Machine Learning Approaches for Equitable Healthcare

by

Irene Y. Chen

A.B., Harvard University (2014)
S.M., Harvard University (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
August 26, 2022

Certified by .....
David Sontag
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee for Graduate Students

Machine Learning Approaches for Equitable Healthcare

by

Irene Y. Chen

Submitted to the Department of Electrical Engineering and Computer Science
on August 26, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in
Electrical Engineering and Computer Science

Abstract

With the proliferation of clinical data and algorithms to improve clinical care, researchers are increasingly concerned about the equity and fairness of the resulting machine learning models. Because the observational data we collect can be noisy, incomplete, and biased, seemingly straight-forward implementation of existing methods for clinical intervention or better understanding human knowledge can lead to inaccurate and inequitable clinical algorithms. To begin to address these challenges, we need new tools to tackle the bias that can arise when modeling data. In this work, we present machine learning approaches for auditing, ameliorating, and preventing bias in the machine learning for healthcare model development process. In particular, we focus on case studies that can provide actionable insights.

In this thesis, we present several examples of machine learning approaches towards equitable healthcare and recommend changes based on the results of the corresponding experiments. Questions of equity and bias can be thought of in terms of the different steps of the model development pipeline. We argue that these model development steps can be made more equitable and unbiased when they 1) mitigate algorithmic bias that may occur from biased data collection or model development, and 2) address known existing systemic health disparities.

We present four case studies of machine learning approaches towards equitable healthcare, and demonstrate these approaches on real clinical tasks. First, we decompose the sources of discrimination and provide empirical estimation techniques. We present results on applying these techniques in the task of intensive care unit mortality prediction and salary prediction. Second, we consider the predictive analytics of health insurance providers, namely predicting the likelihood of hospitalization and the likelihood of high-risk pregnancy. We apply the same discrimination decomposition techniques towards practical steps for mitigating algorithmic discrimination. Third, we study the task of clustering interval-censored time-series data. We develop a deep generative model, called SubLign, to learn the latent delayed entry alignment value for each time-series as well as the heterogeneous progression patterns across the

population. We evaluate our model in the context of synthetically generated data. Following, we study the task of disease subtyping for the improved understanding of disease progression. We present results on clustering clinical patients including heart failure and Parkinson's disease. Finally, we study an example of using machine learning on an understudied problem that affects underserved patients: early detection of intimate partner violence. We develop a model that predicts the likelihood of eventual intimate partner violence self-reporting and radiology injury labeling from radiology reports. We conclude with a discussion about how machine learning can continue to address equity and bias in healthcare.

Thesis Supervisor: David Sontag
Title: Professor of Electrical Engineering and Computer Science

"It was all very well to say 'Drink me,' but the wise little Alice was not going to do that in a hurry. 'No, I'll look first,' she said, 'and see whether it's marked "poison" or not.'"

—Lewis Carroll, *Alice in Wonderland*

Acknowledgments

First and foremost, my greatest thanks go to my advisor David Sontag. I met David the first month of my PhD at an ice cream social. While talking and eating chocolate chip ice cream, I immediately knew that he would be an amazing advisor. His boundless enthusiasm and consistent encouragement have been a guiding light through the PhD. Most notably, he has given me the freedom to make mistakes and find my own research voice—while remaining a committed advisor and providing copious feedback and assistance. Thank you for taking a chance on me.

I would like to thank Marzyeh Ghassemi for being my co-author and thesis reader. Marzyeh is a rare combination of brilliance, creativity, and kindness. Her generosity in time and energy have made an invaluable impact on my PhD and research directions. It is a great honor to be her friend.

I would also like to thank my thesis reader Peter Szolovits. His perceptive eye and unyielding support have been a pleasure to both work with and teach with. A big part of this dissertation is built on top of his research, and his insights and feedback have made this thesis a stronger body of work.

During my PhD, I have met several other mentors who have shaped my work and my journey. I thank my recommendation letter writers Ziad Obermeyer and Sherri Rose for their relentless advocacy, my Microsoft Research internship mentors Solon Barocas and Hal Daumé III for their cheerful discussions, and my academic advisor Una-May O'Reilly for her warmth, especially at the beginning of my PhD. Thank you also to my research qualifying committee David Karger and Hal Abelson for their thoughtful feedback.

Collaboration has been one of the most fun parts of graduate school. I thank my fellow collaborators: Rahul G. Krishnan, Fredrik D. Johansson, Steven Horng, Bharti Khurana, Tristan Naumann, Rajesh Ranganath, Emma Pierson, Shalmali Joshi, Kadijah Ferryman, Laleh Seyyed-Kalantari, Matthew McDermott, Stephanie Gervasi, Aaron Smith-McLallen, May Choi, Willie Boag, and Heather Berlin. In particular, I wrote my first research paper of graduate school with Fredrik. His patience

and standard of excellence have set a high bar for future research and mentorship. Another key collaborator, Stephanie has brought both brightness and competence to our weekly calls. Her passion for high-caliber work and ambition for better and more equitable algorithms have made our Independence Blue Cross collaboration an absolute joy.

I thank the whole Clinical Machine Learning Group, past and present, especially Monica Agrawal, Rebecca Peyser Boiarsky, Michael Oberst, Zeshan Hussain, Christina Ji, Chandler Squires, Hussein Mozannar, Hunter Lang, Alejandro Buendia, Stefan Hegselmann, Nikolaj Thams, Penny Brant, Mercy Asiedu, Ahmed Alaa, Sol Rodriguez, Isaac Lage, Yoni Halpern, Rachel Hodos, Sanjat Kanjilal, and Uri Shalit. Their humor and camaraderie have made the lab feel like a home.

I feel immense gratitude to the teachers and professors who have shaped me before I even stepped foot on MIT campus. Thank you to teachers who nurtured my early love of learning: Rachel Oliver, Paula Watson, Rick Byrd, Lasley Gober, Chris Harrow, Frances Fondren, Funmi Oke, and John Roberts. Thank you for the Harvard professors who instilled the computational fundamentals in me: Margo Levine, Michael Mitzenmacher, Edward Glaeser, Margo Seltzer, Radhika Nagpal, Michael Luca and Ben Edelman. Special thanks to Margo Seltzer who convinced me to take the leap and apply for a PhD. I am appreciative especially to Mike and Ben, my first research mentors.

Although not an educational institution, Dropbox exposed me to large-scale technology that continues to shape how I think about the impact of machine learning on the world (as well as provided a financial cushion that made the PhD stipend more livable). Thank you particularly to Lilian Weng, ChenLi Wang, Zach Kagin, David Kriegman, and Peter Belhumeur.

I have cherished the communities where I have found refuge during my graduate school experience, most especially the Graduate Women in Course 6 and Machine Learning for Health. The EECS Graduate Office, and particularly Leslie Kolodziejski, keeps it and other student organizations alive, with a particular emphasis on students from underrepresented groups. I am immensely grateful for this departmental sup-

port. The Machine Learning for Health workshop and now symposium has introduced me to researchers near in area but far geographically to me. It is a thriving research community, and I am proud to be part of it.

I want to express a more abstract sense of gratitude to the spirit of the Massachusetts Institute of Technology, from whom I have learned much about technological innovation without limits. I will miss my time running along the Charles River in the fading sunlight or huddling in the underground tunnels, warmed by the hot water pipes and the buzz of unbridled intensity for all things.

I have been extremely lucky to be surrounded by many great friends, who have been an infinite source of happiness and perhaps my secret weapon in this PhD. Just to name a few (with apologies for the ones who I have forgotten): Mary Tung, my dear friend for more than half my life, who provides valuable perspective during the stresses of the PhD. Marzyeh, my fearless friend, who challenges me to achieve the unimaginable. Ginny Fahs, my childhood best friend, who reminds me of what it means to be a thoughtful human in this world. Karen Hao, a recent but treasured friend, who remains an inspiration on charting one's life path. Thank you to my trivia gang of Amanda Beck, Nikhil Bhargava, and the incomparable Matthew Brennan. Thank you to my fellow PhD students to whom I've complained and from whom I've drawn strength: Xianglin Flora Meng, Lydia Liu, Maz Abulnaga, Genevieve Flaspohler, Divya Shanmugam, Serena Booth, Sarah Cen, and Willie Boag. Thank you as well to my non-PhD cohort friends Amritha Jayanti, Allison Koenecke, Nora Arkin, Serene Yeo, and Caroline Gwynn for the gift of your friendship.

Lastly and most dearly, I would like to thank my mother, my father, and my brother. They have provided unending support and love for my entire life, as well as providing me access to the best education imaginable. I am proud to be my parents' daughter and to follow in their academic footsteps. This thesis is dedicated to them.

