwise specified, $\|\cdot\|$ denotes the ℓ_2 norm, $\|\cdot\|_1$ the ℓ_1 norm and $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{\dim(\mathbf{x})} |x_i|^p\right)^{1/p}$ the ℓ_p norm, where $p \geq 1$. $\|\cdot\|_0$ denotes the ℓ_0 counting norm. We use ∇ to denote the gradient operator. We use $\mathbb{E}$ and P to denote operators of expectation and probability, respectively. We use $\mathbb{P}$ to denote probability distribution. The notation $\mathbb{E}_{\mathbb{P}}$ denotes the expectation with respect to the probability distribution $\mathbb{P}$. The symbol exp denotes the exponential function, which is a mathematical function denoted by e^x, where e is the base of the natural logarithm. The notation $\mathbb{N}$ represents the set of natural numbers. The notation $\mathbb{R}$ represents the set of all real numbers. The notation $\mathbb{R}^d$ represents the d-dimensional Euclidean space.

Chapter 2

Methods

We use supervised methods to generate predictive models. We utilize both machine learning and statistical methods for this sake. Machine learning methods typically predict the risk or probability of an event of interest without explicitly considering the time until the event occurs. In contrast, survival analysis methods are designed to predict the time until an event of interest, often through the hazard or rate of the event occurring. For machine learning models, we used a variety of supervised classification methods including linear and non-linear algorithms. For survival analysis models, we fit penalized Cox proportional hazards models. For both the machine learning and survival analysis approach, we generated full and sparse models. The full models contain all variables selected after statistical feature selection (SFS) whereas the sparse models contain all variables after both statistical feature selection and univariate feature selection for survival analysis models or recursive feature elimination (RFE) for machine learning models. We evaluated model performance via the Area Under the Curve of a Receiver Operating Characteristic (AUC-ROC), precision and recall metrics, and the weighted-F1 score for machine learning models, and via the concordance index for survival analysis models. These methods are described in more detail below.

2.1 Machine Learning in Healthcare

2.1.1 Classification methods

We explored a variety of supervised machine learning classification methods both linear and non-linear algorithms. Linear classifiers included logistic regression (LR) and linear support vector machines (SVM) (Cortes and Vapnik, 1995), which lead to interpretable predictions. For instance, the LR coefficient of a feature represents the sensitivity of the predicted likelihood to that feature and the absolute value of this coefficient can be interpreted as feature importance. Linear classifiers were fitted with an additional regularization term seeking to prevent the influence of outliers in training or test data (Chen et al., 2020). Regularization prevents overfitting by adding a penalty term to loss function in model training, and promotes simpler and more generalized models. We discuss more on regularization in the Section 2.1.5.1. Non-linear methods such as tree-based learning algorithms included Gradient Boosted Trees (GBM), an ensemble tree-based model that uses a gradient boosting framework (Ke et al., 2017), and Random Forest (RF), a large collection of decision trees which classifies by averaging the decisions of trees (Breiman, 2001). We utilized the feed forward Multilayer Perceptron neural network (MLP) as not only a classification method in our study, but also as a mehtod to improve the performance of the linear models. We discuss more on MLP models in the Section 2.1.6. Non-linear methods are more complex and generally yield better classification performance. These algorithms were chosen because of their extensive usage and their performance superiority demonstrated in the literature (Brisimi et al., 2018) (Hao et al., 2020) (Wang et al., 2020).

We define machine learning algorithms we used in more detail below:

- Logistic Regression (LR): a linear model that predicts the probability of a binary outcome based on input variables by fitting a logistic function to the data.

- Support Vector Machines (SVM): a linear model that aims to find an optimal hyperplane in a high-dimensional space to separate different classes of data by maximizing the margin between the classes.

- Gradient Boosted Decision Trees (GBDT): a non-linear model that is an ensemble learning method that combines multiple decision trees sequentially, training each new tree to correct the mistakes made by the previous trees using gradient information from a chosen loss function, resulting in an accurate and powerful predictive model.

- Random Forest (RF): a non-linear model that is an ensemble learning method that constructs multiple decision trees and combines their predictions through majority voting or averaging, providing robust and accurate predictions by reducing overfitting and capturing complex relationships in the data.

- Perceptron (MLP): a non-linear model that is a feedforward artificial neural network that consists of multiple layers of interconnected nodes, including an input layer, one or more hidden layers, and an output layer, with each node applying a non-linear activation function to produce predictions or classifications based on the input data.

2.1.2 Data pre-processing

We perform several data pre-processing steps:

- First, we convert each categorical variable into a set of indicator variables.

- Second, we handle missing data as follows: For categorical variables with missing data, we set the missing data as the reference category, and for continuous variables, we replace missing values with the median value of available data.

In some cases, based on suggestions of healthcare experts, we fill the missing values with the mode of the variable. We elaborate on this step in each project.

- Third, variables with very low variability (standard deviation < 0.0001) were assessed for removal from the models.
- Fourth, we address potential collinearity issues as follows: for each pair of highly correlated variables (e.g., correlation coefficient > 0.8), we removed the variable that had a lower correlation with the outcome.
- Fifth, we performed statistical feature selection (described in Section 2.1.5.2).
- Last, we standardized each variable by subtracting its mean and dividing by its standard deviation to have a zero mean and unit variance.

By following these steps, the dataset is ready for model training and testing.

2.1.3 Model training-testing

We randomly split the dataset into n equal parts, where n-1 parts are used as the training set, and one part as the test set. We use the training set to tune the model hyperparameters via k-fold cross-validation, explained in more detail in Section 2.1.3.1. We evaluate the performance metrics on the test set. We repeat training and testing t times, each time with a different random split between the training and test sets. The mean and standard deviation of all metrics on the test sets over the t repetitions are reported.

We evaluated model performance using AUC and weighted-F1 score (defined in Section 2.1.4). AUC is more easily interpretable, while the weighted-F1 score is more robust to imbalanced data than AUC (Saito and Rehmsmeier, 2015). We also calculated weighted-precision (i.e., positive predictive value) and weighted-recall (i.e., sensitivity) metrics as follows: we calculated precision and recall among participants with and without the event of interest, and calculated the average scores across groups, weighted by the number individuals in each class.

2.1.3.1 Tuning of hyperparameters

For tuning models parameters, we used 5-fold cross validation. First, we split the training dataset (80% of the full dataset) into five equal parts, or folds. Second, we train the model using four parts as training data. Third, we validate the model on the fifth part. We repeat these three steps for each of the five folds, each time obtaining different values for the model parameters. Finally, we select the values for the model parameters that leads to the model with the best validation performance.

In Logistic Regression (LR) and Support Vector Machine (SVM) models, we consider the inverse of regularization strength as a hyper parameter. We search for the best hyper parameter for example among [0.001, 0.01, 0.1, 1, 10, 100] and choose the one that leads to the best classifier (with the highest AUC). In the artificial neural network (MLP) models, we have one input layer, a number of hidden layers, and one output layer. We tune the number of hidden layers and the number of neurons in the hidden layers. We try different options, for example: (i) one hidden layer with 32, 64, 128, 256, or 512 neurons, (ii) two hidden layers with 16, 32, 64, 128, 256 neurons in the first hidden layer and 2 neurons in the second hidden layer, (iii) two hidden layers with 8, 16, 32, 64, or 128 neurons in the first hidden layer and 4 neurons in the second hidden layer. In the Gradient Boosting Machine (GBM) models, we used LightGBM which uses a leaf-wise tree growth algorithm, which converges faster than the depth-wise growth used by many other tools but can lead to over-fitting if not properly configured. Key parameters to tune for optimal results include: `num_leaves`, which controls tree complexity and should generally be set lower than $2^{(\mathrm{max\_depth})}$ to avoid over-fitting; `min_data_in_leaf`, which prevents over-fitting by ensuring each leaf has enough data points, with hundreds or thousands being suitable for large datasets; and `max_depth`, which explicitly limits tree depth to control complexity (LightGBM-Guide, 2024).

2.1.4 Performance metrics

We define key performance metrics below:

- Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It measures the accuracy of the positive predictions made by the model. Mathematically, it is defined as Equation (2.1):

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}. \tag{2.1}$$

- Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to all observations in the actual class. It measures the model's ability to identify all relevant instances. Mathematically, it is defined as Equation (2.2):

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{2.2}$$

- The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a performance measurement for classification problems at various threshold settings. The ROC is created by plotting the true positive rate (i.e., Recall) against the false positive rate (equal to one minus specificity) at various thresholds. The c-statistic, or the Area Under the ROC Curve (AUC), is used to evaluate prediction performance. It tells how much the model is capable of distinguishing between classes. A perfect predictor has an AUC of 1 and a predictor which makes random guesses has an AUC of 0.5. For example, an AUC of 0.70 implies that, on average, there is a 70% probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This means that if you were to take one positive case and one negative case, 70% of the time, the model will assign a higher probability to the positive case. In other words, the model is likely to correctly identify 70% of the positive cases across different thresholds. Correspondingly, the model will also misclassify some negative instances as positive, reflected by the 30% of cases where it fails to rank a positive instance higher than a negative one. For practical decision-making, we would still need to choose an appropriate threshold to convert these probability estimates into class labels (positive or negative). This threshold choice will depend on the specific context and the relative costs of false positives and false negatives.

- The AUPRC is the area under the curve of precision and recall. The weighted score is the average of the score of each class weighted by the number of participants in each class.

- The F1 score is the harmonic mean of recall and precision. It is particularly useful when the class distribution is imbalanced. The F1 Score is defined as Equation (2.3):

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.3}$$

We calculated a weighted-F1 score to account for imbalance in the proportion of participants with respect to the event of interest. The weighted F1-score is the average of the F1-scores of each class weighted by the number of participants in each class. The weighted-F1 score is between 0 to 1, and a higher value represents a more robust model.

2.1.5 Feature Selection Techniques to Drive Sparse models

Sparsity was motivated by the earlier works (Brisimi et al., 2018) (Brisimi et al., 2019) (Chen and Paschalidis, 2022) (Chen et al., 2019), where it was shown that sparse classifiers can perform almost as well as very sophisticated classification methods. Sparse models are designed to promote sparsity by encouraging most feature weights or coefficients to be close to zero. Sparse models identify a small subset of features that contribute significantly to the prediction or decision-making process. Sparsity enhances computational efficiency, helps with more interpretability, and improves generalization performance of the models. The desirability of sparsity in statistical estimators is particularly pronounced in high-dimensional environments, such as those encountered in biological applications, where interpretability holds paramount importance.

2.1.5.1 Regularization

Regularization is a technique used in machine learning to address overfitting, where a model becomes too specialized to the training data and performs poorly on new data. Two commonly used regularization techniques are ℓ_1 and ℓ_2 regularization. ℓ_1 regularization, also known as lasso regression (Tibshirani, 1996), adds a penalty to the model's loss function proportional to the absolute value of the weights. This penalty encourages some weights to become zero, effectively reducing the number of features used by the model. ℓ_1 regularization is particularly useful when dealing with large numbers of features and the goal is to select only the most relevant ones. Equation (2.4) represents the loss function used in logistic regression with ℓ_1 regularization in which n is the number of samples, y_i represents the actual label of the i-th sample, while $\hat{y}_i$ represents the predicted probability. The absolute value expression $|\beta_j|$ denotes the ℓ_1 norm of the regression coefficient β_j associated with each feature. Also to control the strength of regularization, the regularization parameter λ is introduced.

$$\text{Loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{2.4}$$

On the other hand, ℓ_2 regularization, also known as ridge regression, adds a penalty proportional to the squared value of the weights. Unlike ℓ_1 regularization, ℓ_2 regularization does not push any weights to become exactly zero. Instead, it encourages all weights to be smaller, which can help prevent overfitting and improve model accuracy. Equation (2.5) represents the loss function used in logistic regression with ℓ_2 regularization.

$$\text{Loss} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) + \lambda\sum_{j=1}^{p}\beta_j^2. \tag{2.5}$$

The choice between ℓ_1 and ℓ_2 regularization depends on the specific problem at hand. ℓ_1 regularization is often suitable when dealing with a large number of features and feature selection is important. In contrast, ℓ_2 regularization is beneficial when the number of features is small, and accurate predictions are prioritized over feature reduction. It has been shown that regularization is equivalent to deriving a “robust” model, that is, a model that is robust to the presence of outliers in the training data set.

The book by Chen and Paschalidis (2020, pp. 67-73) offers a comprehensive theoretical basis for understanding how the ℓ_2-regularizer prevents overfitting to training data. This regularization method can be viewed as a means of controlling the level of ambiguity present in the data, thereby shedding light on the reliability of contaminated samples (Chen and Paschalidis, 2018) (Chen et al., 2019) (Chen et al., 2020).

2.1.5.2 Statistical Feature Selection (SFS)

Statistical feature selection (SFS) is a variable selection process. We tested the association between each variable and the outcome and removed variables that were not independently associated with the outcome based on p-value greater than a threshold (e.g. 0.01, 0.05, ...). We used the chi-squared test (Cochran, 1952) for binary predictors and the Kolmogorov-Smirnov test for continuous predictors (Massey Jr, 1951).

2.1.5.3 Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a greedy heuristic algorithm used to solve the combinatorial subset selection problem in feature selection. The objective of RFE is to maximize the performance of the machine learning model while minimizing the number of features. The specific objective function depends on the chosen performance metric. Let's assume we use AUC as our metric of interest. Inputs of this optimization problem are a set of features $F = \{f_1, f_2, ..., f_n\}$, a dataset X consisting of input samples and corresponding output labels, and a model $\mathcal{M}$ that uses a subset of features to make predictions. Constraints to this optimization problem is a condition to terminate the elimination process. This stopping criterion can be defined based on a desired performance threshold or the number of features remaining k. Finally, the output of this optimization problem is $S = \{s_1, s_2, ..., s_m\}$ a subset of features that minimizes the loss function and maximizes the performance of the machine learning model. The RFE problem can be represented in Algorithm (1).

$$\begin{aligned}
&\text{minimize} \quad \text{Loss}(\mathcal{M}(S)) \\
&\text{subject to} \\
&\quad |S| \leq k \quad \text{(Subset Size Constraint)}, \\
&\quad \text{Stopping Criterion:} \\
&\quad \text{e.g., Performance Threshold or Number of Features Remaining.}
\end{aligned}$$

Algorithm 1 Recursive Feature Elimination (RFE)

```
**Input**:
Let $F = \{f_1, f_2, ..., f_n\}$ represent the set of features.
Let $X$ be the input dataset, where each sample $x_i$ is associated with a corresponding output label $y_i$.
Let $\mathcal{M}$ be the machine learning model trained on a subset of features.
Let $k$ be the desired number of features.
**Output**:
$S = \{s_1, s_2, ..., s_m\}$ represent the subset of features selected by RFE.

**procedure** RFE($F$, $X$, $\mathcal{M}$, $k$)
    Step 1: Start with all features $F = \{f_1, f_2, ..., f_n\}$.
    Step 2: Train machine learning model $\mathcal{M}$ and Rank features according to their importance.
    Step 3: Remove the feature with the least importance and obtain $S = \{s_1, s_2, ..., s_m\}$ the subset of features selected by RFE.
    Step 4: Repeat steps 2 and 3 until $k$ features are left or a desired performance is met.
    **return** $S = \{s_1, s_2, ..., s_m\}$
**end procedure**
```

In this formulation, $Loss(\mathcal{M}(S))$ represents the the loss function of the machine learning model $\mathcal{M}$ trained on the subset of features S. The objective is to find the subset S that minimizes the loss function and maximizes the performance while satisfying the subset size constraint and the stopping criterion.

We used Recursive Feature Elimination (RFE) in conjunction with L1-penalized logistic regression (L1LR). We explored different combinations of a regularization parameter and the number of features to select. More specifically, by running L1LR we obtained weights associated with the variables (i.e., the coefficients of the model). We eliminated the variable with the smallest absolute weight and performed L1LR to obtain a new model. We kept iterating in this fashion, eliminating one variable at each iteration, to select a model that maximizes a modified AUC value calculated by subtracting the standard deviation from the mean AUC. This modified AUC value helps assess the stability of the feature selections, providing insights into the reliability of the chosen subsets.

RFE helps to identify the most relevant features in a dataset, which can be useful for reducing dimensionality, improving model performance, and interpreting the underlying relationships in the data (Pedregosa et al., 2011).

2.1.5.4 Comparing RFE technique with a Mixed Integer Linear Programming approach

In Chapter 4, we tried another approach for feature selection using Outer-approximation algorithm represented in (Bertsimas and Dunn, 2019). In the chapter of sparse and robust classification, the authors mentioned that a natural way to induce sparsity is to add a constraint on the number of nonzero coefficients of β and solve Equation (2.6).

$$\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} l\left(y_i, \beta^T x_i\right) + \frac{1}{2\gamma} \|\beta\|_2^2 \quad \text{s.t.} \quad \|\beta\|_0 \leq k. \tag{2.6}$$

The zero norm $\|\beta\|_0$ counts the number of non-zero elements in β. Equation (2.6) is expressed as a convex binary optimization problem in the following theorem.

Theorem *Problem (2.6) is equivalent to*

$$\min_{s\in S_k^p} c(s), \tag{2.7}$$

with $S_k^p = \{s \in \{0,1\}^p,\ e^T s \leq k\}$ and for any $s \in \{0,1\}^p$,

$$c(s) \triangleq \max_{\alpha\in\mathbb{R}^n} f(\alpha, s) \triangleq -\sum_{i=1}^{n} \hat{\ell}(y_i, \alpha_i) - \frac{\gamma}{2}\sum_{j=1}^{n} s_j \alpha^T x_j x_j^T \alpha \quad \text{s.t.} \quad e^T\alpha = 0.$$

In particular, $c(s)$ is convex for $s \in [0,1]^p$. The authors find a solution to Equation (2.6) by iteratively constructing a piece-wise linear lower approximation of c. The solver structure is represented as Outer-approximation algorithm in pseudocode in Algorithm (2) .

Algorithm 2 Outer-approximation algorithm

```
Input: $X \in \mathbb{R}^{n*p}, y \in \{-1,1\}^p, k \in [p]$ and an initial solution $s_1$
Output: An optimal solution $s^*$ to the convex binary optimization problem of Equation (2.6).
$\eta_1 \leftarrow 0$
$t \leftarrow 1$
repeat
    $s_{t+1}, \eta_{t+1} \leftarrow \begin{cases} \arg\min_{s,\eta}\ \eta \\ \quad \text{s.t. } \eta \geq c(s_i) + \nabla c(s_i)^T(s - s_i), \quad \forall i \in [t] \\ \qquad s \in S_k^p \end{cases}$
    $t \leftarrow t + 1$
until $\eta_t \leq c(s_t)$
return $s_t$
```

We utilized the paper associated with the book (Bertsimas et al., 2021) and its github repository in Julia, as well as API from InterpretableAI (IAI)[1], a company associated with MIT that represents a couple of APIs implementing some algorithms including the algorithm we are interested in. We used the method "OptimalFeature-Selection" form the class "iai" via the package "interpretableai".

The models performance using the IAI method compared to the RFE method is not considerably improved. In more detail, with first model we have AUC of 0.7866 with IAI and 0.7842 with RFE; with second model we have AUC of 0.6737 with IAI and 0.6815 with RFE; with third model we have AUC of 0.7424 with IAI and 0.7420

[1] https://www.interpretable.ai/

with RFE. Therefore, considering the expenses of implementing their approach and the fact that, in our specific problem, results with IAI are not better than results with RFE, there is no reason for us to prefer IAI method over the RFE method for feature subset selection.

2.1.6 Our MLP score

In Chapter 4, there was a considerable difference between the AUC of linear models and non-linear models. To improve the performance of our linear models, we utilized nonlinear models, Gradient Boosted Trees (GBT) and multilayer perceptron (MLP), to capture intricate relationships between features.

The multilayer perceptron (MLP) is an Artificial Neural Network (ANN) architecture that has gained popularity for its effectiveness in classification and regression tasks. Composed of interconnected layers of nodes, the MLP applies a linear transformation to the inputs and then applies a non-linear activation function to introduce non-linearity into the network. The input layer of the MLP represents the features of the data, and the output layer provides the predicted output or class probabilities based on the task at hand. The hidden layers in between capture hierarchical representations of the data, allowing the network to learn complex patterns and relationships (Pedregosa et al., 2011).

In our specific work, we aim to utilize an MLP architecture with three hidden layers, each employing the rectified linear unit (ReLU) activation function (Figure 2·1. ReLU is a commonly used activation function that helps address the vanishing gradient problem and allows the network to learn more efficiently. By specifying three hidden layers and using ReLU activation, we seek to enhance the model's capacity to capture intricate relationships and improve its performance in extracting meaningful features from the input features x_1, x_2, x_3, and x_4.

Equation (2.8) shows the mathematical formula associated with our model in

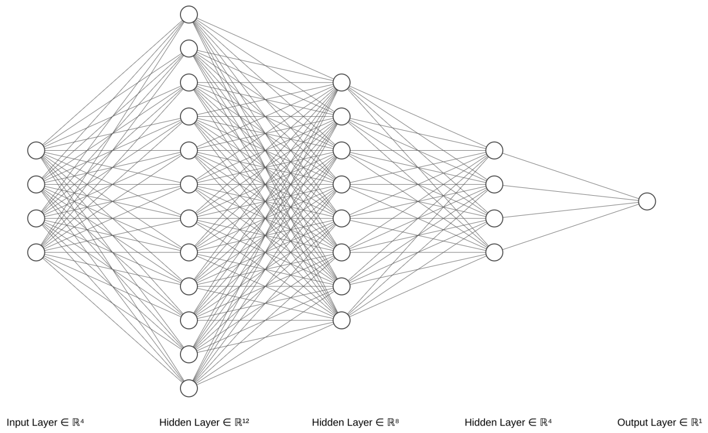

Input Layer ∈ $\mathbb{R}^4$
Hidden Layer ∈ $\mathbb{R}^{12}$
Hidden Layer ∈ $\mathbb{R}^8$
Hidden Layer ∈ $\mathbb{R}^4$
Output Layer ∈ $\mathbb{R}^1$

Figure 2·1: MLP architecture. (Lenail, 2022)

which f is our MLP architecture with three layers, w and b are the trainable parameter of the model, and x are the input features. Also, m, n, p are the number of neurons in each layer, respectively y is the output probability, and the ReLU (Equation (2.9)) is used in our analysis (Pedregosa et al., 2011).

$$y = f\left(\sum_{i=1}^{m} \mathrm{ReLU}\left(\sum_{j=1}^{n} \mathrm{ReLU}\left(\sum_{k=1}^{p} \mathrm{ReLU}\left(\sum_{l=1}^{q} w_l^{(1)} x_l + b_k^{(1)}\right) w_k^{(2)} + b_j^{(2)}\right) w_j^{(3)} + b_i^{(3)}\right)\right). \tag{2.8}$$

$$\mathrm{ReLU}(x) = \max(0, x). \tag{2.9}$$

In our case, we had four features (FSH, LH, SHBG, and estradiol levels) as input features into the MLP model, and we utilized y, the output probability of the MLP model, as a new feature into our predictive models. We only used the training dataset to train this new composite feature.

2.2 Survival Analysis in Healthcare

For survival models, we evaluated performance with the concordance index (defined in Section 2.2.3.3). To develop and evaluate the survival models, we first split the dataset into five random parts of equal size: four parts constituted the training dataset, and the fifth part constituted the testing dataset. We fit the model on the training dataset and evaluated its Concordance Index on the testing dataset. We repeated these calculations (split the data into five random parts, fit the model using the training dataset, evaluate Concordance Index in the testing dataset) five times. Finally, we calculated the mean and standard deviation of the Concordance Index across these five runs.

2.2.1 Survival Analysis Definition and Importance

Survival analysis is a branch of statistics that focuses on the time until an event of interest occurs, such as death or failure. It is essential in healthcare research because it can handle censored data, which occurs when the event has not happened for some subjects during the study. Techniques like the Kaplan-Meier estimator and the Cox proportional hazards model are used to estimate survival probabilities, compare survival between groups, and assess the effects of covariates on survival time. This aids in identifying risk factors and evaluating treatment efficacy, ultimately improving patient care and outcomes (Kleinbaum and Klein, 1996) (Kalbfleisch and Prentice, 2011).

2.2.2 Survival Function Definition and Formulation

The survival function, denoted as $S(t)$, represents the probability that a subject survives beyond time t. Formally, it is defined as Equation (2.10).

$$S(t) = P(T > t). \tag{2.10}$$

Where T is the time to the event of interest. The survival function starts at 1 when $t = 0$ and approaches 0 as t increases. It is crucial for methods like the Kaplan-Meier estimator and the Cox proportional hazards model, which relate survival to covariates. Understanding the survival function helps derive other key quantities, such as the hazard function, facilitating comprehensive analysis of time-to-event data (Kleinbaum and Klein, 1996) (Kalbfleisch and Prentice, 2011).

2.2.3 Cox Proportional Hazard Models

2.2.3.1 Definition

The Cox proportional hazards model, introduced by David Cox in 1972, is a regression model that examines the relationship between survival time and covariates. The Cox proportional hazards model evaluates the effect of covariates on the hazard rate without specifying the baseline hazard function. The Cox regression model conditions the baseline hazard out of the model, allowing it to vary over time. This makes the model significantly more flexible than the traditional Poisson regression approach, where both the rate and rate ratio are assumed to be constant throughout the observation period (Cox, 1972) (Kleinbaum and Klein, 1996).

2.2.3.2 Hazard Function Definition and Formulation

The hazard function, $h(t)$, indicates the instantaneous risk of an event at time t. It is defined as Equation (2.11).

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.11}$$

In the Cox model, it is formulated as Equation (2.12).

$$h(t|x) = h_0(t)\exp(f(x;\beta)). \tag{2.12}$$

where $h(t|x)$ is the hazard function at time t, $h_0(t)$ is the baseline hazard, and $f(x;\beta)$ is the log partial hazard function. The log partial hazard function represents the combined effect of the covariates on the hazard function and is the logarithm of the hazard function portion excluding the baseline hazard. This approach simplifies the estimation process. The model's semi-parametric nature and ability to handle censored data make it crucial for estimating hazard ratios and understanding covariate effects. This formulation separates the baseline hazard from covariate effects, providing flexibility in survival analysis. It is crucial to verify whether the proportional hazards assumption has been violated in Cox regression models. This can be assessed by creating interaction terms between time and exposure, or by plotting Kaplan-Meier curves and log-log survival curves. Additionally, Schoenfeld residuals can be plotted to test this assumption. If the hazard ratio is not constant over time, it is advisable to report the rate ratio results stratified by the time intervals used in the model (e.g., rate ratio for the first 5 years of the study versus the rate ratio for the last 5 years of the study) (Cox, 1972) (Kleinbaum and Klein, 1996).

2.2.3.3 Evaluation Metrics

Evaluation metrics in the context of Cox proportional hazards models include measures such as concordance index (C-index), which assesses the model's ability to correctly rank the survival times based on the predicted risk scores. The concordance index is the fraction or percent of the pair of observations which are concordant and shows a goodness-of-fit statistic for survival analysis. The concordance index is a generalization of the AUC that accounts for event time and loss to follow-up (Longato et al., 2020)(Schmid et al., 2016). Like the AUC, a value of 0.5 indicates that discrimination is no better than random, while a value of 1 would indicate perfect prediction.

2.2.3.4 Univariate feature selection

Univariate feature selection is a variable selection process, applied after statistical feature selection for all survival models. Univariate feature selection evaluates each feature independently based on its relationship with the target. We fit individual Cox proportional hazards models for each variable, such that each model contained only one independent variable, and we recorded the concordance index for each model. We ranked variables based on the associated concordance index and we selected top variables with higher concordance index, as many as we are interested in (in our case: top 10 variables).

2.2.4 Original Cox Models

2.2.4.1 Original Cox Models Log Partial Hazard Function Formulation

In the original Cox proportional hazards model, the log partial hazard function, denoted as $f(x;\beta)$, represents the linear combination of covariates. It is defined as Equation (2.13).

$$f(x;\beta) = \beta^T x. \tag{2.13}$$

Where x is the feature vector and β is the corresponding coefficients. This function quantifies the effect of covariates on the hazard rate, providing a multiplicative effect on the baseline hazard function (Cox, 1972) (Kleinbaum and Klein, 1996).

2.2.4.2 Original Cox Models Hazard Function Formulation

The hazard function in the original Cox proportional hazards model, $h(t|x)$, represents the instantaneous risk of an event occurring at time t given covariates x. It is formulated as Equation (2.14).

$$h(t|x) = h_0(t)\exp(\beta^T x). \tag{2.14}$$

Where $h_0(t)$ is the baseline hazard function, and $\exp(\beta^T x)$ is the exponential of the log partial hazard function. This formulation separates the baseline hazard from the covariate effects, allowing for a flexible and robust analysis of survival data (Cox, 1972) (Kleinbaum and Klein, 1996).

2.2.4.3 Original Cox Models Hazard Ratio

The hazard ratio in the original Cox proportional hazards model is a measure of the effect of covariates on the hazard rate. It is defined as the ratio of the hazard functions for two individuals with different covariate values. For covariates x_i and x_j, the hazard ratio is given by Equation (2.15).

$$\mathrm{HR} = \frac{h(t|x_i)}{h(t|x_j)} = \exp(f(x_i;\beta) - f(x_j;\beta)). \tag{2.15}$$

This ratio indicates the relative risk of the event occurring for the two individuals, with values greater than 1 suggesting higher risk for the individual with covariate values x_i compared to x_j (Cox, 1972) (Kleinbaum and Klein, 1996).

2.2.4.4 Original Cox Models Negative Log Partial Likelihood

The negative log partial likelihood is used to estimate the coefficients β in the original Cox proportional hazards model. The partial likelihood function for n observations is given by Equation (2.16).

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(f(x_i;\beta))}{\sum_{j\in R(t_i)} \exp(f(x_j;\beta))} \right). \tag{2.16}$$

Where $R(t_i)$ is the risk set at time t_i. The negative log partial likelihood is then equation (2.17).

$$-\log L(\beta) = -\sum_{i=1}^{n}\left(f(x_i;\beta) - \log \sum_{j\in R(t_i)} \exp(f(x_j;\beta))\right). \tag{2.17}$$

Minimizing this quantity provides the maximum partial likelihood estimates of the coefficients. This optimization problem can be formulated as Equation (2.18).

$$\min_{\beta}\left\{-\sum_{i=1}^{n}\left(f(x_i;\beta) - \log \sum_{j\in R(t_i)} \exp(f(x_j;\beta))\right)\right\}. \tag{2.18}$$

Solving this optimization problem allows for the assessment of the impact of covariates on survival time (Cox, 1972) (Kleinbaum and Klein, 1996).

2.2.5 Research Gap

The Cox proportional hazards model is a cornerstone in survival analysis, widely used for its ability to relate covariates to the hazard function. Despite its popularity, the model's assumptions about the fixed distribution of covariates and parameters often do not hold true in real-world data. This can lead to significant inaccuracies in survival predictions, especially in the presence of outliers, measurement errors, and dynamic changes in covariate effects over time. To address these issues and enhance the robustness of the Cox model, we propose developing a robust version of the Cox proportional hazards model using Distributionally Robust Optimization (DRO). This approach aims to mitigate the impact of outliers and data perturbations, ensuring more accurate and reliable survival predictions in the presence of real-world data complexities.(Chen et al., 2020)

2.2.6 Distributionally Robust Cox Models

2.2.6.1 Motivation and Advantages of Distributionally Robust Optimization (DRO)

Distributionally Robust Optimization (DRO) is an advanced optimization framework designed to address uncertainties in the distribution of data. Unlike traditional optimization methods that rely on specific assumptions about the data distribution, DRO provides a more flexible and resilient approach by optimizing performance over a set of possible distributions. This makes DRO particularly valuable in scenarios where data is subject to perturbations, outliers, and other real-world complexities.

The key advantages of DRO include robustness to distributional uncertainty, improved handling of outliers and perturbations, flexibility in model assumptions, and enhanced predictive accuracy. DRO accounts for the uncertainty in the distribution of data, providing solutions that are less sensitive to variations and inaccuracies in the assumed data distribution. By considering a range of possible distributions, DRO can mitigate the impact of outliers and data perturbations that can disproportionately affect traditional optimization methods. Moreover, DRO does not require strict assumptions about the exact form of the data distribution, making it applicable to a wider range of real-world problems where such assumptions are often violated. By optimizing over an ambiguity set of distributions, DRO can lead to more reliable and accurate predictions, especially in complex and uncertain environments (Chen et al., 2020). These advantages make DRO a powerful tool for improving the robustness and reliability of predictive models, including the Cox proportional hazards model in survival analysis.

2.2.6.2 Definition and Formulation of Distributionally Robust Optimization (DRO)

Distributionally Robust Optimization (DRO) is an advanced optimization framework that addresses uncertainties in the data distribution by minimizing a worst-case expected loss function over a probabilistic ambiguity set. This ambiguity set is constructed from observed samples and characterized by known properties of the true data-generating distribution. DRO provides robustness against data perturbations, outliers, and modeling inaccuracies, ensuring that the optimization solution remains resilient under various plausible distributions.

The general DRO problem can be formulated as Equation (2.19).

$$\min_{\beta \in \mathbb{R}^d} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}} \left[L(x, \beta) \right]. \tag{2.19}$$

Where β represents the decision variables (model parameters), $\mathbb{R}^d$ is the parameter space, $\mathbb{P}$ is a probability distribution within the ambiguity set $\mathcal{P}$, and $L(x, \beta)$ is the loss function. The ambiguity set $\mathcal{P}$ captures the uncertainty in the distribution of data.

The ambiguity set $\mathcal{P}$ can be defined using various approaches, such as moment constraints, statistical distance measures, or confidence sets derived from historical data. By solving the DRO problem, we obtain a decision β that minimizes the worst-case expected loss, thus providing a robust solution that performs well across a range of plausible distributions (Chen et al., 2020).

2.2.6.3 Definition of Wasserstein Distance

Wasserstein distance, also known as Earth Mover's Distance (EMD), is a measure of the distance between two probability distributions. It quantifies the minimum cost of transforming one distribution into another, considering the ground distance between points. Given two probability distributions P and Q on a metric space $(\mathcal{X}, d)$, the Wasserstein distance of order p (where $p \geq 1$) is defined as Equation (2.20).

$$W_p(P, Q) = \left(\inf_{\gamma \in \Gamma(P,Q)} \int\limits_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\gamma(x, y) \right)^{\frac{1}{p}} . \tag{2.20}$$

Where $\Gamma(P, Q)$ denotes the set of all joint distributions γ on $\mathcal{X} \times \mathcal{X}$ with marginals P and Q. The function $d(x, y)$ represents the ground distance between points x and y.

In the context of DRO, the Wasserstein Distance is used to define the ambiguity set $\mathcal{P}$ as a Wasserstein ball around the empirical distribution $\hat{P}_N$. The ambiguity set can be defined as Equation (2.21).

$$\mathcal{P} = \left\{ Q \in \mathcal{P}(\mathcal{X}) : W_p(Q, \hat{P}_N) \leq \epsilon \right\} . \tag{2.21}$$

Where $\mathcal{P}(\mathcal{X})$ is the space of all probability distributions on $\mathcal{X}$, $\hat{P}_N$ is the empirical distribution based on observed samples, and ϵ is a pre-specified radius.

By incorporating the Wasserstein Distance, DRO effectively handles the distributional uncertainty and ensures robust optimization solutions that account for possible variations in the data distribution (Chen et al., 2020).

2.2.6.4 Challenges

Applying Distributionally Robust Optimization (DRO) to Cox models presents several challenges. The DRO framework requires the loss function to depend solely on individual data points. However, the current individual loss function used in Cox models does not meet this requirement, complicating the direct application of DRO. Additionally, while the loss function $L(.)$ is convex with respect to the covariates x and the model parameters β, it is not convex with respect to the event times y.

This lack of convexity with respect to y poses significant difficulties in ensuring robustness in the presence of distributional uncertainty, as the DRO approach relies on robustifying against the worst-case distribution within the defined ambiguity set. Addressing these challenges is crucial for effectively integrating DRO with Cox models and leveraging its robustness to handle real-world data complexities.

2.2.6.5 Proposed Methodologies

The proposed methodologies include DRO-Cox Sample Splitting and DRO-Cox Global Fixation. The DRO-Cox Sample Splitting method, as outlined by (Hu and Chen, 2022), is designed to enhance fairness in the model. However, it does not achieve superior accuracy under noise compared to the original Cox model. On the other hand, the DRO-Cox Global Fixation method, introduced in (Jin and Paschalidis, 2024), treats all duration data as fixed constants and redefines the individual loss function as Equation (2.22). The data point $i \in [N]$ has feature vectors $x_i \in X$, observed duration time $y_i \geq 0$, and event indicator $\delta_i \in \{0, 1\}$. If $\delta_i = 1$, then the event of interest occurs after duration $t_i = y_i$.

$$\ell_i(x_i, y_i, \delta_i, \beta) = \delta_i \left[\log \left(e^{\beta^T x_i} + \sum_{j: y_j \geq y_i} e^{\beta^T x_j} \right) - \beta^T x_i \right]. \tag{2.22}$$

This redefined loss function aims to improve the robustness of the model under distributional uncertainties by incorporating fixed duration data into the optimization process (Jin and Paschalidis, 2024).

2.2.6.6 DRO-Cox Global Fixation

The N data points $\{(x, y, \delta)_i\}_{i \in [N]}$ form an empirical distribution, which serves as the center of the ambiguity set Ω_ϵ.

The optimization problem is formulated as Equation (2.23).

$$\min_{\beta\in\mathbb{R}^d} \sup_{\mathbb{P}\in\Omega_\epsilon} \mathbb{E}_{\mathbb{P}}\left[\delta\left(\log\left(e^{\beta^T x} + \sum_{j:y_j\geq y} e^{\beta^T x_j}\right) - \beta^T x\right)\right]. \tag{2.23}$$

This optimization problem is inherently complex and cannot be solved directly. Therefore, it is essential to find a tractable form to facilitate estimation (Jin and Paschalidis, 2024).

Theorem (Global Fixation DRO-Cox.) (Jin and Paschalidis, 2024). Suppose the data are sorted in decreasing order with regard to duration y, and the Wasserstein distance is induced by l_p-norm. Let (p, q) be Hölder conjugates, so that $\frac{1}{p} + \frac{1}{q} = 1$. Then the following program provides an upper bound for Equation (2.24).

$$\begin{array}{ll} \min\limits_{\beta,\alpha} & \|\beta, \alpha\|_q \epsilon + \frac{1}{N}\sum_{i=1}^N \delta_i s_i, \\ \text{s.t.} & s_i \geq \log\left(e^{\beta^T x_i} + \sum_{j=1}^k e^{\beta^T x_j}\right) - \beta^T x_i - \alpha(y_i - y_k), \quad \forall 1 \leq i \leq k \leq N \end{array}. \tag{2.24}$$

Chapter 3

Predictive Models of Pregnancy Based on Data from a Preconception Cohort Study

3.1 Introduction

Predictive models could guide clinical care and help couples make informed decisions regarding childbearing. Previous research has identified many individual risk factors for infertility and predictors of fecundability (i.e., the per-cycle probability of conception). Female age and body mass index (BMI), as well as male BMI, have been identified as risk factors for infertility (Best and Bhattacharya, 2015) (Homan et al., 2007) (Sundaram et al., 2017) (Wesselink et al., 2017). In addition, female preconception exposures including alcohol consumption (Fan et al., 2017); sleep quality (Willis et al., 2019); cigarette smoking (Wesselink et al., 2019); use of certain hormonal contraceptives (Yland et al., 2020); dietary factors (Gaskins and Chavarro, 2018); depressive symptoms (Evans-Hoeker et al., 2018) (Nillni et al., 2016); stress (Akhter et al., 2016) (Louis et al., 2011) (Lynch et al., 2014) (Wesselink et al., 2018); and environmental exposures such as air pollution (Conforti et al., 2018) and endocrine disrupting chemicals (Kahn et al., 2021) are associated with reduced fecundability. Other male risk factors include exposure to environmental chemicals (Buck Louis et al., 2016) (Snijder et al., 2012), cigarette smoking (Soares and Melo, 2008), and short sleep duration (Wise et al., 2018). However, few studies have moved beyond individual risk factors to develop predictive models of pregnancy probability, and the predictive power of these models was modest (Collins et al., 1995) (Coppus et al., 2009) (Eimers et al., 1994) (Hunault et al., 2004) (Hunault et al., 2005) (Snick et al., 1997) (Van der Steeg et al., 2007).

In this study, we used supervised machine learning methods to predict the cumulative probability of pregnancy over 6 and 12 menstrual cycles and to predict fecundability (per-cycle probability of conception) in an incident cohort study of pregnancy planners. We considered 163 potential predictors and applied several classification algorithms and variable selection procedures to identify the most accurate models and to evaluate the relative predictive strength of individual risk factors.

3.2 Methods

3.2.1 Study population

Pregnancy Study Online (PRESTO) is a web-based preconception cohort study that examines the extent to which environmental and behavioral factors such as diet, exercise, and medication use influence fertility and pregnancy outcomes (Wise et al., 2015). The study began in 2013 and is ongoing. Eligible female participants are aged 21-45 years, residing in the U.S. or Canada, trying to conceive, and not using fertility treatments. We excluded participants with more than one menstrual cycle of pregnancy attempt time at enrollment because these women may have changed their behaviors in response to difficulties conceiving (Wise et al., 2020). We analyzed data from couples who had not yet tried to conceive and those who had tried for one cycle at study entry together. This is consistent with a report by Joffe et al., which indicated that grouping couples with reports of “zero” and “one” cycle of pregnancy attempt time does not induce bias (Joffe et al., 2005). This study included data from 4,133 participants enrolled during 2013 through 2019.