This doctoral thesis has been examined by a Committee of the
Department of Electrical Engineering and Computer Science as follows:

Professor David Sontag .....
Thesis Superviser
Professor of Electrical Engineering and Computer Science

Professor Peter Szolovits .....
Member, Thesis Committee
Professor of Electrical Engineering and Computer Science

Professor Marzyeh Ghassemi .....
Member, Thesis Committee
Assistant Professor of Electrical Engineering and Computer Science

Contents

1 Introduction **37**

1.1 Motivation and Related Work 37

1.2 Summary of Contributions 39

1.2.1 Chapter 2: Ethical Machine Learning for Healthcare 39

1.2.2 Chapter 3: Mitigating Biased Machine Learning Methods . . . 39

1.2.3 Chapter 4: Auditing Algorithmic Bias in Predictive Algorithms for Health Insurance 40

1.2.4 Chapter 5: Clustering Interval-Censored Multivariate Time-Series Data 40

1.2.5 Chapter 6: Chronic Disease Progression Subtyping 41

1.2.6 Chapter 7: Early Detection of Intimate Partner Violence Using Radiology Reports 41

2 Ethical Machine Learning for Healthcare **43**

2.1 Problem Selection 44

2.1.1 Global Health Injustice 44

2.1.2 Racial Injustice 45

2.1.3 Gender Injustice 45

2.1.4 Diversity of the Scientific Workforce 46

2.2 Data Collection 46

2.2.1 Heterogeneous Data Losses 47

2.2.2 Population-specific Data Losses 49

2.3 Outcome Definition 51

2.3.1 Clinical Diagnosis 51

2.3.2 Health Care Costs . . . 52
2.4 Algorithm Development . . . 53
2.4.1 Understanding Confounding . . . 54
2.4.2 Feature Selection . . . 55
2.4.3 Tuning Parameters . . . 56
2.4.4 Performance Metrics . . . 57
2.4.5 Group Fairness Definition . . . 57
2.5 Post-Deployment Consideration . . . 58
2.5.1 Quantifying Impact . . . 59
2.5.2 Model Generalizability . . . 59
2.5.3 Model and Data Documentation . . . 60
2.5.4 Regulation . . . 61

3 Mitigating Biased Machine Learning Methods 63
3.1 Introduction . . . 63
3.1.1 Contributions . . . 65
3.1.2 Related Work . . . 65
3.2 Decomposing Discrimination . . . 67
3.2.1 Sources of Discrimination . . . 67
3.2.2 Bias-Variance-Noise Decompositions of Discrimination . . . 67
3.3 Methods of Mitigating Algorithmic Discrimination . . . 72
3.3.1 Reducing Discrimination through Data Collection . . . 72
3.3.2 Increasing Training Set Size . . . 72
3.3.3 Measuring Additional Variables . . . 73
3.4 Experiments . . . 74
3.4.1 Income Prediction . . . 74
3.4.2 Intensive Care Unit Mortality Prediction . . . 77
3.4.3 Book Review Ratings . . . 79
3.5 Discussion . . . 81

4 Auditing Algorithmic Bias in Predictive Algorithms for Health Insurance **83**

4.1 Introduction . . . 83

4.1.1 Contributions . . . 84

4.1.2 Related Work . . . 85

4.2 Data . . . 87

4.2.1 Likelihood of Hospitalization . . . 87

4.2.2 High-Risk Pregnancy . . . 88

4.3 Methods . . . 89

4.3.1 Impact of Additional Training Data . . . 89

4.3.2 Bayes Error Estimation . . . 90

4.3.3 Subpopulation Identification Using Topic Modeling . . . 90

4.4 Results . . . 91

4.4.1 Impact of Additional Training Data . . . 91

4.4.2 Bayes Error Estimation . . . 91

4.4.3 Subpopulation Identification Using Topic Modeling . . . 92

4.5 Discussion . . . 97

5 Clustering Interval-Censored Multivariate Time-Series Data **101**

5.1 Introduction . . . 101

5.1.1 Contributions . . . 103

5.1.2 Related Work . . . 103

5.2 SubLign: Subtype & Align . . . 104

5.2.1 Generative Model . . . 104

5.2.2 Inference . . . 106

5.2.3 Remarks . . . 107

5.3 Identifiability Under a Noiseless Model . . . 109

5.3.1 Generative Process . . . 109

5.4 Experiments . . . 114

5.4.1 Datasets . . . 114

5.4.2 Hyperparameters and Baselines . . . 114
5.4.3 Evaluation . . . 117
5.4.4 Statistical Significance . . . 118
5.5 Results . . . 119
5.5.1 Recovering Subtypes with Interval Censoring . . . 119
5.5.2 Recovering Known Alignment Values . . . 119
5.6 Discussion . . . 120

6 Chronic Disease Progression Subtyping 121
6.1 Introductions . . . 121
6.1.1 Contributions . . . 122
6.1.2 Related Work . . . 122
6.2 Subtyping Parkinson's Disease Patients . . . 123
6.2.1 Data . . . 123
6.2.2 Experiments . . . 123
6.2.3 Results . . . 123
6.3 Subtyping Heart Failure Patients . . . 124
6.3.1 Data . . . 124
6.3.2 Methods . . . 125
6.3.3 Experiments . . . 125
6.3.4 Results . . . 127
6.4 Discussion . . . 130

7 Early Detection of Intimate Partner Violence Using Radiology Reports 133
7.1 Introduction . . . 133
7.1.1 Contributions . . . 134
7.1.2 Related Work . . . 135
7.2 Data . . . 136
7.2.1 IPV Patient Selection . . . 136
7.2.2 Control Group Selection . . . 137

7.2.3 Injury Labels . . . 137
7.2.4 Data Cleaning . . . 137
7.2.5 Demographic Data . . . 138
7.3 Methods . . . 140
7.3.1 Experiment Setup . . . 140
7.3.2 Models . . . 140
7.3.3 Evaluation . . . 141
7.4 Results . . . 142
7.4.1 IPV and Injury Prediction Performance and Predictive Features 142
7.4.2 Error Analysis . . . 143
7.4.3 Report-Program Date Gap . . . 144
7.5 Discussion . . . 146

8 Conclusion **147**

A Additional Information for Chapter 3 **151**
A.1 Testing for significant discrimination . . . 151
A.2 Additional experimental details . . . 152
A.2.1 Datasets . . . 152
A.2.2 Synthetic experiments . . . 153
A.2.3 Clinical notes . . . 153
A.3 Exploring model choice . . . 154
A.4 Regression with homoskedastic noise . . . 156
A.5 Bias-variance decomposition. Proof of Theorem 1. . . . 157
A.6 Difference between power law curves . . . 159

B Additional Information for Chapter 4 **161**
B.1 Topic Model for Likelihood of Hospitalization . . . 161
B.2 Topic Model for High-Risk Pregnancy . . . 212

C Additional Information for Chapter 5 **263**
C.1 Identifiability and Inference . . . 264

C.1.1 Identifiability 264
C.1.2 Variational Lower Bound 267
C.2 Experiment Setup Details 267
C.2.1 Optimal hyperparameters for sigmoid and clinical baselines . . 267
C.2.2 Missing values 267
C.2.3 Statistical significance 268
C.3 Additional Experiment Results 268
C.3.1 SubLign and SubNoLign subtype visualization 268
C.3.2 Model Misspecification 268
C.3.3 Missingness Experiments 269
C.4 Quadratic Data Results 271
C.4.1 Setup 271
C.4.2 Optimal hyperparameters for quadratic datasets 272
C.4.3 Results 274

D Additional Information for Chapter 6 **279**
D.1 Data Privacy and Ethics 279
D.2 Clinical dataset biomarkers and baseline features 280

List of Figures

3-1 Scenarios illustrating how properties of the training set and model choice affect perceived discrimination in a binary classification task, under the assumption that outcomes and predictions are *unaware*, i.e. $p(Y \mid X, A) = p(Y \mid X)$ and $p(\hat{Y} \mid X, A) = p(\hat{Y} \mid X)$. Through bias-variance-noise decompositions (see Section 3.2.2), we can identify which of these dominate in their effect on fairness. We propose procedures for addressing each component in Section 3.3.1, and use them in experiments (see Section 3.4) to mitigate discrimination in income prediction and prediction of ICU mortality. 68

3-2 Noise level estimation in income prediction with the Adult dataset. Group differences in false positive rates and false negative rates for a random forest classifier decrease with increasing training set size. . . 75

3-3 Mortality prediction from clinical notes using logistic regression. Best viewed in color. 77

3-4 Goodreads dataset for book rating prediction. Adding training data decreases overall mean squared error (MSE) for both groups while adding training data to only one group has a much bigger impact on reducing $\overline{\Gamma}$. Increasing the number of features reduces MSE but does not reduce $\overline{\Gamma}$. 79

4-1 Logistic regression performance for Likelihood of Hospitalization, measured by **a)** 1 - accuracy and **b)** 1 - positive predictive value (PPV) versus the percentage of total training data. 91

4-2 Logistic regression performance for High-Risk Pregnancy, measured by **a)** 1 - accuracy and **b)** 1 - positive predictive value (PPV) versus the percentage of total training data 92

4-3 Topic weights for **a)** Likelihood of Hospitalization task and **b)** High-Risk Pregnancy task. Prevalent topics include Topic 28 (hypertension) for LOH and Topic 15 for HRP (prenatal care). See supplementary materials Chapter B for full descriptions of the learned topics with representative conditions, medications, specialty visits, and procedures. 97

4-4 Range between maximum and minimum error topic values across race groups for **a)** Likelihood of Hospitalization task and **b)** High-Risk Pregnancy task. 98

5-1 (a) Patient data can be interval-censored, meaning longitudinal data can be missing based on entry to the dataset, e.g., first diagnosis, and the data may lack a common outcome against which to align. Patients may enter the dataset at any stage of the disease. (b) Interval censoring can make clustering patient time-series data challenging because data may be aligned incorrectly, e.g., first diagnosis. We seek to understand disease heterogeneity by inferring subtypes after correcting for misalignment. 102

5-2 Graphical model of SubLign 102

6-1 Subtypes found by SubLign for 619 Parkinson's disease patients and healthy controls. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ with a Benjamini-Hochberg correction are listed. 129

6-2 Subtypes found by SubLign for 423 Parkinson's disease patients only. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ with a Benjamini-Hochberg correction are listed. 129

6-3 Subtypes found by SubLign from heart failure patients using echocardiogram biomarkers. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ with a Benjamini-Hochberg correction are listed. 130

7-1 **Left:** Earliest possible report-program date gap per patient. **Right:** Scatterplot and marginal histograms of report-program date gap (x-axis) and IPV prediction probability (y-axis) for all radiology reports of IPV victims for random forest classifier. See Section 7.3.3 for definition of report-program date gap. 143

A-1 Inverse power-laws (Pow3) fit to generalization error as a function of training set size on synthetic data. Dotted lines are extrapolations from sample sizes indicated by black stars. This illustrates the difficulty of estimating the Bayes error through extrapolation, here at $\overline{N}_0 = 3 \cdot 10^{-4}$ and $\overline{N}_1 = 7 \cdot 10^{-3}$ respectively. 154

A-2 Additional clinical notes experiments highlight the differences in false positive and false negative rates. We also examine the effect of training size on cancer patients in the dataset. 156

C-1 One of three dimensions of learned SubLign and SubNoLign subtypes from sigmoid synthetic data plotted on top of original data generating functions. 269

C-2 Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign outperforms baselines while KMeans+Loss recovers subtypes (ARI metric) better than SubNoLign, but alignment metrics are difficult to recover because of the horizontal subtype . . . 275

C-3 Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign and SubNoLign have near-perfect clustering accuracy (ARI) while alignment metrics (swaps, Pearson) are difficult recover because of the horizontal subtype. 275

C-4 Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign outperforms baselines in clustering and alignment metrics although the task is challenging. 276

C-5 Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign, SubNoLign, and KMeans+Loss perform well on clustering. 276

C-6 Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign learns subtypes and recovers alignment better than baselines. 277

C-7 Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign learns subtypes and alignment values. . . . 277

List of Tables

3.1 Discrimination level estimation in income prediction with the Adult dataset. Estimation of Bayes error lower and upper bounds (E_{low} and E_{up}) for zero-one loss of men and women. Intervals for men and women are non-overlapping for Nearest Neighbors. 75

4.1 Percentage of patient data by race for Likelihood of Hospitalization and High-Risk Pregnancy tasks. 89

4.2 Bayes error noise estimates, including lower and upper bounds, for Likelihood of Hospitalization and High-Risk Pregnancy tasks. Mahalanobis distance Bayes error estimates only contain an upper bound. Other entries with dash lines indicate singular matrix errors from matrix inverse operations. 94

4.3 Most representative labs, procedures, conditions, specialty visits, and medications for Topic 35 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 95

4.4 Most representative labs, procedures, conditions, specialty visits, and medications for topic 44 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 96

4.5 Likelihood of Hospitalization task topic prevalence, difference between maximum and minimum error values across races, and race-specific errors for top 15 topics sorted by difference between maximum and minimume error values across races. Recall that the the dataset has 76.2% White patients, 14.8% Black patients, 4.0% Other patients, and 2.7% Unknown patients (see Table 4.1). 98

4.6 High-Risk Pregnancy task topic prevalence, difference between maximum and minimum error values across races, and race-specific errors for top 15 topics sorted by difference between maximum and minimume error values across races. Recall that the the dataset has 39.0% White patients, 5.7% Black patients, 7.3% Other patients, and 48.0% Unknown patients (see Table 4.1). 99

5.1 Means and standard deviations over 5 trials for synthetic sigmoid dataset with 1000 patients, 3 dimensions, and 4 observations per patient. 118

6.1 Means and standard deviations over 5 trials for 619 patients in the PPMI dataset including 423 Parkinson's disease patients and 196 healthy controls. 128

6.2 Subtypes found by SubLign from Parkinson's disease patients and healthy controls using sparsely collected biomarkers. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ are listed. 130

6.3 Heart Failure KMeans+Loss subtypes (patient counts in parentheses), described by mean baseline features. Only statistically significant features are listed and do not include systolic and diastolic HF, two known phenotypes of HF. 131

7.1 Summary statistics for dataset, with percentages of radiology reports. 139

7.2 Model AUC means and standard deviations over five data splits for IPV and injury prediction using radiology reports. Bold rows indicate best performance for task. 143

7.3 Predictive words for IPV and injury averaged across five trials based on linear coefficients of logistic regression. Underline indicates words consistent with clinical literature. 144

7.4 Error analysis for IPV and injury predictions from random forest classifier. Means and standard deviations of accuracy, sensitivity (TPR), and specificity (TNR) computed over 5 data splits with overall model sensitivity set to 0.95. Bold indicates subgroups with particularly low metrics. 145

A.1 Summary statistics of clinical notes dataset 153

A.2 Top and bottom 5 topics (of 50) based on variance in error rates of groups. Error rates by group and topic $p(\hat{Y} \neq Y|K, A)$ are reported in percentages. 155

B.1 Most representative labs, procedures, conditions, specialty visits, and medications for topic 1 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 162

B.2 Most representative labs, procedures, conditions, specialty visits, and medications for topic 2 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 163

B.3 Most representative labs, procedures, conditions, specialty visits, and medications for topic 3 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 164

B.4 Most representative labs, procedures, conditions, specialty visits, and medications for topic 4 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 165

B.5 Most representative labs, procedures, conditions, specialty visits, and medications for topic 5 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 166

B.6 Most representative labs, procedures, conditions, specialty visits, and medications for topic 6 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 167

B.7 Most representative labs, procedures, conditions, specialty visits, and medications for topic 7 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 168

B.8 Most representative labs, procedures, conditions, specialty visits, and medications for topic 8 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 169

B.9 Most representative labs, procedures, conditions, specialty visits, and medications for topic 9 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 170

B.10 Most representative labs, procedures, conditions, specialty visits, and medications for topic 10 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 171

B.11 Most representative labs, procedures, conditions, specialty visits, and medications for topic 11 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 172

B.12 Most representative labs, procedures, conditions, specialty visits, and medications for topic 12 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 173

B.13 Most representative labs, procedures, conditions, specialty visits, and medications for topic 13 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 174

B.14 Most representative labs, procedures, conditions, specialty visits, and medications for topic 14 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 175

B.15 Most representative labs, procedures, conditions, specialty visits, and medications for topic 15 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 176

B.16 Most representative labs, procedures, conditions, specialty visits, and medications for topic 16 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 177

B.17 Most representative labs, procedures, conditions, specialty visits, and medications for topic 17 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 178

B.18 Most representative labs, procedures, conditions, specialty visits, and medications for topic 18 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 179

B.19 Most representative labs, procedures, conditions, specialty visits, and medications for topic 19 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 180

B.20 Most representative labs, procedures, conditions, specialty visits, and medications for topic 20 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 181

B.21 Most representative labs, procedures, conditions, specialty visits, and medications for topic 21 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 182

B.22 Most representative labs, procedures, conditions, specialty visits, and medications for topic 22 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 183

B.23 Most representative labs, procedures, conditions, specialty visits, and medications for topic 23 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 184

B.24 Most representative labs, procedures, conditions, specialty visits, and medications for topic 24 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 185

B.25 Most representative labs, procedures, conditions, specialty visits, and medications for topic 25 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 186

B.26 Most representative labs, procedures, conditions, specialty visits, and medications for topic 26 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 187

B.27 Most representative labs, procedures, conditions, specialty visits, and medications for topic 27 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 188

B.28 Most representative labs, procedures, conditions, specialty visits, and medications for topic 28 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 189

B.29 Most representative labs, procedures, conditions, specialty visits, and medications for topic 29 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 190

B.30 Most representative labs, procedures, conditions, specialty visits, and medications for topic 30 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 191

B.31 Most representative labs, procedures, conditions, specialty visits, and medications for topic 31 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 192

B.32 Most representative labs, procedures, conditions, specialty visits, and medications for topic 32 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 193

B.33 Most representative labs, procedures, conditions, specialty visits, and medications for topic 33 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 194

B.34 Most representative labs, procedures, conditions, specialty visits, and medications for topic 34 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 195

B.35 Most representative labs, procedures, conditions, specialty visits, and medications for topic 35 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 196

B.36 Most representative labs, procedures, conditions, specialty visits, and medications for topic 36 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 197

B.37 Most representative labs, procedures, conditions, specialty visits, and medications for topic 37 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 198

B.38 Most representative labs, procedures, conditions, specialty visits, and medications for topic 38 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 199

B.39 Most representative labs, procedures, conditions, specialty visits, and medications for topic 39 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 200

B.40 Most representative labs, procedures, conditions, specialty visits, and medications for topic 40 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 201

B.41 Most representative labs, procedures, conditions, specialty visits, and medications for topic 41 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 202

B.42 Most representative labs, procedures, conditions, specialty visits, and medications for topic 42 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 203

B.43 Most representative labs, procedures, conditions, specialty visits, and medications for topic 43 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 204

B.44 Most representative labs, procedures, conditions, specialty visits, and medications for topic 44 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 205

B.45 Most representative labs, procedures, conditions, specialty visits, and medications for topic 45 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 206

B.46 Most representative labs, procedures, conditions, specialty visits, and medications for topic 46 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 207

B.47 Most representative labs, procedures, conditions, specialty visits, and medications for topic 47 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 208

B.48 Most representative labs, procedures, conditions, specialty visits, and medications for topic 48 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 209

B.49 Most representative labs, procedures, conditions, specialty visits, and medications for topic 49 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 210

B.50 Most representative labs, procedures, conditions, specialty visits, and medications for topic 50 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature. 211

B.51 Most representative labs, procedures, conditions, specialty visits, and medications for topic 1 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 213

B.52 Most representative labs, procedures, conditions, specialty visits, and medications for topic 2 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 214

B.53 Most representative labs, procedures, conditions, specialty visits, and medications for topic 3 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 215

B.54 Most representative labs, procedures, conditions, specialty visits, and medications for topic 4 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 216

B.55 Most representative labs, procedures, conditions, specialty visits, and medications for topic 5 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 217

B.56 Most representative labs, procedures, conditions, specialty visits, and medications for topic 6 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 218

B.57 Most representative labs, procedures, conditions, specialty visits, and medications for topic 7 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 219

B.58 Most representative labs, procedures, conditions, specialty visits, and medications for topic 8 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 220

B.59 Most representative labs, procedures, conditions, specialty visits, and medications for topic 9 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 221

B.60 Most representative labs, procedures, conditions, specialty visits, and medications for topic 10 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 222
B.61 Most representative labs, procedures, conditions, specialty visits, and medications for topic 11 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 223
B.62 Most representative labs, procedures, conditions, specialty visits, and medications for topic 12 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 224
B.63 Most representative labs, procedures, conditions, specialty visits, and medications for topic 13 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 225
B.64 Most representative labs, procedures, conditions, specialty visits, and medications for topic 14 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 226
B.65 Most representative labs, procedures, conditions, specialty visits, and medications for topic 15 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 227
B.66 Most representative labs, procedures, conditions, specialty visits, and medications for topic 16 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 228
B.67 Most representative labs, procedures, conditions, specialty visits, and medications for topic 17 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 229
B.68 Most representative labs, procedures, conditions, specialty visits, and medications for topic 18 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 230
B.69 Most representative labs, procedures, conditions, specialty visits, and medications for topic 19 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 231