3.2.2 Data collection

Female participants completed a baseline questionnaire at enrollment, on which they reported data on sociodemographic factors, behavioral factors, medical and reproductive history, and selected male partner characteristics. Ten days after enrollment, participants were invited to complete the diet history questionnaire (DHQ II). The DHQ II was designed by the National Cancer Institute and the first version of the DHQ was validated against 24-hour dietary recalls in a U.S. population (Millen et al., 2006) (Subar et al., 2001). In validation studies, correlations between energy-adjusted, DHQ-reported food servings and 24-hour recall-reported food servings ranged from 0.43 for other starchy vegetables to 0.84 for milk. Based on dietary factors reported via the DHQ II, we assessed overall diet quality using the Healthy Eating Index-2010 (HEI-2010) score (Guenther et al., 2013). Participants completed bimonthly follow-up questionnaires for 12 months, or until reported pregnancy, cessation of pregnancy attempts, study withdrawal, or loss to follow-up, whichever occurred first. Data on menstrual cycle dates, pregnancy attempts, and pregnancy status were obtained via the baseline questionnaire and updated on each follow-up questionnaire. A complete list of the 163 variables included in this analysis is provided in Table 3.1 and 3.2.

3.2.3 Outcomes

We developed three models to predict 1) pregnancy in fewer than 12 menstrual cycles; 2) pregnancy within six menstrual cycles; and 3) the average probability of pregnancy per menstrual cycle. We chose these outcome measures to reflect clinically relevant definitions of infertility, subfertility, and fecundability (Evers, 2002) (Gnoth et al., 2005). For the first two models, we used a dataset with one observation per participant and excluded participants who were lost to follow-up before reaching a study endpoint (for the first model, N = 3,195; for the second model, N=3,476). For the third model

Table 3.1: Complete list of variables included in analysis. Part 1.

Category	**Variables Included in Preliminary Analysis**
Demographic and socioeconomic characteristics	Age, marital status, race, We conceptualized race as a social construct that serves as a rough proxy for exposure to interpersonal and structural racism. ethnicity, region of residence, urbanization of residential area, year at study entry, highest level of education, parents' education level, household income, employment status, hours/week of work, shift work, night shift frequency in the past month.
Lifestyle, behavioral, and wellness factors	Cigarette smoking (if so, number per day); total duration of smoking; history of smoking during pregnancy; use of e-cigarettes (if so, ml/day); frequency of marijuana use; exposure to second-hand smoke; alcohol intake; caffeine consumption; moderate physical activity; vigorous physical activity; sedentary activity; sleep duration; trouble sleeping; perceived stress scale score; major depression inventory score.
Dietary factors and use of supplements	Healthy Eating Index-2010 score; supplemental intake of vitamins A, B1, B2, B3, B4, B5, B6, B7, B12, C, D, E, K; beta-carotene; folic acid; iron; zinc; calcium; magnesium; selenium; omega-3 fatty acids; consumption of whole milk, 2% milk, 1% milk, skim milk, soy milk, other milk, fruit juice, bottled water, tap water, sugar-sweetened soda, diet soda, sugar-sweetened energy drinks, diet energy drinks; use of multivitamins or folic acid supplements.
Early life exposures and family history	Adopted; number of siblings; multiple gestation; born preterm; born with low birthweight; breastfed; delivered via cesarean section; mother's cigarette smoking during pregnancy; mother's age at participant's birth; mother's history of pregnancy complications, miscarriage.

Table 3.2: Complete list of variables included in analysis. Part 2.

Category	**Variables Included in Preliminary Analysis**
Reproductive characteristics and disorders	Age at menarche; menstrual regularity; menstrual period characteristics (typical length, Menstrual cycle length and regularity were assessed via the following questions on the baseline questionnaire: 1) Did your period become regular on its own without the use of hormonal contraceptives like the pill, patch, implants, or injectables (regular in a way so you can usually predict about when the next period will start)? 2) Within the past couple of years, has your menstrual period been regular? Please think about those times you were not using hormonal contraceptives. 3) Thinking about the time(s) when you have not used hormonal contraceptives, what is your typical menstrual cycle length? That is, the number of days from the first day of one menstrual period to the first day of your next menstrual period. number of flow days, flow amount, pain); received human papillomavirus vaccine; abnormal pap smear; ever diagnosed with a thyroid condition, fibroids, polycystic ovarian syndrome, endometriosis, a urinary tract infection, pelvic inflammatory disease, chlamydia, herpes, vaginosis, genital warts; recent use of medications for polycystic ovarian syndrome; gravidity; parity; history of cesarean section; years since last pregnancy; history of unplanned pregnancy; history of subfertility or infertility; history of infertility treatment; history of breastfeeding; number of lifetime sexual partners; doing something to improve pregnancy chances; intercourse frequency; using a fertility app; last method of contraception.
Physical characteristics, non-reproductive medical history, and medication use	Body mass index; waist measure; Ferriman-Gallwey Hirsutism Score; handedness; number of primary care visits last year; high blood pressure; received influenza vaccine last year; ever diagnosed with migraines (if so, recent migraine frequency), asthma, hay fever, depression, anxiety, gastroesophageal reflux disease, diabetes; use of the following medications in the 4 weeks before baseline: pain medications, antibiotics, asthma medications, diabetes medications; use of psychotropic medications.
Environmental exposures (occupational and personal care product use)	Exposed regularly to agricultural pesticides; metal particulates or fumes; solvents, oil-based paints, or cleaning compounds; high temperature environments; chemotherapeutic drugs; engine exhaust; chemicals for hair dyeing, straightening, or curing; chemicals for manicure/pedicure. Use of chemical hair relaxer.
Male partner characteristics	Age, body mass index, education, cigarette smoking (if so, number per day), circumcision status.

(fecundability), we included all participants under observation regardless of follow-up duration (N=4,133).

3.2.4 Predictive Models

3.2.4.1 Pre-processing and statistical feature selection

We performed data pre-processing steps, explauned in more detail in Section 2.1.2 to prepare the dataset for developing predictive models. The threshold for correlation coefficient is considered 0.8. Furthermore, statistical feature selection (SFS), as explained in more detail in Section 2.1.5.2, is done with 0.05 threshold for p-value.

3.2.4.2 Classification methods

For Models I (pregnancy in fewer than 12 menstrual cycles) and II (pregnancy within six menstrual cycles), we explored four supervised classification methods, LR, SVM, MLP, and GBM, to develop predictive models for pregnancy (Hastie et al., 2009) (Jiang et al., 2020), explained in more detail in Section 2.1.1. We considered both an L1-norm (L1LR, L1SVM) and an L2-norm regularizer (L2LR, L2SVM) (Lee et al., 2006) to address overfitting. The former is appropriate if we believe that few variables are predictive of the outcome (sparse model), whereas the latter is appropriate in cases where a dense model is more appropriate. Explained in more detail in Section 2.1.5.1.

We present results for full, sparse, and parsimonious models. The full models (i.e., least parsimonious) contain all variables selected after statistical feature selection (eliminating variables with no statistically significant relationship with the outcome). The sparse models contain variables selected after both statistical feature selection and recursive feature elimination (RFE), explained in more detail in Section 2.1.5.3. The parsimonious models were generated by limiting recursive feature elimination to select a model with up to 15 variables. The parsimonious models are easier to implement and interpret relative to the full models, which have more variables but similar discrimination. To accommodate categorical variables that were recoded as indicator variables in the preprocessing phase, we selected a reference level for each categorical variable and forced every non-reference level to be included in a model if any other (non-reference) level of the categorical variable was selected.

For Model III (fecundability), we fit a discrete-time analog of the Cox proportional hazards model with cycle number as the time scale, allowing for delayed entry into the risk set (i.e., if a participant already had one cycle of pregnancy attempt at enrollment). Participants contributed at-risk cycles to the analysis from enrollment until reported pregnancy or a censoring event, which included initiation of fertility treatment, withdrawal from the study, cessation of pregnancy attempts, loss-to-follow-up, or 12 cycles of pregnancy attempt, whichever occurred first. We present results for the full model after statistical feature selection, as described above, and for a parsimonious model. To derive the parsimonious model, we fit separate Cox models with each individual predictor and then sorted the variables based on each model's concordance index. The concordance index is similar to the AUC (described below) but accounts for event time and loss to follow-up (Longato et al., 2020) (Schmid et al., 2016). We selected the top fifteen variables and forced non-selected levels of polytomous categorical variables into the final model, as described above.

3.2.4.3 Performance metrics

For Models I and II, we evaluated model performance using the AUC and weighted-F1 score, defined in Section 2.1.4. While the AUC is more easily interpretable, the weighted F1-score is more robust to data imbalances (Saito and Rehmsmeier, 2015). We present weighted-precision and weighted-recall metrics. For Model III, we evaluated performance using the concordance index, defined in Section 2.2.3.3.

All analyses were performed with Python statistical functions. Relevant programs can be accessed at github repository. Additional methodological information on how we addressed imbalance in the data and tuning of hyperparameters is as follows. To address the class imbalance in our dataset, we use a class weight (inversely proportional to class size) in the loss function used for training the model. This has the effect of balancing contributions to the loss from both classes. Class weights are used differently depending on the algorithm: for linear models (such as linear SVM or logistic regression), the class weights alter the loss function by weighting the loss of each sample by its class weight. For tree-based algorithms, the class weights are used for reweighting the splitting criterion. However, this rebalancing does not take the weight of samples in each class into account. We tune hyperparameters through cross-validation. In Logistic Regression (LR) and Support Vector Machine (SVM) models, we consider the inverse of regularization strength as a hyper parameter. We search for the best hyper parameter among [0.001, 0.01, 0.1, 1, 10] and choose the one that leads to the best classifier (with the highest AUC). In the artificial neural network (MLP) models, we have one input layer, a number of hidden layers, and one output layer. We tune the number of hidden layers and the number of neurons in the hidden layers. We try different options: (i) one hidden layer with 32, 64, 128, 256, or 512 neurons, (ii) two hidden layers with 16, 32, 64, 128, 256 neurons in the first hidden layer and 2 neurons in the second hidden layer, (iii) two hidden layers with 8, 16, 32, 64, or 128 neurons in the first hidden layer and 4 neurons in the second hidden layer. In the Gradient Boosting Machine (GBM) models, we used LightGBM which is a fast and high-performance GBM framework that grows trees leaf-wise rather than level-wise and incorporates advanced techniques, such as gradient-based one-side sampling and exclusive feature bundling to deal with a large number of data instances and features. We tune a number of hyper parameters such as learning rate, maximum number of leaves in one tree, and minimal number of data in one leaf. More details on the range of numbers used for the tuning of hyper parameters of our LightGBM can be found in the scripts at the github repository[1].

3.2.4.4 Addressing the distribution and bias of positive and negative cases in the data

To address the class imbalance in our dataset, we use a class weight (inversely proportional to class size) in the loss function used for training the model. This has the effect of balancing contributions to the loss from both classes. Class weights are used differently depending on the algorithm: for linear models (such as linear SVM or logistic regression), the class weights alter the loss function by weighting the loss of each sample by its class weight. For tree-based algorithms, the class weights are used for reweighting the splitting criterion. However, this rebalancing does not take the weight of samples in each class into account.

3.2.5 Sensitivity analysis

We restricted our analyses to nulligravid women with no history of infertility to evaluate the robustness of our results in a population that was presumably naïve to their fertility status.

3.3 Results

After excluding participants with incomplete follow-up for Models I and II, we analyzed data from 3,195 and 3,476 participants for Models I and II, respectively, and 16,876 cycles from 4,133 participants for Model III. The study participants were aged 30 years on average and ranged in age from 21 to 44 years. Among the 3,195 participants included in Model I, 2,747 (86%) became pregnant in 12 menstrual cycles. Among the 3,476 participants included in Model II, 2,406 (69%) became pregnant within six menstrual cycles. The distributions of class (i.e., pregnant versus non-

[1] https://github.com/noc-lab/Predictive-models-of-pregnancy

pregnant), overall and by number of menstrual cycles of attempt time at study entry, are presented in Tables 3.3 and 3.4. For each of the three models, the same 163 variables were considered for preprocessing (Table 3.1 and 3.2). After statistical feature selection, 40 variables were selected into the full model predicting pregnancy in 12 menstrual cycles (Model I) and 41 variables were selected into the full model predicting pregnancy within six menstrual cycles (Model II). After recursive feature elimination, 30 and 25 variables were selected for the sparse Models I and II, respectively. The final parsimonious models included 14 and 15 variables for Models I and II, respectively. We present performance statistics for the parsimonious models in Table 3.5. The AUC for Model I was 68-70% for all classification algorithms considered (std: 0.8% to 1.9%). Term std stands for standard deviation. The AUCs for Model II were 65-66% (std: 1.9% to 2.6%). The L2LR and L2SVM algorithms generally yielded the highest AUC. The weighted-F1 scores were similar across each algorithm, and no algorithm consistently yielded the highest score. The weighted-F1 scores obtained with the L2LR algorithm were 81.8 (std: 1.0) for Model I and 67.5 (std: 1.6) for Model II. The parsimonious models performed similarly to the full and sparse models (Table 3.6). The concordance index for Model III was 63.5% for the full model after statistical feature selection (24 variables) and 62.6% for the final parsimonious model. Figure 3·1 presents area under the precision-recall curves for Models I, II, IV, and V.

In order of decreasing magnitude of the regression coefficients (i.e., strongest to weakest predictor), the variables selected into the parsimonious Model I that were positively associated with pregnancy were menstrual cycle length, living in a rural region, daily use of multivitamins or folic acid, using the hormonal intrauterine device (IUD) as one's most recent method of contraception, having previously breastfed an infant, having ever been pregnant, female education, recent influenza vaccination, and

Table 3.3: Distribution of class and number of menstrual cycles of attempt time at study entry.

		Pregnant	Non-pregnant	Cycles[1]	
	N	n (%)	n (%)	0	1
Model I	3195	2747 (86%)	448 (14%)	1348 (42%)	1847 (58%)
Model II	3476	2406 (69%)	1070 (31%)	1462 (42%)	2014 (58%)
Model III	4133	2747 (66%)	1386 (34%)	1737 (42%)	2396 (58%)
Model IV	1571	1320 (84%)	251 (16%)	663 (42%)	908 (58%)
Model V	1722	1139 (66%)	583 (34%)	726 (42%)	996 (58%)
Model VI	1957	1333 (68%)	624 (32%)	819 (42%)	1138 (58%)

[1] Menstrual cycles of attempt time at study entry

Note: Model I predicts pregnancy in <12 menstrual cycles; Model II predicts pregnancy in <7 menstrual cycles; Model III predicts the probability of pregnancy within each menstrual cycle for up to 12 cycles of follow-up; Model IV predicts pregnancy in <12 menstrual cycles among nulligravid women with no history of infertility; Model V predicts pregnancy in <7 menstrual cycles among nulligravid women with no history of infertility; Model VI predicts the probability of pregnancy within each menstrual cycle for up to 12 cycles of follow-up among nulligravid women with no history of infertility.

Table 3.4: Distribution of each class by the number of menstrual cycles of attempt time at study entry.

	No cycles of attempt at study entry		One cycle of attempt at study entry	
	Pregnant	Non-pregnant	Pregnant	Non-pregnant
Model I	1205 (38%)	143 (4%)	1542 (48%)	305 (10%)
Model II	1086 (31%)	376 (11%)	1320 (38%)	694 (20%)
Model III	1213 (29%)	524 (13%)	1557 (38%)	839 (20%)
Model IV	588 (37%)	75 (5%)	732 (46%)	176 (11%)
Model V	518 (30%)	208 (12%)	621 (36%)	375 (22%)
Model VI	591 (30%)	228 (12%)	742 (38%)	396 (20%)

Note: Model I predicts pregnancy in <12 menstrual cycles; Model II predicts pregnancy in <7 menstrual cycles; Model III predicts the probability of pregnancy within each menstrual cycle for up to 12 cycles of follow-up; Model IV predicts pregnancy in <12 menstrual cycles among nulligravid women with no history of infertility; Model V predicts pregnancy in <7 menstrual cycles among nulligravid women with no history of infertility; Model VI predicts the probability of pregnancy within each menstrual cycle for up to 12 cycles of follow-up among nulligravid women with no history of infertility.

Table 3.5: Performance metrics for the parsimonious models, PRESTO 2013-2019.

Algorithm[1]	**Performance Measure (%) (Standard Deviation)**							
	Model I				**Model II**			
	AUC	**Weighted F1 Score**	**Weighted Precision**	**Weighted Recall**	**AUC**	**Weighted F1 Score**	**Weighted Precision**	**Weighted Recall**
L2LR	70.2 (1.6)	81.8 (1.0)	80.8 (1.0)	83.3 (1.3)	66.1 (2.1)	67.5 (1.6)	67.2 (1.5)	69.5 (1.4)
L1LR	69.8 (1.8)	81.6 (0.6)	80.6 (0.8)	83.5 (1.1)	66.0 (1.9)	67.4 (1.3)	66.9 (1.3)	69.3 (1.5)
L1SVM	69.8 (1.9)	81.5 (0.8)	80.6 (0.8)	83.5 (0.8)	66.0 (1.9)	67.4 (1.3)	66.9 (1.3)	69.1 (1.3)
L2SVM	70.0 (1.6)	81.5 (1.1)	80.7 (1.3)	83.3 (1.3)	66.2 (1.4)	67.1 (1.3)	66.9 (1.1)	69.6 (0.9)
MLP	69.9 (0.8)	82.1 (0.9)	80.9 (1.2)	83.9 (1.2)	65.1 (2.1)	65.7 (1.5)	66.5 (1.7)	68.5 (1.7)
LightGBM	68.1 (1.4)	81.6 (0.8)	80.9 (1.0)	82.9 (1.2)	64.9 (2.6)	66.9 (1.3)	66.6 (1.4)	67.6 (1.1)

[1] L2LR, ℓ_2-penalized logistic regression; L1LR, ℓ_1-penalized logistic regression; L1SVM, Support Vector Machine (SVM) with an ℓ_1-norm regularizer; L2SVM, SVM with an ℓ_2-norm regularizer; MLP, Feed Forward Multilayer Perceptron Neural Networks; LightGBM, Light Gradient Boosting Machine.

Note: Model I predicts pregnancy in <12 menstrual cycles (N = 3,195 participants). Model II predicts pregnancy in <7 menstrual cycles (N = 3,476 participants). The parsimonious models contain variables selected after both statistical feature selection and recursive feature elimination, and limiting recursive feature elimination to select a model with up to 15 variables.

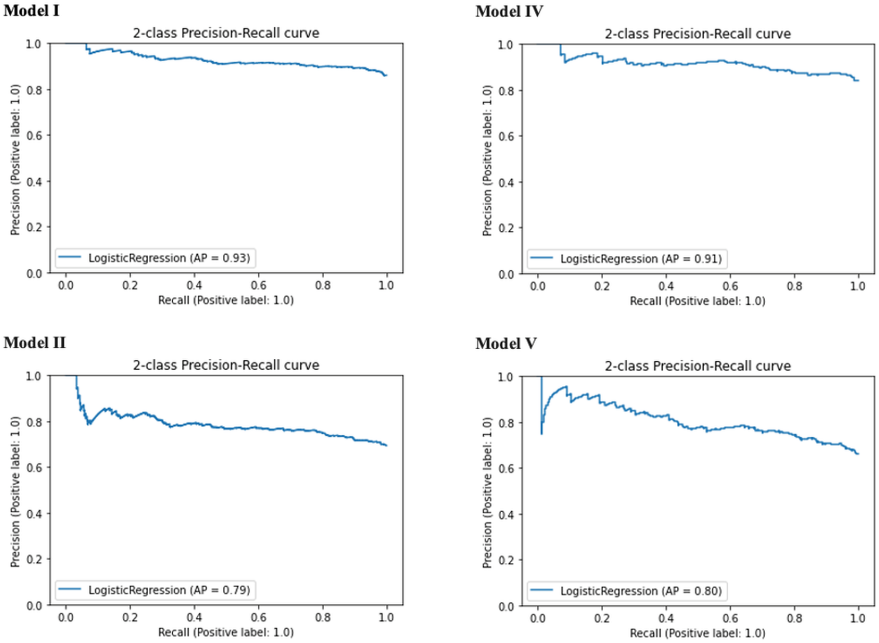


Figure 3·1: Area under the precision-recall curves (AUPRC) for Models I, II, IV, and V. Note: Model I predicts pregnancy in <12 menstrual cycles; Model II predicts pregnancy in <7 menstrual cycles; Model IV predicts pregnancy in <12 menstrual cycles among nulligravid women with no history of infertility; Model V predicts pregnancy in <7 menstrual cycles among nulligravid women with no history of infertility.

Table 3.6: Performance metrics for the full and sparse Models I-II, PRESTO 2013-2019.

Algorithm[1]	**Performance Measure (%) (Standard Deviation)**							
	Model I				**Model II**			
	AUC	**Weighted F1 Score**	**Weighted Precision**	**Weighted Recall**	**AUC**	**Weighted F1 Score**	**Weighted Precision**	**Weighted Recall**
Full Model[2]								
L2LR	70.5 (1.2)	82.0 (0.5)	81.3 (0.6)	82.8 (0.5)	65.9 (1.3)	67.1 (1.4)	66.7 (1.3)	68.4 (2.0)
L1LR	70.3 (1.4)	82.7 (0.7)	81.9 (0.5)	84.1 (1.1)	65.5 (1.5)	66.8 (0.8)	66.4 (0.9)	68.1 (1.6)
L1SVM	70.5 (1.5)	82.6 (0.5)	81.7 (0.4)	83.9 (0.9)	65.2 (1.7)	66.3 (0.9)	65.9 (0.9)	67.0 (1.1)
L2SVM	70.8 (1.4)	82.2 (0.6)	81.5 (0.5)	83.2 (1.1)	65.6 (1.3)	66.8 (1.1)	66.4 (1.1)	67.9 (1.7)
MLP	69.7 (2.1)	81.8 (1.0)	81.2 (0.8)	82.8 (1.8)	63.4 (1.4)	65.5 (1.0)	64.9 (1.2)	66.9 (0.8)
LightGBM	68.8 (0.6)	81.3 (0.6)	80.3 (0.4)	82.8 (1.0)	63.8 (1.9)	65.8 (0.9)	65.3 (0.9)	66.8 (1.0)
Sparse Model[3]								
L2LR	71.2 (1.0)	81.8 (0.6)	81.1 (0.6)	82.8 (0.9)	67.1 (1.5)	67.7 (0.3)	67.3 (0.3)	69.4 (0.9)
L1LR	70.5 (1.7)	81.6 (0.7)	81.7 (0.8)	84.2 (1.0)	66.6 (1.3)	67.1 (0.7)	67.3 (1.0)	68.6 (1.4)
L1SVM	70.7 (1.7)	82.4 (1.2)	81.7 (0.3)	83.7 (1.9)	66.5 (1.2)	67.0 (0.7)	66.7 (0.8)	68.2 (0.9)
L2SVM	71.2 (1.4)	81.5 (0.6)	81.5 (0.4)	83.5 (1.3)	66.8 (1.4)	67.3 (0.5)	66.8 (0.9)	68.4 (0.8)
MLP	70.8 (2.6)	82.5 (1.0)	81.5 (1.2)	84.5 (0.7)	65.3 (0.6)	67.1 (0.6)	66.6 (0.9)	67.7 (1.3)
LightGBM	69.3 (1.3)	80.8 (0.5)	79.9 (0.4)	82.2 (1.2)	65.0 (2.1)	66.3 (1.6)	65.9 (1.5)	66.9 (1.8)

[1] L2LR, ℓ_2-penalized logistic regression; L1LR, ℓ_1-penalized logistic regression; L1SVM, Support Vector Machine (SVM) with an ℓ_1-norm regularizer; L2SVM, SVM with an ℓ_2-norm regularizer; MLP, Feed Forward Multilayer Perceptron Neural Networks; LightGBM, Light Gradient Boosting Machine.

[2] The full models contain all variables selected after statistical feature selection.

[3] The sparse models contain variables selected after both statistical feature selection and recursive feature elimination.

Note: Model I predicts pregnancy in <12 menstrual cycles (N = 3,195 participants). Model II predicts pregnancy in <7 menstrual cycles (N = 3,476 participants).

gravidity (total number of pregnancies) (Table 3.7). The variables that were inversely associated with pregnancy were female age, having a history of infertility, having completed one menstrual cycle of pregnancy attempt time at study entry (versus zero), female BMI, and stress. The distributions of these variables overall, and by pregnancy status, are presented in Table 3.7. Results for parsimonious Models II and III are presented in Tables 3.8 and 3.9, respectively. The variables selected into the parsimonious Model II that were positively associated with pregnancy were daily use of multivitamins or folic acid, having previously breastfed an infant, HEI-2010 score, having a previous unplanned pregnancy, trying to improve one's chances of pregnancy (e.g., charting cycles, ovulation or cervical mucus testing, timing intercourse to the fertile window), and time since the participant's last pregnancy (<1 year). The variables that were inversely associated with pregnancy were female BMI, having a history of infertility, male age, non-use of a fertility app, male BMI, having completed one menstrual cycle of pregnancy attempt time at study entry (versus zero), male partner smoking, female age, and having a history of subfertility or infertility. Results were generally similar for Model III. Variables selected into Model III but neither Models I nor II included intercourse frequency, and menstrual cycle regularity.

Among 1,957 nulligravid women without a history of infertility, we developed models predicting pregnancy in fewer than 12 menstrual cycles (Model IV), predicting pregnancy within six menstrual cycles (Model V), and predicting fecundability (Model VI). We analyzed data from 1,571, 1,722, and 1,957 participants for Models IV, V, and VI, respectively. The performance of these models was slightly lower than the analogous models in the full cohort. The performance statistics for the full and sparse Models IV and V are presented in Table 3.10. Using statistical feature selection, 16 and 12 variables were selected into the full models for Model IV and V, respectively. After recursive feature elimination, 5 and 9 variables were selected

Table 3.7: Variables selected by the parsimonious Model I (predicting pregnancy in 12 cycles) using the L2LR algorithm, PRESTO 2013-2019, n=3,195 participants.

Variable	Coef[1]	Overall		Pregnant		Not pregnant	
		Freq. or Mean	Std.	Freq. or Mean	Std.	Freq. or Mean	Std.
Menstrual cycle length (days)	0.27	29.6	4.0	29.7	4.1	28.7	3.0
Female age at baseline (years)	-0.26	29.8	3.8	29.7	3.6	30.6	4.5
Urbanization of residential area: rural (ref = urbanized area)	0.25	4%	20%	5%	21%	1%	12%
Previously tried to conceive for $\geq$ 12 months: "yes" (ref = "no, tried for < 12 months")	-0.24	5%	21%	4%	19%	10%	30%
One menstrual cycle of attempt time at study entry (ref = 0)	-0.23	58%	49%	56%	50%	68%	47%
Daily use of multivitamins/folic acid (yes/no)	0.22	84%	37%	85%	35%	73%	44%
Last method of contraception: hormonal IUD (yes/no)	0.19	12%	32%	12%	33%	7%	25%
Female BMI (kg/m^2)	-0.19	26.6	6.5	26.3	6.2	28.4	7.8
Ever breastfed an infant (yes/no)	0.18	31%	46%	32%	47%	22%	41%
Ever been pregnant (yes/no)	0.15	50%	50%	52%	50%	42%	49%
Female education (years)	0.14	16.0	1.2	16.1	1.2	15.8	1.4
Received influenza vaccine in the past year (yes/no)	0.13	53%	50%	54%	50%	44%	50%
Stress (Perceived Stress Scale score)	-0.12	15.5	5.8	15.3	5.8	16.3	5.6
Total number of pregnancies	0.12	1.0	1.4	1.0	1.4	0.8	1.4
Urbanization of residential area: Canada (ref = urbanized area)[2]	0.01	18%	39%	18%	39%	19%	39%
Urbanization of residential area: urban cluster (ref = urbanized area)[2]	-0.01	8%	27%	8%	27%	8%	27%
Previously tried to conceive for $\geq$ 12 months: "no, never tried before" (ref = "no, tried for < 12 months")[2]	-0.01	42%	49%	41%	49%	48%	50%

[1] Standardized Regression Coefficient

[2] Variables forced into the model

Table 3.8: Variables selected by the parsimonious Model II (predicting pregnancy within 6 cycles) using the L2LR algorithm, PRESTO 2013-2019, n=3,476 participants.

Variable	**Coef**	**Overall**		**Pregnant**		**Not Pregnant**	
		Freq. or Mean	**Std.**	**Freq. or Mean**	**Std.**	**Freq. or Mean**	**Std.**
Female BMI (kg/m^2)	-0.11	26.8	6.7	26.1	6.1	28.3	7.7
Daily use of multivitamins/folic acid (yes/no)	0.08	84%	37%	86%	35%	78%	42%
Ever breastfed an infant (yes/no)	0.08	30%	46%	33%	47%	24%	43%
Previously tried to conceive for $\geq$ 12 months: "yes" (ref = "no, tried for $<$ 12 months")	-0.08	5%	22%	4%	18%	8%	28%
Healthy Eating Index-2010 score (HEI-2010 score)	0.07	66.0	11.2	66.8	10.9	64.3	11.6
Male age (years)	-0.07	31.8	5.0	31.5	4.6	32.4	5.8
Use of fertility app: "no, but I plan to" (ref = "yes")	-0.07	8%	27%	6%	24%	11%	31%
History of unplanned pregnancy (yes/no)	0.07	34%	47%	37%	48%	27%	44%
Male BMI (kg/m^2)	-0.07	27.7	5.3	27.3	5.1	28.5	5.6
One menstrual cycle of attempt time at study entry (ref = 0)	-0.06	58%	49%	55%	50%	65%	48%
Male cigarette smoking: "yes, on a regular basis" (ref = "no")	-0.06	8%	27%	6%	24%	12%	32%
Female age at baseline (years)	-0.06	29.8	3.8	29.6	3.6	30.3	4.2
Trying to improve chances of pregnancy (yes/no)	0.05	70%	46%	72%	45%	64%	48%
Time since last pregnancy: $<$1 year (ref = nulliparous)	0.05	22%	41%	24%	42%	18%	38%
History of subfertility or infertility (yes/no)	-0.05	10%	30%	9%	28%	13%	34%
Previously tried to conceive for $\geq$ 12 months: "no, never tried before" (ref = "no, tried for $<$ 12 months")[1]	-0.05	42%	49%	40%	49%	46%	50%
Time since last pregnancy: 1-2 years (ref = nulliparous)[1]	0.04	17%	38%	19%	39%	14%	35%
Male cigarette smoking: "yes, occasionally" (ref = "no")[1]	-0.02	4%	20%	4%	19%	5%	22%
Time since last pregnancy: $\geq$ 5 years (ref = nulliparous)[1]	-0.02	6%	24%	5%	22%	8%	27%
Use of fertility app: "no" (ref = "yes")[1]	-0.02	23%	42%	22%	41%	26%	44%
Time since last pregnancy: 3-4 years (ref = nulliparous)[1]	0.02	4%	21%	5%	21%	4%	19%

[1] Variables forced into the model

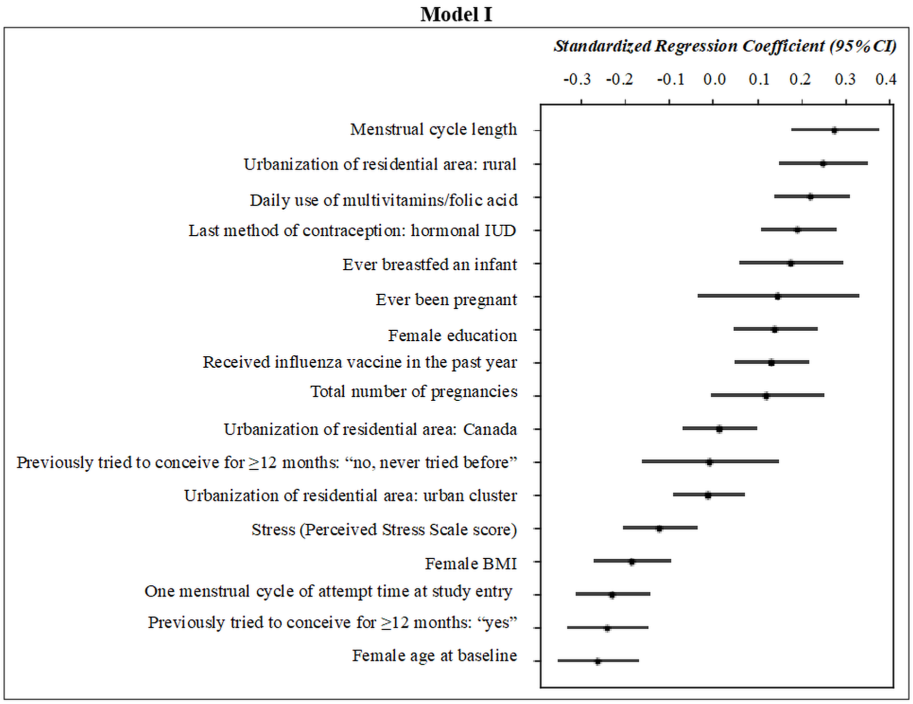


Figure 3·2: Visualization of the Model I coefficients and their 95% confidence intervals. Note: Figures 3·2, 3·3 and 3·4 present the model coefficients with error bands equivalent to 95% confidence intervals. The variables are ordered according to their mean coefficient. All plots are associated with the parsimonious L2LR version of models, consistent with the rest of the Manuscript. Model I predicts pregnancy in < 12 menstrual cycles. Model II predicts pregnancy in > 7 menstrual cycles. Model III predicts the probability of pregnancy within each menstrual cycle for up to 12 cycles of follow-up.

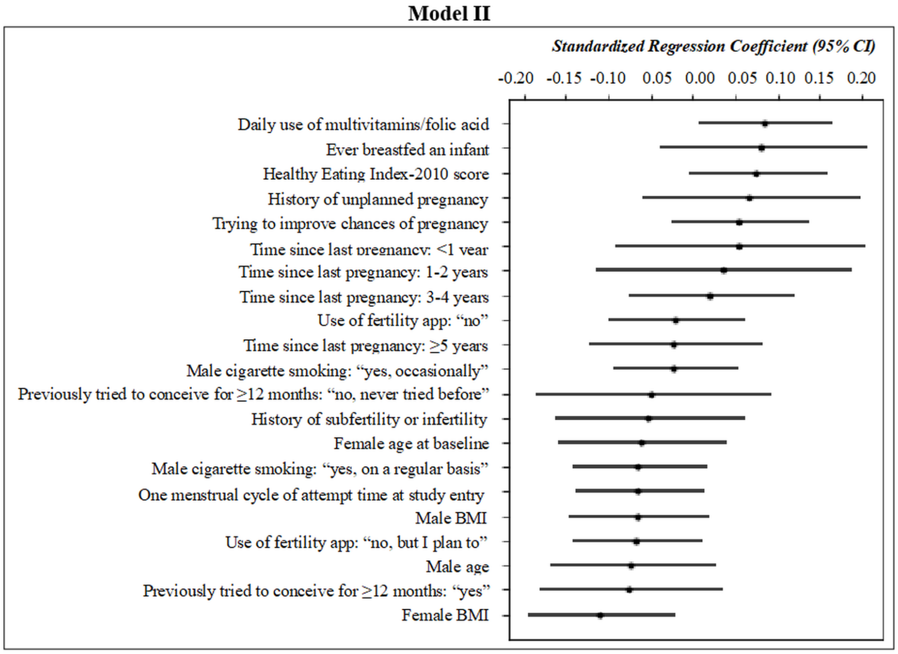


Figure 3·3: Visualization of the Model II coefficients and their 95% confidence intervals.

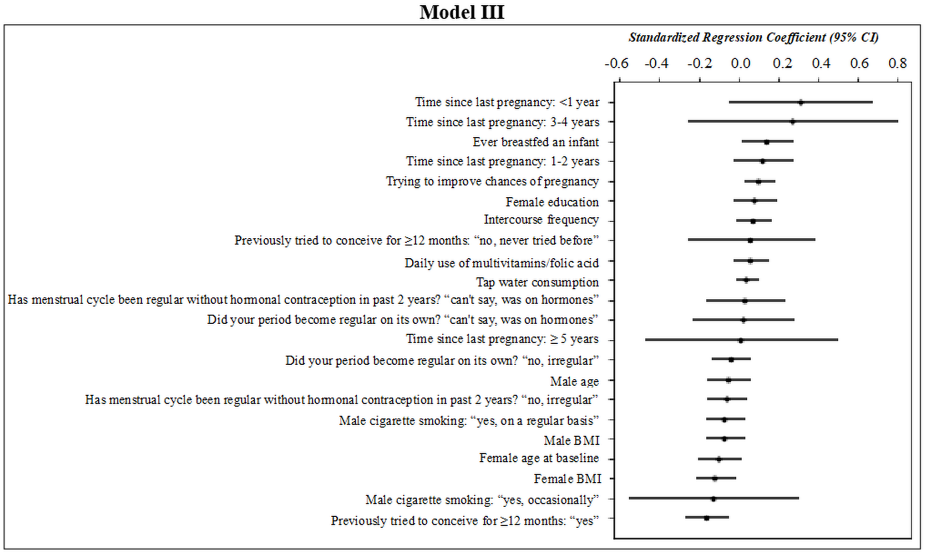


Figure 3·4: Visualization of the Model III coefficients and their 95% confidence intervals.

for the sparse Models IV and V, respectively. Because fewer than 15 features were selected by each of the sparse models, the sparse models were equivalent to the parsimonious models. Consistent with the main analysis, the L2LR algorithm performed best for the sparse models. The AUCs were 69.5% (std: 1.4) for Model IV and 65.6% (std: 2.9) for Model V. The concordance index for Model VI was 60.2%. Variables selected by these models that were positively associated with pregnancy included menstrual cycle length, using a hormonal IUD as one's most recent method of contraception, intercourse frequency, trying to improve one's chances of pregnancy, use of vitamin E supplements, and HEI-2010 score. Variables inversely associated with the probability of pregnancy included having completed one menstrual cycle of pregnancy attempt time at study entry (versus zero), female age, male and female BMI, menstrual cycle irregularity, non-use of a fertility app, stress, depressive symptoms, history of vaginosis, male partner smoking, milk consumption, and sleep characteristics. Occupational exposures including exposure to metal particulates or fumes and exposure to high temperature environments were also selected to Model VI, but with very small coefficients.

3.3.1 Numerical Experiments for Comparison of Original Cox Model and DRO Cox Models

Equation 2.24 introduces $O(N^2)$ constraints, significantly hindering computational efficiency. To address this, we impose constraints only for $i \leq k \leq i + r$, reducing the total number of constraints to $O(rN)$. For example, when we set r=2, only two constraints are added instead of N constraints.

Due to computational complexity, we aim to reduce the dimentianality. For this sake, we perform feature selection, we also randomly resample a portion of each dataset for our experiments. This results in a dataset with 827 participants and 16 features.

To assess the impact of outliers on model performance, we introduce varying proportions of outliers, from 5% to 30%, into a random subsample of the datasets. We evaluate the Original Cox model, the Sample-splitting DRO-Cox, and the Global fixation DRO-Cox, training each with different radii ϵ. The concordance indices of these models are then compared, as shown in Table 3.14.

Table 3.14 reveals that the Global fixation DRO-Cox model consistently outperforms both the Sample-splitting DRO-Cox and the Original Cox model. This comparison underscores the influence of outlier inclusion on the predictive accuracy and robustness of survival analysis models, especially in the presence of outliers (Jin and Paschalidis, 2024).

Table 3.14: Comparison of concordance indices for different ratios of outliers with $\epsilon = 0.05$ in a subset of the pregnancy dataset

Ratio of Outliers	0.1	0.15	0.2	0.25	0.3
Original Cox	0.5340	0.5522	0.5654	0.5491	0.5190
DRO-Cox Sample Splitting	0.4796	0.4827	0.5158	0.5108	0.5132
DRO-Cox Global Fixation	0.5344	0.5602	0.5699	0.5534	0.5207

3.4 Discussion

In this prospective cohort study of 4,133 North American pregnancy planners, we applied several supervised learning methods to predict the probability of pregnancy within three time periods: 12 menstrual cycles, 6 menstrual cycles, and on a per-cycle basis. The L2LR and L2SVM algorithms generally yielded the highest AUC, particularly for the parsimonious models. For all models, discrimination (AUC) was close to 70%. The highest AUCs were 71.2% for Model I, 67.1% for Model II, 69.5% for Model IV, and 65.6% for Model V. These findings demonstrate that it is possible to develop predictive models with reasonable discrimination using self-reported data in the absence of more detailed medical information such as laboratory or imaging tests.

The discrimination of our models is greater than previously published predictive models for pregnancy independent of fertility treatment, which yielded AUC's between 59% and 64% (Coppus et al., 2009). For example, Eimers et al. developed a predictive model for pregnancy among 996 couples consulting for infertility care in the Netherlands between 1974 and 1984 (Eimers et al., 1994). The investigators collected data on patient medical history, laboratory tests including semen analysis and postcoital tests (i.e., an examination of the interaction between sperm and the cervical mucus after intercourse), and a gynecologic physical examination. They used forward stepwise Cox regression to produce a model including female age, duration of infertility, primary versus secondary infertility, history of infertility in the male partner's family, sperm motility, and the postcoital test results. Similar studies were conducted by Collins et al., using data from 1,061 couples seeking infertility care at eleven Canadian University hospitals (Collins et al., 1995), and Snick et al., using data from 402 couples seeking infertility care at a Dutch general hospital (Snick et al., 1997). Hunault et al. pooled the data from the Eimers, Collins, and Snick studies to evaluate the accuracy of these models and to develop two new synthesis models (Hunault et al., 2004). The synthesis models included female age, duration of subfertility, sperm motility, whether the couple had been referred for infertility care by a general physician or a gynecologist, and the results of a postcoital test. These models were externally validated and found to have AUCs of 59-63% (Hunault et al., 2005) (Van der Steeg et al., 2007).

Although previous studies predicted the probability of pregnancy independent of fertility treatment, they were exclusively conducted in populations with subfertility using little or no data on lifestyle, environmental, and sociodemographic factors (Collins et al., 1995) (Coppus et al., 2009) (Eimers et al., 1994) (Hunault et al., 2004) (Hunault et al., 2005) (Snick et al., 1997) (Van der Steeg et al., 2007). Our study may be more generalizable to couples across the fertility spectrum, because we included couples with a wide range of reproductive potential. In addition, we considered a range of potential predictors that may be more easily modified than clinical markers such as semen quality or hormone levels. For example, fertility app use, use of multivitamins or folic acid supplements, and trying to improve one's chances of pregnancy (e.g., charting cycles, ovulation or cervical mucus testing, timing intercourse to the fertile window) are relatively modifiable behaviors. Lifestyle interventions can also be undertaken to modify individual-level behaviors that may increase a couple's chance of conception, such as promoting a healthy BMI, improving diet, and reducing stress. However, many of these behaviors are determined by broader environmental and systemic drivers and thus may be best addressed through macro-level policy interventions that address upstream determinants (e.g., regulation of food supply and marketing). A causal analysis of each risk factor would be worthwhile for future and more targeted work. In this study, there were some variables that appeared to be particularly important predictors of pregnancy. These included female age and BMI, history of infertility, the number of menstrual cycles of pregnancy attempt time at study entry, having previously breastfed an infant, and use of multivitamins or folic acid supplements. These findings are generally consistent with previous studies on individual risk factors for infertility that were conducted in other populations (Cueto et al., 2016) (Homan et al., 2007) (Jensen et al., 1999) (Wise et al., 2011). However, having previously breastfed an infant, which was associated with an increased probability of pregnancy in this study, has not been previously studied as a predictor of fecundability. This may reflect underlying fertility, prolonged effects of hormonal changes during breastfeeding, or higher socioeconomic status among women who breastfeed their infants (Jones et al., 2011) (Odar Stough et al., 2019).

In this study, we developed an additional set of predictive models among nulligravid women with no history of infertility who had been trying to conceive for no more than one menstrual cycle of attempt time at enrollment. The performance of these models was slightly decreased compared with the main analyses. This is likely because having a history of infertility is a strong predictor of future fecundability, and therefore restricting the analytic sample by this variable would limit the predictive ability of the model. This was most obvious in Model V, which predicted pregnancy within six menstrual cycles. In these restricted analyses, the most important predictors of pregnancy across all models were the number of menstrual cycles of pregnancy attempt time at study entry, and female age.

Study limitations include potential misclassification of the predictor variables, given that all data were based on self-reporting. There is limited research on the impact of measurement error on machine learning prediction models (Jiang et al., 2021) (van Doorn et al., 2017), and it is unclear how misclassification of the predictors influenced our study results in terms of accuracy and variable selection. There was also the potential for misspecification of the functional form of the predictor variables, which could have influenced the variable selection process. In addition, there may have been some misclassification of our estimate of time to pregnancy, which relied on self-reported menstrual cycle length and date of the last menstrual period. Given the prospective design of the study, such misclassification is likely to be non-differential with respect to the outcome. Bias may also have been introduced if the length of follow-up varied by the predictors under study, as Models I and II did not account for varying lengths of follow-up. However, results were generally consistent with Model III, which accounted for varying lengths of follow-up. Another potential limitation is our lack of inclusion of important predictors of pregnancy, such as hormone levels, which may have reduced the predictive ability of our models. Other potentially important predictors that we did not measure include environmental exposures (Conforti et al., 2018) (Hipwell et al., 2019) (Kahn et al., 2021), early life adversity (Harville and Boynton-Jarrett, 2013) (Jacobs et al., 2015), occupational stress (Barzilai-Pesach et al., 2006) (Valsamakis et al., 2019), experiences of discrimination (Krieger, 2000), social disadvantage, neighborhood characteristics (Williams and Collins, 2001), and multigenerational exposures (Eskenazi et al., 2021) (Wesselink, 2021). In addition, we lacked comprehensive data on male exposures, which contribute to up to 50% of all subfertility among couples (Irvine, 1998). However, we collected data on several important male characteristics on the female baseline questionnaire, including male age, BMI, education, and smoking status. Overall, we considered a diverse range of 163 potential predictors, which is substantially greater than previous studies in this area (Collins et al., 1995)(Coppus et al., 2009)(Eimers et al., 1994)(Hunault et al., 2004)(Hunault et al., 2005)(Snick et al., 1997)(Van der Steeg et al., 2007). It should be noted that the effect estimates in these models lack causal interpretation, as variables were selected into the final models based on their predictive power, rather than the hypothesized causal structures of the data. Identifying causes of infertility was beyond the scope of this study. Also beyond the scope of this study was the development of models within clinically-relevant subgroups (e.g., age >40 years or infertility-related conditions). Finally, though we validated the models using split sample replication techniques, we were unable to conduct an external validation study.

3.5 Conclusions

In this large prospective cohort, we used machine learning algorithms to develop predictive models of pregnancy, using three distinct, clinically-relevant definitions of infertility, subfertility, and fecundability. Comparing results across the three outcomes facilitates robust triangulation of fertility potential; the relative utility of each outcome may depend on a couple's preferences and risk profile. Our methods can predict pregnancy with discrimination as high as 71.2% by properly weighing a small set of predictive variables that include lifestyle and reproductive characteristics. Overall, the most consistent predictors of the probability of conception were female age, female BMI, male age, male BMI, history of infertility, history of breastfeeding, time since the participant's last pregnancy, daily use of multivitamins or folic acid, trying to improve one's chances of pregnancy (e.g., charting cycles, ovulation or cervical mucus testing, timing intercourse to the fertile window), male partner smoking, and female education. Among nulligravid women without a history of infertility, the most important predictors were female age, female BMI, male BMI, use of a fertility app, and perceived stress. These findings are particularly relevant for couples planning a pregnancy and clinicians providing preconception care to women who are discontinuing contraception in order to conceive. If these models are successfully validated in external populations, they could potentially be implemented as a counseling tool (Yland et al., 2022).

Table 3.9: Variables selected by the parsimonious Model III (fecundability), PRESTO 2013-2019, n=4,133 participants.

Variable	**Hazard Ratio**	**95% Confidence Interval**
Previously tried to conceive for ≥ 12 months: "yes" (ref = "no, tried for < 12 months")	0.85	(0.80, 0.90)
Ever breastfed an infant (yes/no)	1.16	(1.09, 1.23)
Female BMI (kg/m^2)	0.89	(0.84, 0.93)
Time since last pregnancy: 1-2 years (ref = nulliparous)	1.12	(1.04, 1.21)
Female age at baseline (years)	0.90	(0.85, 0.95)
Trying to improve chances of pregnancy (yes/no)	1.11	(1.06, 1.15)
Female education (years)	1.09	(1.03, 1.15)
Intercourse frequency (times/week)	1.08	(1.03, 1.12)
Male BMI (kg/m^2)	0.93	(0.89, 0.98)
Male cigarette smoking: "yes, on a regular basis" (ref = "no")	0.93	(0.89, 0.98)
Has menstrual cycle been regular without hormonal contraception in past 2 years? "no, irregular" (ref = "yes, regular")	0.94	(0.89, 0.99)
Daily use of multivitamins/folic acid (yes/no)	1.06	(1.01, 1.11)
Did your period become regular on its own? "no, irregular" (ref = "yes, regular")	0.96	(0.92, 1.01)
Male age (years)	0.96	(0.91, 1.01)
Tap water consumption (drinks/week)	1.04	(1.01, 1.07)
Time since last pregnancy: <1 year (ref = nulliparous)[1]	1.37	(1.14, 1.64)
Time since last pregnancy: 3-4 years (ref = nulliparous)[1]	1.32	(1.01, 1.71)
Male cigarette smoking: "yes, occasionally" (ref = "no")[1]	0.87	(0.70, 1.08)
Has menstrual cycle been regular without hormonal contraception in past 2 years? "unknown, was using hormonal contraception" (ref = "yes, regular")[1]	1.03	(0.93, 1.14)
Did your period become regular on its own? "unknown, was using hormonal contraception" (ref = "yes, regular")[1]	1.02	(0.89, 1.17)
Time since last pregnancy: ≥5 years (ref = nulliparous)[1]	1.01	(0.79, 1.29)

[1] Variables forced into model

Table 3.10: Performance metrics for the full and sparse Models IV-V, restricting to nulligravid women with no history of infertility, PRESTO 2013-2019.

Algorithm[1]	**Performance Measure (%) (Standard Deviation)**							
	Model IV				**Model V**			
	AUC	**Weighted F1 Score**	**Weighted Precision**	**Weighted Recall**	**AUC**	**Weighted F1 Score**	**Weighted Precision**	**Weighted Recall**
Full Model[2]								
L2LR	68.3 (4.1)	79.2 (1.5)	78.7 (2.5)	81.1 (2.6)	65.1 (3.2)	64.5 (1.7)	65.2 (2.0)	67.5 (2.0)
L1LR	68.4 (4.1)	78.3 (1.7)	77.9 (2.5)	79.7 (2.5)	64.9 (2.7)	65.2 (2.4)	65.2 (2.6)	67.5 (2.0)
L1SVM	68.4 (3.9)	78.6 (1.7)	77.8 (2.4)	80.0 (2.0)	64.9 (2.7)	65.1 (2.3)	65.2 (2.5)	67.0 (2.0)
L2SVM	68.6 (3.7)	79.0 (1.1)	78.4 (2.4)	80.5 (1.9)	64.9 (2.8)	64.6 (2.4)	65.4 (2.7)	67.4 (2.8)
MLP	67.8 (2.4)	79.3 (1.6)	78.6 (1.7)	80.4 (1.9)	63.9 (3.3)	63.9 (2.2)	63.7 (2.2)	65.8 (2.2)
LightGBM	64.3 (3.2)	78.6 (1.4)	77.6 (1.7)	80.1 (1.3)	63.7 (3.0)	63.8 (2.1)	63.4 (2.1)	64.8 (2.0)
Sparse Model[3]								
L2LR	69.5 (1.4)	80.5 (1.0)	79.8 (0.8)	81.8 (1.5)	65.6 (2.9)	66.2 (2.5)	65.9 (2.6)	66.8 (2.4)
L1LR	69.3 (1.3)	80.6 (0.9)	79.9 (0.8)	81.9 (1.5)	64.9 (2.8)	64.7 (2.6)	64.9 (2.9)	67.1 (2.3)
L1SVM	69.4 (1.3)	80.7 (1.1)	80.0 (0.9)	81.9 (1.6)	64.8 (2.8)	64.8 (2.6)	65.0 (3.0)	67.1 (2.3)
L2SVM	69.5 (1.5)	80.6 (1.3)	79.8 (1.1)	81.8 (1.6)	65.2 (2.8)	65.4 (2.2)	65.3 (2.2)	67.3 (2.0)
MLP	68.1 (1.7)	79.6 (1.4)	78.6 (1.9)	81.1 (1.3)	63.9 (2.8)	64.2 (2.8)	64.0 (3.0)	64.8 (2.3)
LightGBM	66.0 (2.4)	79.3 (0.5)	78.4 (0.3)	80.7 (1.2)	63.5 (3.3)	64.3 (2.2)	63.9 (2.2)	65.6 (1.8)

[1] L2LR, ℓ_2-penalized logistic regression; L1LR, ℓ_1-penalized logistic regression; L1SVM, Support Vector Machine (SVM) with an ℓ_1-norm regularizer; L2SVM, SVM with an ℓ_2-norm regularizer; MLP, Feed Forward Multilayer Perceptron Neural Networks; LightGBM, Light Gradient Boosting Machine.

[2] The full models contain all variables selected after statistical feature selection.

[3] The sparse models contain variables selected after both statistical feature selection and recursive feature elimination.

Note: Model IV predicts pregnancy in <12 menstrual cycles among nulligravid women with no history of infertility (N = 1,571 participants). Model V predicts pregnancy in <7 menstrual cycles among nulligravid women with no history of infertility (N = 1,722 participants).

Table 3.11: Variables selected by the parsimonious Model IV (restricted to nulligravid women with no history of infertility) predicting pregnancy within 12 cycles, using the L2LR algorithm, PRESTO 2013-2019, n=1,571 participants.

Variable	**Coef**[1]	**Overall**		**Pregnant**		**Not pregnant**	
		Freq. or Mean	**Std.**	**Freq. or Mean**	**Std.**	**Freq. or Mean**	**Std.**
Menstrual cycle length (days)	0.27	29.6	4.0	29.8	4.2	28.7	2.6
One menstrual cycle of attempt time at study entry (ref=0)	-0.22	58%	49%	55%	50%	70%	46%
Last method of contraception: hormonal IUD (yes/no)[1]	0.22	11%	31%	12%	32%	5%	22%
Stress (Perceived Stress Scale score)	-0.20	15.0	5.5	14.7	5.5	16.1	5.3
Female age at baseline (years)	-0.20	29.2	3.4	29.1	3.3	30.0	3.9

[1] Last methods of contraception were not mutually exclusive and were coded as indicator variables with no reference category.

Table 3.12: Variables selected by the parsimonious Model V (restricted to nulligravid women with no history of infertility) predicting pregnancy within 6 cycles, using the L2LR algorithm, PRESTO 2013-2019, n=1,722 participants.

Variable	**Coef**[1]	**Overall**		**Pregnant**		**Not pregnant**	
		Freq. or Mean	**Std.**	**Freq. or Mean**	**Std.**	**Freq. or Mean**	**Std.**
Female age at baseline (years)	-0.06	29.2	3.4	28.9	3.2	29.8	3.8
Use of fertility app: "no" (ref = "yes")	-0.05	22%	42%	20%	40%	27%	44%
Male BMI (kg/m^2)	-0.04	27.4	5.3	27.0	5.1	28.2	5.5
One menstrual cycle of attempt time at study entry (ref = 0)	-0.04	58%	49%	55%	50%	64%	48%
Did your period become regular on its own? "no, irregular" (ref = "yes, regular")	-0.04	19%	39%	16%	37%	23%	42%
Healthy Eating Index-2010 score (HEI-2010 score)	0.04	67.2	10.8	67.8	10.5	65.8	11.3
Female BMI (kg/m^2)	-0.04	26.3	6.5	25.8	6.0	27.4	7.2
Use of fertility app: "no, but I plan to" (ref = "yes")	-0.03	9%	28%	8%	26%	11%	31%
Stress (Perceived Stress Scale score)	-0.02	15.1	5.6	14.8	5.5	15.6	5.6
Depressive Symptoms (Major Depression Inventory score)	-0.02	9.0	7.0	8.7	6.9	9.6	7.0
Variables forced into the model[1]							
Did your period become regular on its own? "unknown, was using hormonal contraception" (ref = "yes, regular")	0.01	15%	35%	16%	37%	12%	33%

[1] For all models, we selected a reference group for each categorical variable that was recoded as indicator variables in the preprocessing phase and forced every non-reference level to be included in the model if any level of the categorical variable was selected. These variables are listed in addition to the variables selected by the parsimonious model.

Table 3.13: Variables selected by the parsimonious Model VI (restricted to nulligravid women with no history of infertility) predicting pregnancy within 6 cycles, using the L2LR algorithm, PRESTO 2013-2019, n=1,957 participants.

Variable	**Hazard Ratio**	**95% Confidence Interval**
Female age at baseline (years)	0.90	(0.84, 0.96)
Has menstrual cycle been regular without hormonal contraception in past 2 years? “no, irregular” (ref = “yes, regular”)	0.90	(0.84, 0.97)
Use of fertility app: “no” (ref = “yes”)	0.90	(0.85, 0.97)
Ever diagnosed with vaginosis (yes/no)	0.91	(0.85, 0.97)
Intercourse frequency (times/week)	1.10	(1.03, 1.17)
Male cigarette smoking: “yes, on a regular basis” (ref = “no”)	0.91	(0.85, 0.97)
Female BMI (kg/m^2)	0.91	(0.85, 0.98)
Exposed regularly to metal particulates or fumes (yes/no)	1.08	(1.02, 1.15)
Male BMI (kg/m^2)	0.94	(0.88, 1.01)
Trying to improve chances of pregnancy (yes/no)	1.06	(0.99, 1.13)
2% milk consumption (drinks/week)	0.95	(0.89, 1.01)
Use of vitamin E supplements (yes/no)	1.05	(0.99, 1.12)
Nightly sleep duration (hours)	0.96	(0.90, 1.02)
Night shift work (number of shifts in past month)	0.96	(0.90, 1.03)
Has menstrual cycle been regular without hormonal contraception in past 2 years? “unknown, was using hormonal contraception” (ref = “yes, regular”)	0.98	(0.86, 1.12)
Exposed regularly to environments with high temperature (yes/no)	1.02	(0.95, 1.08)
Use of fertility app: “no, but I plan to” (ref = “yes”)[1]	0.76	(0.60, 0.96)
Male cigarette smoking: “yes, occasionally” (ref = “no”)[1]	0.86	(0.62, 1.19)
Has menstrual cycle been regular without hormonal contraception in past 2 years? “unknown, was using hormonal contraception” (ref = “yes, regular”)[1]	0.98	(0.86, 1.12)

[1] Variables forced into model

Chapter 4

Predicting Polycystic Ovary Syndrome (PCOS) with Machine Learning Algorithms from Electronic Health Records

4.1 Introduction

Polycystic ovary syndrome (PCOS) is the most common type of ovulation disorder and endocrinopathy among reproductive age women. PCOS is a diagnosis of exclusion after other endocrinopathies known to affect ovulation have been evaluated including thyroid, adrenal, and pituitary related disease. Based on the Rotterdam criteria, PCOS is diagnosed when two of the three following criteria are exhibited: clinical or biochemical hyperandrogenism, oligo-anovulation, and polycystic ovary morphology (PCOM) on transvaginal or transabdominal ultrasound. PCOS has a population prevalence of 5-15%, depending on the diagnostic criteria used (Azziz et al., 2009).

PCOS is associated with multiple health issues and increased morbidity and mortality, including a high chronic disease burden that is also very costly for individuals with PCOS and insurers (Riestenberg et al., 2022). PCOS is the leading cause of anovulatory infertility in reproductive-aged women. In fact, over 90% of anovulatory women who present to infertility clinics have PCOS (Sirmans and Pate, 2013). PCOS patients have an increased risk of endometrial hyperplasia and endometrial cancer (Barry et al., 2014) due to anovulatory cycles leading to long periods of exposure to the effects of unopposed estrogen. PCOS has been associated with the development of metabolic syndrome (Lim et al., 2019), diabetes (Anagnostis et al., 2018), cerebrovascular disease and hypertension (Wekker et al., 2020), compared to women without PCOS. Despite these serious health consequences, PCOS frequently goes undiagnosed due to the wide range of symptom severity on presentation, leading to delayed treatment and potentially more severe clinical sequelae due to lack of preventive care, health management, and counseling (Barry et al., 2014). Even when PCOS is diagnosed, it is often very delayed. One study found that over one-third of women with PCOS waited over two years and were seen by three or more providers before finally receiving the diagnosis (Gibson-Helm et al., 2017).

Predictive models can play a significant role in aiding earlier diagnosis of PCOS, though several include only those women presenting for fertility care. One model used serum anti-Müllerian hormone (AMH) and androstenedione levels, menstrual cycle length, and BMI to predict the development of PCOS in Chinese women (Xu et al., 2022). Another model used only AMH and BMI to predict a diagnosis of PCOS or other ovulatory dysfunction disorders (Vagios et al., 2021). Other studies have created predictive models for certain outcomes among women with PCOS such as pregnancy outcomes (Kuang et al., 2015) (Jiang et al., 2022) and insulin resistance (Gennarelli et al., 2000). In this study, we use clinical and socioeconomic variables among 30,601 women aged 18 to 45 years within the electronic health records (EHR) to develop predictive model utilizing machine learning algorithms with the goal of earlier detection and treatment of PCOS.

4.2 Materials and Methods

4.2.1 Data acquisition

The dataset was created by querying de-identified patient data from female patients aged 18 to 45 years who had or were considered at risk for PCOS diagnosis by having had any one of the three diagnostic procedures for PCOS in their EHR. Included within the initial sample were those patients who had any visit to Boston Medical Center (BMC) for primary care, obstetrics and gynecology, endocrinology, family medicine, or general internal medicine and received: 1) a pelvic/transvaginal ultrasound for any reason, 2) androgen lab assessment, or had clinical symptoms of androgen excess, 3) an ICD-9 label for irregular periods, or 4) a PCOS diagnosis, between October 2003 to December 2016 within the BMC Clinical Data Warehouse (CDW). The start-date was selected to reflect the first day that ICD-9 codes were used and recorded at BMC. The end date reflected cessation of use of the ICD-9 codes and transition to ICD-10 codes within BMC. To avoid misidentifying an ovulation disorder caused by another endocrinopathy, exclusion criteria included diagnosis of concurrent endocrinopathy, such as thyroid disorders, hyperaldosteronism, Cushing's syndrome, other adrenal gland disorders, or malignancy based on ICD-9 codes as listed in Table 4.1.

4.2.2 Ethical approval

The study was approved by the Institutional Review Board of Boston University School of Medicine and the Harvard T.H. Chan School of Public Health (Protocol # H35708) and is considered non-human subjects research.

4.2.3 Reference label definitions

4.2.3.1 Individual predictors

Time-varying predictor variables with a date stamp before that of the outcome of interest were included in our models. We considered the following predictor variables: Socioeconomic and lifestyle demographic variables: age, race (White/ Caucasian, Black/ African American, Hispanic/ Latina, Asian, Native Hawaiian/ Pacific Islander, Middle Eastern, Other/ Unknown), smoking status (yes/no), marital status

Table 4.1: Exclusion ICD-9 Coding Associations.

Endocrinopathies	**ICD-9 Codes**
Goiter, specified as simple	240
Goiter, unspecified	240.9
Nontoxic uninodular goiter	241
Nontoxic multinodular goiter	241.1
Thyrotoxicosis with or without goiter	242
Congenital hypothyroidism	243
Acquired hypothyroidism	244
Thyroiditis	245
Other disorders of thyroid	246
Cushing's syndrome	255
Hyperaldosteronism	255.1
Adrenogenital disorders	255.2
Other corticoadrenal overactivity	255.3
Corticoadrenal insufficiency	255.4
Other adrenal hypofunction	255.5
Medulloadrenal hyperfunction	255.6
Other specified disorders of adrenal glands	255.8
Unspecified disorder of adrenal glands	255.9
Other ovarian dysfunction	256.8
Malignancy	**ICD-9 Codes**
Malignant neoplasm of corpus uteri, except isthmus	182

(single, married, separated, divorced, widowed, other), homelessness (yes/no), and highest level of education (8th grade or less, some high school, high school graduate, some college/ technical/ vocational training, graduated college/technical school/ vocational training, declined to answer, other). Anthropometrics: Body mass index (BMI, kg/m^2) was either calculated from height and weight or abstracted as the listed BMI variable associated with each visit. BMI was then categorized into three categories: normal (BMI $<$ 25 kg/m^2); overweight (BMI between 25-30 kg/m^2); and obese (BMI $>$ 30 kg/m^2). To further capture the obesity population in the absence of height/weight/BMI data, the obese category also included any patient with an ICD-9 code for unspecified obesity (278.00), morbid obesity (278.01), localized adiposity (278.1), and/or a history of gastric bypass. BMI $<$ 18.5, typically considered underweight, represented a small fraction of the total study population (1.5%) and thus did not have sufficient participants to create a separate category. Furthermore, a model to predict PCOS created by Xu et al. based on age, menstrual cycle length, BMI, AMH, testosterone, androstenedione, and follicle count did not find a significant difference in predictive effect when comparing BMI 18.5-24 to <18.5 (Xu et al., 2022). Cardiovascular health: To include blood pressure as a predictor variable, we defined a categorical hypertension variable by using systolic (SBP) and diastolic (DBP) blood pressure readings and ICD-9 diagnostic codes for unspecified essential hypertension (401.9), benign essential hypertension (401.1), and essential primary hypertension (401.0). Blood pressure was categorized into three groups: normal, defined by no ICD-9 codes for hypertension recorded and SBP $<$ 120 mmHg, and DBP $<$ 80 mmHg; elevated, defined by no ICD-9 codes for hypertension recorded and SBP was 120-129 mmHg or DBP $<$ 80 mmHg; hypertension, defined by any ICD-9 code for hypertension recorded or SBP $>$ 140 mmHg or DBP $>$ 90 mmHg.

Reproductive endocrine predictive variables: beta human chorionic gonadotropin (bHCG) level (negative bHCG $<$ 5 mIU/mL, positive bHCG $>$ 5 mIU/mL), HIV status (negative/positive), age at menarche, pelvic inflammatory disease diagnosis (614.9), history of hysterosalpingogram, and gravidity (history of present or prior pregnancy within obstetric history). Endocrine and metabolic lab values included: TSH, glycosylated hemoglobin (A1c) as a marker for diabetes, low-density lipoprotein (LDL), high density lipoprotein (HDL), and diagnosis of hypercholesterolemia (272.0). Of note, our model did not include androgen precursors such as DHEA or androstenedione as, according to Monash guidelines, these values provide limited additional information in the diagnosis of PCOS (Villarroel et al., 2015) (Teede et al., 2018).

4.2.3.2 Combined predictors

Expecting a nonlinear relationship between many reproductive hormones and a PCOS diagnosis, we used a multilayer perceptron (MLP) neural network to map follicle-stimulating hormone (FSH), luteinizing hormone (LH), sex hormone binding globulin (SHBG), and estradiol (E2) values to a composite metric we call MLP score. The MLP score was repetitively trained and the hyperparameters were tuned to generate a predictive probability associated with PCOS diagnosis for each predictive model, as described with further detail below.

4.2.3.3 Outcomes

Defining PCOS: PCOS diagnosis was assigned for any patient who had an ICD-9 code for PCOS (256.4) or met the Rotterdam criteria (ESHRE et al., 2004), according to which a positive diagnosis is made in the presence of two out of the following three features: (i) irregular menses (IM) as defined by rare menses, oligo-ovulation, or anovulation; (ii) hyperandrogenism (HA) as defined by clinical or biochemical androgen excess; and (iii) polycystic ovarian morphology (PCOM) noted on transabdominal or transvaginal ultrasound. Based on these three criteria, we defined three auxiliary variables IM, HA, and PCOM to use in the definition of our labels. PCOM was captured through diagnostic radiology text reports from ovarian ultrasound imaging for the subset that had ultrasound imaging (Cheng and Mahalingaiah, 2019). Defining Irregular Menstruation (IM): IM was defined with the following ICD-9 codes: absence of menstruation (626.0), scanty or infrequent menstruation (626.1), irregular menstrual cycle (626.4), unspecified disorders of menstruation and abnormal bleeding from female genital tract (626.9), and infertility, female associated with anovulation (628.0) (Sirmans and Pate, 2013). Defining Hyperandrogenism (HA): HA was assigned to a patient if any of the androgen lab testing for bioavailable testosterone, free testosterone, or total testosterone was greater than clinical thresholds of 11 ng/dL, 5 pg/mL, 45 ng/dL, respectively. In addition, HA was assigned if ICD-9 codes for hirsutism (704.1) or acne (706.1 or 706.0) were recorded for a patient. Defining ultrasound characteristics for polycystic ovarian morphology (PCOM): Among those with an ultrasound in this dataset, PCOM was identified on ultrasound reports using natural language processing (NLP) with complete methods detailed by Cheng and Mahalingaiah (Cheng and Mahalingaiah, 2019), to report PCOM as identified (PCOM present), unidentified (PCOM absent), or indeterminate (PCOM unidentifiable based on source report data). We considered four models to predict the following: Model I: patients with ICD-9 diagnosis of PCOS (256.4) within the EHR; Model II: patients with IM and HA without a ICD-9 PCOS code; Model III: patients with at least two out of the three conditions IM/HA/PCOM and without a ICD-9 PCOS code; Model IV: all patients meeting inclusion criteria for Model I or Model III. ICD-9 codes were abstracted from the billing code and diagnosis code associated with each encounter within the EHR. Model I included all patients who were diagnosed with PCOS. Model II and its superset Model III, which additionally includes PCOM findings on ultrasound, was composed of patients who did not have a PCOS diagnosis code but met diagnostic criteria of PCOS based on Rotterdam criteria, representing the patient population with undiagnosed PCOS. Model IV essentially captures all women who were diagnosed or met criteria for PCOS within our population. Table 4.2 details model definitions and includes the count and percent of patients in each category. The date of diagnosis was assigned by the date of PCOS ICD-9 code (256.4) for Model I, the date of the latest diagnostic criteria met for Model II and III, and the earlier date associated with Model I and Model III, for Model IV.

4.2.4 Predictive models

4.2.4.1 Classification methods

We Utilized four supervised classification methods, LR, SVM, GBM, and RF, to develop predictive models, explained in more detail in Section 2.1.1. We considered both an L1-norm and an L2-norm regularizer, explained in more detail in Section 2.1.5.1. We tuned GBM's hyperparameters through cross-validation. Explained in

Table 4.2: Definitions for Predictive PCOS models generated from the Boston Medical Center Clinical Data Warehouse.

PCOS (ICD-9 256.4)	**IM**	**HA**	**PCOM**	**Model I**	**Model II**	**Model III**	**Model IV**	**Count**	**Percent**
0	0	0	0	0	0	0	0	17,834	58.3
1	0	0	0	1	exc	exc	1	313	1
0	1	0	0	0	0	0	0	5,012	16.4
0	1	1	0	0	0	0	0	5,035	16.5
0	0	1	0	0	0	0	0	275	0.8
1	1	0	0	1	exc	exc	1	248	0.8
1	0	1	0	1	exc	exc	1	376	1.2
1	1	1	0	1	exc	exc	1	5	0
0	1	0	1	1	exc	exc	1	1,037	3.4
0	1	1	1	1	exc	exc	1	35	0.1
0	0	1	1	1	exc	exc	1	25	0.1
1	1	0	1	1	exc	exc	1	368	1.2
1	0	1	1	1	exc	exc	1	4	0
1	1	1	1	1	exc	exc	1	9	0
0	1	1	1	1	exc	exc	1	19	0.1
1	0	1	1	1	exc	exc	1	6	0
Positive Label				1,329	1,056	1,116	2,445		
Total Patients								29,485	100

exc = exclude This table shows all possible combinations of presence/absence of variables and how they were included in each model. 0 = absent, 1 = present, exclude = excluded from model. IM = irregular menstruation; HA = hyperandrogenism; PCOM = polycystic ovary morphology.

more detail in Section 2.1.3.1.

4.2.4.2 Performance metrics

To assess model performance, we use AUC-ROC and weighted F1-score. Explained in more detail in Section 2.1.4.

4.2.4.3 Pre-processing and Statistical feature selection (SFS)

We performed data pre-processing steps, explauned in more detail in Section 2.1.2 to prepare the dataset for developing predictive models. A summary of the missing variables for each model is provided in Table 4.3. The threshold for correlation coefficient is considered 0.8. Furthermore, statistical feature selection (SFS), as explained in more detail in Section 2.1.5.2, is done with 0.01 threshold for p-value. Representative aggregated patient-level statistics for each model are shown in Table 4.4 and 4.5. Highly correlated variables and the retained variable are provided in Table 4.6.

Table 4.3: Summary of missingness in continuous variables.

Variable	**Model I**		**Model II**		**Model III**		**Model IV**	
	# of missing	Imputed value	# of missing	Imputed value	# of missing	Imputed value	# of missing	Imputed value
Testosterone	27365	29.8	27201	29.0	27251	29.0	28115	32.0
Free Testosterone	26975	3.0	26843	3.0	26887	3.0	27682	3.3
Bioavailable Testosterone	26987	6.05	26857	6.2	26901	6.0	27698	6.7
Gravidity	11282	2.0	11141	2.0	11148	2.0	11537	2.0
Age at Menarche	13120	12.0	12984	12.0	12996	12.0	13430	12.0
Total Cholesterol	17905	172.0	17688	172.0	17710	172.0	18350	172.0
HDL	18104	51.0	17887	51.0	17909	51.0	18562	51.0
LDL	18955	101.0	18746	101.0	18772	101.0	19508	101.0
TSH	20604	1.22	20303	1.21	20339	1.21	21296	1.2
A1C	23139	5.4	22949	5.4	22995	5.4	23954	5.4
FSH	24338	5.0	24205	5.1	24236	5.1	24893	5.0
LH	25941	6.3	25823	6.1	25862	6.1	26641	6.5
SHBG	27181	39.0	27049	41.0	27094	41.0	27942	38.0
Estradiol	27331	58.0	27072	58.0	27120	58.0	28235	58.0

HDL = high-density lipoprotein; LDL = low-density lipoprotein; TSH = thyroid stimulating hormone; FSH = follicle stimulating hormone; LH = luteinizing hormone; SHBG = sex hormone-binding globulin.

Table 4.4: Variables before statistical feature selection (SFS) for each model. Part 1.

Variable	**Model I**			**Model II**			**Model III**			**Model IV**		
	Y1-mean	Y0-mean	p-value	Y1-mean	Y0-mean	p-value	Y1-mean	Y0-mean	p-value	Y1-mean	Y0-mean	p-value
Gravidity	1.28	2.08	4.55E-78	1.42	2.08	2.35E-63	1.46	2.08	2.47E-63	1.36	2.08	2.17E-135
LH	8.03	6.53	3.65E-69	7.72	6.35	5.07E-35	7.69	6.35	2.81E-36	8.04	6.70	4.47E-94
SHBG	38.48	39.90	1.87E-58	40.03	41.75	4.89E-32	40.01	41.75	1.87E-32	37.84	38.94	8.51E-78
FSH	4.94	5.24	9.46E-44	5.28	5.35	3.04E-30	5.26	5.34	1.42E-32	5.04	5.22	1.55E-68
Obesity	0.51	0.27	1.38E-81	0.34	0.27	9.60E-06	0.34	0.27	2.73E-06	0.43	0.27	2.86E-66
Positive bHCG	0.05	0.23	1.50E-48	0.10	0.23	4.14E-21	0.11	0.23	3.59E-20	0.08	0.23	2.23E-65
Age	31.34	33.79	1.70E-25	31.01	33.79	2.26E-31	31.16	33.79	5.91E-30	31.26	33.79	1.91E-52
Obese BMI	0.45	0.25	7.99E-57	0.30	0.25	1.08E-03	0.30	0.25	7.69E-04	0.38	0.25	1.03E-44
HDL	50.21	51.58	1.15E-14	52.13	51.59	1.03E-05	52.04	51.59	4.06E-12	51.04	51.58	3.65E-25
Negative bHCG	0.26	0.23	1.99E-22	0.37	0.23	1.44E-22	0.37	0.23	2.20E-25	0.31	0.23	2.29E-65
Total Cholesterol	174.77	173.12	1.70E-12	173.37	173.14	2.82E-10	173.13	173.12	7.72E-11	174.14	173.11	1.48E-22
Hypertension	0.31	0.21	6.02E-14	0.25	0.21	7.25E-06	0.21	0.21	8.32E-02	0.31	0.21	3.63E-12
LDL	102.51	101.51	2.10E-03	101.28	101.01	1.66E-06	101.15	101.01	6.31E-07	101.91	101.51	4.06E-27
Hispanic/Latina Race	0.07	0.10	1.82E-01	0.06	0.10	2.69E-06	0.06	0.10	2.00E-03	0.07	0.10	2.34E-05
Estradiol	60.70	59.37	7.32E-03	61.40	59.39	9.91E-04	61.58	59.39	3.78E-04	61.11	59.38	8.21E-05
Education – Some College	0.18	0.15	3.44E-02	0.19	0.15	3.32E-03	0.19	0.15	4.65E-03	0.18	0.15	1.55E-04
Smoker	0.09	0.14	6.62E-05	0.12	0.14	4.11E-01	0.12	0.14	5.73E-01	0.11	0.14	3.00E-04
TSH	1.31	1.26	5.24E-02	1.28	1.26	8.51E-03	1.28	1.26	3.51E-03	1.30	1.26	8.49E-03
Elevated BP	0.12	0.10	2.85E-02	0.11	0.10	4.29E-01	0.11	0.10	3.35E-01	0.12	0.10	9.50E-03
Marital Status: Single	0.77	0.76	6.82E-01	0.81	0.76	1.31E-03	0.81	0.76	2.78E-03	0.79	0.76	9.87E-03
Gastric Bypass History	0.00	0.01	6.50E-02	0.00	0.01	3.31E-01	0.00	0.01	2.68E-01	0.00	0.01	1.25E-02
Age at Menarche	12.10	12.23	8.95E-02	12.18	12.23	3.10E-01	12.18	12.23	1.88E-01	12.13	12.23	1.46E-02
Overweight ICD-9 278.02	0.04	0.03	2.58E-01	0.04	0.03	2.96E-01	0.04	0.03	1.50E-01	0.04	0.03	3.30E-02
Normal BMI	0.15	0.26	3.57E-16	0.34	0.26	8.57E-07	0.33	0.26	6.76E-06	0.23	0.26	4.80E-02

Table 4.5: Variables before statistical feature selection (SFS) for each model. Part 2.

Variable	**Model I**			**Model II**			**Model III**			**Model IV**		
	Y1-mean	Y0-mean	p-value	Y1-mean	Y0-mean	p-value	Y1-mean	Y0-mean	p-value	Y1-mean	Y0-mean	p-value
A1C	5.43	5.42	1.19E-01	5.42	5.42	2.45E-01	5.42	5.42	3.19E-01	5.43	5.42	5.26E-01
Education: Declined	0.03	0.04	6.85E-01	0.02	0.04	5.52E-01	0.02	0.04	5.69E-02	0.03	0.04	6.20E-02
Marital Status: Separated	0.01	0.01	8.11E-01	0.01	0.01	4.24E-01	0.01	0.01	4.77E-01	0.01	0.01	7.94E-01
Black/African American Race	0.40	0.40	1.00E+00	0.45	0.40	1.84E-02	0.46	0.40	2.03E-03	0.43	0.40	1.05E-01
Marital Status: Married	0.19	0.20	8.72E-01	0.16	0.20	3.69E-02	0.17	0.20	6.70E-02	0.18	0.20	1.30E-01
Race: Other	0.04	0.03	2.54E-01	0.04	0.03	4.69E-01	0.04	0.03	6.82E-01	0.04	0.03	1.64E-01
Marital Status: Widowed	0.00	0.00	6.47E-01	0.00	0.00	7.26E-01	0.00	0.00	7.08E-01	0.00	0.00	3.85E-01
Normal BP	0.46	0.51	1.74E-02	0.60	0.51	1.37E-07	0.60	0.51	3.94E-08	0.53	0.51	3.92E-01
White/Caucasian Race	0.26	0.27	9.77E-01	0.25	0.27	8.19E-01	0.25	0.27	5.44E-01	0.25	0.27	6.43E-01
Education: high school graduate	0.25	0.24	9.12E-01	0.25	0.24	8.19E-01	0.26	0.24	7.47E-01	0.25	0.24	6.63E-01
Human immuno virus ICD-9 042	0.00	0.01	2.42E-01	0.01	0.01	9.97E-01	0.00	0.01	9.80E-01	0.00	0.01	7.04E-01
Education: Some high school	0.17	0.20	2.40E-01	0.20	0.20	9.97E-01	0.20	0.20	9.75E-01	0.19	0.20	7.09E-01
Education: I did not attend school	0.03	0.04	6.51E-01	0.04	0.04	9.65E-01	0.04	0.04	9.95E-01	0.04	0.04	7.48E-01
Marital Status: Divorced	0.01	0.01	1.00E+00	0.01	0.01	4.77E-01	0.01	0.01	5.34E-01	0.01	0.01	7.68E-01
Education: Pre-registration	0.01	0.01	2.33E-01	0.01	0.01	9.16E-01	0.01	0.01	8.72E-01	0.01	0.01	8.21E-01
Education: 8th grade or less	0.01	0.01	9.70E-01	0.01	0.01	8.63E-01	0.01	0.01	8.46E-01	0.01	0.01	8.23E-01
Education: other	0.01	0.01	9.97E-01	0.01	0.01	8.14E-01	0.01	0.01	7.51E-01	0.01	0.01	8.50E-01
FemPelv.Inflamm.DisNOS ICD-9 614.9	0.00	0.01	1.60E-01	0.01	0.01	9.84E-01	0.01	0.01	6.27E-01	0.01	0.01	1.00E+00
American Indian/Native American	0.00	0.01	9.67E-01	0.00	0.01	9.88E-01	0.00	0.01	9.72E-01	0.00	0.01	9.24E-01
Homeless Indicator No	0.99	0.98	9.17E-01	0.98	0.98	1.00E+00	0.98	0.98	9.97E-01	0.99	0.98	9.33E-01
Homeless Indicator Yes	0.01	0.02	9.23E-01	0.02	0.02	1.00E+00	0.02	0.02	9.98E-01	0.01	0.02	9.40E-01
Middle Eastern Race	0.01	0.01	1.00E+00	0.01	0.01	7.67E-01	0.01	0.01	8.45E-01	0.01	0.01	9.42E-01
Education: graduated college	0.20	0.21	9.08E-01	0.21	0.21	1.00E+00	0.21	0.21	1.00E+00	0.20	0.21	9.59E-01
Asian Race	0.04	0.05	7.03E-01	0.05	0.05	7.76E-01	0.05	0.05	9.39E-01	0.04	0.05	9.79E-01
Native Hawaiian/Pacific Islander Race	0.00	0.00	8.49E-01	0.00	0.00	9.78E-01	0.00	0.00	9.85E-01	0.00	0.00	9.86E-01
Pure Hypercholesterolemia ICD-9 272.0	0.01	0.01	1.00E+00	0.01	0.01	9.94E-01	0.01	0.01	9.74E-01	0.01	0.01	9.86E-01
Overweight BMI	0.18	0.19	4.58E-01	0.22	0.19	3.32E-01	0.19	0.19	1.62E-01	0.19	0.19	9.90E-01
History of Hysterosalpingogram	0.05	0.06	1.00E+00	0.06	0.06	9.86E-01	0.06	0.06	9.93E-01	0.05	0.06	1.00E+00
Marital Status: Other	0.02	0.02	8.25E-01	0.02	0.02	8.48E-01	0.02	0.02	7.48E-01	0.02	0.02	1.00E+00

Table 4.6: Highly correlated variables and the retained variable.