B.70 Most representative labs, procedures, conditions, specialty visits, and medications for topic 20 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 232
B.71 Most representative labs, procedures, conditions, specialty visits, and medications for topic 21 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 233
B.72 Most representative labs, procedures, conditions, specialty visits, and medications for topic 22 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 234
B.73 Most representative labs, procedures, conditions, specialty visits, and medications for topic 23 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 235
B.74 Most representative labs, procedures, conditions, specialty visits, and medications for topic 24 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 236
B.75 Most representative labs, procedures, conditions, specialty visits, and medications for topic 25 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 237
B.76 Most representative labs, procedures, conditions, specialty visits, and medications for topic 26 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 238
B.77 Most representative labs, procedures, conditions, specialty visits, and medications for topic 27 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 239
B.78 Most representative labs, procedures, conditions, specialty visits, and medications for topic 28 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 240
B.79 Most representative labs, procedures, conditions, specialty visits, and medications for topic 29 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 241

B.80 Most representative labs, procedures, conditions, specialty visits, and medications for topic 30 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 242
B.81 Most representative labs, procedures, conditions, specialty visits, and medications for topic 31 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 243
B.82 Most representative labs, procedures, conditions, specialty visits, and medications for topic 32 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 244
B.83 Most representative labs, procedures, conditions, specialty visits, and medications for topic 33 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 245
B.84 Most representative labs, procedures, conditions, specialty visits, and medications for topic 34 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 246
B.85 Most representative labs, procedures, conditions, specialty visits, and medications for topic 35 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 247
B.86 Most representative labs, procedures, conditions, specialty visits, and medications for topic 36 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 248
B.87 Most representative labs, procedures, conditions, specialty visits, and medications for topic 37 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 249
B.88 Most representative labs, procedures, conditions, specialty visits, and medications for topic 38 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 250
B.89 Most representative labs, procedures, conditions, specialty visits, and medications for topic 39 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 251

B.90 Most representative labs, procedures, conditions, specialty visits, and medications for topic 40 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 252
B.91 Most representative labs, procedures, conditions, specialty visits, and medications for topic 41 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 253
B.92 Most representative labs, procedures, conditions, specialty visits, and medications for topic 42 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 254
B.93 Most representative labs, procedures, conditions, specialty visits, and medications for topic 43 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 255
B.94 Most representative labs, procedures, conditions, specialty visits, and medications for topic 44 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 256
B.95 Most representative labs, procedures, conditions, specialty visits, and medications for topic 45 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 257
B.96 Most representative labs, procedures, conditions, specialty visits, and medications for topic 46 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 258
B.97 Most representative labs, procedures, conditions, specialty visits, and medications for topic 47 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 259
B.98 Most representative labs, procedures, conditions, specialty visits, and medications for topic 48 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 260
B.99 Most representative labs, procedures, conditions, specialty visits, and medications for topic 49 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 261

B.100 Most representative labs, procedures, conditions, specialty visits, and medications for topic 50 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature. 262

C.1 Model misspecification experiment means and standard deviations using 5 cubic splines datasets. 269

C.2 Experiments on synthetic data with 50% of the data missing. Baselines include SuStaIn [313], BayLong [153], PAGA [307], SPARTan [237], clustering using Soft-DTW [74], and clustering using Kernel-DTW [81]. Imputation methods include MICE [304] and MRNN [312]. 270

C.3 Experiments on synthetic data with 25% of the data missing. Baselines include SuStaIn [313], BayLong [153], PAGA [307], SPARTan [237], clustering using Soft-DTW [74], and clustering using Kernel-DTW [81]. Imputation methods include MICE [304] and MRNN [312]. 271

C.4 Experiments on synthetic data with 0% of the data missing. 272

C.5 Quadratic dataset subtype generating functions and corresponding figure numbers 273

Chapter 1

Introduction

1.1 Motivation and Related Work

Machine learning has an opportunity to fundamentally change healthcare. As of June 2021, the Food and Drug Administration (FDA) reports that 343 artificial intelligence (AI) and machine learning (ML)-enabled medical devices are now approved [105], with more surely to follow. The promise of ML for healthcare cannot be overestimated, with researchers pursuing medical tasks including but not limited to: early diagnosis [196], risk stratification [23], triage [149], treatment selection [154], clinical trials [301], disease progression [101], and enabling patient interaction with the healthcare system. [297] Part of this advancement is predicated on the availability of data through the expansion of health data including electronic health records (EHRs), signal data, genomic data, and wearable app data. Computational advances in deep learning have leveraged these large datasets and accelerated the process, with AI algorithms outperforming humans in scoped tasks including melanoma classification [139], lung computed tomography [15], and chest radiograph reading. [283]

As these models proliferate into parts of the medical system, there remain technological challenges in place. For example, electronic health records can be missing large amounts of data [122] as patients come in and out of the health system. Treatment variation can differ across clinicians [182], hospital systems [270], or patient insurance types. [65] Complicating the matter is the fact that medical knowledge is not set in

stone. In fact, 13% of medical practice papers are medical reversals. [248]

When considering questions of equity and bias, the technical challenges deepen. The healthcare system as it currently exists has known systemic health inequities in areas including maternal mortality [73] and access to care. [186] These disparities can affect the data that is collected. Even large population health datasets like genome-wide association studies have over 96% participants of European descent. [222] These differences in data collection can have big effects, especially when the conditional data distributions are different. In one example, heart attacks can manifest differently for men and women [104], which may affect the care patients receive and how early people enter the healthcare system. [85]

For my thesis, I study techniques to address bias that may occur due to the use of algorithms in medical contexts. Although scientific fields may define concepts like bias and equity in a variety of ways, we define bias and equity related to the impacts on protected groups including race and gender. When we consider the machine learning development process from problem selection to post-deployment monitoring, it is possible to target bias at different stages. One possibility is to examine algorithms near- or post-deployment using a bias audit that assesses the comparative performances across known patient subpopulations. Quantifiable algorithmic bias at this stage would correspond to disparate impact on the different patient subpopulations. Solutions to address algorithmic bias found through bias audits would directly correspond to resource allocation to ameliorate this bias in deployed settings. Another possibility is to build considerations for the system health disparities that feed into the observational health data directly into the algorithm.

The remaining document is split up as follows.

In Chapter 3, I present a method for addressing post-deployment algorithmic bias auditing, specifically decomposing sources of unfairness in supervised learning using a bias-variance-noise decomposition. In Chapter 4, I show empirical results using these techniques to audit predictive algorithms used by a large health insurer. Questions of equity and bias can also be considered in the context of algorithm development. In Chapter 5, I present a method for correcting for known health disparities in access to

care to correct confounding bias in disease phenotyping algorithms. This algorithm is validated on real-world chronic condition data from lupus, heart failure, and Parkinson's patients in Chapter 6. Lastly, in Chapter 7 I detail an example of using machine learning to tackle patient populations and health conditions that may be previously overlooked, namely the application of early detection of intimate partner violence.

1.2 Summary of Contributions

1.2.1 Chapter 2: Ethical Machine Learning for Healthcare

Originally published at Annual Reviews of Biomedical Data Science as [57] The use of machine learning (ML) in healthcare raises numerous ethical concerns, especially as models can amplify existing health inequities. Here, we outline ethical considerations for equitable ML in the advancement of healthcare. Specifically, we frame ethics of ML in healthcare through the lens of social justice. We describe ongoing efforts and outline challenges in a proposed pipeline of ethical ML in health, ranging from problem selection to postdeployment considerations. We close by summarizing recommendations to address these challenges.

1.2.2 Chapter 3: Mitigating Biased Machine Learning Methods

Originally published at NeurIPS as [51] Recent attempts to achieve fairness in predictive models focus on the balance between fairness and accuracy. In sensitive applications such as healthcare or criminal justice, this trade-off is often undesirable as any increase in prediction error could have devastating consequences. In this work, we argue that the fairness of predictions should be evaluated in context of the data, and that unfairness induced by inadequate samples sizes or unmeasured predictive variables should be addressed through data collection, rather than by constraining the model. We decompose cost-based metrics of discrimination into bias, variance, and noise, and propose actions aimed at estimating and reducing each term. Finally,

we perform case-studies on prediction of income, mortality, and review ratings, confirming the value of this analysis. We find that data collection is often a means to reduce discrimination without sacrificing accuracy.

1.2.3 Chapter 4: Auditing Algorithmic Bias in Predictive Algorithms for Health Insurance

Partially published in Health Affairs as [118], partially original work Health insurance providers use predictive algorithms for many uses, including case management prioritization. Because of the large reach of these algorithms, it is essential to rigorously audit and understand the impact. In partnership with Independence Blue Cross (IBC), a health insurer based in Philadelphia, we examine two algorithms that form subcomponents of case management processes: 1) Likelihood of Hospitalization (LOH), which predicts future acute and non-elective hospitalization in the next six months, and 2) Baby Blueprints (BBP), which predicts high-risk pregnancy for women of childbearing age. We present results examining these algorithms for algorithmic bias. We open source our code as a package called `omop-fairness` for general use on all OMOP-standard data.

1.2.4 Chapter 5: Clustering Interval-Censored Multivariate Time-Series Data

Originally published at AAAI as [55] Unsupervised learning is often used to uncover clusters in data. However, different kinds of noise may impede the discovery of useful patterns from real-world time-series data. In this work, we focus on mitigating the interference of interval censoring in the task of clustering for disease phenotyping. We develop a deep generative model, called SubLign, that clusters time-series while correcting for censorship time. We provide conditions under which clusters and the amount of delayed entry may be identified from data under a noiseless model. On synthetic data, we demonstrate accurate, stable, and interpretable results that outperform several benchmarks.