Highly correlated pairs	**The variable we kept among highly correlated pairs**
Marital Status – Married Marital Status – Single	Marital Status – Married
Homeless indicator – No Homeless indicator – Yes	Homeless Indicator – Yes
Obesity Obese BMI	Obesity
Total Cholesterol	Total Cholesterol kept in Model I and Model IV
LDL	LDL kept in Model II and Model III

BMI = body mass index (kg/m^2); LDL = low-density lipoprotein.

4.2.4.4 Training-test splitting

We split the dataset into five random parts, where four parts were used as the training set, and the remaining part was used for testing. We used the training set to tune the model hyperparameters via 5-fold cross-validation, and we evaluated the performance metrics on the testing set. We repeated training and testing five times, each time with a different random split into training/test sets. The mean and standard deviation of the metrics on the test sets over the five repetitions are reported. Explained in more detail in Section 2.1.3.

4.2.5 Development of the MLP score

For every model, there was a considerable difference between the AUC of linear models and non-linear models. To improve the performance of our linear models, we utilized nonlinear models to capture intricate relationships between features. We utilized Gradient Boosted Trees (GBT) to find which features most commonly appeared together among decision trees. We found FSH, LH, SHBG, and estradiol levels to be a meaningful group of features which are all reproductive hormones and continuous variables that appeared together among trees for all our models. We subsequently used these four features as input features into a multilayer perceptron (MLP) neural network model with three hidden layers, each employing the rectified linear unit (ReLU) activation function. The neural network was trained using the training set to classify PCOS. We used the output probability of the MLP model, which we called "MLP score," as a new feature into our original predictive models.

4.2.6 Recursive feature elimination (RFE)

We also used a recursive feature elimination approach with L1-penalized logistic regression (L1-regularized RFE) to extract the most informative features and develop parsimonious models. Explained in more detail in Section 2.1.5.3.

4.2.7 Final predictive models

We computed the performance of the following models: L1-penalized logistic regression (LR-L1), support vector machine (SVM-L1), random forest (RF), and gradient boosted machine. We calculated each variable's LR coefficient with a 95% confidence interval (β [95%CI]), the correlation of the variable with the outcome (Y-correlation), the p-value of each variable (p-value), the mean of the variable (Y1-mean) in the PCOS labeled patients, the mean of the variable (Y0-mean) in the patients without the PCOS label, and the mean and standard deviation of the variable over all patients (All-mean and All-SD). Ranking predictor variables by the absolute value of their coefficients in the logistic regression model amounts to ranking these variables by how much they affect the predicted probability of the outcome. A positive coefficient implies that the larger the value of the variable within the range specified by the data, the higher the chance of having a PCOS diagnosis as defined by the model outcome.

4.3 Results

4.3.1 Results of data acquisition and data pre-processing

After inclusion and exclusion criteria were applied to all 65,431 women within the initial data pool, 30,601 patient records were available for this analysis and defined populations are included in Figure 4·1. There were 1,329 patients (4.5%) with a PCOS ICD-9 diagnosis code (Model I). 1,465 patients had records with PCOM results as present, absent, or unidentifiable. There were 1,056 patients (3.6%) with no ICD 256.4 indication and presence of IM and HA (Model II). There were 1,116 (3.8%) patients with no ICD 256.4 indication and presence of at least two out of three criteria of IM, HA, or PCOM (Model III). Finally, there were 2,445 PCOS patients (8.0%) in the combined analysis of Model I and III (Model IV). A tabulation of which subjects were included in each model based on all possible combinations of the presence or absence of each variable is shown in Table 2. In the total cohort, the patients were predominantly Black/African American (40.3%) and White (26.5%), with an average age of 33.6 years (SD = 6.6). Complete demographic characteristics are described in Table 4.7.

There were 43 categorical variables and 12 continuous variables retained as predictors after the data pre-processing procedures. There were four pairs of highly correlated variables and one variable from each correlated pair included in the final model as noted in Table 4.6. Table 4.4 and 4.5 describes all 51 variables used by the predictive models.

4.4 Model Performance

Tables 4.7, 4.8, 4.9 and4.10 display the parsimonious models that use the MLP score (LR-L2-MLP score) and show the most significant variables in the prediction of the outcome for Models I, II, III, and IV, respectively. All p-values were less than 0.05, which was set as the significance level. Feature importance graphs based on logistic regression coefficients (± 95% confidence interval) are visualized in Figure 4·2.

For Model I, the parsimonious predictive model achieved an AUC (SD) of 82.3% (1.7). The MLP score (β = 0.71) and obesity (β = 0.45) were positively correlated with PCOS diagnosis. Pregnancy (gravidity β = -0.53; positive pregnancy test β = -0.50), normal BMI (β = -0.24), smoking (β = -0.18), age (β = -0.16), and Hispanic race (β = -0.10) were inversely correlated with PCOS diagnosis as shown in Table 4.8.

For Model II, the parsimonious predictive model achieved an AUC (SD) of 77.6% (1.3). The MLP score (β = 0.61), obesity (β = 0.21), normal BMI (β = 0.15), normal blood pressure (β = 0.16), negative pregnancy test (β = 0.12), and normal HDL (β = 0.08) were positively correlated with undiagnosed PCOS. Age (β = -0.27), pregnancy (gravidity β = -0.26; positive pregnancy test β = -0.19), and Hispanic race (β = -0.18) were inversely correlated with undiagnosed PCOS as show in Table 4.9.

For Model III, the parsimonious predictive model achieved an AUC (SD) of 77.4% (1.6). The MLP score (β = 0.60), obesity (β = 0.19), normal blood pressure (β = 0.17), normal BMI (β = 0.14), Black race (0.13), negative pregnancy test (β = 0.12), and normal HDL (β = 0.09) were positively correlated with undiagnosed PCOS. Age (β = -0.25), pregnancy (gravidity β = -0.24; positive pregnancy test β = -0.20), and Hispanic race (β = -0.15) were inversely correlated with undiagnosed PCOS as show in Table 4.10. For Model IV, the parsimonious predictive model achieved an AUC (SD) of 79.1% (1.1). The MLP score (β = 0.7), obesity (β = 0.31), normal BMI (β = 0.15), hypertension (β = 0.07) and some higher degree of education, such as college or vocational/technical school (β = 0.06) were positively correlated with PCOS diagnosis. Age (β = -0.21), pregnancy (gravidity β = -0.37; positive pregnancy test β = -0.34; negative pregnancy test β = -0.05), Hispanic race (β = -0.12), and smoking (β = -0.08) were inversely correlated with PCOS diagnosis as shown in Table 4.11.

GBT models had the highest performance. Predictions of PCOS in a test set of patients not used during algorithm training achieved 85%, 81%, 80%, and 82% AUC for Models I, II, III, and IV, respectively. We also report the performance with the logistic regression model (LR-L1) after SFS and the performance when using our developed MLP score alongside variables selected via recursive feature elimination (LR-L2-MLP score). Table 4.12 and Table 4.13 displays features for each model, associated with LR-L1 algorithm after SFS. As we hypothesized, developing models using the MLP score (LR-L2-MLP score) leads to improvement of the performance of linear models (LR-L1) for Models I, II, III, and IV, respectively from 79%, 72%, 73%, and 75% AUC to 82%, 78%, 77%, and 79% AUC. Table 4.14 details the models with the best performance (highest AUC) using all 51 features before and after statistical feature selection (SFS). In Table 4.14, the means and standard deviations of AUC and weighted-F1 scores on the test set over the five repetitions are listed. One of these repetitions in shown in Figure 4·3, which shows the ROC curves pertaining to the parsimonious models utilizing the MLP score (LR-L2-MLP score). Table 4.15 displays the performance of all models and all algorithms, before and after statistical feature selection (SFS). The feature importance in GBT models after SFS is visualized in Figure 4·4. Of note, it quantifies the extent to which a feature is used for making decisions within the ensemble of decision trees but does not show directionality.

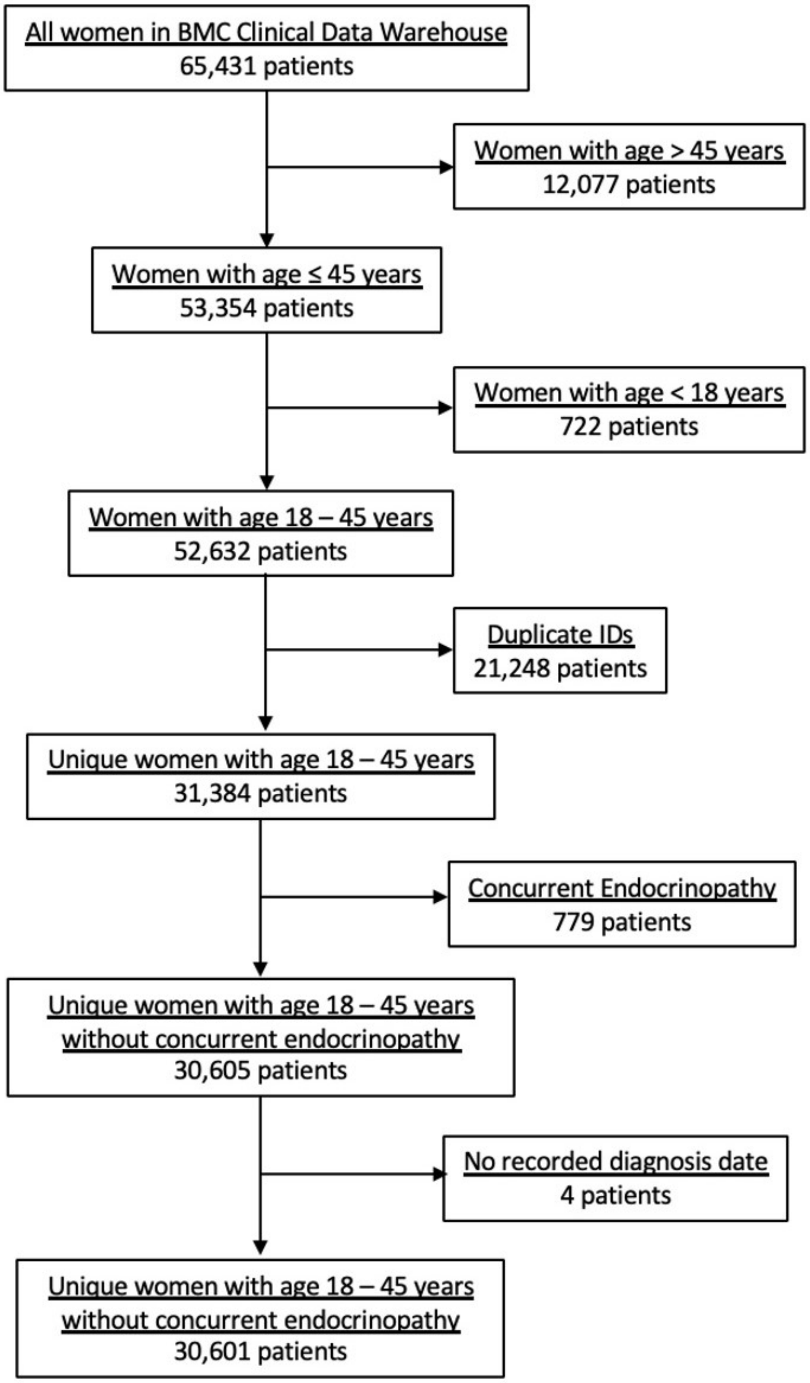

All women in BMC Clinical Data Warehouse
65,431 patients
Women with age > 45 years
12,077 patients
Women with age ≤ 45 years
53,354 patients
Women with age < 18 years
722 patients
Women with age 18 – 45 years
52,632 patients
Duplicate IDs
21,248 patients
Unique women with age 18 – 45 years
31,384 patients
Concurrent Endocrinopathy
779 patients
Unique women with age 18 – 45 years
without concurrent endocrinopathy
30,605 patients
No recorded diagnosis date
4 patients
Unique women with age 18 – 45 years
without concurrent endocrinopathy
30,601 patients

Figure 4·1: Flow of patients from the BMC CDW into the dataset used by the study.

Table 4.7: Demographic characteristics of the study population and by model.

Variable	Model I	Model II	Model III	Model IV
Age, Mean years (SD)	33.6 (6.6)	33.7 (6.6)	33.7 (6.6)	33.6 (6.6)
Race, n (%)				
Black/African American	11881 (40.3)	11824 (40.5)	11861 (40.5)	12395 (40.5)
White/Caucasian	7812 (26.5)	7733 (26.5)	7741 (26.4)	8086 (26.4)
Hispanic/Latina	2858 (9.7)	2837 (9.7)	2841 (9.7)	2929 (9.6)
Asian	1350 (4.6)	1354 (4.6)	1354 (4.6)	1406 (4.6)
Middle Eastern	175 (0.6)	176 (0.6)	176 (0.6)	184 (0.6)
American Indian/Native American	163 (0.6)	162 (0.6)	162 (0.6)	168 (0.5)
Native Hawaiian/Pacific Islander	17 (0.1)	18 (0.1)	18 (0.1)	18 (0.1)
Other	979 (3.3)	966 (3.3)	966 (3.3)	1023 (3.3)
Unknown	4250 (14.41)	4146 (14.19)	4153 (14.19)	4392 (14.4)
Marital Status				
Single	22325 (75.7)	22155 (75.8)	22199 (75.8)	23224 (75.9)
Married	5833 (19.8)	5753 (19.7)	5767 (19.7)	6018 (19.7)
Separated	392 (1.3)	391 (1.3)	392 (1.3)	401 (1.3)
Divorced	388 (1.3)	379 (1.3)	380 (1.3)	397 (1.3)
Widowed	35 (0.1)	35 (0.1)	35 (0.1)	35 (0.1)
Other	502 (1.7)	489 (1.7)	489 (1.7)	516 (1.7)
Unknown	10 (0.03)	10 (0.03)	10 (0.03)	10 (0.03)
Body Mass Index (BMI), kg/m^2				
Normal (BMI $<$ 25)	7534 (25.6)	7685 (26.3)	7697 (26.3)	7902 (25.8)
Overweight (BMI between 25-30)	5694 (19.3)	5689 (19.5)	5707 (19.5)	5941 (19.4)
Obese (BMI $\geq$ 30)	7645 (25.9)	7369 (25.2)	7387 (25.2)	7985 (26.1)
Unknown	8612 (29.2)	8469 (29.0)	8481 (29.0)	8773 (28.7)

Model I, PCOS ICD-9 diagnosis within the EHR; Model II, irregular menstruation and hyperandrogenism without ICD-9 PCOS code; Model III, at least two out of the three conditions (irregular menstruation, hyperandrogenism, or polycystic ovary morphology on ultrasound) and without ICD-9 PCOS code; Model IV, meets inclusion criteria for Model I or Model III.

Table 4.8: Most significant variables for PCOS diagnosis prediction in Model I.

Variables	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$ - %95 CI	Y-correlation	p-value	Y1-mean	Y0-mean	All-mean	All-std
MLP Score	0.71	0.028	0.33	6.80E-197	0.17	0.04	0.05	0.08
Intercept	-0.68	–	–	–	–	–	–	–
Gravidity	-0.53	0.018	-0.12	4.55E-78	1.28	2.08	2.04	1.39
Positive bHCG	-0.50	0.019	-0.09	1.50E-48	0.05	0.23	0.22	0.42
Obesity	0.45	0.017	0.11	1.38E-81	0.51	0.27	0.28	0.45
Normal BMI	-0.24	0.017	-0.05	3.57E-16	0.15	0.26	0.26	0.44
Smoker	-0.18	0.017	-0.03	6.62E-05	0.09	0.14	0.14	0.34
Age	-0.16	0.016	-0.08	1.70E-25	31.34	33.79	33.68	6.61
Hispanic/Latina Race	-0.10	0.016	-0.02	1.82E-03	0.07	0.10	0.10	0.30

bHCG, beta-human chorionic gonadotropin.

Table 4.9: Most significant variables for PCOS diagnosis prediction in Model II.

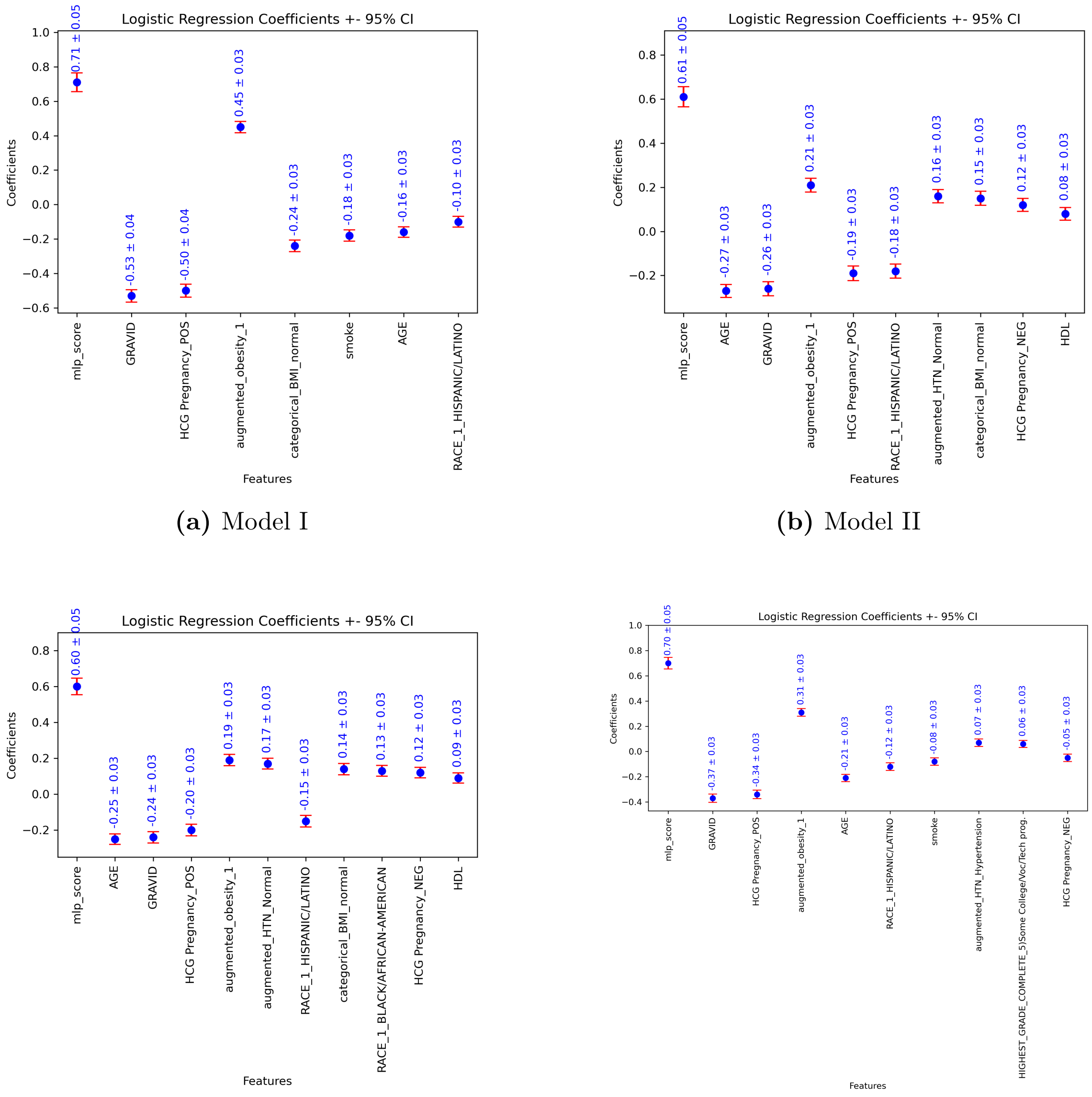
Variables	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$ - %95 CI	Y-correlation	p-value	Y1-mean	Y0-mean	All-mean	All-std
MLP Score	0.61	0.023	0.26	2.13E-142	0.12	0.04	0.04	0.06
Intercept	-0.44	–	–	–	–	–	–	–
Age	-0.27	0.015	-0.08	2.26E-31	31.01	33.79	33.69	6.61
Gravidity	-0.26	0.016	-0.09	2.35E-63	1.42	2.08	2.06	1.39
Obesity	0.21	0.016	0.03	9.60E-06	0.34	0.27	0.27	0.44
Positive bHCG	-0.19	0.017	-0.06	4.14E-21	0.10	0.23	0.23	0.42
Hispanic/Latina Race	-0.18	0.016	-0.02	2.69E-03	0.06	0.10	0.10	0.30
Normal BP	0.16	0.015	0.03	1.37E-07	0.60	0.51	0.51	0.50
Normal BMI	0.15	0.016	0.03	8.57E-07	0.34	0.26	0.26	0.44
Negative bHCG	0.12	0.015	0.06	1.44E-22	0.37	0.23	0.23	0.42
HDL	0.08	0.015	0.01	1.03E-10	52.13	51.59	51.61	7.86

BP, blood pressure; BMI, body mass index (kg/m^2); HDL, high-density lipoprotein; bHCG, beta-human chorionic gonadotropin.

Table 4.10: Most significant variables for PCOS diagnosis prediction in Model III.

Variables	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$ - %95 CI	Y-correlation	p-value	Y1-mean	Y0-mean	All-mean	All-std
MLP Score	0.60	0.023	0.26	7.41E-142	0.10	0.04	0.04	0.05
Intercept	–	–	–	–	–	–	–	–
Age	-0.25	0.015	-0.08	5.91E-30	31.16	33.79	33.69	6.61
Gravidity	-0.24	0.016	-0.09	2.47E-63	1.46	2.08	2.06	1.39
Positive bHCG	-0.20	0.017	-0.06	3.59E-20	0.11	0.23	0.23	0.42
Obesity	0.19	0.016	0.03	2.73E-06	0.34	0.27	0.27	0.44
Normal BP	0.17	0.015	0.04	3.94E-08	0.60	0.51	0.51	0.50
Hispanic/Latina Race	-0.15	0.016	-0.02	2.00E-03	0.06	0.10	0.10	0.30
Normal BMI	0.14	0.016	0.03	6.76E-06	0.33	0.26	0.26	0.44
Black/African American Race	0.13	0.015	0.02	2.03E-03	0.46	0.40	0.41	0.49
Negative bHCG	0.12	0.015	0.06	2.20E-25	0.37	0.23	0.23	0.42
HDL	0.09	0.015	0.01	4.06E-12	52.04	51.59	51.61	7.86

BP, blood pressure; BMI, body mass index (kg/m^2); HDL, high-density lipoprotein; bHCG, beta-human chorionic gonadotropin.



(a) Model I

(b) Model II

(c) Model III

(d) Model IV

Figure 4·2: Feature importance graphs based on logistic regression coefficients (± 95% confidence interval), associated with parsimonious models utilizing the MLP score (LR-L2-MLP score). The absolute value of the logistic regression coefficients shows how much the variable affects the predicted probability of the outcome. A positive/negative coefficient implies that the larger the absolute value of the variable, the higher/lower the chance of having a PCOS diagnosis as defined by the model outcome.

Table 4.11: Most significant variables for PCOS diagnosis prediction in Model IV.

Variables	$\boldsymbol{\beta}$	$\boldsymbol{\beta}$ - %95 CI	Y-correlation	p-value	Y1-mean	Y0-mean	All-mean	All-std
MLP Score	0.70	0.024	0.36	0.00E-01	0.20	0.07	0.08	0.10
Intercept	-0.44	–	–	–	–	–	–	–
Gravidity	-0.37	0.017	-0.14	2.17E-135	1.36	2.08	2.02	1.39
Positive bHCG	-0.34	0.017	-0.10	2.23E-65	0.08	0.23	0.22	0.41
Obesity	0.31	0.015	0.10	2.86E-66	0.43	0.27	0.28	0.45
Age	-0.21	0.015	-0.10	1.91E-52	31.26	33.79	33.59	6.62
Hispanic/Latina Race	-0.12	0.015	-0.03	2.34E-06	0.07	0.10	0.10	0.29
Smoker	-0.08	0.015	-0.02	3.00E-04	0.11	0.14	0.14	0.34
Hypertension	0.07	0.015	0.04	3.63E-12	0.28	0.21	0.22	0.41
Education – Some College/Technical /Vocational School	0.06	0.014	0.03	1.55E-04	0.18	0.15	0.15	0.36
Negative bHCG	-0.05	0.015	0.05	2.29E-16	0.31	0.23	0.24	0.42

bHCG, beta-human chorionic gonadotropin.

Table 4.12: Features for each model, associated with LR-L1 algorithm after SFS.

			Model I				
Variables	β	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
Intercept	-0.62	-	-	-	-	-	-
FSH	-0.6	-0.03	9.46E-44	4.94	5.24	5.23	2.30
LH	0.54	0.12	3.65E-69	8.03	6.53	6.40	2.70
Positive bHCG	-0.54	-0.09	1.50E-48	0.05	0.23	0.22	0.42
Gravidity	-0.53	-0.12	4.55E-78	1.28	2.08	2.04	1.39
Obesity	0.49	0.11	1.38E-81	0.51	0.27	0.28	0.45
Age	-0.25	-0.08	1.70E-25	31.34	33.79	33.68	6.61
Normal BMI	-0.22	-0.05	3.57E-16	0.15	0.26	0.26	0.44
Smoker	-0.18	-0.03	6.62E-05	0.09	0.14	0.14	0.34
Total Cholesterol	0.11	0.02	1.70E-06	174.77	173.12	173.19	20.85
Hispanic/Latina Race	-0.11	-0.02	1.82E-03	0.07	0.10	0.10	0.30
Estradiol	0.08	0.02	7.32E-03	60.70	59.37	59.43	16.54
HDL	-0.08	-0.04	1.15E-14	50.21	51.58	51.52	7.78
Hypertension	0.08	0.05	6.02E-14	0.31	0.21	0.22	0.41
SHBG	0.01	-0.03	1.87E-58	38.48	39.90	39.83	9.94
			Model II				
Variables	β	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
Age	-0.31	-0.08	2.26E-31	31.01	33.79	33.69	6.61
Intercept	-0.31	-	-	-	-	-	-
LH	0.29	0.09	5.07E-35	7.72	6.35	6.40	2.73
Gravidity	-0.29	-0.09	2.35E-63	1.42	2.08	2.06	1.39
Obesity	0.29	0.03	9.60E-06	0.34	0.27	0.27	0.44
Positive bHCG	-0.2	-0.06	4.14E-21	0.10	0.23	0.23	0.42
Negative bHCG	0.19	0.06	1.44E-22	0.37	0.23	0.23	0.42
FSH	-0.16	0.00	3.04E-30	5.28	5.35	5.34	2.51
Normal BMI	0.15	0.03	8.57E-07	0.34	0.26	0.26	0.44
Hispanic/Latina Race	-0.13	-0.02	2.69E-03	0.06	0.10	0.10	0.30
Normal BP	0.13	0.03	1.37E-07	0.60	0.51	0.51	0.50
SHBG	-0.08	-0.03	4.89E-32	40.03	41.75	41.68	9.23
HDL	0.06	0.01	1.03E-10	52.13	51.59	51.61	7.86
Education – Some College	0.06	0.02	3.32E-03	0.19	0.15	0.15	0.36
Estradiol	0.04	0.02	9.91E-04	61.40	59.39	59.46	16.82
LDL	0.01	0.00	1.66E-06	101.28	101.51	101.50	15.34
TSH	0.01	0.01	8.51E-03	1.28	1.26	1.26	0.41

Education – Some College = Education – Some College/Technical/Vocational School
HDL = high-density lipoprotein; LDL = low-density lipoprotein; TSH = thyroid stimulating hormone; FSH = follicle stimulating hormone; LH = luteinizing hormone; SHBG = sex hormone binding globulin; BMI = body mass index; b-HCG = beta-human chorionic gonadotropin; BP = blood pressure.

Table 4.13: Features for each model, associated with LR-L1 algorithm after SFS.

Model III							
Variables	β	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
Intercept	-0.29	-	-	-	-	-	-
LH	0.29	0.09	2.81E-36	7.69	6.35	6.40	2.73
Age	-0.28	-0.08	5.91E-30	31.16	33.79	33.69	6.61
Gravidity	-0.27	-0.086	4.00E+00	1.46E+00	2.08	2.06	1.39
Obesity	0.27	0.03	2.73E-06	0.34	0.27	0.27	0.44
Negative HCG	0.22	0.06	2.20E-25	0.37	0.23	0.23	0.42
Positive HCG	-0.2	-0.06	3.59E-20	0.11	0.23	0.23	0.42
FSH	-0.17	-0.01	1.42E-32	5.26	5.34	5.34	2.49
Normal BP	0.15	0.04	3.94E-08	0.60	0.51	0.51	0.50
Normal BMI	0.13	0.03	6.76E-06	0.33	0.26	0.26	0.44
Black/African American Race	0.12	0.02	2.03E-03	0.46	0.40	0.41	0.49
Hispanic/Latina Race	-0.1	-0.02	2.00E-03	0.06	0.10	0.10	0.30
SHBG	-0.07	-0.04	1.87E-32	40.01	41.75	41.68	9.23
HDL	0.06	0.01	4.06E-12	52.04	51.59	51.61	7.86
Education – Some College	0.06	0.02	4.65E-03	0.19	0.15	0.15	0.36
Estradiol	0.05	0.02	3.78E-04	61.58	59.39	59.47	16.88
TSH	0.04	0.01	3.51E-03	1.28	1.26	1.26	0.41
LDL	0.01	0.00	6.31E-07	101.37	101.51	101.50	15.34

Model IV							
Variables	β	**Y-correlation**	**p-value**	**Y1-mean**	**Y0-mean**	**All-mean**	**All-std**
LH	0.39	0.13	4.47E-94	8.04	6.70	6.81	2.78
Gravidity	-0.39	-0.14	2.17E-135	1.36	2.08	2.02	1.39
Obesity	0.38	0.10	2.86E-66	0.43	0.27	0.28	0.45
Intercept	-0.35	-	-	-	-	-	-
Positive HCG	-0.35	-0.10	2.23E-65	0.08	0.23	0.22	0.41
FSH	-0.3	-0.02	1.55E-68	5.04	5.22	5.20	2.09
Age	-0.28	-0.10	1.91E-52	31.26	33.79	33.59	6.62
Hispanic/Latina Race	-0.11	-0.03	2.34E-06	0.07	0.10	0.10	0.29
Hypertension	0.1	0.04	3.63E-12	0.28	0.21	0.22	0.41
Smoker	-0.09	-0.02	3.00E-04	0.11	0.14	0.14	0.34
Estradiol	0.06	0.03	3.49E-06	61.11	59.38	59.52	17.10
Education – Some College	0.06	0.03	1.55E-04	0.18	0.15	0.15	0.36
Total Cholesterol	0.06	0.01	1.48E-15	174.10	173.12	173.19	21.00
Elevated BP	0.06	0.02	9.50E-03	0.12	0.10	0.10	0.30
Negative HCG	0.05	0.05	2.29E-16	0.31	0.23	0.24	0.42
HDL	-0.01	-0.02	3.65E-25	51.04	51.58	51.54	7.88
SHBG	-0.01	-0.03	8.51E-78	37.84	38.94	38.85	10.03
TSH	0.01	0.02	8.49E-03	1.30	1.26	1.27	0.42

Education – Some College = Education – Some College/Technical/Vocational School

HDL = high-density lipoprotein; LDL = low-density lipoprotein; TSH = thyroid stimulating hormone; FSH = follicle stimulating hormone; LH = luteinizing hormone; SHBG = sex hormone binding globulin; BMI = body mass index; b-HCG = beta-human chorionic gonadotropin; BP = blood pressure.

Table 4.14: Model performance over the test set, in the format of mean percentage (SD percentage) over 5 repetitions.

	Model I		Model II		Model III		Model IV	
	AUC	**F1-weighted**	**AUC**	**F1-weighted**	**AUC**	**F1-weighted**	**AUC**	**F1-weighted**
Best full models before SFS								
	GBM (51 features)		**GBM (51 features)**		**GBM (51 features)**		**GBM (51 features)**	
	85.2 (1.8)	94.5 (0.2)	80.6 (0.5)	95.1 (0.2)	80.4 (0.7)	94.8 (0.1)	81.8 (1.4)	91.1 (0.4)
Best full models after SFS								
	GBM (14 features)		**GBM (16 features)**		**GBM (17 features)**		**GBM (17 features)**	
	83.6 (1.7)	94.5 (0.2)	80.5 (0.7)	95.1 (0.2)	79.8 (1.1)	94.8 (0.1)	81.1 (1.3)	90.9 (0.3)
	LR-L1 (14 features)		**LR-L1 (16 features)**		**LR-L1 (17 features)**		**LR-L1 (17 features)**	
	79.2 (1.9)	93.9 (0.2)	71.7 (0.9)	94.7 (0.1)	72.9 (2.1)	94.4 (0.1)	74.8 (1.1)	89.7 (0.3)
	LR-L2-MLP(8 features)		**LR-L2-MLP(10 features)**		**LR-L2-MLP(11 features)**		**LR-L2-MLP(10 features)**	
	82.3 (1.7)	94.5 (0.1)	77.6 (1.3)	95.1 (0.1)	77.4 (1.6)	94.9 (0.1)	79.1 (1.1)	90.8 (0.3)

AUC, area under the receiver operator characteristic curve; GBM, gradient boosted machine; LR-L1, L1-penalized logistic regression; LR-L2-MLP score, parsimonious models logistic regression with MLP score.

Table 4.15: Performance of all models and all algorithms, before and after SFS. The means and standard deviations of AUC and weighted-F1 scores on the test set over the five repetitions are listed in the format of mean percentage (SD percentage).

	Model I		**Model II**		**Model III**		**Model IV**	
	AUC	**F1-weighted**	**AUC**	**F1-weighted**	**AUC**	**F1-weighted**	**AUC**	**F1-weighted**
Models before SFS (51 features)								
LR-L1	79.4 (2.0)	93.9 (0.1)	72.9 (0.9)	94.8 (0.1)	73.9 (1.9)	94.5 (0.2)	75.3 (1.1)	89.8 (0.5)
SVM-L1	79.3 (2.0)	93.9 (0.1)	73.1 (0.7)	94.7 (0.1)	74.1 (1.9)	94.4 (0.2)	75.3 (1.1)	89.7 (0.3)
GBM	85.2 (1.8)	94.5 (0.2)	80.6 (0.5)	95.1 (0.2)	80.4 (0.7)	94.8 (0.1)	81.8 (1.4)	91.1 (0.4)
RF	84.3 (1.7)	94.5 (0.2)	79.6 (0.4)	95.1 (0.2)	80.3 (0.7)	94.8 (0.1)	81.3 (1.6)	90.9 (0.3)
Models after SFS (14, 16, 17, 17 features respectively for Model I, II, III, IV)								
LR-L1	79.2 (1.9)	93.9 (0.2)	71.7 (0.9)	94.7 (0.1)	72.9 (2.1)	94.4 (0.1)	74.8 (1.1)	89.7 (0.3)
SVM-L1	79.1 (1.9)	93.8 (0.2)	71.7 (0.9)	94.7 (0.1)	72.8 (2.0)	94.4 (0.1)	74.7 (1.0)	89.6 (0.3)
GBM	83.6 (1.7)	94.5 (0.2)	80.5 (0.7)	95.1 (0.2)	79.8 (1.1)	94.8 (0.1)	81.1 (1.3)	90.9 (0.3)
RF	82.8 (2.0)	94.5 (0.1)	79.5 (0.5)	95.0 (0.3)	79.7 (0.8)	94.9 (0.2)	80.4 (1.3)	90.8 (0.3)
Parsimonious models (LR-L2-MLP score (8, 10, 11, 10 features respectively for Model I, II, III, IV))								
LR-L2-MLP	82.3 (1.7)	94.5 (0.1)	77.6 (1.3)	95.1 (0.1)	77.4 (1.6)	94.9 (0.1)	79.1 (1.1)	90.8 (0.3)

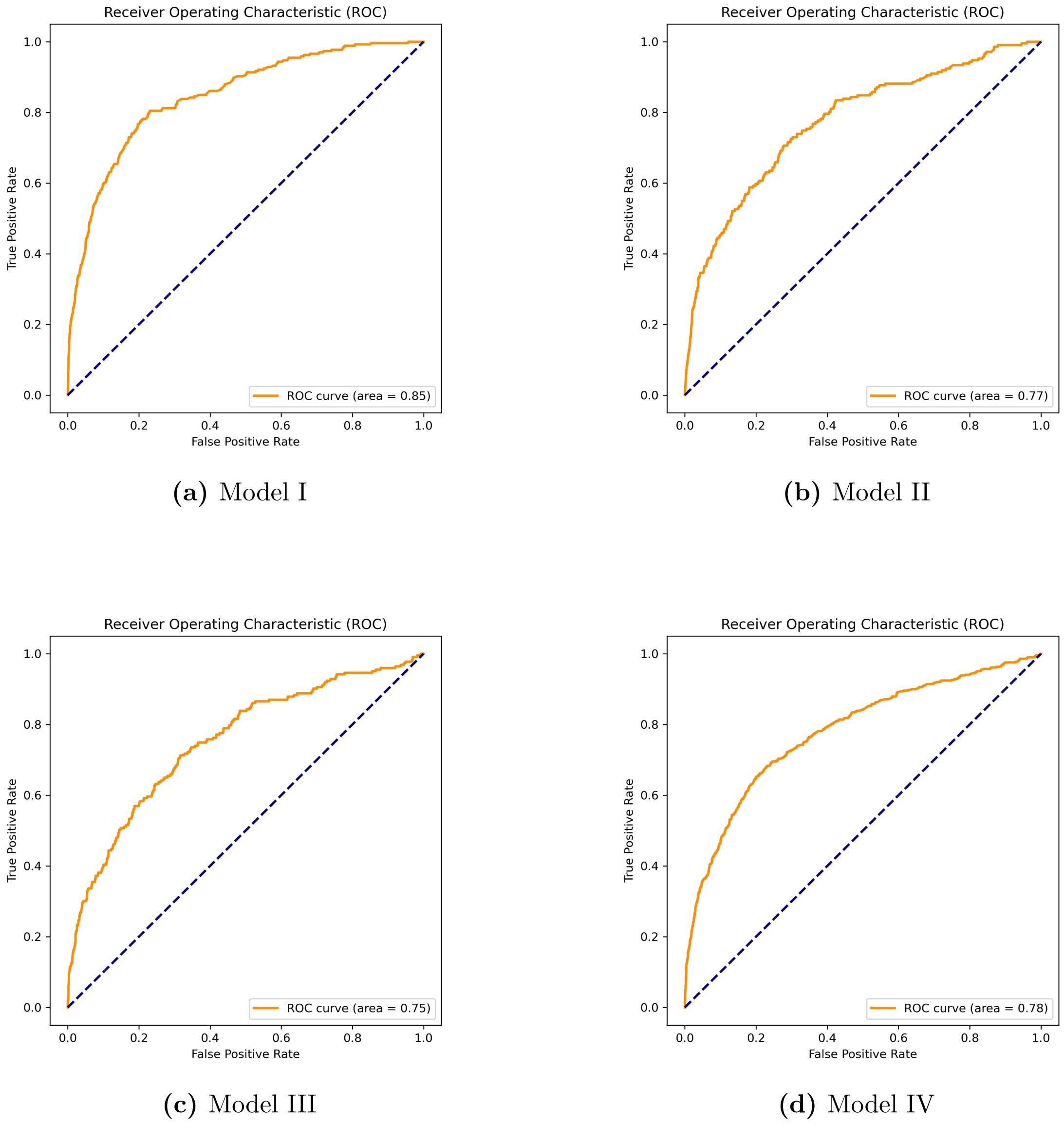
AUC = area under the receiver operator characteristic curve; LR-L1 = L1-penalized logistic regression; SVM-L1 = support vector machine; GBM = gradient boosted machine; RF = random forest; LR-L2-MLP = parsimonious models logistic regression with MLP score.

4.5 Discussion

Evaluating an at-risk population for PCOS is essential for early diagnosis and initiating multi-disciplinary care with the goal of reducing health risks (endometrial hyperplasia/cancer), infertility and pregnancy complications, and chronic disease burden including cardiometabolic disorders associated with PCOS. Retrospective analysis of the at-risk population within an urban health center allows for assessment of factors predictive of diagnosis. Of note, the study sample represents a population of patients who had any visit to BMC for primary care, obstetrics and gynecology, endocrinology, family medicine, or general internal medicine and does not represent a random sample. While this is not a population level assessment, our model is applicable to patients with high suspicion for PCOS who interact with the healthcare system.

The ranked list of variables, from the most predictive to the least predictive of the PCOS outcome, informed the main drivers of the predictive models. For example, non-gravidity, high levels of LH, low levels of FSH, obesity, and higher BMI increase the likelihood of PCOS. These variables are consistent with key variables from other models and in the pathophysiology of PCOS. The overall predictive accuracy was high for all models, suggesting that a predictive model may assist in early detection of PCOS within those at risk in an electronically interfaced medical record. Furthermore, we found that non-linear models had superior predictive capacity compared to linear models for all four model outcomes, potentially allowing for inclusion of non-linear reproductive hormone relationships.

When assessing patients who received a diagnosis of PCOS (Model I), the most predictive factors related to diagnosis were hormone levels (as captured by the MLP score) and obesity, a clinical factor in supporting a PCOS diagnosis. Specifically, there is a non-linear relationship between reproductive hormones such as FSH, LH, and estradiol. Often these hormonal lab tests are obtained randomly in those with



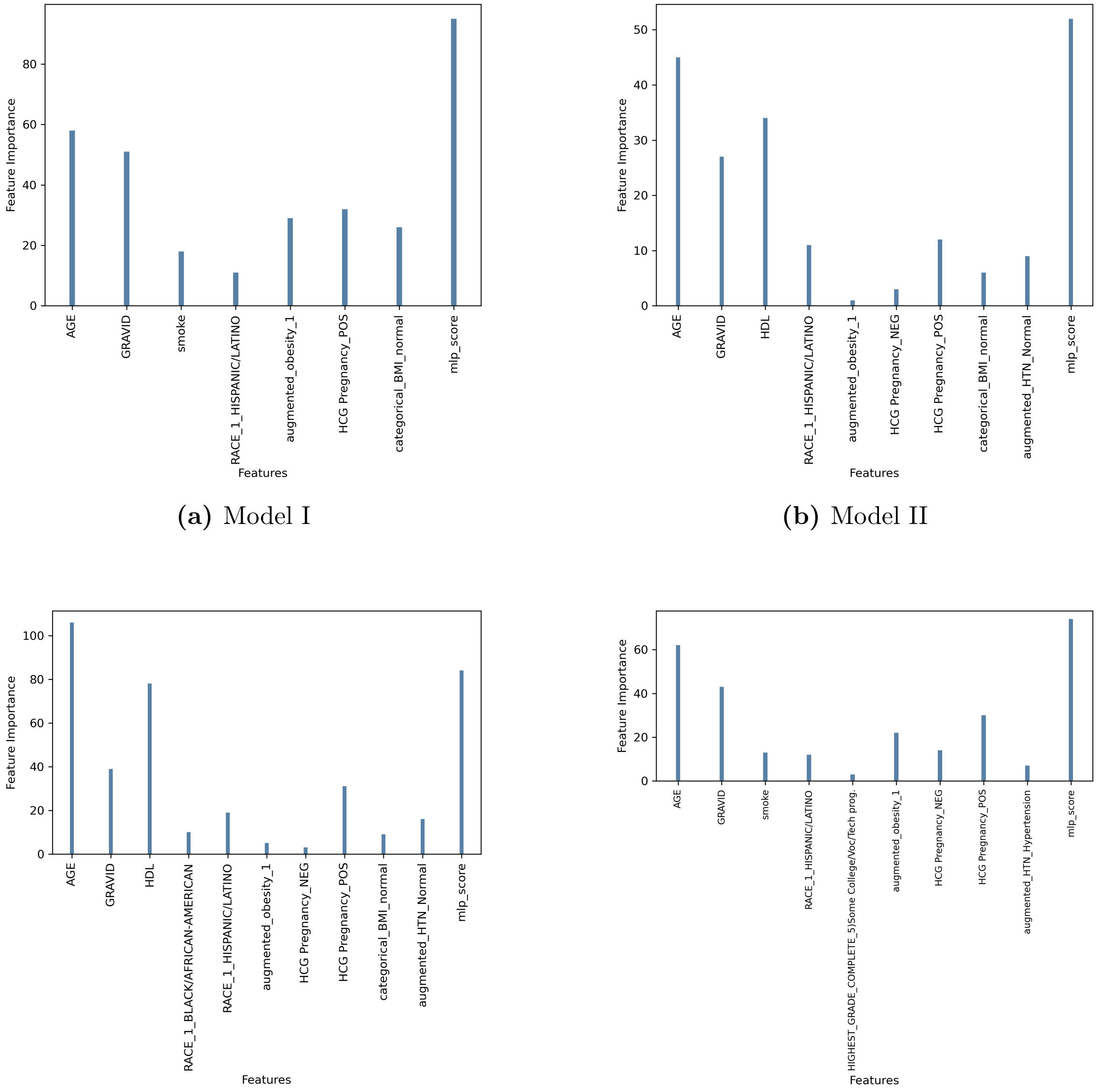
(a) Model I

(b) Model II

(c) Model III

(d) Model IV

Figure 4·3: Example of receiver operator characteristic (ROC) curves associated with parsimonious logistic regression models utilizing the MLP score (LR-L2-MLP score).



(a) Model I

(b) Model II

(c) Model III

(d) Model IV

Figure 4·4: Example of receiver operator characteristic (ROC) curves associated with parsimonious logistic regression models utilizing the MLP score (LR-L2-MLP score).

oligomenorrhea, and it is also common to find an elevated FSH to LH ratio. A concern may also be the misclassification of hypothalamic amenorrhea into the group classified as PCOS where the FSH and LH levels would be low or suppressed, or in the setting of premature ovarian insufficiency, notable by an elevated FSH and low estradiol. The MLP score allows for the diversity of relationships of these hormone levels and was trained using a neural network to appropriately classify PCOS. Additionally, prior pregnancy (gravidity) and a positive pregnancy test were negatively associated with a diagnosis of PCOS, consistent with the underlying increased risk of infertility due to oligo-ovulation. Normal BMI and smoking, a known ovarian toxicant, were negatively associated with the presence of a PCOS diagnosis, which may indicate patient characteristics that increase risk of a delayed PCOS diagnosis. These identified variables demonstrate the robustness of the model towards predicting phenotypic traits of patients with PCOS, which is aligned with the performance accuracy. While the significant factors such as hormone levels, gravidity, bHCG, and obesity identified in the model are already known to be associated with PCOS, the true impact of our model lies within the implementation of such a tool within the EHR. For example, a real-world application of this model in the clinical setting would entail integration of our model into the electronic health record system that would provide the probability of PCOS diagnosis or set a threshold for suspicion for each patient to aid a provider's evaluation. Though substantial system modifications may be required, integration of our model into the EHR system would lead to more timely diagnosis and optimize referrals for downstream follow-up for known clinical sequelae associated with PCOS.

When assessing patients who met diagnostic criteria without the ICD-9 label of PCOS (Models II and III), predictive factors both supported the underlying PCOS diagnosis and alluded towards factors that may contribute to missing the diagnosis despite meeting Rotterdam criteria. Similar to Model I, gravidity and a positive pregnancy test were negatively associated with Models II and III diagnosis, while obesity was positively associated with Models II and III diagnosis, consistent with Model I. Interestingly, distinct positive predictors among Models II and III were normal BMI, normal blood pressure, and normal HDL. These patients may present as the "lean" phenotype of PCOS or those with mild features, leading to underdiagnosis of PCOS. Diagnosing "lean" PCOS can be more nuanced, potentially delaying diagnosis or requiring more specialized consultation (Toosy et al., 2018). Within our cohort, 1,116 individuals were identified by the model without the ICD-9 code that met Rotterdam PCOS diagnostic criteria (Model III), suggesting the predictive value of our models to identify at risk groups within a large health system and reduce delays in diagnosis. Given that women often wait over two years and see numerous health professionals before receiving a diagnosis of PCOS, the integration of high-quality AI-based diagnostic tools with the EHR could significantly contribute to more timely diagnosis (Gibson-Helm et al., 2017).

Consistent with Models I, II, and III, positive pregnancy test and gravidity were both negatively associated with PCOS diagnosis in Model IV while obesity and presence of hypertension were both positively associated with the Model IV combined PCOS outcome. Some higher degree of education, such as college or vocational/technical school, was also positively associated with the outcomes of undiagnosed PCOS and combined PCOS (Models II, III, and IV), which may suggest that education status and patient's self-advocacy for seeking care within a medical system may be implicated specifically in under-diagnosed individuals. Of note, we dropped insurance status after finding that the null was a strong predictor of PCOS, though it is interesting to note that 83% of 331 patients in this dataset with missing insurance have PCOS. Insurance status alludes to socioeconomic barriers such as access to care, which can result in a delay in timely diagnosis through either inability to seek evaluation or follow through with testing. While the implications of insurance status and social determinants of health are beyond the scope of this paper, it is important to note that persistence in seeking treatment within a fractionated health care system can be challenging financially and psychologically, as patients may need multiple evaluation or specialist's consultation to reach the right diagnosis.

A recent systematic review investigated the utility of artificial intelligence and machine learning in the diagnosis or classification of PCOS (Barrera et al., 2023). Their search ultimately included 31 studies with sample sizes ranging from 9 to 2,000 patients with PCOS. Methods employed by these models included support vector machine, K-nearest neighbor, regression models, random forest, and neural networks. Only 19% of included studies performed all major steps of training, testing, and validating their model. Furthermore, only 32% of included studies used standardized diagnostic criteria such as the Rotterdam criteria or NIH criteria. The authors found that the ROC of included studies ranged from 73-100%. While it is difficult to directly compare our models' performance to those included in the review due to possible differences in model training and diagnostic criteria, our models' AUCs fell within the range reported in the systematic review of 73 to 100%. Only one study sourced their data from electronic health records to build their model (Castro et al., 2015). Despite the lack of standardized model training and diagnostic criteria used in these studies, the review concluded that artificial intelligence and machine learning provide promise in detecting PCOS, allowing for an avenue for early diagnosis.

Outside of the machine learning models included in the systematic review, other predictive models have been created for earlier detection of PCOS as well as for predicting long-term health outcomes among women with a diagnosis of PCOS. One such model was created from 11,720 ovarian stimulation cycles at Peking University Third Hospital. The model used serum antimullerian hormone (AMH) and androstenedione levels, BMI, and menstrual cycle length to predict a diagnosis of PCOS and achieved an AUC of 85%. The algorithm was then developed into an online platform that is able to calculate one's risk of PCOS given certain indicators that are inputted into the model, allowing for better screening abilities in the clinic (Xu et al., 2022). Another study created a similar model, taking into account AMH and BMI to predict a diagnosis of PCOS or other ovulatory dysfunction disorders among 2,322 women (Vagios et al., 2021). They found that in women with higher BMIs and lower AMH levels could be used to predict PCOS compared to normal-weight or underweight women. Deshmukh et al. created a simple four-variable model which included free androgen index (FAI), 17-hydroxyprogesterone, AMH, and waist circumference for predicting risk of PCOS in a cross-sectional study involving 111 women with PCOS and 67 women without PCOS (Deshmukh et al., 2019). Lastly, Joo et al. used polygenic and phenotypic risk scores to develop a PCOS risk prediction algorithm (Joo et al., 2020). They found high degrees of association between PCOS and various metabolic and endocrine disorders including obesity, type 2 diabetes, hypercholesterolemia, disorders of lipid metabolism, hypertension, and sleep apnea (Joo et al., 2020).

In addition to the goal of improved screening for PCOS, models have been created to predict long-term clinical outcomes in women with PCOS, such as ovulation, conception, and live birth (Kuang et al., 2015) (Jiang et al., 2021). Given the increased risk of insulin resistance in women with PCOS, Gennarelli et al. created a mathematical model to predict insulin sensitivity based on variables such as BMI, waist and hip circumferences, truncal-abdominal skin folds, and serum concentrations of androgens, SHBG, triglycerides, and cholesterol (Gennarelli et al., 2000). Models to predict non-alcoholic fatty liver disease risk among young adults with PCOS have also been generated (Carreau et al., 2019). Combining earlier detection with more accurate risk stratification of clinical sequalae through predictive modeling can significantly improve the long-term health outcomes of women with PCOS. Application of our models to predict other downstream health risks after the diagnosis of PCOS is a future area of research.

Beyond the long-term health impacts of PCOS, the condition also carries a significant economic cost for our healthcare system. A study by Riestenberg et al (2022) recently estimated the total economic burden of PCOS, as well as the cost specifically for pregnancy-related complications and long-term health morbidities (Riestenberg et al., 2022). The authors estimated the annual economic burden of PCOS to be $8 billion as of 2020 in the United States. Furthermore, the excess cost of pregnancy-related comorbidities such as gestational hypertension, gestational diabetes, and preeclampsia attributable to PCOS totals $375 million USD annually. Outside of pregnancy, the cost of long-term comorbidities associated with PCOS including stroke and type 2 diabetes mellitus was estimated at $3.9 billion USD. Meanwhile, the cost for diagnostic evaluation of PCOS was less than 2% of the total economic burden. This estimated financial burden suggests that predictive models aiding earlier diagnosis could not only reduce long-term health consequences of PCOS but also alleviate significant healthcare costs associated with the condition.

Given the high prevalence, significant healthcare burden, and heterogeneity in clinical presentation of PCOS, AI-based tools are well suited for earlier diagnosis of PCOS. Our study had many strengths. First, our machine learning models, which were highly accurate and robust in PCOS diagnosis prediction, were created using the largest sample size to date (Barrera et al., 2023). Second, our model was tested and trained on a diverse Safety-Net hospital-sourced population not restricted to the context of fertility care. Third, it is the only model that incorporated three data streams (ICD-9 codes, clinical laboratory findings, and radiologic findings) and an MLP score.

Fourth, the parsimonious and interpretable models were very close in achieving full model predictive accuracy, performing relatively closely to the best-performing nonlinear models. Essentially, our parsimonious models "isolate" nonlinearities in hormone levels (captured by the MLP score) and linearly combine that score with other variables. Most models evaluate reproductive hormones (FSH, estradiol, LH, and SHBG) as individual variables within linear models, which does not account for the high inter- and intra-patient variability. By using non-linear mapping of the hormone values, we were able to generate a composite variable allowing for a linear function that correlates with the likelihood of an accurate prediction. Last, our variables are easily accessible in an electronic health dataset, rendering the models helpful for clinical prediction. Our study did not evaluate AMH as a predictive variable because it was not widely utilized during the time window of this data extraction corresponding with ICD-9 codes.

Despite these strengths, our model is not without limitations. First, it is only directly applicable to those who interact with the medical system and those deemed "at-risk" for a PCOS diagnosis, which would not facilitate population-based prediction. More specifically, the models' data are sourced from an urban, hospital-based population which may limit the generalizability of these findings to other patient populations such as those living in rural areas. Additional studies need to be conducted in other patient populations or unselected community-based populations to externally validate the use of these models, especially expanding to the entire population within a health system to evaluate the accuracy of our models (Azziz et al., 2004). Second, we must interpret our data within the limitations of informative presence in EHR data. Informative presence is defined as data that is present and informed with respect to the health outcome, in this case PCOS, as well as behavioral patterns of interaction with healthcare institutions which may be additionally impacted by marginalization

(Harton et al., 2022). This is an important consideration as a potential source of bias for interpreting predictive models using EHR data (Harton et al., 2022) (Sisk et al., 2021). Nevertheless, we were able to extract over 1000 patients who were undiagnosed with PCOS among the population, suggesting the predictive value of the modelling in identifying diagnosis gaps among specific populations within a large health system. Third, it is possible that additional examination of the medical record beyond ICD-9 diagnosis may allow for more clarification of risk in the presumed PCOS group. Fourth, while our model considers numerous important variables, other potentially relevant predictors such as diet, genetic factors, and lifestyle factors, were not incorporated and represents as an area of growth for future predictive models. Last, our exclusion of concurrent endocrinopathies was chosen to avoid incorrectly including ovulation disorders caused by other endocrinopathies, but it is possible that this was an overly strict exclusion criterion.

In conclusion, this novel machine learning algorithm incorporates three data streams from a large EHR dataset to assess PCOS risk. This model can be integrated into the EHR to aid clinicians in earlier diagnosis of PCOS and connect patients to interventions and healthcare providers across their reproductive lifespan with the goal of health optimization and risk reduction (Zad et al., 2024b).

Chapter 5

Predictive Models of Miscarriage Based on Data From a Preconception Cohort Study

5.1 Introduction

Miscarriage, or pregnancy loss before 20 completed weeks of gestation, affects approximately 20% of recognized pregnancies (Rossen et al., 2018). The strongest identified predictors of miscarriage are older parental age and history of miscarriage (Wilcox et al., 1988). Other confirmed risk factors include low and high body mass index (Arck et al., 2008)(Hahn et al., 2014), caffeine consumption (Savitz et al., 2008)(Weng et al., 2008)(Hahn et al., 2014), alcohol intake (Klonoff-Cohen et al., 2003)(Henriksen et al., 2004)(Andersen et al., 2012)(Feodor Nilsson et al., 2014), and smoking (Venners et al., 2004)(George et al., 2006)(Nielsen et al., 2006), though the etiology of miscarriage remains poorly understood.

Several studies have developed predictive models of pregnancy loss among individuals receiving treatment with assisted reproduction technology (ART) (Choong et al., 2003)(Yi et al., 2016)(Liu et al., 2020), individuals with recurrent miscarriage (Quenby and Farquharson, 1993)(Caetano et al., 2006)(Dai et al., 2022)(du Fossé et al., 2022), and individuals with threatened miscarriage (Huang et al., 2022). Most of these studies have relied on clinical assessments such as early pregnancy ultrasound measurements and laboratory values. Other studies have attempted to predict miscarriage based on early pregnancy characteristics (e.g., parental age, ultrasound measurements, and laboratory values) in general populations (DeVilbiss et al., 2020)(Li et al., 2022). However, no study has derived a predictive model of miscarriage using prospectively collected data on the couple during the preconception period. Predicting primary (i.e., first-time) miscarriage is of great importance, given the high rate of miscarriage and the impact of miscarriage on mental health and fertility outcomes. Moreover, primary miscarriage likely shares many risk factors with recurrent miscarriage (Cramer and Wise, 2000).

In this North American prospective preconception cohort study, we predicted risk of miscarriage using 189 self-reported variables describing a variety of preconception sociodemographic, lifestyle, dietary, and anthropometric factors. We used supervised machine learning methods with several classification algorithms and variable selection procedures.

5.2 Materials and Methods

5.2.1 Study population

Pregnancy Study Online (PRESTO) is an ongoing web-based preconception cohort study that collects data on a variety of environmental and behavioral factors in addition to pregnancy outcomes (Wise et al., 2015). At enrollment, eligible participants were female, aged 21-45 years, residents of the United States (US) or Canada, and trying to conceive without the use of fertility treatment. Participants were followed for up to 12 months of pregnancy attempts, during which time they could have initiated fertility treatment. Participants who conceived were followed through pregnancy and postpartum.

During 2013 through 2022, 16,631 female participants enrolled in PRESTO and completed a baseline questionnaire. We excluded 37 participants who were not from the US or Canada, 120 who were already pregnant at study entry, 203 who completed the baseline questionnaire <11 weeks before analysis (and therefore had no opportunity for follow-up), and 41 who completed the baseline questionnaire >2 months after the screening questionnaire. Approximately 36% of participants were lost to follow-up. Among those who were lost to follow-up, we successfully collected information on pregnancy for 25% of participants via email or phone contact, or by searching for baby registries and birth announcements online; for 5% by linking to birth registries in selected states (CA, FL, MA, MI, NY, OH, PA, TX); and for 5% by linking to FertilityFriend.com data (a mobile computing fertility-tracking app).

In total, 8,739 participants became pregnant during follow-up (we included only the first observed pregnancy per participant in these analyses). We excluded 19 participants with missing data on categorical variables (handling of missing data is described in the Supplementary Material), retaining a total of 8,720 participants in the dataset used for our analysis. The institutional review board at Boston University Medical Campus approved the study protocol.

5.2.2 Data collection

Female participants completed a baseline questionnaire and follow-up questionnaires every eight weeks until pregnancy. Those who conceived completed an early pregnancy questionnaire at a median of 9 weeks' gestation and a late pregnancy questionnaire at approximately 32 weeks' gestation. On baseline, follow-up, and pregnancy questionnaires, we collected data on pregnancy status, sociodemographic factors, lifestyle and behavioral factors, anthropometrics, medical and reproductive history, and selected male partner characteristics. Reproductive history included gravidity, parity, and history of miscarriage (i.e., miscarriages that occurred prior to enrolling in PRESTO), among other variables. Participants were also invited to complete the web-based Diet History Questionnaire (DHQ II: 2013-2019; DHQ III: 2020-2022) ten days after enrollment. The DHQ was designed by the National Cancer Institute and the first version of the DHQ was validated against 24-hour dietary recalls in a US population (Subar et al., 2001) (Millen et al., 2006). We used DHQ data to calculate the Healthy Eating Index-2010 (HEI-2010) score, a measure of diet quality (Guenther et al., 2013). For time-varying characteristics, we prioritized data collected most recently before conception to avoid bias due to conditioning on future information (Suissa and Dell'Aniello, 2020). Table 5.1 and 5.2 provide a complete list of the 160 variables included in this analysis and when they were ascertained. Ninety variables were binary, 58 were continuous, and 12 were categorical. Table 5.3 describes the percentage of missingness for each predictor variable and the Methods Supplement provides an overview of how missing data were handled.

5.2.3 Assessment of miscarriage

We defined miscarriage as pregnancy loss before 20 completed weeks of gestation (including blighted ovum and chemical pregnancy but excluding ectopic pregnancy and induced abortion). On follow-up questionnaires, participants reported the date of their last menstrual period, whether they were currently pregnant, and whether they had experienced a miscarriage since completing their previous questionnaire. Participants who reported a miscarriage were asked how many weeks the pregnancy lasted and on what date the pregnancy ended. Pregnant participants reported the due date of their current pregnancy and the date of their first positive pregnancy test. Pregnant participants were asked to report the method(s) used to confirm their pregnancy (e.g., home pregnancy test, urine or blood test in doctor's office, ultrasound). More than 95% of participants reported using a home pregnancy test to identify their pregnancy.

For participants who reported a miscarriage, we used the participant's reported gestational weeks at loss when available (defined as weeks since the last menstrual period). Among participants who did not report their gestational week at loss but

Table 5.1: Complete list of variables included in analysis to generate predictive models of miscarriage in PRESTO, 2013-2022. Part 1.

Category	**Variables Included in Preliminary Analysis**
Demographic and socioeconomic characteristics	Age*, marital status, region of residence, urbanization of residential area, highest level of education, parents' education level, household income, employment status, hours/week of work, shift work, night shift frequency in the past month.
Lifestyle, behavioral, and wellness factors	Years in a steady relationship, cigarette smoking (if so, number per day)*, total duration of smoking; history of smoking during pregnancy; use of e-cigarettes (if so, ml/day)*; frequency of marijuana use*; exposure to second-hand smoke*; alcohol intake*; caffeine consumption*; moderate physical activity; vigorous physical activity; sedentary activity; sleep duration*; trouble sleeping*; perceived stress scale score*; major depression inventory score*.
Dietary factors and use of supplements	Healthy Eating Index-2010 score; supplemental intake of vitamins A, B1, B2, B3, B5, B6, B7, B12, C, E, K; beta-carotene; folic acid; iron; zinc; calcium; magnesium; selenium; omega-3 fatty acids; consumption of whole milk, 2% milk, 1% milk, skim milk, soy milk, other milk, fruit juice, sugar-sweetened soda*, diet soda*, sugar-sweetened energy drinks*, diet energy drinks*; use of multivitamins or folic acid supplements*.
Early life exposures and family history	Adopted; number of siblings; multiple gestation; born preterm; born with low birthweight; breastfed; delivered via cesarean section; mother's cigarette smoking during pregnancy; mother's age at participant's birth; mother's history of pregnancy complications; mother's history of miscarriage.

Table 5.2: Complete list of variables included in analysis to generate predictive models of miscarriage in PRESTO, 2013-2022. Part 2.

Category	**Variables Included in Preliminary Analysis**
Reproductive characteristics and disorders	Use of fertility treatment to conceive the study pregnancy (if yes, type of treatment); history of miscarriage; age at menarche; menstrual regularity; menstrual period characteristics (typical length, number of flow days, flow amount, pain)*; received human papillomavirus vaccine; abnormal pap smear; ever diagnosed with a thyroid condition*; fibroids, polycystic ovarian syndrome, endometriosis, a urinary tract infection, pelvic inflammatory disease, chlamydia, herpes, vaginosis, genital warts; Ferriman-Gallwey Hirsutism Score; recent use of medications for polycystic ovarian syndrome*; gravidity; parity; history of cesarean section; years since last pregnancy; history of unplanned pregnancy; history of subfertility or infertility; history of infertility treatment*; history of breastfeeding; number of lifetime sexual partners; last method of contraception; number of menstrual cycles to conceive the study pregnancy.
Physical characteristics, non-reproductive medical history, and medication use	Body mass index; waist circumference; handedness; number of primary care visits last year; high blood pressure; received influenza vaccine last year*; ever diagnosed with migraines (if so, recent migraine frequency), asthma, hay fever, depression*, anxiety*, gastroesophageal reflux disease, diabetes; use of the following medications in the 4 weeks before baseline: pain medications*, antibiotics*, asthma medications*, diabetes medications*; use of psychotropic medications*.
Environmental exposures (occupational and personal care product use)	Exposed regularly to agricultural pesticides; metal particulates or fumes; solvents, oil-based paints, or cleaning compounds; high temperature environments; chemotherapeutic drugs; engine exhaust; chemicals for hair dyeing, straightening, or curing; chemicals for manicure/pedicure; use of chemical hair relaxer.
Male partner characteristics	Age*, body mass index, education, cigarette smoking (if so, number per day), circumcision status.
*These variables were considered time-varying characteristics and were updated on follow-up questionnaires completed after the baseline questionnaire but before conception.	

Table 5.3: Missing data among predictor variables in PRESTO.

Variable Name	**Participants with missing data**	**% (out of 8,739)**
Categorical Variables		
Handedness	14	<1%
Female smoking status	9	<1%
Male smoking status	9	<1%
Menstrual cycle regularity (initial)	7	<1%
Tried to get pregnant for 12 months or more	5	<1%
Menstrual cycle regularity (recent)	5	<1%
Binary variables		
Mother's history of miscarriage	1998	22.9%
Mother's history of pregnancy problems	1394	16.0%
Male partner circumcision status	1055	12.1%
Secondhand smoking status (current, at home)	820	9.4%
Last method of contraception, barrier methods	819	9.4%
Last method of contraception, natural methods	819	9.4%
Participant was born preterm	803	9.2%
Secondhand smoking status (current, at work)	742	8.5%
Conceived through fertility treatment	729	8.3%
Participant was born with low birth weight	680	7.8%
History of an abnormal Pap smear	566	6.5%
Secondhand smoking status (age 0-10, at home)	550	6.3%
Ever visited a physician for difficulty getting pregnant	527	6.0%
Mother's history of C-section for participant's birth	341	3.9%
Participant was a twin/triplet	161	1.8%
Working rotating shifts	157	1.8%
Continuous		
Ferriman-Gallwey score	4334	49.6%
Waist measure	3304	37.8%
Current e-cigarettes (ml/day)	3107	35.6%
Duration participant was breastfed	2926	33.5%
Number of lifetime sexual partners	2376	27.2%
Mother's smoking history while pregnant (num. of cigs)	1331	15.2%
Male BMI	1012	11.6%
Father's education	439	5.0%
Household income	230	2.6%
Mother's education	205	2.3%
Night shift frequency in past month	198	2.3%
Male education	170	1.9%
Job hours/week	126	1.4%

Note: All categorical variables are presented in this table; however, we only present continuous and binary variables with >1% missingness here.

who reported a due date (11%), we estimated gestational age as: (pregnancy end date – (pregnancy due date – 280 days))/7 (on Obstetric Practice et al., 2013). Among participants who reported neither their gestational week at loss nor their due date (21%), we estimated week at loss as: (pregnancy end date – last menstrual period date)/7. Approximately 97% of miscarriages were identified via study questionnaires; the remaining 3% were identified via the study withdrawal form, via email or phone contact, by linking to birth registries, or by linking to FertilityFriend.com data.

5.2.4 Statistical analysis

We used supervised machine learning methods to generate predictive models of miscarriage. We generated both static and survival models. For all analyses, we first performed several pre-processing steps including statistical feature selection, explained in more detail in Section 2.1.2 and Section 2.1.5.2. For static models, we used a variety of supervised classification methods including linear (e.g., logistic regression) and non-linear (e.g., Gradient Boosted Trees) algorithms, explained in more detail in Section 2.1.1. For survival models, we fit Cox proportional hazards models. For both static and survival models, we generated full and sparse models. The full models contain all variables selected by statistical feature selection, whereas the sparse models contain all variables selected by both statistical feature selection and univariate feature selection for survival models or recursive feature elimination for static models. We evaluated model performance via the area under the receiver operating characteristic curve (AUC), precision and recall metrics, and the weighted-F1 score for static models, and via the concordance index for survival models. Metrics are explained in more detail in Section 2.1.4.

5.2.5 Sensitivity analysis

We repeated all analyses among primigravid participants to generate models predictive of primary miscarriage, which may have different predictors from secondary or recurrent miscarriage. We also restricted the dataset to ≥8 gestational weeks to assess the extent to which predictors differed for later losses, which are less likely to be attributable to random chromosomal aberrations (Savitz et al., 2002). All analyses were performed with Python packages. Relevant programs can be accessed in github repository[1].

5.3 Results

We analyzed data from 8,720 pregnant participants, among whom 1,775 (20.4%) experienced miscarriage during the 12-month study period. Miscarriages were reported as early as 3 gestational weeks (median=6; interquartile range: 5-8 gestational weeks). We observed 567 late miscarriages (32% occurring ≥8 gestational weeks). The distribution of gestational weeks at miscarriage is presented in Table 5.4. Mean age was 30 years for female participants and 32 years for male partners. Mean BMI of female participants was 27 kg/m^2 and 28 kg/m^2 for male partners. Approximately one quarter of couples resided in the Northeast US, while 22% resided in the South, 22% in the Midwest, 16% in the West, and 16% in Canada. Approximately one quarter of participants had a previous miscarriage, 35% had had an unplanned pregnancy before enrolling in PRESTO, and about half were parous. Almost 14% of female participants reported any history of subfertility or infertility, and 7% of study pregnancies were conceived via fertility treatment.

5.3.1 Survival models

After statistical feature selection, 17 variables remained in the dataset. The variables selected into the full survival model are presented in Table 5.5. The variables selected into the sparse survival model are presented in Table 5.6. The strongest two predictors in the sparse survival model were female age at conception (HR=1.19; 95% CI: 1.11, 1.27) and history of miscarriage (HR=1.10; 95% CI: 1.03, 1.17), which were both

[1] https://github.com/noc-lab/Predictive-models-of-miscarriage/

Table 5.4: Distribution of gestational age at miscarriage in PRESTO, 2013-2022.

Gestational week at miscarriage	**N(%)**
Total	N=1,775
3	53 (3.0)
4	358 (20.2)
5	346 (19.5)
6	305 (17.2)
7	146 (8.2)
8	143 (8.1)
9	137 (7.7)
10	123 (6.9)
11	67 (3.8)
12	49 (2.8)
13	15 (0.8)
14	9 (0.5)
15	8 (0.5)
16	5 (0.3)
17	6 (0.3)
18	3 (0.2)
19	2 (0.1)

positively associated with miscarriage (Table 2). All other variables selected into the sparse model had very small or null associations with miscarriage. Variables that were very slightly positively associated with miscarriage were use of omega-3 or fish oil supplements (HR=1.04; 95% CI: 0.99, 1.10), number of prior pregnancies (HR=1.04; 95% CI: 0.99, 1.10), history of subfertility or infertility (HR=1.04; 95% CI: 0.97, 1.11), male partner age at conception (HR=1.03; 95% CI: 0.97, 1.10), and having a history of unplanned pregnancy (HR=1.01; 95% CI: 0.94, 1.09). Variables that were very slightly inversely associated with miscarriage included having been pregnant before (HR=0.95; 95% CI: 0.87, 1.05) and being vaccinated against human papillomavirus (HPV) (HR=0.98; 95% CI: 0.93, 1.04). The concordance index of the final sparse survival model, applied to the testing dataset, was 55.4%, indicating poor-to-moderate discrimination (i.e., ability of the model to discriminate between individuals with and without miscarriage).

When we restricted the incident period to ≥8 gestational weeks (n=6,993; 32% of all miscarriages), 4 variables remained after statistical feature selection. The strongest

Table 5.5: Variables selected by the full survival model predicting miscarriage in PRESTO, 2013-2022.

Variable	**Hazard Ratio^1 (95% CI)**
Female age at conception (years)	1.20 (1.12, 1.29)
Female smoking: current regular smoker (ref = never smoker)	0.90 (0.84, 0.96)
History of miscarriage (yes/no)	1.11 (1.04, 1.18)
Geographic region of residence: Northeast US (ref = South US)	0.93 (0.87, 0.99)
Use of vitamin B7 (yes/no)	1.07 (0.99, 1.14)
Healthy Eating Index-2010 score (HEI-2010 score)	0.94 (0.89, 0.99)
Use of vitamin B6 (yes/no)	1.04 (0.99, 1.10)
Ever pregnant before (yes/no)	0.96 (0.87, 1.05)
Number of prior pregnancies	1.04 (0.98, 1.10)
Use of vitamin B1 (yes/no)	0.96 (0.89, 1.04)
Use of omega-3 or fish oil supplements (yes/no)	1.04 (0.99, 1.09)
Male age at conception (years)	1.04 (0.97, 1.10)
History of subfertility or infertility (yes/no)	1.03 (0.97, 1.11)
Ever received HPV vaccine	0.99 (0.94, 1.04)
Use of vitamin C (yes/no)	1.01 (0.96, 1.06)
History of unplanned pregnancy (yes/no)	0.99 (0.92, 1.07)
Previously tried to conceive for $\geq$ 12 months: "no, never tried before" (ref = "no")	1.00 (0.92, 1.08)
Variables forced into the model^2	
Female smoking: former smoker (ref = never smoker)	0.98 (0.93, 1.03)
Geographic region of residence: Canada (ref = South US)	0.98 (0.93, 1.04)
Geographic region of residence: West US (ref = South US)	1.01 (0.95, 1.07)
Female smoking: current occasional smoker (ref = never smoker)	0.99 (0.94, 1.05)
Geographic region of residence: Midwest US (ref = South US)	1.00 (0.94, 1.07)
Previously tried to conceive for $\geq$ 12 months: "yes" (ref = "no")	1.00 (0.94, 1.07)

Abbreviations: CI, confidence interval; HPV, human papillomavirus; US, United States.

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

[2] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable in the preprocessing phase and forced every non-reference level to be included in the model if any level of the categorical variable was selected. These variables are listed in addition to the variables selected by the sparse model.

predictors of miscarriage were female age at conception, male partner age at conception, and history of unplanned pregnancy, each of which was positively associated

Table 5.6: Variables selected by sparse survival model to predict miscarriage in PRESTO, 2013-2022.

Variable	**Hazard Ratio[1] (95% CI)**
Female age at conception (years)	1.19 (1.11, 1.27)
History of miscarriage (yes/no)	1.10 (1.03, 1.17)
Ever pregnant before (yes/no)	0.95 (0.87, 1.05)
Use of omega-3 or fish oil supplements (yes/no)	1.04 (0.99, 1.10)
Number of prior pregnancies	1.04 (0.99, 1.10)
History of subfertility or infertility[2] (yes/no)	1.04 (0.97, 1.11)
Male age at conception (years)	1.03 (0.97, 1.10)
Ever received HPV vaccine	0.98 (0.93, 1.04)
History of unplanned pregnancy (yes/no)	1.01 (0.94, 1.09)
Previously tried to conceive for ≥ 12 months[2]: "no, never tried before" (ref = "no")	1.00 (0.93, 1.08)
Variables forced into the model[3]	
Previously tried to conceive for ≥ 12 months: "yes" (ref = "no")	0.99 (0.77, 1.26)

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

[2] History of subfertility or infertility is derived from participants' responses to questions about their reproductive history and was defined as having previously tried to conceive for ≥ 6 months for any prior pregnancy; previously tried to conceive for ≥ 12 months was participants' response to the question, "have you ever tried for ≥ 12 months without conceiving?"

[3] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable in the preprocessing phase and forced every non-reference level to be included in the model if any level of the categorical variable was selected. These variables are listed in addition to the variables selected by the sparse model.

Abbreviations: CI, confidence interval; HPV, human papillomavirus.

with miscarriage (Table 5.7). The Healthy Eating Index-2010 score was also selected into this model and was inversely associated with miscarriage. The concordance index for this model was 55.6%.

When we restricted to primigravid participants (n=4,267), 9 variables remained after statistical feature selection. In this model, variables that were positively associated with miscarriage included female age at conception, male age at conception, use of omega-3 or fish oil supplements, recent use of psychotropic medications, and female BMI; variables that were inversely associated with miscarriage included being

Table 5.7: Variables selected by the sparse survival model predicting miscarriage after restricting to ≥8 gestational weeks in PRESTO, 2013-2022.

Variable	**Hazard Ratio[1] (95% CI)**
Female age at conception (years)	1.17 (1.05, 1.30)
Male age at conception (years)	1.09 (0.98, 1.20)
History of unplanned pregnancy (yes/no)	1.07 (0.98, 1.16)
Healthy Eating Index-2010 score (HEI-2010 score)	0.96 (0.88, 1.04)

Note: The Sparse and Full models were equivalent.

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

married, use of oral contraceptives as the most recent contraceptive method, residence in the Northeast US, and the Healthy Eating Index-2010 score (Table 5.8). The concordance index for this model was 57.4%. Among primigravid participants who contributed ≥8 gestational weeks to the analysis (n=3,488), only female and male partner age remained after statistical feature selection, and the concordance index was 53.3% (Table 5.9).

5.3.2 Static models

Variables selected into the full static models are presented in Table S3. After recursive feature elimination, there were 9 variables in the sparse model (Table 5.10). Performance metrics for all static models are presented in Table 5.11). The weighted-F1 score ranged from 72.6% for the LR-L1 model to 73.5% for the RF model. The two most important variables selected into the sparse static model were female age at conception and history of miscarriage, which were both positively associated with miscarriage.

When we restricted the incident period to ≥8 gestational weeks (6,993 pregnancies), 4 features remained after statistical feature selection, and 2 remained after recursive feature elimination (Table 5.12)). Female and male age at conception were the final two variables selected into the sparse model, with a weighted-F1 score of

Table 5.8: Variables selected by the sparse survival model predicting miscarriage among primigravid participants in PRESTO, 2013-2022.

Variable	**Hazard Ratio[1] (95% CI)**
Married (yes/no)	0.94 (0.88, 0.99)
Female age at conception (years)	1.07 (1.00, 1.14)
Last method of contraception was oral contraceptives (yes/no)	0.94 (0.88, 1.00)
Geographic region of residence: Northeast US (ref = South US)	0.94 (0.88, 1.01)
Male age at conception (years)	1.05 (0.98, 1.13)
Use of omega-3 or fish oil supplements (yes/no)	1.05 (0.99, 1.11)
Recent use of psychotropic medications (yes/no)	1.04 (0.98, 1.10)
Female BMI (kg/m^2)	1.04 (0.97, 1.10)
Healthy Eating Index-2010 score (HEI-2010 score)	0.97 (0.91, 1.03)
Variables forced into the model[2]	
Geographic region of residence: West US (ref = South US)	1.05 (0.99, 1.12)
Geographic region of residence: Canada (ref = South US)	0.99 (0.92, 1.05)
Geographic region of residence: Midwest US (ref = South US)	1.01 (0.95, 1.08)

Abbreviations: BMI, body mass index; CI, confidence interval; US, United States.

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

[2] For all models, we selected a reference group for each categorical variable that was recoded as an indicator variable in the preprocessing phase and forced every non-reference level to be included in the model if any level of the categorical variable was selected. These variables are listed in addition to the variables selected by the sparse model.