1.2.5 Chapter 6: Chronic Disease Progression Subtyping

Published at AAAI as [55] We study how to find subtypes using longitudinal data from patients with chronic conditions. When alignment information across time-series dataset isn't known, we use SubLign as outlined in Chapter 5. On a dataset of Parkinson's disease patients, we study how motor skills data reveals clinically confirmed differences between Parkinson's disease patients and control patients, while correcting for interval censorship. On a dataset of heart failure patients, we recover known subtypes of diastolic and systolic heart failure and confirm recent clinical findings of differences of progression for obese patients and women, while correcting for interval censorship.

1.2.6 Chapter 7: Early Detection of Intimate Partner Violence Using Radiology Reports

Originally published at PSB2020 as [53] Intimate partner violence (IPV) is an urgent, prevalent, and under-detected public health issue. We present machine learning models to assess patients for IPV and injury. We train the predictive algorithms on radiology reports with 1) IPV labels based on entry to a violence prevention program and 2) injury labels provided by emergency radiology fellowship-trained physicians. Our dataset includes 34,642 radiology reports and 1479 patients of IPV victims and control patients. Our best model predicts IPV a median of 3.08 years before violence prevention program entry with a sensitivity of 64% and a specificity of 95%. We conduct error analysis to determine for which patients our model has especially high or low performance and discuss next steps for a deployed clinical risk model.

Chapter 2

Ethical Machine Learning for Healthcare

As machine learning algorithms are increasingly used for clinical care, the ethical concerns proliferate in a variety of clinical applications. Excitement about human-level performance [283] of machine learning for health is balanced against ethical concerns, such as the potential for these tools to exacerbate existing health disparities. [34] For instance, recent work has demonstrated that state-of-the-art clinical prediction models underperform on women, ethnic and racial minorities, and those with public insurance. [58] Even more worrisome, healthcare models designed to optimize referrals to long-term case management programs for millions of patients have been found to exclude Black patients with similar health conditions compared to white patients from case management programs. [227]

I argue that we can leverage bioethics principles to inform the machine learning process. In contrast to previous work designed to inform clinical care practices, these ethical considerations span the entire model development pipeline from problem specification to post-deployment considerations. I focus primarily on differences between groups induced by, or related to, the model development pipeline, drawing on both the bioethics principle of justice and the established social justice centering of public health ethics. Unjust differences in quality and outcomes of healthcare between groups often reflect existing societal disparities for disadvantaged groups. This frame-

work aligns with other bioethics principles such as beneficence and nonmaleficence; however, the focus is primarily on groups of patients rather than on individuals.

2.1 Problem Selection

There are many factors that influence the selection of a research problem, from interest to available funding. However, problem selection can also be a matter of justice if the research questions that are proposed, and ultimately funded, focus on the health needs of advantaged groups. Below we provide examples of how disparities in research teams and funding priorities exacerbate existing socioeconomic, racial, and gender injustices.

2.1.1 Global Health Injustice

The "10/90" gap refers to the fact that the vast majority of health research dollars are spent on problems that affect a small fraction of the global population. [296, 244] Diseases that are most common in lower-income countries receive far less funding than diseases that are most common in high-income countries [298] (relative to the number of individuals they affect). As an example, 26 poverty-related diseases account for 14% of the global disease burden, but receive only 1.3% of global health-related research and development expenditure. Nearly 60% of the burden of poverty-related neglected diseases occurs in Western and Eastern sub-Saharan Africa as well as South Asia. Malaria, tuberculosis, and HIV/AIDS all have shares of global health-related research and development expenditure that are at least five times smaller than their share of global disease burden. [298] This difference in rates of funding represents an injustice because it further exacerbates the disadvantages faced by Global South populations. While efforts like the "All Of Us" Project [229] and the 23andMe's Call for Collaboration [292] seek to collect more inclusive data, these efforts have come under criticism for not reflecting global health concerns, particularly among Indigenous groups. [287]

2.1.2 Racial Injustice

Racial bias affects which health problems are prioritized and funded. For example, sickle cell disease and cystic fibrosis are both genetic disorders of similar severity, but sickle cell disease is more common in Black patients, while cystic fibrosis is more common in white patients. In the United States (US), however, cystic fibrosis receives 3.4 times more funding per affected individual from the US National Institutes of Health (NIH), the largest funder of US clinical research, and hundreds of times more private funding. [94] The disparities in funding persist despite the 1972 Sickle Cell Anemia Control Act, which recognizes that sickle cell has been neglected by the wider research community. Further, screening for sickle cell disease is viewed by some as unfair targeting [233], and Black patients with the disease who seek treatment are often maligned as drug abusers. [257]

2.1.3 Gender Injustice

Women's health conditions like endometriosis are poorly understood; as a consequence, even basic statistics like the prevalence of endometriosis remain unknown, with estimates ranging from 1% to 10% of the population. [48, 91] Similarly, the menstrual cycle is stigmatized and understudied [48, 243], producing a dearth of understanding that undermines the health of half the global population. Basic facts about the menstrual cycle — including which menstrual experiences are normal and which are predictive of pathology — remain unknown. [48] This lack of focus on the menstrual cycle propagates into clinical practice and data collection despite evidence that it affects many aspects of health and disease. [145, 12] Menstrual cycles are also not often recorded in clinical records and global health data. [48] In fact, the NIH did not have an R01 grant, the NIH's original and historically oldest grant mechanism, relating to the influence of sex and gender on health and disease until 2019. [2] Notably, recent work has moved to target such understudied problems via ambulatory women's health-tracking mobile apps. These crowd-sourcing efforts stand to accelerate women's health research by collecting data from cohorts that are orders of

magnitude larger than those used in previous studies. [48]

2.1.4 Diversity of the Scientific Workforce

The diversity of the scientific workforce profoundly influences the problems studied, and contributes to the biases in problem selection. [168] Research shows that scientists from underrepresented racial and gender groups tend to prioritize different research topics. They produce more novel research, but their innovations are taken up at lower rates. [147] Female scientists tend to study different scientific subfields, even within the same larger field — for example, within sociology, they have been historically better-represented on papers about sociology of the family or early childhood [303] — and express different opinions about ethical dilemmas in computer science. [242] Proposals from white researchers in the US are more likely to be funded by the NIH than proposals from Black researchers [150, 123], which in turns affects what topics are given preference. For example, a higher fraction of NIH proposals from Black scientists study community and population-level health. [150] Overall, this evidence suggests that diversifying the scientific workforce will lead to problem selection that more equitably represents the interests and needs of the population as a whole.

2.2 Data Collection

The role of health data is ever-expanding, with new data sources routinely being integrated into decision-making around health policy and design. This wealth of high-quality data, coupled with advancements in ML models, has played a significant role in accelerating the use of computationally informed policy and practice to strengthen health care and delivery platforms. Unfortunately, data can be biased in ways that have (or can lead to) disproportionate negative impacts on already marginalized groups. First, data on group membership can be completely absent. For instance, countries such as Canada and France do not record race and ethnicity in their nationalized health databases [3, 188], making it impossible to study race-based disparities and hypotheses around associations of social determinants of health. Second,

data can be imbalanced. Recent work on acute kidney injury achieved state-of-the-art prediction performance in a large dataset of 703,782 adult patients using 620,000 features; however, they note that model performance was lower in female patients since female patients comprised 6.38% of patients in the training data. [282] Other work has indicated that this issue can not be simply addressed by "pre-training" a model in a more balanced data setting prior to fine-tuning on an imbalanced dataset. [212] This indicates that a model cannot be "initialized" with a balanced baseline representation which ameliorates issues of imbalance in downstream tasks, and suggests we must solve this problem at the root, be it with more balanced and comprehensive data, specialty learning algorithms, or combinations therein. Finally, while some sampling biases can be recognized and possibly corrected, others may be difficult to correct. For example, work in medical imaging has demonstrated that models may overlook unforeseen stratification of conditions, like rare manifestations of diseases, which can result in harm in clinical settings. [226, 293]

In this section, we discuss common biases in data collection. We consider two types of processes that result in a loss of data. First, processes that affect what kind of information is collected, or heterogenous data loss, across varying input types. For example, clinical trials with aggressive inclusion criteria or social media data that reflects those with access to devices. Second, we examine processes that affect whether an individual's information is collected, or population-specific data losses, where individuals are impacted by their population type, often across data input categories. For example, undocumented immigrants may fear deportation if they participate in health care systems.

2.2.1 Heterogeneous Data Losses

Some data loss is specific to the data type, due to assumptions about noise that may have been present during the collection process. However, data noise and missingness can cause unjust inequities that impact populations in different ways. We cover four main data types: randomized controlled trials (RCTs), electronic health care records (EHR), administrative health data, and social media data.

Randomized Controlled Trials Randomized controlled trials are often run specifically to gather "unbiased" evidence of treatment effects. However, RCTs have notoriously aggressive exclusion (or inclusion) criteria [256], which create study cohorts that are not representative of general patient populations. [70] In one study of RCTs used to define asthma treatment, an estimated 94% of the adult asthmatic population would not have been eligible for the trials. [284] There is a growing methodological literature designing methods to generalize RCT treatment effects to other populations. [272] However, current empirical evidence indicates that such generalizations can be challenging given available data or may require strong assumptions in practice.

Electronic Health Records Much recent work in ML also leverages large electronic health records data. EHR data are a complex reflection of patient health, health care systems, and providers, where data missingness is a known, and meaningful, problem. [302] As one salient example, a large study of laboratory tests to model three-year survival found that health care process features had a stronger predictive value than the patient's physiological feaures. [7] Further, not all treatments investigated in RCTs can be easily approximated in EHR. [22]

Biases in EHR data may arise due to differences in patient populations, access to care, or the availability of EHR systems. [99] As an example, the widely-used MIMIC-III EHR dataset includes most patients who receive care at the intensive care units in Beth Israel Deaconess Medical Center (BIDMC), but this sample is obviously limited by which individuals have access to care at BIDMC, which has a largely white patient population. [51] In the United States, uninsured Black and Hispanic or Latin(o/x) patients, as well as Hispanic or Latin(o/x) Medicaid patients, are less likely to have primary care providers with EHR systems, as compared to white patients with private insurance. [146] Other work has shown that gender discrimination in health care access has not been systematically studied in India, primarily due to a lack of reliable data. [166]

Administrative Health Records In addition to RCTs and EHR, health care billing claims data, clinical registries, and linked health survey data are also common

data sources in population health and health policy research [134, 305], with known biases in which populations are followed, and who is able to participate. Translating such research into practice is a crucial part of maintaining health care quality, and limited participation of minority populations by sexual orientation and gender identity [42], race and ethnicity [201], and language [177] can lead to health interventions and policies that are not inclusive, and can create new injustices for these already marginalized groups.