Table 5.9: Variables selected by the sparse survival model predicting miscarriage after restricting to ≥8 gestational weeks among primigravid participants in PRESTO, 2013-2022.

Variable	**Hazard Ratio[1] (95% CI)**
Female age at conception (years)	1.07 (0.98, 1.18)
Male age at conception (years)	1.06 (0.96, 1.16)

[1] Continuous variables were standardized; the effect estimate is the hazard ratio for a one-unit increase in the z-score for that variable.

88.0%. Among primigravid participants (n=4,267), 9 features remained after statistical feature selection, and all of these remained after recursive feature elimination. The weighted-F1 score of the sparse model was 73.8%, and the two most important variables selected into the model were residing in the Northeast US (negatively associated with miscarriage) and female age at conception (positively associated with miscarriage) (Table 5.13). Among primigravid participants with pregnancies lasting ≥8 gestational weeks (n=3,488), 2 features remained after statistical feature selection and only 1 remained in the final sparse model: male age at conception (Table 5.14). The weighted-F1 score for this model was 88.5%.

5.3.3 Numerical Experiments for Comparison of Original Cox Model and DRO Cox Models

Equation 2.24 introduces $O(N^2)$ constraints, significantly hindering computational efficiency. To address this, we impose constraints only for $i \leq k \leq i + r$, reducing the total number of constraints to $O(rN)$. For example, when we set r=2, only two constraints are added instead of N constraints.

Due to computational complexity, we aim to reduce the dimentianality. For this sake, we perform feature selection, we also randomly resample a portion of each dataset for our experiments. This results in a dataset with 872 participants and 20 features.

To assess the impact of outliers on model performance, we introduce varying proportions of outliers, from 5% to 30%, into a random subsample of the datasets. We evaluate the Original Cox model, the Sample-splitting DRO-Cox, and the Global fixation DRO-Cox, training each with different radii ϵ. The concordance indices of these models are then compared, as shown in Table 5.15.

Table 5.15 reveals that the Global fixation DRO-Cox model consistently outperforms both the Sample-splitting DRO-Cox and the Original Cox model. This comparison underscores the influence of outlier inclusion on the predictive accuracy and robustness of survival analysis models, especially in the presence of outliers.

5.4 Discussion

In this prospective cohort study of North American pregnancy planners, we developed predictive models for miscarriage based on self-reported preconception data. Previous studies have identified few confirmed causes of miscarriage, and the strongest identified risk factors in these studies were age and history of miscarriage (Wilcox et al., 1988). In the present study, we generated models with moderate predictive power: the weighted-F1 score ranged from 73-89% for static models and the concordance index ranged from 53-56% for survival models. However, the AUC was <60% for all static models. Consistent with previous studies, our findings indicate that advancing female and male partner age are the most important predictors of miscarriage, and that female age is generally more predictive than male age. After age, history of miscarriage appeared to be the strongest predictor of miscarriage. These factors were consistently predictive of miscarriage across a variety of models and settings.

Our study identified several preconception dietary factors as predictors of miscarriage, albeit most associations were very small and consistent with the null. Specifically, a healthier diet as measured by the Healthy Eating Index-2010 score (e.g., greater intake of fruits and vegetables, whole grains, dairy, seafood & plant proteins, and unsaturated fats) was associated with a slightly lower rate of miscarriage. In addition, use of omega-3 or fish oil supplements was associated with a slightly increased rate of miscarriage and several B-vitamins were selected with inconsistent associations. Several studies have investigated the relation between dietary factors and miscarriage (Hsiao et al., 2019)(Gaskins et al., 2014)(Laursen et al., 2022)(Gaskins et al., 2019)(Karayiannis et al., 2018)(Twigt et al., 2012)(Wesselink et al., 2022). One study – with a similar design to PRESTO – reported an inverse association between adherence to Nordic dietary guidelines (which emphasize fish consumption) and risk of miscarriage (Laursen et al., 2022). Another study evaluated the association between pre-pregnancy adherence to three dietary patterns – the Healthy Eating Index 2010, the Alternative Mediterranean Diet, and the Fertility Diet (FD) – and risk of miscarriage among 15,950 pregnancies in the Nurses' Health Study II (Gaskins et al., 2014). The authors reported no association between these dietary patterns and miscarriage. The role of dietary factors remains debated, and the predictive ability of these variables in our study was small.

An unexpected finding in our study was the selection of smoking status into the sparse static model and the full survival model in the full study population (i.e., not restricted by gravidity or gestational week). However, the overall prevalence of smoking was quite low in this study population (4%), and this variable was not consistently selected into all models. Moreover, the detrimental health effects of smoking tobacco are well documented, and several studies have identified a positive association between current smoking and miscarriage risk (George et al., 2006) (Nielsen et al., 2006).

The following variables were selected into models developed among primigravid participants but not among those who were previously pregnant: marital status, pregravid use of oral contraceptives, recent use of psychotropic medications, and female BMI. Being married was associated with a lower rate of miscarriage, which could be related to higher socioeconomic position, greater social and emotional support, and lower stress levels. However, factors such as perceived stress scores and household income were not selected as important predictors of miscarriage during the statistical feature selection process.

Some (Hahn et al., 2015)(Sackoff et al., 1994)(Rothman, 1977) but not all (Risch et al., 1988)(Jellesen et al., 2008) studies reported that pregravid use of oral contraceptives was associated with a lower risk of miscarriage compared with non-use, in agreement with the present study. However, a recently published paper conducted in

PRESTO reported that pregravid use of oral contraceptives was not associated with miscarriage (Yland et al., 2020). This contrast may be due to differences in model selection, as the previous publication aimed to estimate potential causal effects of contraceptive use. The potential association between use of psychotropic medications and miscarriage has been debated. However, a recent study reported that use of antidepressants was not associated with miscarriage after controlling for depression diagnosis (Kjaersgaard et al., 2013). High BMI has previously been associated with an increased risk of miscarriage (Arck et al., 2008)(Savitz et al., 2008). Among 5,132 couples who conceived in a Danish preconception cohort study, the adjusted HR for miscarriage among women with BMI $\geq$30 kg/m^2 relative to those with BMI 20-24 kg/m^2 was 1.23 (95% CI: 0.98, 1.54) (Savitz et al., 2008).

We attempted to isolate predictors of later miscarriage, as earlier miscarriages ($<$8 weeks' gestation) are more likely to be due to chromosomal abnormalities than later losses (Pflueger, 2005). However, the predictive ability of models restricted to $\geq$8 gestational weeks was no better than those generated in the entire dataset, and the list of variables selected for these models was similar to those based on full spectrum of gestational ages (all miscarriages).

Previous studies have developed models to predict miscarriage in special populations, such as couples with recurrent miscarriage (Quenby and Farquharson, 1993) (Caetano et al., 2006) (Dai et al., 2022) (du Fossé et al., 2022) or those using ART (Choong et al., 2003) (Yi et al., 2016) (Liu et al., 2020). These studies largely relied upon ultrasound measurements (e.g., gestational sac size, crown-rump length, fetal heart rate) or laboratory values (e.g., beta-human chorionic gonadotropin, progesterone levels) during early pregnancy. One study in the Netherlands attempted to predict pregnancy outcome among 526 couples with unexplained recurrent miscarriage (du Fossé et al., 2022). Data on previous miscarriages and fertility treatment;

and male and female age, BMI, and smoking status were included, and all were identified as potential predictors of miscarriage, with an AUC of 0.66. The present study greatly expands on the breadth of potential predictors assessed. Moreover, our findings might be useful for couples who wish to understand their risk for miscarriage before trying to conceive spontaneously.

Study limitations include bias due to missingness or misclassification of predictor variables. All data were self-reported, and certain variables such as dietary factors or medication use may be more vulnerable to misclassification than others. The impact of misclassification on our findings is challenging to quantify, as there is little research on the impact of measurement error on machine learning prediction models (Jiang et al., 2021) (van Doorn et al., 2017). Outcome misclassification is also possible but unlikely: more than 95% of participants reported using at-home-pregnancy tests and we ascertained miscarriages as early as 3 weeks' gestation. In addition, although we evaluated a wide range of variables, we were unable to include environmental exposures (e.g., phthalates, phenols, pesticides, etc.) as potential predictors. Moreover, we did not evaluate interactions between the independent variables, such as depressive symptoms and use of psychotropic medications. Finally, though we validated the models using split-sample replication techniques, we were unable to conduct an external validation study. Given that more than 93 of PRESTO participants had spontaneous conceptions, our results may not generalize to ART-conceived conceptions.

5.5 Conclusions

In this study, we used a variety of supervised machine learning methods to generate predictive models of miscarriage based on self-reported preconception data. We considered 160 potential predictors of miscarriage and analyzed data from nearly 9,000 pregnancies. Female age, male age, and history of miscarriage were the most important predictors of miscarriage, consistent with existing knowledge. The overall performance of our models was moderate. Our findings suggest that predictability of miscarriage is limited based on preconception lifestyle characteristics, including reproductive and medical factors (Yland et al., 2024).

Table 5.10: Variables selected by sparse static model (L2LR)

Variable	OR (95% CI)	β	Correlation with outcome	Overall frequency/mean (std.)	Mean of Miscarriage	Mean of No miscarriage
Female age at conception (years)	1.23 (1.20, 1.27)	0.21	0.09	30.2 (3.9)	30.9	30.0
History of miscarriage (yes/no)	1.16 (1.13, 1.20)	0.15	0.07	26% (44%)	32%	24%
Female smoking: current regular smoker (ref = never smoker)	0.89 (0.86, 0.91)	-0.12	-0.04	4% (19%)	3%	4%
Geographic region of residence: Northeast US (ref = South US)	0.90 (0.87, 0.92)	-0.11	-0.04	24% (43%)	21%	25%
Healthy Eating Index-2010 score (HEI-2010 score)	0.92 (0.90, 0.95)	-0.08	-0.02	66.8 (9.2)	66.4	66.8
Use of omega-3 or fish oil supplements (yes/no)	1.06 (1.04, 1.09)	0.06	0.04	19% (39%)	22%	18%
Use of vitamin B6 (yes/no)	1.05 (1.02, 1.08)	0.05	0.04	5% (21%)	6%	4%
Ever pregnant before (yes/no)	0.96 (0.93, 0.99)	-0.04	-0.04	51% (50%)	55%	50%
Use of vitamin C (yes/no)	1.04 (1.01, 1.07)	0.04	0.03	7% (25%)	8%	6%
Geographic region of residence: Canada (ref = South US)[1]	0.97 (0.94, 1.00)	-0.03	0.00	16% (37%)	15%	16%
Female smoking: former smoker (ref = never smoker)[1]	0.97 (0.95, 0.99)	-0.03	0.00	12% (33%)	13%	12%
Geographic region of residence: Midwest US (ref = South US)[1]	0.99 (0.96, 1.02)	-0.01	0.01	22% (41%)	22%	22%
Female smoking: current occasional smoker (ref = never smoker)[1]	0.99 (0.97, 1.01)	-0.01	0.00	3% (16%)	3%	3%
Geographic region of residence: West US (ref = South US)[1]	1.00 (0.97, 1.03)	0.00	0.02	16% (37%)	18%	16%

[1] Variables forced into models

Table 5.11: Performance metrics for the static models predicting miscarriage in PRESTO, 2013-2022.

Algorithm	**Performance Measure (%) (Standard Deviation)**				
	AUC	**Accuracy**	**Weighted-F1 Score**	**Weighted Precision Score**	**Weighted Recall Score**
Full population					
LR-L1	56.8 (1.0)	75.8 (0.6)	72.6 (0.3)	70.9 (0.3)	75.8 (0.6)
SVM-L1	56.9 (1.0)	76.3 (0.9)	72.8 (0.2)	71.2 (0.4)	76.3 (0.9)
GBT	60.5 (0.7)	77.1 (0.9)	73.0 (0.5)	72.0 (0.5)	77.1 (0.9)
RF	59.3 (1.1)	77.5 (1.0)	73.5 (0.7)	72.2 (1.2)	77.5 (1.0)
LR-L2 RFE	57.6 (0.6)	76.3 (0.9)	72.7 (0.5)	71.0 (0.6)	76.3 (0.9)
Subset: ≥8 Gestational Weeks					
LR-L1	55.6 (2.8)	91.4 (0.2)	88.2 (0.2)	86.4 (0.6)	91.4 (0.2)
SVM-L1	55.6 (2.8)	91.4 (0.2)	88.2 (0.2)	86.4 (0.6)	91.4 (0.2)
GBT	58.5 (2.8)	91.3 (0.4)	88.2 (0.2)	86.7 (0.7)	91.4 (0.2)
RF	57.0 (3.0)	90.4 (0.6)	87.8 (0.4)	86.3 (0.9)	90.4 (0.4)
LR-L2 RFE	56.4 (2.5)	91.2 (0.6)	88.0 (0.2)	85.9 (0.8)	91.2 (0.6)
Subset: Primigravid					
LR-L1	57.3 (1.1)	78.6 (0.6)	73.8 (0.7)	71.7 (1.2)	78.6 (0.6)
SVM-L1	57.2 (1.1)	78.5 (0.7)	73.8 (0.7)	71.7 (1.2)	78.5 (0.7)
GBT	57.6 (1.9)	77.7 (1.7)	74.0 (1.3)	71.6 (1.2)	77.7 (1.7)
RF	56.0 (2.1)	75.2 (2.0)	71.7 (1.3)	70.1 (1.3)	75.2 (2.0)
LR-L2 RFE	57.3 (1.1)	78.6 (0.6)	73.8 (0.7)	71.7 (1.0)	78.6 (0.6)
Subset: Primigravid ≥8 Gestational Weeks					
LR-L1	53.4 (4.1)	92.0 (0.2)	88.7 (0.3)	86.0 (1.5)	92.0 (0.2)
SVM-L1	53.4 (4.1)	92.0 (0.2)	88.7 (0.3)	86.0 (1.5)	92.0 (0.2)
GBT	51.2 (4.8)	90.9 (0.4)	88.5 (0.3)	85.6 (0.5)	90.9 (0.4)
RF	55.1 (4.5)	91.6 (0.6)	88.5 (0.7)	85.0 (0.7)	91.6 (0.4)
LR-L2 RFE	55.5 (3.5)	91.7 (0.5)	88.2 (0.6)	85.5 (0.4)	91.7 (0.5)

Abbreviations: LR-L1=logistic regression with an ℓ_1-norm regularization term; SVM-L1=support vector machines with an ℓ_1-norm regularization term; GBT=Gradient Boosted Trees; RF=Random Forest; LR-L2 RFE=logistic regression with an ℓ_2-norm regularization term.

Table 5.12: Variables selected by the sparse static model (logistic regression with an ℓ_2-norm regularization term) predicting miscarriage after restricting to ≥8 gestational weeks in PRESTO, 2013-2022.

Variable	**OR (95% CI)**	β	**Correlation with outcome**	**Overall mean (std.)**	**Mean, by outcome status**[1]	
					Miscarriage	**No miscarriage**
Female age at conception (years)	1.19 (1.14, 1.23)	0.17	0.07	30.1 (3.8)	30.9	30.0
Male age at conception (years)	1.09 (1.06, 1.13)	0.09	0.06	31.9 (4.9)	32.9	31.9

Abbreviations: β, regression coefficient; CI, confidence interval; LR-L2, logistic regression model with an L2 penalty; OR, odds ratio (exp[β]); RFE, recursive feature elimination; std, standard deviation.

[1] These cells should be interpreted as the mean for each variable among individuals with or without miscarriage. For example, the average age of female participants who experienced a miscarriage was 30.9 years.

Table 5.13: Variables selected by the sparse static model (logistic regression with an ℓ_2-norm regularization term) predicting miscarriage among primigravid participants in PRESTO, 2013-2022.

Variable	**OR (95% CI)**	β	**Corr. with outcome**	**Overall mean (std.)**	**Mean,by outcome status**[1]	
					Miscarriage	**No miscarriage**
Geographic region of residence: Northeast US (ref = South US)	0.88 (0.84, 0.92)	-0.13	-0.05	26% (44%)	22%	27%
Female age at conception (years)	1.12 (1.07, 1.17)	0.11	0.05	29.5 (3.5)	29.9	29.4
Married (yes/no)	0.90 (0.87, 0.94)	-0.10	-0.05	94% (24%)	91%	95%
Last method of contraception was oral contraceptives (yes/no)	0.90 (0.87, 0.94)	-0.10	-0.04	29% (45%)	25%	30%
Use of omega-3 or fish oil supplements (yes/no)	1.09 (1.06, 1.13)	0.09	0.04	18% (39%)	22%	18%
Recent use of psychotropic medications (yes/no)	1.09 (1.06, 1.13)	0.09	0.05	13% (33%)	16%	12%
Female BMI (kg/m^2)	1.05 (1.01, 1.09)	0.05	0.03	26.3 (6.2)	26.7	26.2
Male age at conception (years)	1.04 (0.99, 1.09)	0.04	0.05	31.3 (4.6)	31.8	31.2
Healthy Eating Index-2010 score (HEI-2010 score)	0.96 (0.93, 1.00)	-0.04	-0.02	67.5 (9.0)	67.2	67.5
Variables forced into the model						
Geographic region of residence: West US (ref = South US)	1.05 (1.01, 1.10)	0.05	0.04	15% (36%)	19%	15%
Geographic region of residence: Canada (ref = South US)	0.95 (0.91, 0.99)	-0.05	-0.01	17% (38%)	16%	18%
Geographic region of residence: Midwest US (ref = South US)	0.98 (0.94, 1.02)	-0.02	0.01	20% (40%)	21%	20%

[1] These cells should be interpreted as the mean or percentage for each variable among individuals with or without miscarriage. For example, the average age of female participants who experienced a miscarriage was 29.9 years.

Table 5.14: Variables selected by the sparse static model (logistic regression with an ℓ_2-norm regularization term) predicting miscarriage after restricting to ≥8 gestational weeks among primigravid participants in PRESTO, 2013-2022.

Variable	**OR (95% CI)**	β	**Corr. with outcome**	**Overall mean (std.)**	**Mean,by outcome status[1]**	
					Miscarriage	**No miscarriage**
Male age at conception (years)	1.25 (1.20, 1.30)	0.22	0.06	31.3 (4.6)	32.3	31.2

Abbreviations: β, regression coefficient; CI, confidence interval; LR-L2, logistic regression model with an L2 penalty; OR, odds ratio (exp[β]); RFE, recursive feature elimination; std, standard deviation.

[1] These cells should be interpreted as the mean for each variable among individuals with or without miscarriage. For example, the average age of male participants who experienced a miscarriage was 32.3 years.

Ratio of Outliers	0.1	0.15	0.2	0.25	0.3
Original Cox	0.5828	0.5762	0.5608	0.5815	0.6035
DRO-Cox Sample Splitting	0.5147	0.5185	0.5196	0.5013	0.5167
DRO-Cox Global Fixation	0.5830	0.5776	0.5652	0.5924	0.6099

Table 5.15: Comparison of concordance indices for different ratios of outliers with $\epsilon = 0.005$ in a subset of the miscarriage dataset.

Chapter 6

Predictive Models of Deep Molecular Response to Imatinib Treatment in Chronic Myeloid Leukemia patients

6.1 Introduction

Chronic Myeloid Leukemia (CML) is a myeloproliferative neoplasm characterized by the accumulation of circulating leukemic cells with the Philadelphia chromosome (Ph), resultant of translocation t(9;22) (Nowell and Hungerford, 1960). This translocation gives rise to the fusion gene BCR::ABL1, which encodes a constitutively active tyrosine kinase protein (Shitvelman et al., 1985). The actionable gene fusion led to the development of specific and effective tyrosine kinase inhibitors (TKI) and CML has become one of the major cases of success in cancer history (Mughal et al., 2016). Two decades following the introduction of imatinib (IM), the management of CML has evolved significantly (Michel et al., 2019)(Michel et al., 2019). The focal point of optimal CML patient care currently revolves around achieving and sustaining a prolonged Deep Molecular Response (DMR), which, when accomplished, renders patients eligible for treatment-free remission (TFR) (Schiffer, 2019). Beyond enhancing patient well-being, this holds implications for both patient outcomes and pharmacoeconomics, particularly relevant within the economic landscape of low or middle-income countries such as Brazil.

Monitoring the response to TKI treatment is carried out by quantitative polymerase chain reaction (qPCR) at determined timepoints, as recommended by the European Leukemia Net (ELN) (Hochhaus et al., 2020). Despite the well-established framework for molecular treatment surveillance in CML, existing predictive models for critical factors such as resistance, risk of progression, and the enduring molecular response post-therapy remain limited. This study aims to fill this gap by developing predictive models for DMR achievement in CML patients undergoing IM therapy. Leveraging comprehensive clinical data and early stages BCR::ABL1/ABL1IS quantification, our goal is to predict the likelihood of DMR achievement in future phases.

6.2 Materials and methods

6.2.1 Data description

The proposed predictive models encompass both linear and non-linear classification methodologies, meticulously crafted based on a dataset comprising 219 CML patients from the Brazilian National Cancer Institute (INCA). Only patients who received exclusively IM treatment were included in this cohort in Figure 6·1. We retrospectively analyzed de-identified molecular, clinical and laboratory data from all recruited patients in Figure 6·2.

A detailed description of the dataset is provided in Section 6.2.1.1; and the inclusion/exclusion of features is described in Tables 6.1, 6.2, and 6.3. As mentioned, in this study molecular response was defined by the BCR::ABL1/ABL1% ratio measured by qPCR in International Scale (IS) at three-month intervals, following ELN recommendations (Hochhaus et al., 2020). Time 0 was defined at the start of IM; we then simplified our model considering ELN monitoring milestones of 3, 6, 12, 18, 24 and 60 months.

Because missing values precluded accurate estimation, we imputed missing values with the median for all variables, except for BCR::ABL1/ABL1% and cytogenetic

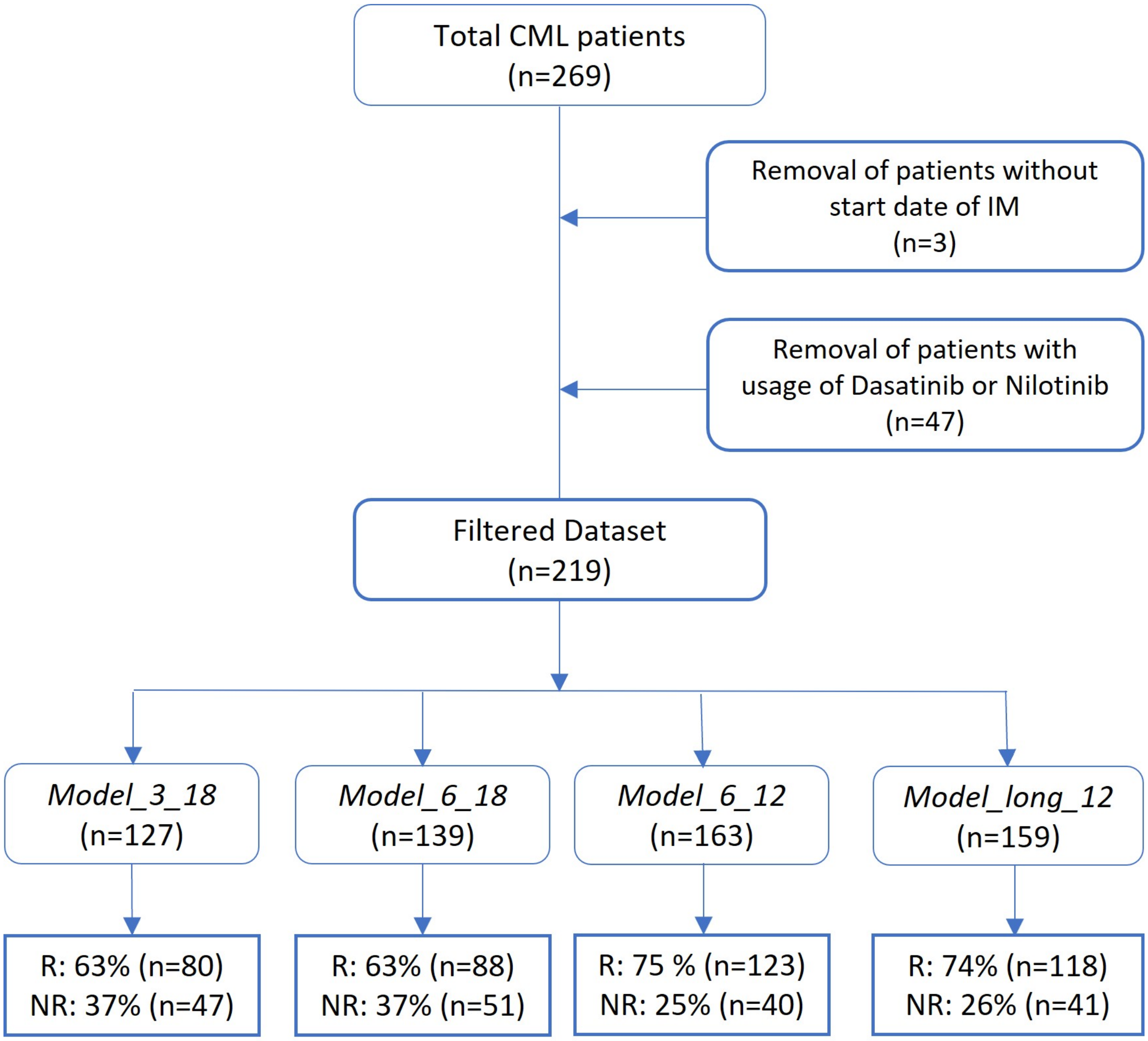
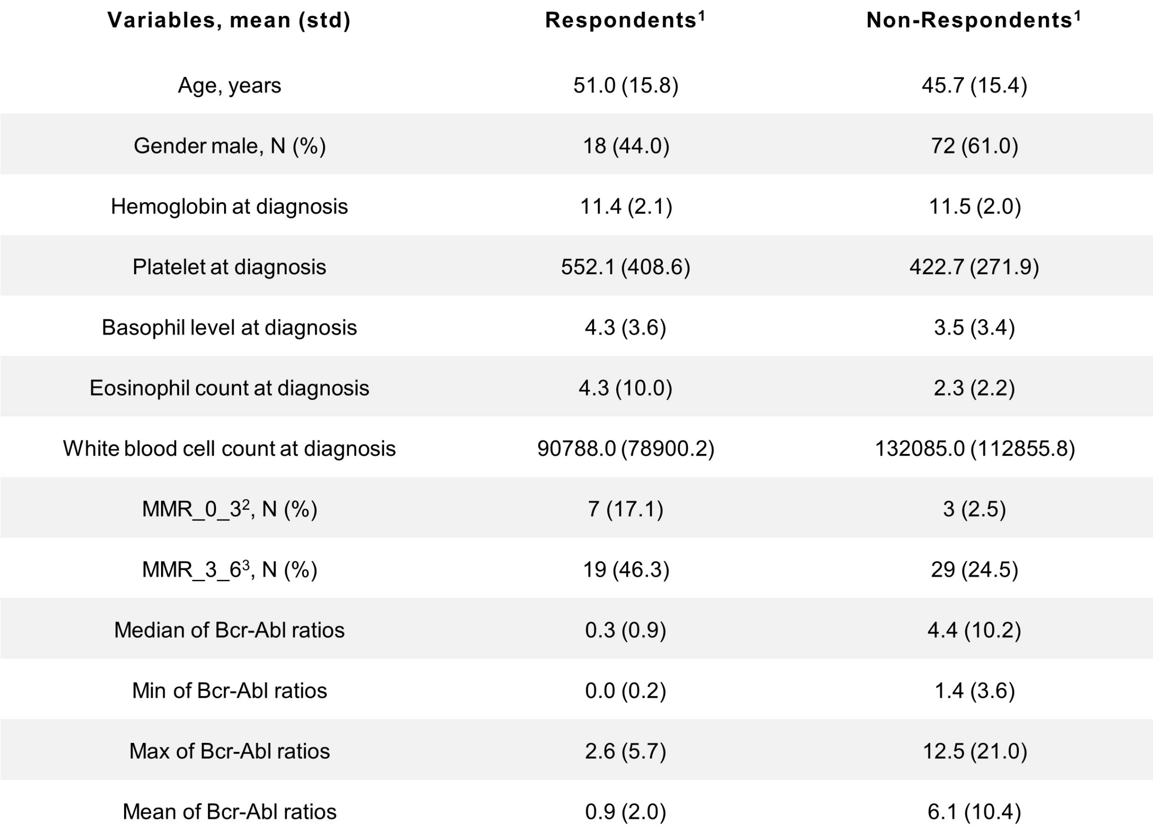


Figure 6·1: The exclusion criteria of total CML patients, the number of patients in each model, and the number/percentage of patients who achieve DMR (responders, represented as [R]) and do not achieve DMR (non-responders, represented as [NR]) at the time of each model's outcome.



Variables, mean (std)	Respondents[1]	Non-Respondents[1]
Age, years	51.0 (15.8)	45.7 (15.4)
Gender male, N (%)	18 (44.0)	72 (61.0)
Hemoglobin at diagnosis	11.4 (2.1)	11.5 (2.0)
Platelet at diagnosis	552.1 (408.6)	422.7 (271.9)
Basophil level at diagnosis	4.3 (3.6)	3.5 (3.4)
Eosinophil count at diagnosis	4.3 (10.0)	2.3 (2.2)
White blood cell count at diagnosis	90788.0 (78900.2)	132085.0 (112855.8)
MMR_0_3[2], N (%)	7 (17.1)	3 (2.5)
MMR_3_6[3], N (%)	19 (46.3)	29 (24.5)
Median of Bcr-Abl ratios	0.3 (0.9)	4.4 (10.2)
Min of Bcr-Abl ratios	0.0 (0.2)	1.4 (3.6)
Max of Bcr-Abl ratios	2.6 (5.7)	12.5 (21.0)
Mean of Bcr-Abl ratios	0.9 (2.0)	6.1 (10.4)

Figure 6·2: The baseline characteristics of patients.

[1] The respondent and non-respondent definition is based on Model_long_12 where respondents are the patients who reach DMR at 24 months after the start of the treatment and maintain this status until 60 months after treatment initiation.

[2] MR_0_3 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the first three-month interval after start of IM.

[3] MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from 3 to 6 months after the start of IM.

response (CyR) ratios. For these two ratios, we computed summary statistics such as the median, mean, min, max, and standard deviation (std) of the recorded values for each patient, and we used these as features instead of the original recorded values of BCR::ABL1/ABL1% and cytogenetic response (CyR) ratios. We call these new features aggregated features.

Table 6.1: Description of features included in the Original Dataset. Part 1.

Record time	Features Category	Features
At diagnosis	Demographic Blood characteristics Cytogenetics	’SEX’, ’AGE_AT_DIAGNOSIS’, ’CML_PHASE_AT_DIAGNOSIS’, ’SPLEEN_SIZE_AT_DIAGNOSIS_(CM)’, ’WHITE_BLOOD_CELL_COUNT_AT_DIAGNOSIS’, ’HEMOGLOBIN_AT_DIAGNOSIS’, ’PLATELET_AT_DIAGNOSIS’, ’BASOPHIL_LEVEL_AT_DIAGNOSIS’, ’EOSINOPHIL_COUNT_AT_DIAGNOSIS’, ’BLAST_COUNT_AT_DIAGNOSIS’, ’SOKAL_INDEX’, ’SOKAL_CLASSIFICATION’, ’TRANSCRIPTS_AT_DIAGNOSIS’, ’CYTOGENETICS_AT_DIAGNOSIS’
At start of IM	Blood characteristics Cytogenetics	’WHITE_BLOOD_CELL_COUNT_AT_START_IM’, ’PLATELET_COUNT_AT_START_IM’, ’HEMOGLOBIN_COUNT_AT_START_IM’, ’BLAST_COUNT_AT_START_IM’, ’EOSINOPHIL_COUNT_AT_START_IM’, ’BASOPHIL_COUNT_AT_START_IM’, ’CYTOGENETICS_START_IM(%)’
After start of IM in three-month intervals and so on	Treatment-related Treatments: Hydroxyurea (Hydrea) Interferon (INF) Hematopoietic stem cell transplantation (BMT)	Three-month interval from 0 (start of IM) to 3 ’Hydrea_0_3’, ’INF_0_3’, ’BMT_0_3’, ’IM_0_3’, ’CHANGE_DOSAGE_0_3’
		Three-month interval from 3 to 6 ’Hydrea_3_6’, ’INF_3_6’, ’BMT_3_6’, ’IM_3_6’, ’CHANGE_DOSAGE_3_6’
		Three-month interval from 6 to 9 ’Hydrea_6_9’, ’INF_6_9’, ’BMT_6_9’, ’IM_6_9’, ’CHANGE_DOSAGE_6_9’
		Three-month interval from 9 to 12 ’Hydrea_9_12’, ’INF_9_12’, ’BMT_9_12’, ’IM_9_12’, ’CHANGE_DOSAGE_9_12’
		Interval from 12 to 18 ’Hydrea_12_18’, ’INF_12_18’, ’BMT_12_18’, ’IM_12_18’, ’CHANGE_DOSAGE_12_18’
		After 18 months after start of IM ’Hydrea_after_18_mo’, ’INF_after_18_mo’, ’BMT_after_18_mo’, ’IM_after_18’, ’CHANGE_DOSAGE_after_18’

Table 6.2: Description of features included in the Original Dataset. Part 2.

Record time	**Features Category**	**Features**
After start of IM in three-month intervals and so on	Responses: Hematologic Response (HR), Major Molecular Response (MR)	Three-month interval from 0 (start of IM) to 3 ’HR_0_3’, ’HR_LOSS_0_3’, ’MR_0_3’, ’MR_LOSS_0_3’
		Three-month interval from 3 to 6 ’HR_3_6’, ’HR_LOSS_3_6’, ’MR_3_6’, ’MR_LOSS_3_6’
		Three-month interval from 6 to 9 ’HR_6_9’, ’HR_LOSS_6_9’, ’MR_6_9’, ’MR_LOSS_6_9’
		Three-month interval from 9 to 12 ’HR_9_12’, ’HR_LOSS_9_12’, ’MR_9_12’, ’MR_LOSS_9_12’
	Aggregated features: BCR-ABL1 ratios: ’Mol_median’, ’Mol_min’, ’Mol_max’, ’Mol_mean’, ’Mol_std’ Cytogenetic Response: ’CyR_median’, ’CyR_min’, ’CyR_max’, ’CyR_mean’	’Mol_3m_IM’, ’Mol_6m_IM’, ’Mol_9m_IM’, ’Mol_12m_IM’, ’Mol_15m_IM’, ’Mol_18m_IM’, ’Mol_21m_IM’, ’Mol_24m_IM’, ’Mol_30m_IM’, ’Mol_60m_IM’, ’CYTOGENETIC_RESPONSE_3mo_IM’, ’CYTOGENETIC_RESPONSE_6mo_IM’, ’CYTOGENETIC_RESPONSE_12mo_IM’, ’CYTOGENETIC_RESPONSE_18mo_IM’

Table 6.3: Description of features included in the new dataset (combination of Original Dataset and Additional Dataset)

Record time	**Features Category**	**Features**
At diagnosis	Demographic Blood characteristics Cytogenetics	’SEX’, ’AGE_AT_DIAGNOSIS’, ’WHITE_BLOOD_CELL_COUNT_AT_DIAGNOSIS’, ’HEMOGLOBIN_AT_DIAGNOSIS’, ’PLATELET_AT_DIAGNOSIS’, ’BASOPHIL_LEVEL_AT_DIAGNOSIS’, ’EOSINOPHIL_COUNT_AT_DIAGNOSIS’
Responses: Major Molecular Response (MMR)		Three-month interval from 0 (start of IM) to 3 ’MR_0_3’, ’MR_LOSS_0_3’
		Three-month interval from 3 to 6 ’MR_3_6’, ’MR_LOSS_3_6’
	Aggregated features: BCR-ABL1 ratios: ’Mol_median’, ’Mol_min’, ’Mol_max’, ’Mol_mean’, ’Mol_std’	’Mol_3m_IM’, ’Mol_6m_IM’, ’Mol_12m_IM’, ’Mol_18m_IM’, ’Mol_24m_IM’, ’Mol_60m_IM’

6.2.1.1 Detailed dataset description

At the start of this study, we had a cohort of 144 patients, and we built our original models based on this dataset (Original Dataset). The preliminary results showed high standard deviation in evaluation metrics due to the small size of the dataset and its imbalanced nature. We subsequently gained access to other patients' information (Additional Dataset), and we tried to increase the size of the dataset as well as to decrease its imbalance by adding more respondent patients. However, their available information was not as inclusive as our Original Dataset; hence, we extracted patients who at least have the information required by our parsimonious models (models including features selected by feature selection). In our results, we display two kinds of tables: (i) Tables 6.1 and 6.2: associated with our Original Dataset of 144 patients for whom we have a wide range of information, and the primary feature selection is done based on this cohort. (ii) Table 6.3: associated with the new dataset (combination of Original Dataset and Additional Dataset). Figure 6·3 shows an overview of the steps in model development.

We considered two types of information: (i) variables evaluated only at diagnosis and/or start of IM treatment; (ii) variables with values available at three-month intervals after the start of IM. Qualitative data were defined as binary variables to indicate use (1) or no use (0) of treatments with hydroxyurea, interferon, and hematopoietic stem cell transplantation in each three-month time interval. Some of the earlier patients in our dataset were submitted to changes in IM dosage during treatment, so to account for that in our models, we defined a continuous variable equal to the average of imatinib dosages in each three-month interval. Because missing values precluded accurate estimation, we imputed the missing values with the median for all variables, except for BCR-ABL1 and cytogenetic response (CyR) ratios. For these two ratios, we computed summary statistics such as the median, mean, min, max, and standard

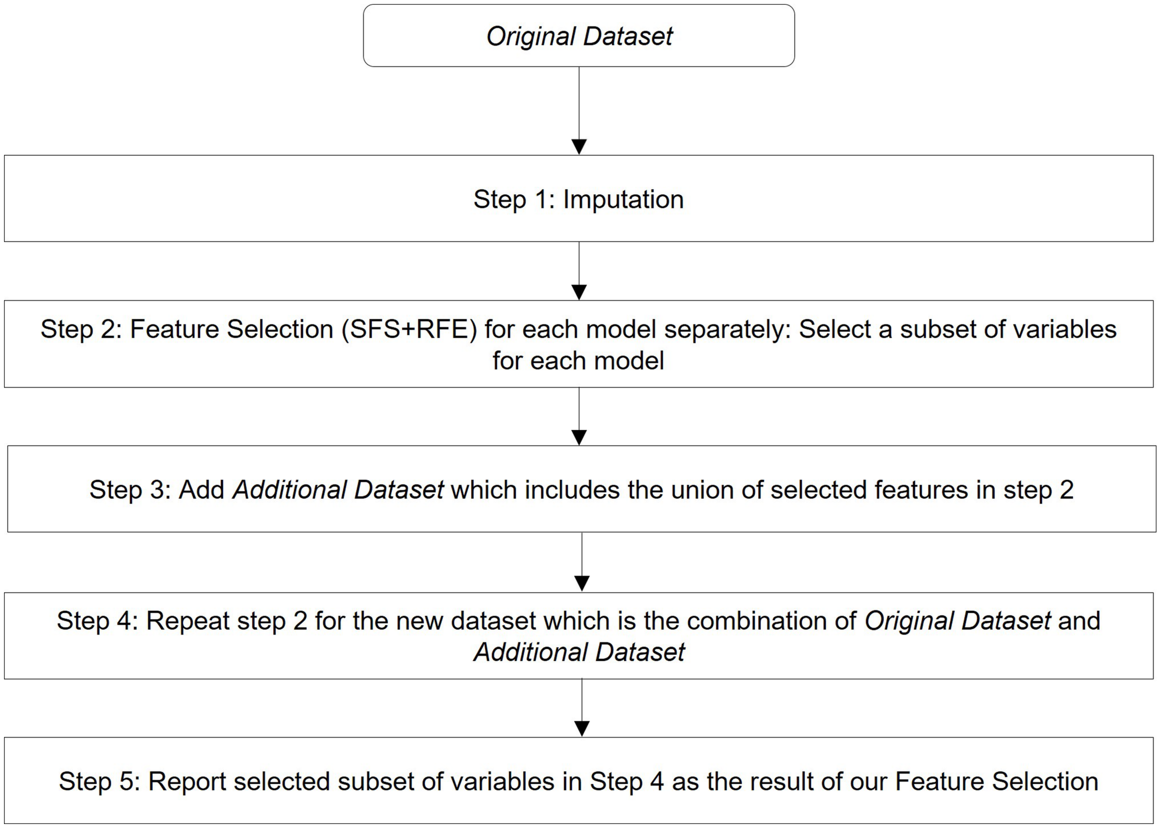

Original Dataset
Step 1: Imputation
Step 2: Feature Selection (SFS+RFE) for each model separately: Select a subset of variables for each model
Step 3: Add Additional Dataset which includes the union of selected features in step 2
Step 4: Repeat step 2 for the new dataset which is the combination of Original Dataset and Additional Dataset
Step 5: Report selected subset of variables in Step 4 as the result of our Feature Selection

Figure 6·3: An overview of the steps in model development.

deviation (std) of the recorded values for each patient, and we used these as features instead of the original recorded values of BCR::ABL1/ABL1% and cytogenetic response (CyR) ratios. We call these new features aggregated features. In our final models, some of the aggregated features are not used because they have high correlation with each other, and we retain only one variable among each highly correlated pair. Table 6.4 reports the percentage of missing values in the Original Dataset. Table 6.5 reports the percentage of missing values in the new dataset (combination of Original Dataset and Additional Dataset). Additionally, these tables present the count of non-missing entries for each variable considered. It is pertinent to note that within the scope of variables analyzed, "White Blood Cell count at diagnosis" is the sole predictor employed in our proposed models, as shown in the third column of the tables. Moreover, we utilize aggregated features rather than the raw values recorded for BCR::ABL1/ABL1% ratios. The model's reliance on selected variables suggests that the impact of missing values is mitigated, especially for variables not directly utilized as predictors. Hence, a variable's absence does not significantly detract from our analysis, provided that a decent number of non-missing data remains available for model training and validation. This approach ensures the robustness of our findings despite the inherent challenges posed by incomplete datasets. As an important clarification, it's essential to note that our models exclude patients lacking data for variables critical to the models' outcomes. For instance, in Model_3_18, patients are omitted from analysis if they do not have recorded values for the BCR-ABL ratio at 18 months post-initiation of IM treatment. Consequently, the absence of this specific data point does not impact the performance or integrity of Model_3_18. Furthermore, this particular value is not employed as a predictive variable in any of our other models.