Social Media Data Data from social media platforms and search-based research by nature consists only of individuals with internet access. [86] Even small choices like limiting samples to those from desktop versus mobile platforms are a problematic distinction in non-North American contexts. [4] Beyond concerns about access to resources or geographic limitations, data collection and scraping pipelines for most social media cohorts do not yield a random sample of individuals. Further, the common practice of limiting analysis to those satisfying a specified threshold of occurrence can lead to skewed data. As an example, when processing the large volume of Twitter data (7.6 billion tweets) researchers may first restrict to users who can be mapped to a US county (1.78 billion), then to those Tweets that contain only English (1.64 billion tweets), and finally remove users who made less than 30 posts (1.53 billion). [124]

2.2.2 Population-specific Data Losses

As with data types, the modern data deluge does not apply equally to all communities. Historically underserved groups are often underrepresented, misrepresented, or entirely missing from health data that inform consequential health policy decisions. When individuals from disadvantaged communities appear in observational datasets, they are less likely to be accurately captured due to errors in data collection and systemic discrimination. Larger genomics datasets often target European populations, producing genetic risk scores which are more accurate in individuals of European ancestry than other ancestries. [208] We note four specific examples of populations that are commonly impacted: low- and middle-income nationals, transgender and gender non-conforming individuals, undocumented migrants, and pregnant women.

Low- and Middle-Income Nationals Health data are infrequently collected due to resource constraints, and even basic disease statistic data such as prevalence of mortality rates can be challenging to find for low- and middle-income nations. [4] When data are collected, it is not digitized, and often contains errors. In 2001, the World Health Organization found that only 9 out of the 46 member states in Sub-Saharan Africa could produce death statistics for a global assessment of the burden of disease, with data coverage often less than 60% in these countries. [157]

Transgender and Gender Non-conforming Individuals The health care needs and experiences of transgender and gender non-conforming individuals are not well-documented in datasets [156] because documented sex, not gender identity, is what is usually available. However, documented sex is often discordant with gender identity for transgender and gender non-conforming individuals. Apart from health documentation concerns, transgender people are often concerned about their basic physical safety when reporting their identities. In the US, it was only in 2016, with the release of the US Transgender Survey that there was a meaningfully sized dataset — 28,000 respondents — to enable significant analysis and quantification of discrimination and violence that transgender people face. [156]

Undocumented Immigrants Safety concerns are important in data collection for undocumented migrants, where socio-political environments can lead to individuals feeling unsafe during reporting opportunities. When immigration policies limit access to public services for immigrants and their families, these restrictions lead to spillover effects on clinical diagnoses. As one salient example, autism diagnoses for Hispanic children in California fell following aggressive federal anti-immigrant policies requiring citizenship verification at hospitals. [107]

Pregnant Women Despite pregnancy being neither rare nor an illness, the US continues to experience rising maternal mortality and morbidity rates. In the US, the maternal mortality rate has more than doubled from 9.8 per 100,000 live births in 2000 to 21.5 in 2014. [8] Importantly, disclosure protocols recommend suppression of information in nationally available datasets when the number of cases or events in a data "cell" is low, to reduce the likelihood of a breach of confidentiality. For example,

the US Centers for Disease Control suppresses numbers for counties with fewer than 10 deaths for a given disease. [279] Although these data omissions occur because of patient privacy, such censoring on the *dependent* variable introduces particularly pernicious statistical bias and, as a result, much remains to be understood about what community, health facility, patient, and provider-level factors drive high mortality rates.

2.3 Outcome Definition

The next step in the model pipeline is to define the outcome of interest for a health care task. Even seemingly straightforward tasks like defining whether a patient *has* a disease can be skewed by how prevalent diseases are, or how they manifest in some patient populations. For example, a model predicting if a patient will develop heart failure will need labeled examples both of patients who have heart failure, and patients without heart failure. Choosing these patients can rely on parts of the EHR that may be skewed due to lack of access to care, or abnormalities in clinical care: e.g., economic incentives may alter diagnosis code logging [170], clinical protocol affects the frequency and observation of abnormal tests [7], historical racial mistrust may delay care and affect patient outcomes [32], and naive data collection can yield inconsistent labels in chest X-rays. [226] Such biased labels, and the resulting models, may cause clinical practitioners to allocate resources poorly.

We discuss social justice considerations in two examples of commonly modelled health care outcomes: clinical diagnosis and health care costs. In each example, it is essential that model developers choose a reliable proxy and account for noise in the outcome labels as these choices can have a large impact on performance and equity of the resulting model.

2.3.1 Clinical Diagnosis

Clinical diagnosis is a fundamental task for clinical prediction models, e.g., models for computer-aided diagnosis from medical imaging. In clinical settings, researchers

often select patient disease occurrence as the prediction label for models. However, there are many options for the choice of a disease occurrence label. For example, the outcome label for developing cardiovascular disease could be defined through the occurrence of specific phrases in clinical notes. However, women can manifest symptoms of acute coronary syndrome differently [46] and receive delayed care as a result [38], which may then manifest in diagnosis labels derived from the clinical notes being gender-skewed. Because differences in label noise results in disparities in model impact, researchers have the responsibility to choose and improve disease labels, so that these inequalities do not further exacerbate disparities in health.

Additionally, it is important to consider the health care system in which disease labels are logged. For example, health care providers leverage diagnosis codes for billing purposes, not clinical research. As a result, diagnosis codes can create ambiguities because of overlap and hierarchy in codes. Moreover, facilities have incentives to under-report [170] and over-report [255, 117] outcomes, yielding differences in model representations.

Recent advances in improving disease labels target statistical corrections based on estimates of the label noise. For instance, a positive label may be reliable, but the omission of a positive label could either indicate a negative label (i.e., no disease) or merely a missed positive label. Methods to address the positive-unlabeled setting use estimated noise rates [221] or hand-curated labels that are strongly correlated with positive labels, known also as "silver-standard" labels, from clinicians. [133] Clinical analysis of sources of error in disease labels can also guide improvements [225] and identify affected groups. [226]

2.3.2 Health Care Costs

Developers of clinical models may choose to predict health care costs, meaning the ML model seeks to predict which patients will cost the health care provider more in the future. Some model developers may use health care costs as a proxy for future health needs to guide accurate targeting of interventions [227], with the underlying assumption that addressing patients with future health need will limit future costs.

Others may explicitly want to understand patients who will have high health care cost to reduce the total cost of health care. [276] However, because socioeconomic factors affect both access to health care *and* access to financial resources, these models may yield predictions that exacerbate inequities.

For model developers seeking to optimize for health need, health care costs can deviate from health need on an individual level because of patient socioeconomic factors. For instance, in a model used to allocate care management program slots to high-risk patients, the choice of future health care costs as a predictive outcome led to racial disparities in patient allocation to the program. [227] Health care costs can differ from health need on an institutional level due to underinsurance and undertreatment within the patient population. [66] After defining health disparities as all differences except those due to clinical need and preferences, researchers have found racial disparities in mental health care. Specifically, white patients had higher rates of initiation of treatment for mental health compared to Black and Hispanic or Latin(o/x) patients. Because the analysis controls for health need, the disparities are solely a result of differences in health care access and systemic discrimination. [67]

Addressing issues that arise from the use of health care costs depends on the setting of the ML model. In cases where health need is of highest importance, a natural solution is to choose another outcome definition besides health care costs, e.g., the number of chronic diseases as a measure of health need. If a model developer is most concerned with cost, it is possible to correct for health disparities in predicting health care costs by building fairness considerations directly into the predictive model objective function. [319] Building these types of algorithmic procedures is further discussed in Section 2.4.

2.4 Algorithm Development

Algorithm development considers the construction of the underlying computation for the ML model and presents a major vulnerability and opportunity for ethical ML in health care. Just as data are not neutral, algorithms are not neutral. A

disproportionate amount of power lies with research teams who, after determining the research questions, make decisions about critical components of an algorithm such as the loss function. [130, 168] In the case of loss functions, common choices like the L_1 absolute error loss and L_2 squared error loss do not target the same conditional functions of the outcome, instead minimizing the error in the median and mean respectively. Using a surrogate loss (e.g., hinge loss for the error rate) can provide computational efficiency, but it may not reflect the ethical criteria that we truly care about. Recent work has shown that models trained with a surrogate loss may exhibit "approximation errors" that disproportionately affect undersampled groups in the training data. [200] Similarly, one might choose to optimize the worst-case error across groups as opposed to the average overall error. Such choices may seem purely technical, but reflect value statements about what should be optimized, potentially leading to differences in performance among marginalized groups. [262]

In this section, we review several crucial factors in model development that potentially impact ethical deployment capacity: understanding (and accounting for) confounding, feature selection, tuning parameters, and defining "fairness" itself.

2.4.1 Understanding Confounding

Developing models that use sensitive attributes without a clear causal understanding of their relationship to outcomes of interest can significantly affect model performance and interpretation. This is relevant to algorithmic problems focused on prediction, not just causal inference. Independent variables are defined as variables whose variation does not depend on that of another whereas dependent variables are variables whose value may depend on that of another. "Confounding" features — i.e., those features that influence both the independent variables and the dependent variable — require careful attention. The vast majority of models learn patterns based on observed correlations between training data, even when such correlations do not occur in test data. For instance, recent work has demonstrated that classification models designed to detect hair color learn gender-biased decision boundaries when trained on confounded data, i.e., if women are primarily blond in training data, the model

incorrectly associates gender with the hair label in test samples. [160]

As ML methods are increasingly used for clinical decision support, it is critical to account for confounding features. In one canonical example, asthmatic patients presenting with pneumonia are given aggressive interventions that ultimately improve their chances of survival over non-asthmatic patients. [47] When the hospital protocol assigned additional treatment to patients with asthma, those patients had improved outcomes. Thus the treatment policy was a confounding factor that altered the data in a seemingly straight-forward prediction task such that patients with asthma were erroneously predicted by models to have *lower* risk of dying from pneumonia.