6.2.2 Predictive Models

We performed data pre-processing steps, explauned in more detail in Section 2.1.2 to prepare the dataset for developing predictive models. The threshold for correlation coefficient is considered 0.75. To reduce the less informative features and simplify the models, we applied Statistical Feature Selection (SFS), explained in more detail in Section 2.1.5.2, with 0.05 threshold for p-value. Variables with no variability (std<0.0001) were removed

When predicting DMR at later months using patient characteristics at diagnosis and early treatment information, we considered the following models: (a) prediction of DMR at 18 months using information up to 3 months (Model_3_18), (b) prediction of DMR at 18 months using information up to 6 months (Model_6_18), and (c) prediction of DMR at 12 months using information up to 6 months (Model_6_12).

Table 6.4: The percentage of missing values and the count of non-missing values in the Original Dataset with 144 patients.

Feature	**Percentage of missing values (%)**	**Number of non-missing values**	**Used as a Most Important Predictor**
CYTOGENETIC RESPONSE 3mo IM	81	28	No
CYTOGENETIC RESPONSE 18mo IM	76	35	No
BCR-ABL ratio at 60 months after start of IM	72	41	No
BCR-ABL ratio at 21 months after start of IM	69	44	No
CYTOGENETIC RESPONSE 24mo IM	67	47	No
BCR-ABL ratio at 9 months after start of IM	64	52	No
BCR-ABL ratio at 15 months after start of IM	62	55	No
CYTOGENETIC RESPONSE 12mo IM	55	65	No[1]
CYTOGENETIC RESPONSE 6mo IM	50	72	No[1]
BCR-ABL ratio at 30 months after start of IM	43	82	No
BCR-ABL ratio at 24 months after start of IM	41	85	No
BCR-ABL ratio at 6 months after start of IM	38	90	No
BCR-ABL ratio at 18 months after start of IM	35	94	No
CYTOGENETICS START IM	28	103	No
BCR-ABL ratio at 3 months after start of IM	24	110	No[1]
BCR-ABL ratio at 12 months after start of IM	16	121	No
SPLEEN SIZE AT DIAGNOSIS	8	132	No
BLAST COUNT AT DIAGNOSIS	7	134	No
BASOPHIL LEVEL AT DIAGNOSIS	7	134	No
PLATELET_AT_DIAGNOSIS	7	134	No
WHITE BLOOD CELL COUNT AT DIAGNOSIS	6	136	Yes
EOSINOPHIL COUNT AT DIAGNOSIS	6	135	No
HEMOGLOBIN AT DIAGNOSIS	6	135	No
HEMOGLOBIN COUNT AT START IM	4	138	No
SOKAL CLASSIFICATION	4	138	No
WHITE BLOOD CELL COUNT AT START IM	4	138	No
TRANSCRIPTS_AT_DIAGNOSIS	4	138	No
BASOPHIL COUNT AT START IM	4	138	No
EOSINOPHIL COUNT AT START IM	4	138	No
SOKAL INDEX	4	138	No
BLAST COUNT AT START IM	4	138	No
PLATELET_COUNT_AT_START_IM	4	138	No

[1] We utilize *aggregated features* rather than the raw values recorded for BCR::ABL1/ABL1% ratios.

Table 6.5: The percentage of missing values and the count of non-missing values in the new dataset (combination of Original Dataset and Additional Dataset)

Feature	Percentage of missing values (%)	Number of non-missing values	Used as a Most Important Predictor
BCR-ABL ratio at 60 months after start of IM	46	118	No
BCR-ABL ratio at 24 months after start of IM	27	160	No
BCR-ABL ratio at 6 months after start of IM	25	165	No[1]
BCR-ABL ratio at 18 months after start of IM	23	168	No
BCR-ABL ratio at 3 months after start of IM	15	186	No[1]
BCR-ABL ratio at 12 months after start of IM	11	196	No[1]
BASOPHIL LEVEL AT DIAGNOSIS	5	209	No
PLATELET AT DIAGNOSIS	5	209	No
WHITE BLOOD CELL COUNT AT DIAGNOSIS	4	211	Yes
EOSINOPHIL COUNT AT DIAGNOSIS	4	210	No
HEMOGLOBIN AT DIAGNOSIS	4	210	No

[1] We utilize *aggregated features* rather than the raw values recorded for BCR::ABL1/ABL1% ratios.

Additionally, we attempted to optimize the decision-making process regarding TFR by forecasting the probability of attaining a sustained DMR, specifically achieving DMR at least 24 months post-treatment initiation and maintaining this status through the 60-month mark. This led to another model that predicted the long-term DMR using information up to 12 months after the start of IM (Model_long_12). Figure 6·1 provides the number of patients in each model, and the number/percentage of patients who achieve DMR (responders) and those who do not achieve DMR (non-responders) at the time of the model's outcome. Tables 6.5 and 6.6 represent each model's feature statistics including p-value associated with the null hypothesis of each variable having the same distribution in the two cohorts (responders and non-responders).

Table 6.6: Cohort statistics of the new dataset (combination of Original Dataset and Additional Dataset).

Variable	**Model_3_18**			**Model_6_18**			**Model_6_12**			**Model_long_12**		
	yes	no	p-value	yes	no	p-value	yes	no	p-value	yes	no	p-value
Age at diagnosis	50.1	45.0	0.09	50.0	45.0	0.15	51.8	45.1	0.10	51	45.7	0.05
Sex (male)	47%	63%	0.40	47%	61%	0.44	48%	60%	0.58	44%	61%	0.30
Hemoglobin count at diagnosis	11.3	11.4	0.07	11.4	11.5	0.06	12.1	11.2	<0.001	11.4	11.5	<0.001
Platelet count at diagnosis	534.5	418.6	0.02	527.8	415.0	0.01	547.5	425.7	0.12	552.1	422.7	0.04
Basophil level at diagnosis	3.7	4.0	0.96	3.5	3.9	1.00	4.0	3.9	0.96	4.3	3.5	0.36
Eosinophil count at diagnosis	2.7	3.3	0.38	2.6	3.1	0.19	2.8	2.9	0.88	4.3	2.3	0.83
WBC count at diagnosis	91580.4	133164.6	0.05	92386.7	127800.6	0.11	78037.3	139953.8	0.02	90788.0	132085.0	0.11
MR_0_3[1]	14.9%[2]	3.8%[3]	0.17	13.7%	3.4%	0.16	2.3%	3.3%	<0.001	17.1%	2.5%	0.01
MR_3_6[4]	-	-	-	51.0%	19.3%	<0.001	60.0%	22.8%	<0.001	46.3	24.5	0.08
Median of Bcr-Abl ratios	2.1	16.6	<0.001	1.3	11.0	<0.001	1.1	9.6	<0.001	0.3	4.4	<0.001
Min of Bcr-Abl ratios	0.7	6.0	<0.001	0.6	1.3	0.03	0.0	1.1	<0.001	0.0	1.4	<0.001
Max of Bcr-Abl ratios	2.1	16.6	<0.001	1.9	16.0	<0.001	1.9	13.5	<0.001	2.6	12.5	<0.001
Mean of Bcr-Abl ratios	2.1	16.6	<0.001	1.3	11.0	<0.001	1.1	9.6	<0.001	0.9	6.1	<0.001

[1] MR_0_3 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the first three-month interval after start of IM.

[2] This means that among responders in **Model_3_18**, 14.9% of them achieved MR_0_3.

[3] This means that among non-responders in **Model_3_18**, 3.8% of them achieved MR_0_3.

[4] MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from 3 to 6 months after the start of IM.

We explored supervised classification methods, LR, SVM, MLP, RF, and GBM , explained in more detail in Section 2.1.1. We considered both an L1-norm (L1LR, L1SVM) and an L2-norm regularizer (L2LR, L2SVM) (Lee et al., 2006) to address overfitting. Explained in more detail in Section 2.1.5.1.

We randomly split the dataset into three equal parts, where two parts were used as the training set, and the third part as the test set. The training set is used to tune the model hyperparameters via 3-fold cross-validation. We evaluated all performance metrics on the test set. We repeated training and testing five times, each time with a different random split between the training and test sets. The mean and standard deviation of all metrics on the test sets over the five repetitions are reported. Explained in more detail in Section 2.1.3. Since all molecular data are internationally uniformed using the International Scale12, an external validation cohort is less critical for our study.

To assess model performance, we use AUC-ROC and weighted F1-score. Explained in more detail in Section 2.1.4.

We also used a recursive feature elimination approach with L1-penalized logistic regression (L1-regularized RFE) to extract the most informative features and develop parsimonious models. Explained in more detail in Section 2.1.5.3. We will be referring to the resulting model as the parsimonious model. Figure 6·3 presents an overview of the steps taken in model development.

6.2.2.1 Feature statistics in Detail

Statistics of the new dataset (combination of Original Dataset and Additional Dataset) of patients who achieved DMR and those who did not are represented in Table 6.6 for each model. The columns “yes” and “no” show the mean of each variable among patients who did and did not achieve DMR, respectively. The statistics of the Original Dataset of patients who achieved DMR and those who did not are represented in Tables 6.7 and 6.8 for each model.

Table 6.7: Cohort statistics of Original Dataset. Part 1.

Variable	**Model_3_18**			**Model_6_18**			**Model_6_12**			**Model_long_12**		
	yes	no	p-value	yes	no	p-value	yes	no	p-value	yes	no	p-value
Sex (male)	48%	68%	0.41	45%	66%	0.40	38%	63%	0.40	38%	62%	0.46
Age at diagnosis	46.0	44.6	0.81	45.7	44.5	0.77	52.0	45.1	0.41	48.1	45.2	0.21
CML phase at diagnosis	0%	3%	0.87	0%	3%	0.87	0%	3%	0.93	0%	2%	0.96
Spleen size at diagnosis (CM)	1.0	4.2	0.08	1.1	4.0	0.07	0.5	1.0	0.14	1.4	3.2	0.40
WBC count at diagnosis	86850	131104	0.03	85938	128118	0.02	59113	134597	0.01	63507	124123	0.02
Hemoglobin count at diagnosis	11.6	11.5	0.70	11.6	11.5	0.52	12.3	11.4	0.27	12.0	11.6	0.40
Platelet count at diagnosis	509.0	433.1	0.13	519.3	436.1	0.13	439.2	440.0	0.93	473.8	447.8	0.33
Basophil level at diagnosis	3.2	4.4	0.19	3.3	4.4	0.28	3.4	4.3	0.88	4.2	4.0	1.00
Eosinophil count at diagnosis	2.1	3.4	0.38	2.0	3.3	0.27	1.9	3.1	0.72	4.4	2.5	0.89
Blast count at diagnosis	1.5	1.4	0.98	1.4	1.4	0.90	1.0	1.5	0.98	0.8	1.7	0.57
Sokal index	0.9	0.9	0.51	0.8	0.9	0.61	0.9	0.8	0.23	0.8	0.9	0.33
Sokal classification	2.5	2.4	0.91	2.6	2.4	0.91	2.5	2.4	1.00	2.8	2.3	0.51
Transcripts at diagnosis	1.8	1.8	1.00	1.8	1.8	1.00	1.9	1.8	0.83	1.5	1.8	0.78
WBC count at start of IM	12691	13997	0.92	12352	13589	0.80	11035	13176	0.78	8011	11896	0.25
Platelet count at start of IM	426.1	343.0	0.46	412.0	346.8	0.72	322.0	370.7	0.32	281.7	382.9	0.73
Hemoglobin count at start of IM	11.5	11.8	0.88	11.5	11.9	0.87	12.1	11.6	0.93	11.5	11.8	0.92
Blast count at start of IM	0.1	0.3	1.00	0.1	0.3	1.00	0.1	0.3	1.00	0.2	0.2	1.00
Eosinophil count at start of IM	6.5	5.4	0.54	6.2	5.0	0.65	9.2	5.7	0.59	9.7	4.7	0.49
Basophil count at start of IM	15.4	12.9	0.63	14.7	12.4	0.40	12.2	13.6	0.41	13.2	12.4	0.65

Table 6.8: Cohort statistics of Original Dataset. Part 2.

Variable	Model_3_18			Model_6_18			Model_6_12			Model_long_12		
	yes	no	p-value	yes	no	p-value	yes	no	p-value	yes	no	p-value
IM 0_3[1]	400.0	416.7	1.00	400.0	415.4	1.00	400.0	415.4	1.00	400.0	412.2	1.00
HR 0_3[2]	100%	92%	0.60	100%	93%	0.82	100%	98%	0.82	100%	98%	0.96
MR_0_3[3]	14%	3%	0.36	14%	3%	0.34	23%	2%	0.01	8%	4%	0.93
IM 3_6[4]	400.0	418.4	1.00	400.0	416.48	1.00	400.0	414.6	1.00	400.0	414.6	1.00
HR 3_6[5]	100%	92%	0.60	100%	93%	0.82	100%	98%	0.82	100%	98%	0.96
MR_3_6[6]	-	-	-	59%	25%	0.03	54%	29%	0.34	38%	32%	0.97
Median of Bcr-Abl ratios	2.9	19.2	<0.001	1.8	13.0	<0.001	2.6	11.4	<0.001	0.2	5.2	<0.001
Min of Bcr-Abl ratios	2.9	19.2	<0.001	0.8	7.3	<0.001	0.8	7.0	<0.001	0.0	1.7	<0.001
Max of Bcr-Abl ratios	2.9	19.2	<0.001	2.8	18.6	<0.001	4.5	15.9	0.02	4.8	15.9	0.02
Mean of Bcr-Abl ratios	2.9	19.2	<0.001	1.8	13.0	<0.001	2.6	11.4	<0.001	1.5	7.1	0.00
Std of Bcr-Abl ratios	2.9	19.2	<0.001	1.8	13.0	<0.001	2.6	11.4	<0.001	1.5	7.1	0.00
Median of CyR	92.4	93.4	1.00	83.6	90.4	0.79	92.3	89.4	1.00	74.4	77.0	0.88
Min of CyR	79.3	78.1	1.00	53.5	52.8	1.00	90.0	58.5	0.36	12.8	31.2	0.36
Max of CyR	100.0	98.3	1.00	100.0	98.5	1.00	100.0	98.9	1.00	100.0	97.6	1.00
Mean of CyR	90.6	89.9	1.00	71.4	80.7	0.04	76.2	79.3	0.83	63.2	68.7	0.48

[1] IM_0_3 shows the imatinib dosage a patient receives in the first three-month interval after start of IM.

[2] HR_0_3 indicates achieving hematologic response in the first three-month interval after start of IM.

[3] MR_0_3 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the first three-month interval after start of IM.

[4] IM_3_6 shows the imatinib dosage a patient receives in the 3-month interval from 3 to 6 months after the start of IM.

[5] HR_3_6 indicates achieving hematologic response in the 3-month interval from 3 to 6 months after the start of IM.

[6] MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from 3 to 6 months after the start of IM.

6.3 Results

The results associated with parsimonious versions of Model_3_18 and Model_long_12 are reported in Figures 6·4, 6·5, and 6·6. Figure 6·4 shows the feature importance based on Logistic Regression coefficients in the L2LR model. The median of the patient's recorded BCR::ABL1/ABL1% ratios, white blood cell count at diagnosis, and the achievement of MMR are among the most important predictive variables. Figure 6·5 shows ROC curves associated with the L2LR parsimonious models. Figure 6·6 presents the mean and std of performance metrics on the test set. The best AUCs achieved over all models are between 82% and 86%, while we use only two variables. Detailed results including results associated with Model_6_12 and Model_6_18 are given in Tables 6.9 and 6.10, and Figures 6·7 and 6·8. Figure 6·9 displays the feature importance in parsimonious models based on Random Forest feature importance. The results suggest moderate to strong predictive power for DMR shorter term (Model_3_18, Model_6_18, and Model_6_12) but also longer-term (Model_long_12), informing physicians on how to optimize treatment but also recommending discontinuing treatment upon achieving DMR.

6.4 Discussion

The median of BCR::ABL1/ABL1% ratio measurements is consistently included among the most important predictive variables in all our models. Moreover, other important predictive variables in our models are the white blood cell count at the time of diagnosis and the achievement of MMR at 3 and 6 months marks (Table 6.10), in agreement with clinical observations described in literature (Hochhaus et al., 2020) (Wang et al., 2019) (Hehlmann et al., 2014) (Bonifacio et al., 2019) (Hasford et al., 2011).

Table 6.9: The most important predictors selected by feature selection (SFS and RFE) in parsimonious format of all models. We list the LR coefficients of each variable (**Coef**), the correlation of the variable with the outcome (**Y_corr**), the mean of the variable (**Y1_mean**) in the patients achieving DMR at the time corresponding to each model, and the mean of the variable (**Y0_mean**) in the remaining (non-DMR) patients.

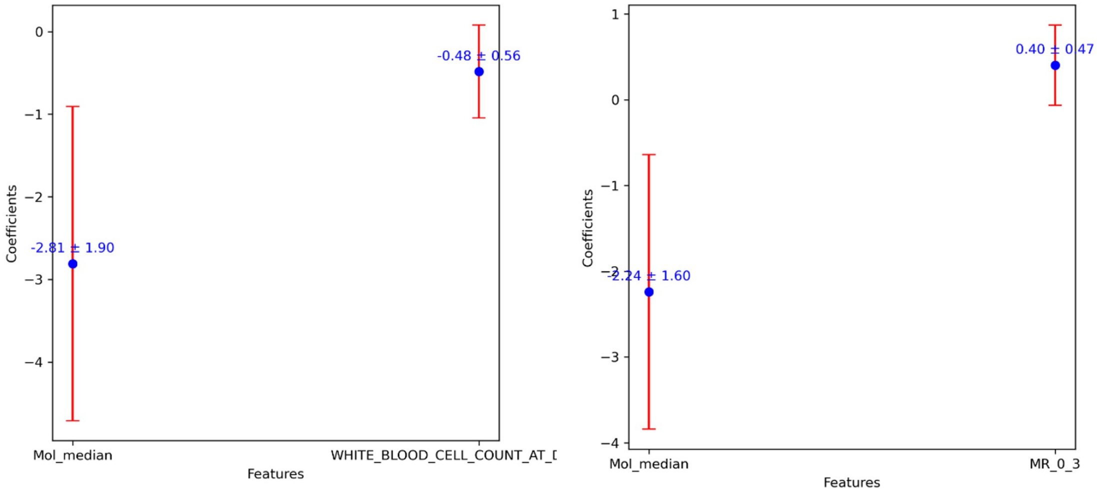
Coef	Variable	All_mean	All_std	Y1_mean	Y1_std	Y0_mean	Y0_std	p-value	Y_corr
Model_3_18									
-2.81	Mol_median	11.24	19.28	2.07	4.03	16.62	22.45	1.94E-11	-0.37
-0.48	WBC count at diagnosis	117775	101378	91580	77334	133164	110732	4.73E-02	-0.20
Model_6_18									
-3.59	Mol_median	7.47	14.01	1.3	2.8	11.04	16.48	5.97E-13	-0.34
0.65	MR_3_6[1]	0.31	0.46	0.51	0.5	0.19	0.4	1.69E-03	0.33
Model_6_12									
-1.42	Mol_median	7.54	14.14	1.1	3.99	9.64	15.57	6.33E-15	-0.26
0.70	MR_3_6[1]	0.32	0.47	0.6	0.5	0.23	0.42	2.41E-04	0.34
Model_long_12									
-2.24	Mol_median	3.37	8.95	0.33	0.90	4.42	10.17	1.19E-08	-0.20
0.40	MR_0_3[2]	0.06	0.24	0.17	0.38	0.03	0.16	1.23E-02	0.26

[1] MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from time 3 to 6 months after the start of IM.

[2] MR_0_3 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from time 0 to 3 months after the start of IM.

Table 6.10: All models' evaluation metrics on the test set.

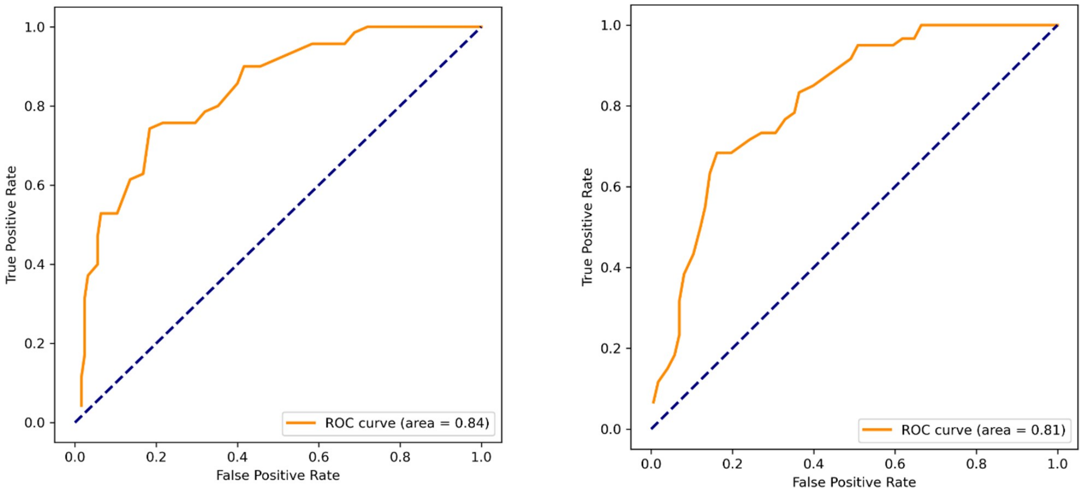
	AUC		Weighted F1		Weighted AUPRC		Weighted Precision		Weighted Recall	
	Mean (%)	Std (%)	Mean (%)	Std (%)	Mean (%)	Std (%)	Mean (%)	Std (%)	Mean (%)	Std (%)
Model_3_18										
L2LR	83.7	3.0	70.8	4.2	74.6	7.2	73.3	2.4	72.3	3.8
L1LR	83.6	3.1	70.8	4.2	74.5	7.4	73.3	2.4	72.3	3.8
L1SVM	83.1	3.2	72.0	7.0	73.3	7.3	74.2	4.8	73.8	4.9
L2SVM	83.5	3.1	72.8	5.9	74.5	7.2	74.8	4.5	74.4	4.4
RF	80.8	5.2	75.3	6.0	73.8	6.9	77.0	6.1	76.4	4.9
NN	78.5	6.4	67.9	7.8	64.8	9.2	70.8	7.2	70.8	5.3
LGB	79.9	7.6	77.3	6.2	73.4	8.2	78.4	7.2	77.9	5.6
Model_6_18										
L2LR	82.5	3.5	74.8	3.4	72.5	4.4	75.3	3.3	74.8	3.6
L1LR	83.0	3.1	76.1	1.7	73.1	4.7	76.4	2.1	76.2	1.7
L1SVM	82.7	2.8	75.7	2.6	72.8	4.4	76.0	2.8	75.7	2.6
L2SVM	82.2	3.2	74.8	3.1	72.2	4.1	75.3	3.3	74.8	3.2
RF	83.4	4.9	76.9	5.3	71.1	9.2	78.0	4.7	76.7	5.7
NN	81.4	2.8	71.9	1.5	70.5	4.6	74.6	2.6	72.4	4.3
LGB	84.0	4.8	78.8	4.3	71.2	6.7	79.1	4.2	79.0	4.3
Model_6_12										
L2LR	86.0	2.2	82.3	3.3	68.2	3.7	83.2	1.9	83.7	2.0
L1LR	86.6	2.1	82.3	3.3	68.8	3.3	83.2	1.9	83.7	2.0
L1SVM	84.4	1.7	82.3	3.3	66.7	3.6	83.2	1.9	83.7	2.0
L2SVM	84.8	1.1	82.3	3.3	67.1	3.0	83.2	1.9	83.7	2.0
RF	85.9	5.9	86.2	1.6	64.5	5.8	86.4	2.0	86.4	1.2
NN	81.5	4.3	76.4	8.1	57.6	15.4	78.8	6.0	76.9	10.1
LGB	84.0	8.1	83.1	1.6	65.4	6.9	83.4	1.5	83.7	2.0
Model_long_12										
L2LR	82.1	5.1	77.6	7.4	61.0	13.1	80.4	4.9	76.7	8.4
L1LR	80.3	6.1	77.1	7.8	59.9	13.3	79.7	5.7	76.3	8.8
L1SVM	80.2	4.8	77.0	7.0	59.5	10.9	79.5	4.7	76.3	7.9
L2SVM	80.2	4.8	77.0	7.0	59.5	10.9	79.5	4.7	76.3	7.9
RF	79.4	4.7	77.0	7.2	64.0	8.3	77.5	7.6	77.5	7.4
NN	82.1	5.1	77.6	7.4	61.0	13.1	80.4	4.9	76.7	8.4
LGB	81.2	5.1	81.7	5.5	66.3	8.5	82.7	4.7	81.7	6.3



Model_3_18

Model_long_12

Figure 6·4: The feature importance in L2LR parsimonious models based on Logistic Regression coefficients +- 95% CI. The median of the patient's recorded BCR-ABL1 ratios, white blood cell count at diagnosis, and the achievement of Major Molecular Response (MMR, defined as ≥MR3.0) are among the most important predictive variables. MR_0_3 indicates achieving MMR in the 3-month interval from time 0 to 3 months after the start of IM.



Model_3_18

Model_long_12

Figure 6·5: ROC curves associated with the L2LR parsimonious version of each model. The Area Under the ROC Curve (AUC), is used to evaluate prediction performance. A perfect predictor has an AUC of 1 and a predictor which makes random guesses has an AUC of 0.5.

	Model_3_18		**Model_long_12**	
	AUC	**Weighted F1 score**	**AUC**	**Weighted F1 score**
	mean% (std%)			
L2LR	83.7 (3.0)	70.8 (4.2)	82.1 (5.1)	77.6 (7.4)
L1LR	83.6 (3.1)	70.8 (4.2)	80.3 (6.1)	77.1 (7.8)
L1SVM	83.1 (3.2)	72.0 (7.0)	80.2 (4.8)	77.0 (7.0)
L2SVM	83.5 (3.1)	72.8 (5.9)	80.2 (4.8)	77.0 (7.0)
RF	80.8 (5.2)	75.3 (6.0)	79.4 (4.7)	77.0 (7.2)
NN	78.5 (6.4)	67.9 (7.8)	82.1 (5.1)	77.6 (7.4)
LGB	79.9 (7.6)	77.3 (6.2)	81.2 (5.1)	81.7 (5.5)

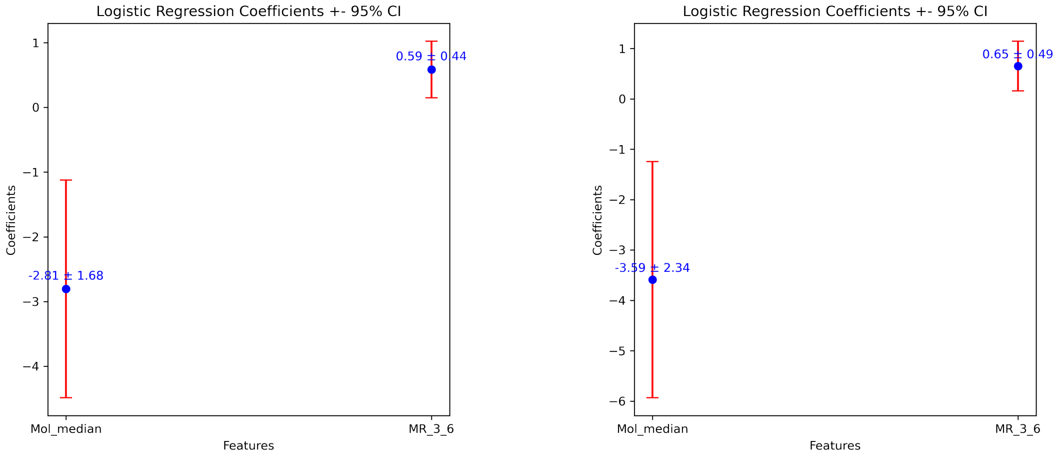
Figure 6·6: Test set performance of the parsimonious models using LR, SVM, RF, MLP, and LightGBM.

In summary, our results are in agreement with recent data which indicates that patients with an early molecular response have a higher probability of achieving and maintaining DMR (Shanmuganathan et al., 2021), since those patients will have smaller BCR::ABL1/ABL1%IS median values and will achieve MMR before 6 months of treatment.

Table 6.11: Probabilities calculated by our models alongside different categories of patients from Original Dataset.

	Model Output Probability (Median)			
	Model_3_18	**Model_6_18**	**Model_6_12**	**Model_long_12**
High risk (Sokal Score)	0.24	0.13	0.20	0.23
Intermediate risk (Sokal Score)	0.39	0.38	0.41	0.44
Low risk (Sokal Score)	0.42	0.41	0.40	0.49
3-months BCR::ABL1 > 10%	0.01	0.03	0.07	0.19
3-months BCR::ABL1 <= 10%	0.59	0.54	0.48	0.52
12-months BCR::ABL1 > 0.01%	0.35	0.36	0.38	0.43
12-months BCR::ABL1 <= 0.01%	0.77	0.87	0.78	0.56

Stopping TKI treatment can be considered a safe option that especially benefits patients with comorbidities and the young, providing a higher quality of life and reducing costs. Results shown here reveal that the probability of reaching DMR can be predicted with high accuracy. To our knowledge, this is the first study to model DMR using real world data from a cohort of CML patients in Brazil. Our models can help hematologists to inform decisions regarding TKI discontinuity to patients.



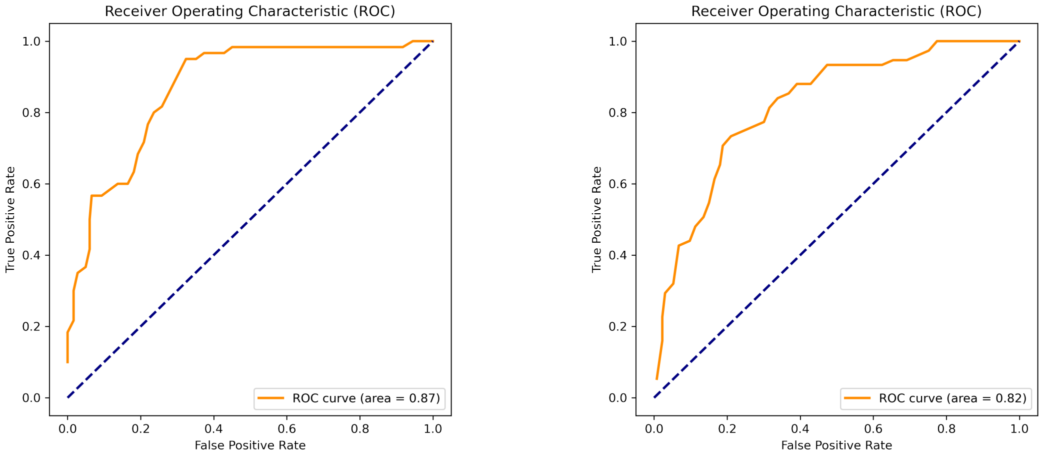
(a) Model_6_12

(b) Model_6_18

Figure 6·7: The feature importance in L2LR parsimonious models based on Logistic Regression coefficients +- 95% CI. MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from time 3 to 6 months after the start of IM.

We currently implemented our models as on-line modules publicly available to the medical and scientific community at github repository[1](Zad et al., 2024a).

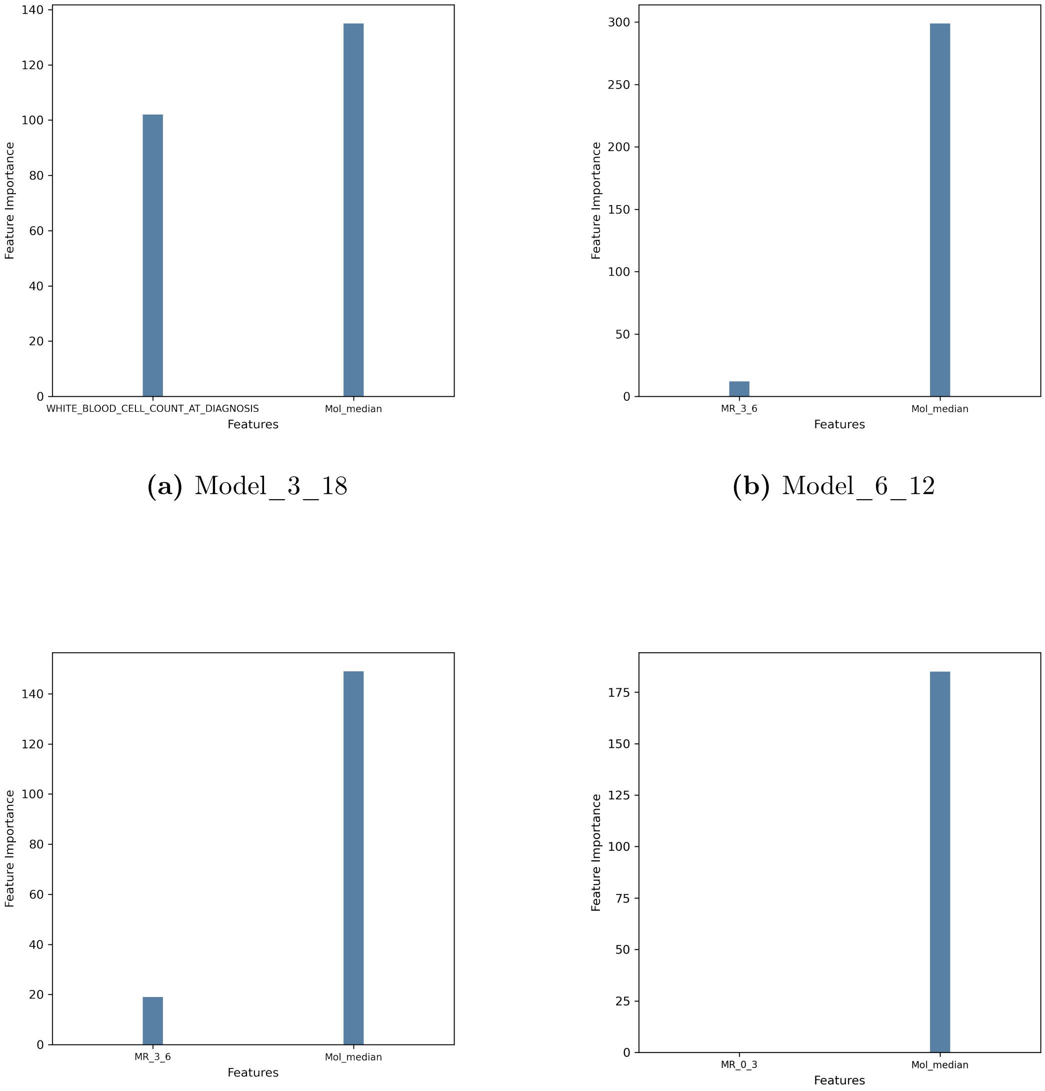
[1] https://github.com/noc-lab/Predictive-models-of-DMR-in-CML-patients



(a) Model_6_12

(b) Model_6_18

Figure 6·8: ROC curves associated with the parsimonious version of each model. The Area Under the ROC Curve (AUC) is used to evaluate prediction performance. A perfect predictor has an AUC of 1 and a predictor which makes random guesses has an AUC of 0.



(a) Model_3_18

(b) Model_6_12

(c) Model_6_18

(d) Model_long_12

Figure 6·9: The feature importance in parsimonious models based on Random Forest feature importance. MR_0_3 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from time 0 to 3 months after the start of IM. MR_3_6 indicates achieving Major Molecular Response (MMR, defined as ≥MR3.0) in the 3-month interval from time 3 to 6 months after the start of IM.

Chapter 7

Conclusions

7.1 Summary

This dissertation has explored the development and application of explainable and sparse predictive models in two critical areas of healthcare: reproductive health and oncology. Through the use of advanced machine learning techniques and survival analysis, we aimed to enhance predictive accuracy and provide actionable insights to improve patient outcomes.

In reproductive health, we developed machine learning models to predict the probability of conception using self-reported health data from a North American preconception cohort study. Factors such as sociodemographics, lifestyle, medical history, diet quality, and specific male partner characteristics were analyzed, resulting in models that demonstrated improved discrimination and potential clinical utility.

We also applied machine learning algorithms to electronic health record data to identify predictor variables associated with polycystic ovarian syndrome (PCOS) diagnosis. Gradient boosted trees and feed-forward multilayer perceptron classifiers were employed to create a scoring system, enhancing model performance and providing a valuable tool for early detection and intervention.

For miscarriage risk prediction, we utilized static and survival analysis, including Cox proportional hazard models, to develop predictive models assessing miscarriage risk among female participants who conceived during the study period. Our study revealed that most miscarriages were due to random genetic errors during early pregnancy, indicating limited predictability based on preconception sociodemographic and lifestyle characteristics.

In managing Chronic Myeloid Leukemia (CML), we developed predictive models to forecast whether patients would achieve deep molecular response (DMR) at later treatment stages and maintain this status up to 60 months post-treatment initiation. These models offer insights into treatment effectiveness and patient management, supporting clinical decision-making and improving long-term patient outcomes.

The models developed in this dissertation emphasize explainability, ensuring that the predictions are interpretable and actionable for healthcare professionals. The findings underscore the potential of predictive modeling to improve outcomes in reproductive health and oncology, demonstrating the value of machine learning algorithms in healthcare research and the prediction of critical health events.

7.2 Future work

Future research should aim to integrate diverse data sources, including genetic, environmental, and real-time patient monitoring data, to enhance the robustness and accuracy of predictive models. This integration will provide a more comprehensive understanding of the factors influencing health outcomes.

External validation is essential to ensure the generalizability and reliability of predictive models across different populations and settings. Confirming that models perform well not only in the original study cohort but also in broader, more diverse populations is crucial for increasing their applicability and trustworthiness.

While this dissertation emphasizes model explainability, there is always room for improvement. Future work should explore advanced techniques for enhancing the transparency and interpretability of machine learning models. This will ensure that healthcare professionals can fully understand and trust the predictions.

Leveraging deep learning techniques can improve the integration and analysis of complex, high-dimensional data, enhancing the accuracy, robustness, and scalability of predictive models in healthcare. Deep learning can handle vast and intricate datasets, uncovering subtle patterns and interactions that traditional methods might miss, leading to more precise and reliable predictions.

The methodologies developed in this dissertation can be extended to other health conditions beyond reproductive health and oncology. Future research could apply these techniques to areas such as cardiovascular diseases, mental health, and infectious diseases, broadening the impact of predictive modeling in healthcare.

Implementing BERT models to handle and analyze large volumes of textual data in electronic health records, clinical notes, and medical literature can improve feature extraction and contextual understanding for better predictive performance. BERT's advanced natural language processing capabilities can extract nuanced insights from unstructured text, contributing to more informed and accurate predictive models.

To transition from research to clinical practice, future work should focus on the implementation and validation of predictive models in real-world clinical settings. Collaborating with healthcare providers to test and refine these models can ensure their practical utility and effectiveness in improving patient care. Longitudinal studies are crucial for understanding the long-term impact of predictive models on patient outcomes. Future research should also explore adaptive models that can update and improve over time as new data becomes available, ensuring that predictions remain accurate and relevant.

By addressing these areas, future work can continue to advance the field of predictive modeling in healthcare, ultimately contributing to better patient outcomes and more informed clinical decision-making.

References

Akhter, S., Marcus, M., Kerber, R. A., Kong, M., and Taylor, K. C. (2016). The impact of periconceptional maternal stress on fecundability. *Annals of epidemiology*, 26(10):710–716.

American Cancer Society (2022). Key statistics for chronic myeloid leukemia. Website. Accessed: 2022-03-07, https://www.cancer.org/cancer/chronic-myeloid-leukemia/about/statistics.html.

Amini, S., Hao, B., Yang, J., Karjadi, C., Kolachalama, V. B., Au, R., and Paschalidis, I. C. (2024). Prediction of alzheimer's disease progression within 6 years using speech: A novel approach leveraging language models. *Alzheimer's & Dementia*.

Amini, S., Hao, B., Zhang, L., Song, M., Gupta, A., Karjadi, C., Kolachalama, V. B., Au, R., and Paschalidis, I. C. (2023). Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimer's & Dementia*, 19(3):946–955.