Simply controlling for confounding features by including them as features in classification or regression models may be insufficient to learn reliable models because features can have a mediating or moderating effect (post-treatment effect on outcomes of interest) and have to be incorporated differently into model design. [141]

Modern ML and causal discovery techniques can identify sources of confounding at scale [125], although validation of such methods can be challenging because of the lack of interventional data. ML methods have also been proposed to estimate causal effects from observational data. [294, 60] In practice, when potential hidden confounding is suspected, either mediating features or proxies can be leveraged [214, 141] or sensitivity analysis methods can be used to determine potential sources of errors in effect estimates. [108] Data-augmentation and sampling methods may also be used to mitigate effects of model confounding. For example, augmenting X-ray images with rotated and translated variants can help train a model that is not sensitive to orientation of an image. [194]

2.4.2 Feature Selection

With large-scale digitization of EHR and other sources, sensitive attributes like race and ethnicity may be increasingly available (although prone to misclassification and missingness). However, blindly incorporating factors like race and gender in a predictive model may exacerbate inequities for a wide range of diagnostics and treatments. [300] These resulting inequities can lead to unintended and permanent em-

bedding of biases in algorithms used for clinical care. For example, vaginal birth after cesarean (VBAC) scores are used to predict success of "trial of labor" of pregnant women with a prior cesarean section; however, these scores explicitly include a race component as an input which reduces the chance of VBAC success for Black and Hispanic women. Although researchers found that previous observational studies showed correlation between racial identity and success of trial of labor [129], the underlying cause of this association is not well-understood. Such naïve inclusion of race information could exacerbate disparities in maternal mortality. This ambiguity calls race-based 'correction' in scores like VBAC into question. [300]

Automation in feature selection does not eliminate the need for contextual understanding. For example, stepwise regression is commonly used and taught as a technique for feature selection despite known limitations. [278] While specific methods have varying initialization (e.g., start with an empty set of features or full set of features) and processing steps (e.g., deletion vs. addition of features), most rely on p-values, R^2, or other global fit metrics to select features. Weaknesses of stepwise regressions include the misleading nature of p-values and the final set depending on if and when features were considered. [137] In ML, penalized regressions like lasso regression are popular for automated feature selection, but the lasso trades potential increases in estimation bias for reductions in variance by shrinking some feature coefficients to zero. Features selected by lasso may be co-linear with other features not selected. [155] Over-interpretation of the selected features in any automated procedures should therefore be avoided in practice given these pitfalls. Researchers should also consider the humans-in-the-loop framework where incorporation of automated procedures is blended with investigator knowledge. [179]

2.4.3 Tuning Parameters

There are many tuning parameters that may be set a priori or selected via cross-validation. [155] These range from the learning rate in a neural network to the minimum size of the terminal leaves in a random forest. In the latter example, default settings in R for classification will allow trees to grow until there is just one observa-

tion in a terminal leaf. This can lead to overfitting the model to the training data and a loss of generalizability to the target population. Lack of generalizability is a central concern for ethical ML given the previously discussed issues in data collection and study inclusion. When data lack diversity and are not representative of the target population where the model would be deployed, overfitting algorithms to this data has the potential to disproportionately harm marginalized groups. [263] Using cross-validation to select tuning parameters does not automatically solve these problems as cross-validation still operates with respect to an a priori chosen optimization target.

2.4.4 Performance Metrics

There are many commonly used performance metrics for model evaluation such as area under the receiver-operating characteristic (AUC), precision-recall curves (AUPRC), and calibration. [103] However, the appropriate metrics to optimize depend on intended use case and relative value of true positives, false positives, true negatives, and false negatives. Not only can AUC be misleading when considering other global fit metrics (e.g., high AUC masking weak true positive rate), but it does not describe the impact of the model across selected groups. Further, even "objective" metrics and scores can be deeply flawed, and lead to over or under-treatment of minorities if blindly applied. [299] Note that robust reporting of results should include an explicit statement of other non-optimized metrics, including the original intended use case, the training cohort and case, or level of model uncertainty.

2.4.5 Group Fairness Definition

The specific definition of fairness for a given application often impacts the choice of a loss function, and therefore the underlying algorithm. Individual fairness imposes classifier performance requirements that operate over pairs of individuals, e.g., similar individuals should be treated similarly. [88] Group fairness operates over "protected groups" (based on some sensitive attribute) by requiring that a classifier performance metric be balanced across those groups. [89, 64] For instance, a model may be par-

tially assessed by calculating the true positive rate separately among rural and urban populations to ensure risk score similarity. Regressions subject to group fairness constraints or penalties optimizing toward joint global and group fit considerations have also been developed. [40, 315, 319]

Recent work has focused on identifying and mitigating violations of fairness definitions in health care settings. While most of these algorithms have emerged outside the field of health care, researchers have designed penalized and constrained regressions to improve the performance of health insurance plan payment. This payment system impacts tens of millions of lives in the United States and is known to undercompensate insurers for individuals with certain health conditions, including mental health and substance use disorders, in part because billing codes do not accurately capture diagnoses. [218] Undercompensation creates incentives for insurers to exclude individuals with these health conditions from enrollment, limiting their access to care. Regressions subject to group fairness constraints or penalties were successful in removing nearly all undercompensation for a single group with negligible impacts on global fit. [319] Subsequent work incorporating multiple groups into the loss function also saw improvements in undercompensation for the majority of groups not included. [213]

2.5 Post-Deployment Consideration

Often the goal of model training is to ultimately deploy it in a clinical, epidemiological, or policy service. However, deployed models can have lasting ethical impact beyond the model performance measured in development. For example, in the inclusion of race in the clinical risk scores described earlier that may lead to chronic over- or undertreatment. [300] Here we outline considerations for robust deployment by highlighting the need for careful performance reporting, auditing generalizability, documentation, and regulation.

2.5.1 Quantifying Impact

Unlike in other settings with high-stakes decisions, e.g., aviation, clinical staff performance is not audited by an external body. [140] Instead, clinicians are often a self-governing body, relying on clinicians themselves to determine when a colleague is underperforming or in breach of ethical practice principles, e.g., through such tools as surgical morbidity and mortality conferences. [220] Clinical staff can also struggle to keep abreast of what current best practice recommendations are, as these can change dramatically over time; one study found that more than 400 previously routine practices were later contradicted in leading clinical journals. [142]

Hence, it is important to measure and address the downstream impact of models though audits for bias and examination of clinical impact. [58] Regular "auditing" post-deployment, i.e., detailed inspection of model performance on various groups and outcomes, may reveal the impact of models on different populations [227] and identify areas of potential concern. Some recent work has targeted causal models in dynamic systems in order to reduce the severity of bias. [72] Others have targeted bias reduction through model construction with explicit guarantees about balanced performance [293], or specifying groups which must have equal performance. [224] Additionally, there is the possibility that models may help to de-bias current clinical care by reducing known biases against minorities [285] and disadvantaged majorities. [236]

2.5.2 Model Generalizability

As has been raised in previous sections, a crucial concern with model deployment is generalization. Any shifts in data distributions can significantly impact model performance when the settings for development and for deployment differ. For example, chest X-ray diagnosis models can have high performance on test data drawn from the same hospital but degrade rapidly on data from another hospital. [316] Other work in gender bias on chest X-ray data has demonstrated both that small proportions of female chest x-rays degrades diagnostic performance accuracy in female patients [185], and that this is not simply addressed in all cases by adding in more female

X-rays. [268] Even within a single hospital, models trained on data from an initial EHR system data deteriorated significantly when tested on data from a new EHR system. [223] Finally, data artifacts that induce strong priors in what patterns ML models are sensitive to have the potential to perpetrate harms when used without awareness. [28] For example, patients with dark skin can have morphological variation and disease manifestations that are not easily detected under the defaults that are set by predominantly white-skinned patients. [181]

Several algorithms have recently been proposed to account for distribution shift in data. [273] However, these algorithms have significant limitations as they typically require assumptions about the nature or amount of distributional shift an algorithm can accommodate. Some, like [273], may require a clear indication of which distributions in a health care pipeline are expected to change, and develop models for prediction accordingly. Many of these assumptions may be verifiable. If not, periodically monitoring for data shifts [77], and potentially retraining models when performance deteriorates due to such shifts is an imperative deployment consideration with significant ethical implications.