Amini, S., Zhang, L., Hao, B., Gupta, A., Song, M., Karjadi, C., Lin, H., Kolachalama, V. B., Au, R., and Paschalidis, I. C. (2021). An artificial intelligence-assisted method for dementia detection using images from the clock drawing test. *Journal of Alzheimer's Disease*, 83(2):581–589.

Anagnostis, P., Tarlatzis, B. C., and Kauffman, R. P. (2018). Polycystic ovarian syndrome (pcos): Long-term metabolic consequences. *Metabolism*, 86:33–43.

Andersen, A.-M. N., Andersen, P. K., Olsen, J., Grønbæk, M., and Strandberg-Larsen, K. (2012). Moderate alcohol intake during pregnancy and risk of fetal death. *International journal of epidemiology*, 41(2):405–413.

Arck, P. C., Rücke, M., Rose, M., Szekeres-Bartho, J., Douglas, A. J., Pritsch, M., Blois, S. M., Pincus, M. K., Bärenstrauch, N., Dudenhausen, J. W., et al. (2008). Early risk factors for miscarriage: a prospective cohort study in pregnant women. *Reproductive biomedicine online*, 17(1):101–113.

Azziz, R., Carmina, E., Dewailly, D., Diamanti-Kandarakis, E., Escobar-Morreale, H. F., Futterweit, W., Janssen, O. E., Legro, R. S., Norman, R. J., Taylor, A. E., and Witchel, S. F. (2009). The androgen excess and pcos society criteria for the polycystic ovary syndrome: the complete task force report. *Fertility and sterility*, 91(2):456–488.

Azziz, R., Woods, K. S., Reyna, R., Key, T. J., Knochenhauer, E. S., and Yildiz, B. O. (2004). The prevalence and features of the polycystic ovary syndrome in an unselected population. *The Journal of Clinical Endocrinology & Metabolism*, 89(6):2745–2749.

Barrera, F. J., Brown, E. D., Rojo, A., Obeso, J., Plata, H., Lincango, E. P., Terry, N., Rodríguez-Gutiérrez, R., Hall, J. E., and Shekhar, S. (2023). Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review. *Frontiers in Endocrinology*, 14:1106625.

Barry, J. A., Azizia, M. M., and Hardiman, P. J. (2014). Risk of endometrial, ovarian and breast cancer in women with polycystic ovary syndrome: a systematic review and meta-analysis. *Human reproduction update*, 20(5):748–758.

Barzilai-Pesach, V., Sheiner, E. K., Sheiner, E., Potashnik, G., and Shoham-Vardi, I. (2006). The effect of women's occupational psychologic stress on outcome of fertility treatments. *Journal of occupational and environmental medicine*, pages 56–62.

Bertsimas, D. and Dunn, J. (2019). *Machine learning under a modern optimization lens*. Dynamic Ideas LLC Charlestown, MA.

Bertsimas, D., Pauphilet, J., and Van Parys, B. (2021). Sparse classification: a scalable discrete optimization perspective. *Machine Learning*, 110:3177–3209.

Best, D. and Bhattacharya, S. (2015). Obesity and fertility. *Hormone molecular biology and clinical investigation*, 24(1):5–10.

Bonifacio, M., Tiribelli, M., Binotto, G., Miggiano, M. C., Basso, M., Calistri, E., Scaffidi, L., Stella, R., Frison, L., Sartori, R., et al. (2019). Generic versus branded imatinib as frontline therapy in chronic-phase chronic myeloid leukemia patients in italy: A case-control study. *Blood*, 134:5909.

Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003). Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436.

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

Brisimi, T. S., Xu, T., Wang, T., Dai, W., Adams, W. G., and Paschalidis, I. C. (2018). Predicting chronic disease hospitalizations from electronic health records: an interpretable classification approach. *Proceedings of the IEEE*, 106(4):690–707.

Brisimi, T. S., Xu, T., Wang, T., Dai, W., and Paschalidis, I. C. (2019). Predicting diabetes-related hospitalizations based on electronic health records. *Statistical methods in medical research*, 28(12):3667–3682.

Buck Louis, G., Barr, D. B., Kannan, K., Chen, Z., Kim, S., and Sundaram, R. (2016). Paternal exposures to environmental chemicals and time-to-pregnancy: overview of results from the life study. *Andrology*, 4(4):639–647.

Caetano, M. R., Couto, E., Passini Junior, R., Simoni, R. Z., and Barini, R. (2006). Gestational prognostic factors in women with recurrent spontaneous abortion. *Sao Paulo Medical Journal*, 124:181–185.

Carreau, A.-M., Pyle, L., Garcia-Reyes, Y., Rahat, H., Vigers, T., Jensen, T., Scherzinger, A., Nadeau, K. J., and Cree-Green, M. (2019). Clinical prediction score of nonalcoholic fatty liver disease in adolescent girls with polycystic ovary syndrome (pcos-hs index). *Clinical endocrinology*, 91(4):544–552.

Castro, V., Shen, Y., Yu, S., Finan, S., Pau, C. T., Gainer, V., Keefe, C. C., Savova, G., Murphy, S. N., Cai, T., et al. (2015). Identification of subjects with polycystic ovary syndrome using electronic health records. *Reproductive Biology and Endocrinology*, 13:1–8.

Chandra, A., Copen, C. E., and Stephen, E. H. (2013). *Infertility and impaired fecundity in the United States, 1982-2010: data from the National Survey of Family Growth.* US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250.

Chen, R. and Paschalidis, I. C. (2018). A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13).

Chen, R. and Paschalidis, I. C. (2022). Robust grouped variable selection using distributionally robust optimization. *Journal of Optimization Theory and Applications*, 194(3):1042–1071.

Chen, R., Paschalidis, I. C., et al. (2020). Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243.

Chen, R., Paschalidis, I. C., Hatabu, H., Valtchinov, V. I., and Siegelman, J. (2019). Detection of unwarranted ct radiation exposure from patient and imaging protocol meta-data using regularized regression. *European Journal of Radiology Open*, 6:206–211.

Cheng, J. J. and Mahalingaiah, S. (2019). Data mining polycystic ovary morphology in electronic medical record ultrasound reports. *Fertility Research and Practice*, 5(1):1–7.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR.

Choong, S., Rombauts, L., Ugoni, A., and Meagher, S. (2003). Ultrasound prediction of risk of spontaneous miscarriage in live embryos from assisted conceptions. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 22(6):571–577.

Cochran, W. G. (1952). The $\chi 2$ test of goodness of fit. *The Annals of mathematical statistics*, pages 315–345.

Collett, D. (2023). *Modelling survival data in medical research.* Chapman and Hall/CRC.

Collins, J. A., Burrows, E. A., and Willan, A. R. (1995). The prognosis for live birth among untreated infertile couples. *Fertility and sterility*, 64(1):22–28.

Conforti, A., Mascia, M., Cioffi, G., De Angelis, C., Coppola, G., De Rosa, P., Pivonello, R., Alviggi, C., and De Placido, G. (2018). Air pollution and female fertility: a systematic review of literature. *Reproductive Biology and Endocrinology*, 16(1):1–9.

Coppus, S., van der Veen, F., Opmeer, B., Mol, B., and Bossuyt, P. (2009). Evaluating prediction models in reproductive medicine. *Human reproduction*, 24(8):1774–1778.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cramer, D. W. and Wise, L. A. (2000). The epidemiology of recurrent pregnancy loss. In *Seminars in reproductive medicine*, volume 18, pages 331–340. Copyright© 2000 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel.:+ 1 (212) 584-4662.

Cueto, H., Riis, A., Hatch, E., Wise, L., Rothman, K., Sørensen, H., and Mikkelsen, E. (2016). Folic acid supplementation and fecundability: a danish prospective cohort study. *European journal of clinical nutrition*, 70(1):66–71.

Dai, Y.-F., Lin, L.-Z., Lin, N., He, D.-Q., Guo, D.-H., Xue, H.-L., Li, Y., Xie, X., Xu, L.-P., and He, S.-Q. (2022). Apa scoring system: a novel predictive model based on risk factors of pregnancy loss for recurrent spontaneous abortion patients. *Journal of Obstetrics and Gynaecology*, 42(6):2069–2074.

Deshmukh, H., Papageorgiou, M., Kilpatrick, E. S., Atkin, S. L., and Sathyapalan, T. (2019). Development of a novel risk prediction and risk stratification score for polycystic ovary syndrome. *Clinical Endocrinology*, 90(1):162–169.

DeVilbiss, E. A., Mumford, S. L., Sjaarda, L. A., Connell, M. T., Plowden, T. C., Andriessen, V. C., Perkins, N. J., Hill, M. J., Silver, R. M., and Schisterman, E. F. (2020). Prediction of pregnancy loss by early first trimester ultrasound characteristics. *American journal of obstetrics and gynecology*, 223(2):242–e1.

du Fossé, N. A., van der Hoorn, M.-L. P., de Koning, R., Mulders, A. G., van Lith, J. M., le Cessie, S., and Lashley, E. E. (2022). Toward more accurate prediction of future pregnancy outcome in couples with unexplained recurrent pregnancy loss: taking both partners into account. *Fertility and Sterility*, 117(1):144–152.

Eimers, J. M., te Velde, E. R., Gerritse, R., Vogelzang, E. T., Looman, C. W., and Habbema, J. D. F. (1994). The prediction of the chance to conceive in subfertile couples. *Fertility and sterility*, 61(1):44–52.

ESHRE, T. R., Group, A.-S. P. C. W., et al. (2004). Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertility and sterility*, 81(1):19–25.

Eskenazi, B., Ames, J., Rauch, S., Signorini, S., Brambilla, P., Mocarelli, P., Siracusa, C., Holland, N., and Warner, M. (2021). Dioxin exposure associated with fecundability and infertility in mothers and daughters of seveso, italy. *Human Reproduction*, 36(3):794–807.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

Evans-Hoeker, E. A., Eisenberg, E., Diamond, M. P., Legro, R. S., Alvero, R., Coutifaris, C., Casson, P. R., Christman, G. M., Hansen, K. R., Zhang, H., et al. (2018). Major depression, antidepressant use, and male and female fertility. *Fertility and sterility*, 109(5):879–887.

Evers, J. L. (2002). Female subfertility. *The lancet*, 360(9327):151–159.

Fan, D., Liu, L., Xia, Q., Wang, W., Wu, S., Tian, G., Liu, Y., Ni, J., Wu, S., Guo, X., et al. (2017). Female alcohol consumption and fecundability: a systematic review and dose-response meta-analysis. *Scientific reports*, 7(1):13815.

Feodor Nilsson, S., Andersen, P., Strandberg-Larsen, K., and Nybo Andersen, A.-M. (2014). Risk factors for miscarriage from a prevention perspective: a nationwide follow-up study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 121(11):1375–1385.

Gaskins, A. J. and Chavarro, J. E. (2018). Diet and fertility: a review. *American journal of obstetrics and gynecology*, 218(4):379–389.

Gaskins, A. J., Nassan, F. L., Chiu, Y.-H., Arvizu, M., Williams, P. L., Keller, M. G., Souter, I., Hauser, R., Chavarro, J. E., Team, E. S., et al. (2019). Dietary patterns and outcomes of assisted reproduction. *American journal of obstetrics and gynecology*, 220(6):567–e1.

Gaskins, A. J., Rich-Edwards, J. W., Hauser, R., Williams, P. L., Gillman, M. W., Penzias, A., Missmer, S. A., and Chavarro, J. E. (2014). Prepregnancy dietary patterns and risk of pregnancy loss. *The American journal of clinical nutrition*, 100(4):1166–1172.

Gennarelli, G., Holte, J., Berglund, L., Berne, C., Massobrio, M., and Lithell, H. (2000). Prediction models for insulin resistance in the polycystic ovary syndrome. *Human Reproduction*, 15(10):2098–2102.

George, L., Granath, F., Johansson, A. L., Annerén, G., and Cnattingius, S. (2006). Environmental tobacco smoke and risk of spontaneous abortion. *Epidemiology*, 17(5):500–505. DOI: https://doi.org/10.1097/01.ede.0000229984.53726.33.

Gibson-Helm, M., Teede, H., Dunaif, A., and Dokras, A. (2017). Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. *The Journal of Clinical Endocrinology & Metabolism*, 102(2):604–612.

Gnoth, C., Godehardt, E., Frank-Herrmann, P., Friol, K., Tigges, J., and Freundl, G. (2005). Definition and prevalence of subfertility and infertility. *Human reproduction*, 20(5):1144–1147.

Guenther, P. M., Casavale, K. O., Reedy, J., Kirkpatrick, S. I., Hiza, H. A., Kuczynski, K. J., Kahle, L. L., and Krebs-Smith, S. M. (2013). Update of the healthy eating index: Hei-2010. *Journal of the Academy of Nutrition and Dietetics*, 113(4):569–580.

Hahn, K., Wise, L., Rothman, K., Mikkelsen, E., Brogly, S., Sørensen, H., Riis, A., and Hatch, E. (2015). Caffeine and caffeinated beverage consumption and risk of spontaneous abortion. *Human reproduction*, 30(5):1246–1255.

Hahn, K. A., Hatch, E. E., Rothman, K. J., Mikkelsen, E. M., Brogly, S. B., Sørensen, H. T., Riis, A. H., and Wise, L. A. (2014). Body size and risk of spontaneous abortion among danish pregnancy planners. *Paediatric and perinatal epidemiology*, 28(5):412–423.

Hao, B., Sotudian, S., Wang, T., Xu, T., Hu, Y., Gaitanidis, A., Breen, K., Velmahos, G. C., and Paschalidis, I. C. (2020). Early prediction of level-of-care requirements in patients with covid-19. *Elife*, 9:e60519.

Harton, J., Mitra, N., and Hubbard, R. A. (2022). Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *Journal of the American Medical Informatics Association*, 29(7):1191–1199.

Harville, E. W. and Boynton-Jarrett, R. (2013). Childhood social hardships and fertility: a prospective cohort study. *Annals of epidemiology*, 23(12):784–790.

Hasford, J., Baccarani, M., Hoffmann, V., Guilhot, J., Saussele, S., Rosti, G., Guilhot, F., Porkka, K., Ossenkoppele, G., Lindoerfer, D., et al. (2011). Predicting complete cytogenetic response and subsequent progression-free survival in 2060 patients with cml on imatinib treatment: the eutos score. *Blood, The Journal of the American Society of Hematology*, 118(3):686–692.

Hashemi, N., Hao, B., Ignatov, M., Paschalidis, I. C., Vakili, P., Vajda, S., and Kozakov, D. (2023). Improved prediction of mhc-peptide binding using protein language models. *Frontiers in Bioinformatics*, 3.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hehlmann, R., Müller, M. C., Lauseker, M., Hanfstein, B., Fabarius, A., Schreiber, A., Proetel, U., Pletsch, N., Pfirrmann, M., Haferlach, C., et al. (2014). Deep molecular response is reached by the majority of patients treated with imatinib, predicts survival, and is achieved more quickly by optimized high-dose imatinib: results from the randomized cml-study iv. *Journal of clinical oncology*, 32(5):415–423.

Henriksen, T. B., Hjollund, N. H., Jensen, T. K., Bonde, J. P., Andersson, A.-M., Kolstad, H., Ernst, E., Giwercman, A., Skakkebæk, N. E., and Olsen, J. (2004). Alcohol consumption at the time of conception and spontaneous abortion. *American journal of epidemiology*, 160(7):661–667.

Hipwell, A. E., Kahn, L. G., Factor-Litvak, P., Porucznik, C. A., Siegel, E. L., Fichorova, R. N., Hamman, R. F., Klein-Fedyshin, M., Harley, K. G., and for Environmental Influences on Child Health Outcomes, P. C. (2019). Exposure to non-persistent chemicals in consumer products and fecundability: a systematic review. *Human Reproduction Update*, 25(1):51–71.

Hochhaus, A., Baccarani, M., Silver, R., et al. (2020). Recommandations européennes leukemianet 2020 pour le traitement de la leucémie myéloïde chronique. *Leukemia*, pages 966–984.

Homan, G., Davies, M., and Norman, R. (2007). The impact of lifestyle factors on reproductive performance in the general population and those undergoing infertility treatment: a review. *Human reproduction update*, 13(3):209–223.

Hsiao, P. Y., Fung, J. L., Mitchell, D. C., Hartman, T. J., and Goldman, M. B. (2019). Dietary quality, as measured by the alternative healthy eating index for pregnancy (ahei-p), in couples planning their first pregnancy. *Public health nutrition*, 22(18):3385–3394.

Hu, S. and Chen, G. H. (2022). Distributionally robust survival analysis: A novel fairness loss without demographics. In *Machine Learning for Health*, pages 62–87. PMLR.

Huang, J., Lv, P., Lian, Y., Zhang, M., Ge, X., Li, S., Pan, Y., Zhao, J., Xu, Y., Tang, H., et al. (2022). Construction of machine learning tools to predict threatened miscarriage in the first trimester based on aea, progesterone and β-hcg in china: a multicentre, observational, case-control study. *BMC Pregnancy and Childbirth*, 22(1):1–8.

Hunault, C., Habbema, J., Eijkemans, M., Collins, J., Evers, J., and Te Velde, E. (2004). Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models. *Human Reproduction*, 19(9):2019–2026.

Hunault, C. C., Laven, J. S., van Rooij, I. A., Eijkemans, M. J., te Velde, E. R., and Habbema, J. D. F. (2005). Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples. *Human Reproduction*, 20(6):1636–1641.

Irvine, D. S. (1998). Epidemiology and aetiology of male infertility. *Human reproduction*, 13(suppl_1):33–44.

Jacobs, M. B., Boynton-Jarrett, R. D., and Harville, E. W. (2015). Adverse childhood event experiences, fertility difficulties and menstrual cycle characteristics. *Journal of Psychosomatic Obstetrics & Gynecology*, 36(2):46–57.

Jellesen, R., Strandberg-Larsen, K., Jørgensen, T., Olsen, J., Thulstrup, A. M., and Andersen, A.-M. N. (2008). Maternal use of oral contraceptives and risk of fetal death. *Paediatric and perinatal epidemiology*, 22(4):334–340.

Jensen, T. K., Scheike, T., Keiding, N., Schaumburg, I., and Grandjean, P. (1999). Fecundability in relation to body mass and menstrual cycle patterns. *Epidemiology*, 10(4):422–428.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4).

Jiang, T., Gradus, J. L., Lash, T. L., and Fox, M. P. (2021). Addressing measurement error in random forests using quantitative bias analysis. *American Journal of Epidemiology*, 190(9):1830–1840.

Jiang, T., Gradus, J. L., and Rosellini, A. J. (2020). Supervised machine learning: a brief primer. *Behavior Therapy*, 51(5):675–687.

Jiang, X., Liu, R., Liao, T., He, Y., Li, C., Guo, P., Zhou, P., Cao, Y., and Wei, Z. (2022). A predictive model of live birth based on obesity and metabolic parameters in patients with pcos undergoing frozen-thawed embryo transfer. *Frontiers in Endocrinology*, 12:799871.

Jin, Y. and Paschalidis, I. C. (2024). A novel approach for distributionally robust learning in survival analysis. Working paper.

Joffe, M., Key, J., Best, N., Keiding, N., Scheike, T., and Jensen, T. K. (2005). Studying time to pregnancy by use of a retrospective design. *American journal of epidemiology*, 162(2):115–124.

Jones, J. R., Kogan, M. D., Singh, G. K., Dee, D. L., and Grummer-Strawn, L. M. (2011). Factors associated with exclusive breastfeeding in the united states. *Pediatrics*, 128(6):1117–1125.

Joo, Y. Y., Actkins, K., Pacheco, J. A., Basile, A. O., Carroll, R., Crosslin, D. R., Day, F., Denny, J. C., Velez Edwards, D. R., Hakonarson, H., et al. (2020). A polygenic and phenotypic risk prediction for polycystic ovary syndrome evaluated by phenome-wide association studies. *The Journal of Clinical Endocrinology & Metabolism*, 105(6):1918–1936.

Kahn, L. G., Harley, K. G., Siegel, E. L., Zhu, Y., Factor-Litvak, P., Porucznik, C. A., Klein-Fedyshin, M., Hipwell, A. E., and program collaborators for Environmental Influences on Child Health Outcomes Program (2021). Persistent organic pollutants and couple fecundability: a systematic review. *Human Reproduction Update*, 27(2):339–366.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

Karayiannis, D., Kontogianni, M. D., Mendorou, C., Mastrominas, M., and Yiannakouris, N. (2018). Adherence to the mediterranean diet and ivf success rate among non-obese women attempting fertility. *Human Reproduction*, 33(3):494–502.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Kjaersgaard, M. I. S., Parner, E. T., Vestergaard, M., Sørensen, M. J., Olsen, J., Christensen, J., Bech, B. H., and Pedersen, L. H. (2013). Prenatal antidepressant exposure and risk of spontaneous abortion–a population-based study. *PLoS One*, 8(8):e72095.

Klein, J. P., Moeschberger, M. L., et al. (2003). *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer.

Kleinbaum, D. G. and Klein, M. (1996). *Survival analysis a self-learning text*. Springer.

Klonoff-Cohen, H., Lam-Kruglick, P., and Gonzalez, C. (2003). Effects of maternal and paternal alcohol consumption on the success rates of in vitro fertilization and gamete intrafallopian transfer. *Fertility and sterility*, 79(2):330–339.

Krieger, N. (2000). Discrimination and health. *Social epidemiology*, 1:36–75.

Kuang, H., Jin, S., Hansen, K. R., Diamond, M. P., Coutifaris, C., Casson, P., Christman, G., Alvero, R., Huang, H., Bates, G. W., et al. (2015). Identification and replication of prediction models for ovulation, pregnancy and live birth in infertile women with polycystic ovary syndrome. *Human Reproduction*, 30(9):2222–2233.

Laursen, A. S. D., Johannesen, B. R., Willis, S. K., Hatch, E. E., Wise, L. A., Wesselink, A. K., Rothman, K. J., Sørensen, H. T., and Mikkelsen, E. M. (2022). Adherence to nordic dietary patterns and risk of first-trimester spontaneous abortion. *European Journal of Nutrition*, 61(6):3255–3265.

Lee, E. T. and Wang, J. (2003). *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons.

Lee, S.-I., Lee, H., Abbeel, P., and Ng, A. Y. (2006). Efficient l_1 regularized logistic regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 21, pages 401–408. https://cdn.aaai.org/AAAI/2006/AAAI06-064.pdf.

Lenail, A. (2022). Nn-svg. https://alexlenail.me/NN-SVG/.

Li, S., Li, D., and Ma, Y. (2022). A mathematical model to predict the probability of a successful pregnancy. *Journal of Obstetrics and Gynaecology Research*, 48(7):1632–1640.

LightGBM-Guide (2024). Lightgbm parameters tuning guide.

Lim, S.-S., Kakoly, N. S., Tan, J. W. J., Fitzgerald, G., Bahri Khomami, M., Joham, A. E., Cooray, S. D., Misso, M. L., Norman, R. J., Harrison, C. L., and Ranasinha, S. (2019). Metabolic syndrome in polycystic ovary syndrome: a systematic review, meta-analysis and meta-regression. *Obesity reviews*, 20(2):339–352.

Liu, L., Jiao, Y., Li, X., Ouyang, Y., and Shi, D. (2020). Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. *Computer Methods and Programs in Biomedicine*, 196:105624.

Longato, E., Vettoretti, M., and Di Camillo, B. (2020). A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, 108:103496.

Louis, G. M. B., Lum, K. J., Sundaram, R., Chen, Z., Kim, S., Lynch, C. D., Schisterman, E. F., and Pyper, C. (2011). Stress reduces conception probabilities across the fertile window: evidence in support of relaxation. *Fertility and sterility*, 95(7):2184–2189.

Lynch, C., Sundaram, R., Maisog, J., Sweeney, A., and Buck Louis, G. (2014). Preconception stress increases the risk of infertility: results from a couple-based prospective cohort study—the life study. *Human reproduction*, 29(5):1067–1075.

Macaluso, M., Wright-Schnapp, T. J., Chandra, A., Johnson, R., Satterwhite, C. L., Pulver, A., Berman, S. M., Wang, R. Y., Farr, S. L., and Pollack, L. A. (2010). A public health focus on infertility prevention, detection, and management. *Fertility and sterility*, 93(1):16–e1.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Michel, C., Burchert, A., Hochhaus, A., Saussele, S., Neubauer, A., Lauseker, M., Krause, S. W., Kolb, H.-J., Hossfeld, D. K., Nerl, C., et al. (2019). Imatinib dose reduction in major molecular response of chronic myeloid leukemia: results from the german chronic myeloid leukemia-study iv. *Haematologica*, 104(5):955.

Millen, A. E., Midthune, D., Thompson, F. E., Kipnis, V., and Subar, A. F. (2006). The national cancer institute diet history questionnaire: validation of pyramid food servings. *American Journal of Epidemiology*, 163(3):279–288.

Mughal, T. I., Radich, J. P., Deininger, M. W., Apperley, J. F., Hughes, T. P., Harrison, C. J., Gambacorti-Passerini, C., Saglio, G., Cortes, J., and Daley, G. Q. (2016). Chronic myeloid leukemia: reminiscences and dreams. *Haematologica*, 101(5):541.

Nielsen, A., Gerd Hannibal, C., Eriksen Lindekilde, B., Tolstrup, J., Frederiksen, K., Munk, C., Bergholt, T., Buss, L., Ottesen, B., Grønbaek, M., et al. (2006). Maternal smoking predicts the risk of spontaneous abortion. *Acta obstetricia et gynecologica Scandinavica*, 85(9):1057–1065.

Nillni, Y. I., Wesselink, A. K., Gradus, J. L., Hatch, E. E., Rothman, K. J., Mikkelsen, E. M., and Wise, L. A. (2016). Depression, anxiety, and psychotropic medication use and fecundability. *American journal of obstetrics and gynecology*, 215(4):453–e1.

Nowell, P. C. and Hungerford, D. A. (1960). Chromosome studies on normal and leukemic human leukocytes. *Journal of the National Cancer Institute*, 25(1):85–109.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216.

Odar Stough, C., Khalsa, A. S., Nabors, L. A., Merianos, A. L., and Peugh, J. (2019). Predictors of exclusive breastfeeding for 6 months in a national sample of us children. *American Journal of Health Promotion*, 33(1):48–56.

on Obstetric Practice, A. C. et al. (2013). Definition of term pregnancy. *Obstetrics and Gynecology*, 122:1139–1140.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pflueger, S. M. V. (2005). Cytogenetics of spontaneous abortion. In Gersen, S. L. and Keagle, M. B., editors, *The Principles of Clinical Cytogenetics*, pages 323–345. Humana Press, Totowa, NJ. Accessed September 26, 2022.

Quenby, S. M. and Farquharson, R. G. (1993). Predicting recurring miscarriage: what is important? *Obstetrics & Gynecology*, 82(1):132–138.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G.-Z. (2016). Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21.

Riestenberg, C., Jagasia, A., Markovic, D., Buyalos, R. P., and Azziz, R. (2022). Health care-related economic burden of polycystic ovary syndrome in the united states: pregnancy-related and long-term health consequences. *The Journal of Clinical Endocrinology & Metabolism*, 107(2):575–585.

Risch, H. A., WEISS, N. S., Aileen Clarke, E., and MILLER, A. B. (1988). Risk factors for spontaneous abortion and its recurrence. *American journal of epidemiology*, 128(2):420–430.

Rossen, L. M., Ahrens, K. A., and Branum, A. M. (2018). Trends in risk of pregnancy loss among us women, 1990–2011. *Paediatric and perinatal epidemiology*, 32(1):19–29.

Rothman, K. J. (1977). Fetal loss, twinning and birth weight after oral-contraceptive use. *New England journal of medicine*, 297(9):468–471.

Rudin, C. and Radin, J. (2019). Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2):10–1162.

Sackoff, J., Kline, J., and Susser, M. (1994). Previous use of oral contraceptives and spontaneous abortion. *Epidemiology*, 5(4):422–428.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

Savitz, D. A., Chan, R. L., Herring, A. H., Howards, P. P., and Hartmann, K. E. (2008). Caffeine and miscarriage risk. *Epidemiology*, 19(1):55–62. DOI: https://doi.org/10.1097/ede.0b013e31815c09b9.

Savitz, D. A., Hertz-Picciotto, I., Poole, C., and Olshan, A. F. (2002). Epidemiologic measures of the course and outcome of pregnancy. *Epidemiologic reviews*, 24(2):91–101.

Schiffer, C. A. (2019). Discontinuation of tyrosine kinase inhibitors in patients with chronic myelogeneous leukemia–you can do this at home if you read the instructions. *Haematologica*, 104(8):1508.

Schmid, M., Wright, M. N., and Ziegler, A. (2016). On the use of harrell's c for clinical risk prediction via random survival forests. *Expert Systems with Applications*, 63:450–459.

Shanmuganathan, N., Pagani, I. S., Ross, D. M., Park, S., Yong, A. S., Braley, J. A., Altamura, H. K., Hiwase, D. K., Yeung, D. T., Kim, D.-W., et al. (2021). Early bcr-abl1 kinetics are predictive of subsequent achievement of treatment-free remission in chronic myeloid leukemia. *Blood*, 137(9):1196–1207.

Shickel, B., Tighe, P. J., Bihorac, A., and Rashidi, P. (2017). Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Shitvelman, E., Lifshitz, B., Gale, R., and Canaani, E. (1985). Fused transcript of abl and bcr genes in chronic myelogenous leukemia. *Nature*, 315(6020):550–554.

Sirmans, S. M. and Pate, K. A. (2013). Epidemiology, diagnosis, and management of polycystic ovary syndrome. *Clinical Epidemiology*, 5:1–13. DOI: https://doi.org/10.2147/clep.s37559.

Sisk, R., Lin, L., Sperrin, M., Barrett, J. K., Tom, B., Diaz-Ordaz, K., Peek, N., and Martin, G. P. (2021). Informative presence and observation in routine health data: a review of methodology for clinical risk prediction. *Journal of the American Medical Informatics Association*, 28(1):155–166.

Snick, H., Snick, T., Evers, J., and Collins, J. (1997). The spontaneous pregnancy prognosis in untreated subfertile couples: the walcheren primary care study. *Human reproduction (Oxford, England)*, 12(7):1582–1588.

Snijder, C. A., te Velde, E., Roeleveld, N., and Burdorf, A. (2012). Occupational exposure to chemical substances and time to pregnancy: a systematic review. *Human reproduction update*, 18(3):284–300.

Soares, S. R. and Melo, M. A. (2008). Cigarette smoking and reproductive function. *Current Opinion in Obstetrics and Gynecology*, 20(3):281–291.

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative validation of the block, willett, and national cancer institute food frequency questionnaires: the eating at america's table study. *American journal of epidemiology*, 154(12):1089–1099.

Suissa, S. and Dell'Aniello, S. (2020). Time-related biases in pharmacoepidemiology. *Pharmacoepidemiology and drug safety*, 29(9):1101–1110.

Sundaram, R., Mumford, S. L., and Buck Louis, G. M. (2017). Couples' body composition and time-to-pregnancy. *Human reproduction*, 32(3):662–668.

Talaei, M. and Izadi, I. (2024a). Adaptive differential privacy in federated learning: A priority-based approach. *arXiv preprint arXiv:2401.02453.*

Talaei, M. and Izadi, I. (2024b). Comments on" federated learning with differential privacy: Algorithms and performance analysis". *arXiv e-prints*, pages arXiv–2406.

Talaei, M. and Izadi, I. (2024c). Enhancing federated learning with adaptive differential privacy and priority-based aggregation. *arXiv preprint arXiv:2406.18491.*

Talaei, M., Rikos, A. I., Olshevsky, A., White, L. F., and Paschalidis, I. C. (2024). Network-based epidemic control through optimal travel and quarantine management. *arXiv preprint arXiv:2407.19133*.

Teede, H., Misso, M., Costello, M., Dokras, A., Laven, J., Moran, L., Piltonen, T., Norman, R., et al. (2018). International evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Melbourne, Australia: Monash University*. https://www.monash.edu/__data/assets/pdf_file/0003/3379521/Evidence-Based-Guidelines-2023.pdf.

Therneau, T. M. (1997). Extending the cox model. In *Proceedings of the first Seattle symposium in biostatistics: survival analysis*, pages 51–84. Springer.

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

Toosy, S., Sodi, R., and Pappachan, J. M. (2018). Lean polycystic ovary syndrome (pcos): an evidence-based practical approach. *Journal of Diabetes & Metabolic Disorders*, 17:277–285.

Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. Hachette UK.

Twigt, J., Bolhuis, M., Steegers, E., Hammiche, F., Van Inzen, W., Laven, J., and Steegers-Theunissen, R. (2012). The preconception diet is associated with the chance of ongoing pregnancy in women undergoing ivf/icsi treatment. *Human reproduction*, 27(8):2526–2531.

Vagios, S., James, K. E., Sacha, C. R., Hsu, J. Y., Dimitriadis, I., Bormann, C. L., and Souter, I. (2021). A patient-specific model combining antimüllerian hormone and body mass index as a predictor of polycystic ovary syndrome and other oligo-anovulation disorders. *Fertility and Sterility*, 115(1):229–237.

Valsamakis, G., Chrousos, G., and Mastorakos, G. (2019). Stress, female reproduction and pregnancy. *Psychoneuroendocrinology*, 100:48–57.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477.

Van der Steeg, J., Steures, P., Eijkemans, M., Habbema, J., Hompes, P., Broekmans, F., van Dessel, H., Bossuyt, P., van der Veen, F., and Mol, B. (2007). Cecerm study group (collaborative effort for clinical evaluation in reproductive medicine). pregnancy is predictable: a large-scale prospective external validation of the prediction of spontaneous pregnancy in subfertile couples. *Human Reproduction*, 22:536–42.

van Doorn, S., Brakenhoff, T. B., Moons, K. G., Rutten, F. H., Hoes, A. W., Groenwold, R. H., and Geersing, G. J. (2017). The effects of misclassification in routine healthcare databases on the accuracy of prognostic prediction models: a case study of the cha2ds2-vasc score in atrial fibrillation. *Diagnostic and prognostic research*, 1:1–9.

Venners, S. A., Wang, X., Chen, C., Wang, L., Chen, D., Guang, W., Huang, A., Ryan, L., O'Connor, J., Lasley, B., et al. (2004). Paternal smoking and pregnancy loss: a prospective study using a biomarker of pregnancy. *American Journal of Epidemiology*, 159(10):993–1001.

Villarroel, C., López, P., Merino, P. M., Iñiguez, G., Sir-Petermann, T., and Codner, E. (2015). Hirsutism and oligomenorrhea are appropriate screening criteria for polycystic ovary syndrome in adolescents. *Gynecological Endocrinology*, 31(8):625–629.

Wang, R., Cong, Y., Li, C., Zhang, C., and Lin, H. (2019). Predictive value of early molecular response for deep molecular response in chronic phase of chronic myeloid leukemia. *Medicine*, 98(15).

Wang, T., Paschalidis, A., Liu, Q., Liu, Y., Yuan, Y., Paschalidis, I. C., et al. (2020). Predictive models of mortality for hospitalized patients with covid-19: retrospective cohort study. *JMIR medical informatics*, 8(10):e21788.

Wekker, V., Van Dammen, L., Koning, A., Heida, K., Painter, R., Limpens, J., Laven, J., Roeters van Lennep, J., Roseboom, T., and Hoek, A. (2020). Long-term cardiometabolic disease risk in women with pcos: a systematic review and meta-analysis. *Human reproduction update*, 26(6):942–960.

Weng, X., Odouli, R., and Li, D.-K. (2008). Maternal caffeine consumption during pregnancy and the risk of miscarriage: a prospective cohort study. *American journal of obstetrics and gynecology*, 198(3):279–e1.

Wesselink, A. K. (2021). Multigenerational effects of environmental exposures. *Human Reproduction*, 36(3):539–542.

Wesselink, A. K., Hatch, E. E., Rothman, K. J., Mikkelsen, E. M., Aschengrau, A., and Wise, L. A. (2019). Prospective study of cigarette smoking and fecundability. *Human Reproduction*, 34(3):558–567.

Wesselink, A. K., Hatch, E. E., Rothman, K. J., Weuve, J. L., Aschengrau, A., Song, R. J., and Wise, L. A. (2018). Perceived stress and fecundability: a preconception cohort study of north american couples. *American journal of epidemiology*, 187(12):2662–2671.

Wesselink, A. K., Rothman, K. J., Hatch, E. E., Mikkelsen, E. M., Sørensen, H. T., and Wise, L. A. (2017). Age and fecundability in a north american preconception cohort study. *American journal of obstetrics and gynecology*, 217(6):667–e1.

Wesselink, A. K., Willis, S. K., Laursen, A. S. D., Mikkelsen, E. M., Wang, T. R., Trolle, E., Tucker, K. L., Rothman, K. J., Wise, L. A., and Hatch, E. E. (2022). Protein-rich food intake and risk of spontaneous abortion: a prospective cohort study. *European Journal of Nutrition*, 61(5):2737–2748.

Wilcox, A. J., Weinberg, C. R., O'Connor, J. F., Baird, D. D., Schlatterer, J. P., Canfield, R. E., Armstrong, E. G., and Nisula, B. C. (1988). Incidence of early loss of pregnancy. *New England Journal of Medicine*, 319(4):189–194.

Williams, D. R. and Collins, C. (2001). Racial residential segregation: a fundamental cause of racial disparities in health. *Public Health Reports*, 116(5):404–416. DOI: https://doi.org/10.1093/phr/116.5.404.

Willis, S. K., Hatch, E. E., Wesselink, A. K., Rothman, K. J., Mikkelsen, E. M., and Wise, L. A. (2019). Female sleep patterns, shift work, and fecundability in a north american preconception cohort study. *Fertility and sterility*, 111(6):1201–1210.

Wise, L. A., Mikkelsen, E. M., Rothman, K. J., Riis, A. H., Sørensen, H. T., Huybrechts, K. F., and Hatch, E. E. (2011). A prospective cohort study of menstrual characteristics and time to pregnancy. *American journal of epidemiology*, 174(6):701–709.

Wise, L. A., Rothman, K. J., Mikkelsen, E. M., Stanford, J. B., Wesselink, A. K., McKinnon, C., Gruschow, S. M., Horgan, C. E., Wiley, A. S., Hahn, K. A., et al. (2015). Design and conduct of an i nternet-based preconception cohort study in n orth a merica: P regnancy s tudy o nline. *Paediatric and perinatal epidemiology*, 29(4):360–371.

Wise, L. A., Rothman, K. J., Wesselink, A. K., Mikkelsen, E. M., Sorensen, H. T., McKinnon, C. J., and Hatch, E. E. (2018). Male sleep duration and fecundability in a north american preconception cohort study. *Fertility and sterility*, 109(3):453–459.

Wise, L. A., Wesselink, A. K., Hatch, E. E., Weuve, J., Murray, E. J., Wang, T. R., Mikkelsen, E. M., Sørensen, H. T., and Rothman, K. J. (2020). Changes in behavior with increasing pregnancy attempt time: a prospective cohort study. *Epidemiology (Cambridge, Mass.)*, 31(5):659.

Xu, H., Feng, G., Alpadi, K., Han, Y., Yang, R., Chen, L., Li, R., and Qiao, J. (2022). A model for predicting polycystic ovary syndrome using serum amh, menstrual cycle length, body mass index and serum androstenedione in chinese reproductive aged population: A retrospective cohort study. *Frontiers in Endocrinology*, 13.

Yi, Y., Lu, G., Ouyang, Y., Gong, F., Li, X., et al. (2016). A logistic model to predict early pregnancy loss following in vitro fertilization based on 2601 infertility patients. *Reproductive Biology and Endocrinology*, 14(1):1–7.

Yland, J. J., Bresnick, K. A., Hatch, E. E., Wesselink, A. K., Mikkelsen, E. M., Rothman, K. J., Sørensen, H. T., Huybrechts, K. F., and Wise, L. A. (2020). Pregravid contraceptive use and fecundability: prospective cohort study. *BMJ: British Medical Journal*, 371:m3966. https://doi.org/10.1136/bmj.m3966.

Yland, J. J., Wang, T., Zad, Z., Willis, S. K., Wang, T. R., Wesselink, A. K., Jiang, T., Hatch, E. E., Wise, L. A., and Paschalidis, I. C. (2022). Predictive models of pregnancy based on data from a preconception cohort study. *Human Reproduction*, 37(3):565–576.

Yland, J. J., Zad, Z., Wang, T. R., Wesselink, A. K., Jiang, T., Hatch, E. E., Paschalidis, I. C., and Wise, L. A. (2024). Predictive models of miscarriage on the basis of data from a preconception cohort study. *Fertility and Sterility*.

Zad, Z., Bonecker, S., Wang, T., Zalcberg, I., Stelzer, G. T., Sabioni, B., Gutiyama, L. M., Fleck, J. L., and Paschalidis, I. C. (2024a). Prediction of deep molecular response in chronic myeloid leukemia using supervised machine learning models. *Leukemia research*, 141:107502.

Zad, Z., Jiang, V. S., Wolf, A. T., Wang, T., Cheng, J. J., Paschalidis, I. C., and Mahalingaiah, S. (2024b). Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. *Frontiers in Endocrinology*, 15:1298628.

CURRICULUM VITAE

[redacted]

[redacted]