2.5.3 Model and Data Documentation

Clear documentation enables insight into the model development and data collection. Good model documentation should include clinically specific features of model development that can be assessed and recorded beforehand, such as logistics within the clinical setting, potential unintended consequences, and trade-offs between bias and performance. [264] In addition to raising ethical concerns in the pipeline, the process of co-designing “checklists" with clinical practitioners formalizes ad-hoc procedures and empowers individual advocates. [204] Standardized reporting of model performance—such as the one-page summary “model cards” for model reporting [217]—can empower clinical practitioners to understand model limitations and future model developers to identify areas of improvement. Similarly, better documentation of the data supporting initial model training can help expose sources of discrimination in the collected data. Modelers could use “datasheets” for datasets to detail the conditions of data

collection. [115]

2.5.4 Regulation

In the United States, the Food and Drug Administration (FDA) has responsibility for the regulation of health care ML models. As there does not exist comprehensive guidance for health care model research and subsequent deployment, the opportunity is ripe to create a comprehensive framework to audit and regulate models. Currently, the FDA's proposed ML-specific modifications to the software as a medical device (SaMD) regulations draw a distinction between models that are trained and then frozen prior to clinical deployment and models that continue to learn on observed outcomes. Although models in the latter class can leverage larger, updated datasets, they also face additional risk due to model drift and may need additional audits. [106] Such frameworks should explicitly account for health disparities across the stages of ML development in health, and ensure health equity audits as part of postmarket evaluation. [98] We also note that there are many potential legal implications, e.g., in malpractice and liability suits, that will require new solutions. [274]

Researchers have proposed additional frameworks to guide clinical model development, which could inspire future regulation. ML model regulation could draw from existing regulatory frameworks: a randomized controlled trial for ML models would assess patient benefit compared to a control cohort of standard clinical practice [198], and a drug development pipeline for ML models would define a protocol for adverse events and model recalls. [68] The clinical interventions accompanying the clinical ML model should be analyzed to contextualize the use of the model in the clinical setting. [232]

Chapter 3

Mitigating Biased Machine Learning Methods

3.1 Introduction

As machine learning algorithms increasingly affect decision making in society, many have raised concerns about the fairness and biases of these algorithms, especially in applications to healthcare or criminal justice, where human lives are at stake. [13, 21] It is often hoped that the use of automatic decision support systems trained on observational data will remove human bias and improve accuracy. However, factors such as data quality and model choice may encode unintentional discrimination, resulting in systematic disparate impact.

We study fairness in prediction of outcomes such as recidivism, annual income, or patient mortality. Fairness is evaluated with respect to *protected groups* of individuals defined by attributes such as gender or ethnicity. [258] Following previous work, we measure discrimination in terms of differences in prediction cost across protected groups. [41, 88, 96] Correcting for issues of data provenance and historical bias in labels is outside of the scope of this work. Much research has been devoted to constraining models to satisfy cost-based fairness in prediction, as we expand on below. The impact of data collection on discrimination has received comparatively little attention.

Fairness in prediction has been encouraged by adjusting models through regularization [24, 164], constraints [163, 314], and representation learning. [317] These attempts can be broadly categorized as model-based approaches to fairness. Others have applied data preprocessing to reduce discrimination [132, 96, 43], with empirical examples. [109] Inevitably, however, restricting the model class or perturbing training data to improve fairness may harm predictive accuracy. [69]

A *tradeoff* of predictive accuracy for fairness is sometimes difficult to motivate when predictions influence high-stakes decisions. In particular, post-hoc correction methods based on randomizing predictions [136, 245] are unjustifiable for ethical reasons in clinical tasks such as severity scoring. Moreover, as pointed out by [308], post-hoc correction may lead to suboptimal predictive accuracy compared to other equally fair classifiers.

Disparate predictive accuracy can often be explained by insufficient or skewed sample sizes or inherent unpredictability of the outcome given the available set of variables. With this in mind, we propose that fairness of predictive models should be analyzed in terms of model bias, model variance, and outcome noise *before* they are constrained to satisfy fairness criteria. This exposes and separates the adverse impact of inadequate data collection and the choice of the model on fairness. The cost of fairness need not always be one of predictive accuracy, but one of investment in data collection and model development. In high-stakes applications, the benefits often outweigh the costs.

In this work, we use the term "discrimination" to refer to specific kinds of differences in the predictive power of models when applied to different protected groups. In some domains, such differences may not be considered discriminatory, and it is critical that decisions made based on this information are sensitive to this fact. For example, in prior work, researchers showed that causal inference may help uncover which sources of differences in predictive accuracy introduce unfairness. [183] In this work, we assume that observed differences are considered discriminatory and discuss various means of explaining and reducing them.

3.1.1 Contributions

Our work was done in collaboration with Fredrik D. Johannson and David Sontag. We give a procedure for analyzing discrimination in predictive models with respect to cost-based definitions of group fairness, emphasizing the impact of data collection. First, we propose the use of bias-variance-noise decompositions for separating sources of discrimination. Second, we suggest procedures for estimating the value of collecting additional training samples. Finally, we propose the use of clustering for identifying subpopulations that are discriminated against to guide additional variable collection. We use these tools to analyze the fairness of common learning algorithms in three tasks: predicting income based on census data, predicting mortality of patients in critical care, and predicting book review ratings from text. We find that the accuracy in predictions of the mortality of cancer patients vary by as much as 20% between protected groups. In addition, our experiments confirm that discrimination level is sensitive to the quality of the training data.

3.1.2 Related Work

We study fairness in prediction of an outcome $Y \in \mathcal{Y}$. Predictions are based on a set of covariates $X \in \mathcal{X} \subseteq \mathbb{R}^k$ and a *protected attribute* $A \in \mathcal{A}$. In mortality prediction, X represents the medical history of a patient in critical care, A the self-reported ethnicity, and Y mortality. A model is considered fair if its errors are distributed similarly across protected groups, as measured by a cost function γ. Predictions learned from a training set d are denoted $\hat{Y}_d := h(X, A)$ for some $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ from a class $\mathcal{H}$. The protected attribute is assumed to be binary, $\mathcal{A} = \{0, 1\}$, but our results generalize to the non-binary case. A dataset $d = \{(x_i, a_i, y_i)\}_{i=1}^n$ consists of n samples distributed according to $p(X, A, Y)$. When clear from context, we drop the subscript from $\hat{Y}_d$.

A popular cost-based definition of fairness is the *equalized odds* criterion, which states that a binary classifier $\hat{Y}$ is fair if its false negative rates (FNR) and false positive rates (FPR) are equal across groups. [136] We define FPR and FNR with

respect to protected group $a \in \mathcal{A}$ by

$$\begin{aligned} \mathrm{FPR}_a(\hat{Y}) &\coloneqq \mathbb{E}_X[\hat{Y} \mid Y = 0, A = a], \\ \mathrm{FNR}_a(\hat{Y}) &\coloneqq \mathbb{E}_X[1 - \hat{Y} \mid Y = 1, A = a] \ . \end{aligned}$$

Exact equality, $\mathrm{FPR}_0(\hat{Y}) = \mathrm{FPR}_1(\hat{Y})$, is often hard to verify or enforce in practice. Instead, we study the *degree* to which such constraints are violated. More generally, we use differences in *cost functions* γ_a between protected groups $a \in \mathcal{A}$ to define the *level of discrimination* Γ,

$$\Gamma^{\gamma}(\hat{Y}) := \left| \gamma_0(\hat{Y}) - \gamma_1(\hat{Y}) \right| \ . \tag{3.1}$$

In this work we study cost functions $\gamma_a \in \{\mathrm{FPR}_a, \mathrm{FNR}_a, \mathrm{ZO}_a\}$ in binary classification tasks, with $\mathrm{ZO}_a(\hat{Y}) := \mathbb{E}_X[\mathbb{1}[\hat{Y} \neq Y] \mid A = a]$ the *zero-one loss*. In regression problems, we use the group-specific *mean-squared error* $\mathrm{MSE}_a := \mathbb{E}_X[(\hat{Y} - Y)^2 \mid A = a]$. According to (3.1), predictions $\hat{Y}$ satisfy equalized odds on d if $\Gamma^{\mathrm{FPR}}(\hat{Y}) = 0$ *and* $\Gamma^{\mathrm{FNR}}(\hat{Y}) = 0$.

Calibration and impossibility A score-based classifier is *calibrated* if the prediction score assigned to a unit equals the fraction of positive outcomes for all units assigned similar scores. It is impossible for a classifier to be calibrated in every protected group and satisfy multiple cost-based fairness criteria at once, unless accuracy is perfect or base rates of outcomes are equal across groups. [62] A relaxed version of this result [176] applies to the discrimination level Γ. Inevitably, both constraint-based methods and our approach are faced with a choice between which fairness criteria to satisfy, and at what cost.

3.2 Decomposing Discrimination

3.2.1 Sources of Discrimination

There are many potential sources of discrimination in predictive models. In particular, the choice of hypothesis class $\mathcal{H}$ and learning objective has received a lot of attention. [41, 317, 100] However, data collection—the chosen set of predictive variables X, the sampling distribution $p(X, A, Y)$, and the training set size n—is an equally integral part of deploying fair machine learning systems in practice, and it should be guided to promote fairness. Below, we tease apart sources of discrimination through bias-variance-noise decompositions of cost-based fairness criteria. In general, we may think of noise in the outcome as the effect of a set of unobserved variables U, potentially interacting with X. Even the optimal achievable error for predictions based on X may be reduced further by observing parts of U. In Figure 3-1, we illustrate three common learning scenarios and study their fairness properties through bias, variance, and noise.

To account for randomness in the sampling of training sets, we redefine discrimination level (3.1) in terms of the *expected* cost $\overline{\gamma}_a(\hat{Y}) := \mathbb{E}_D[\gamma_a(\hat{Y}_D)]$ over draws of a random training set D.

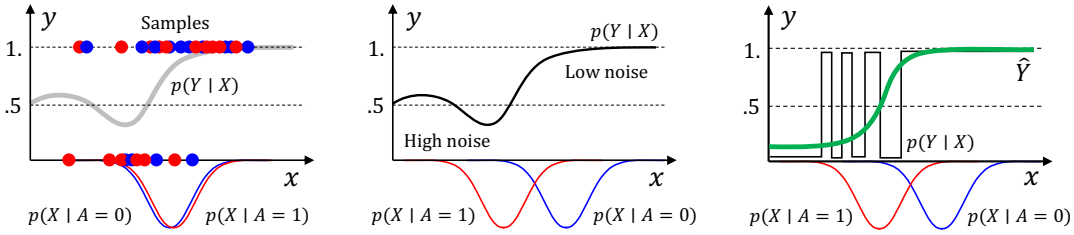
Definition 1. The *expected discrimination level* $\overline{\Gamma}(\hat{Y})$ of a predictive model $\hat{Y}$ learned from a random training set D, is

$$\overline{\Gamma}(\hat{Y}) := \left|\mathbb{E}_D\left[\gamma_0(\hat{Y}_D) - \gamma_1(\hat{Y}_D)\right]\right| = \left|\overline{\gamma}_0(\hat{Y}) - \overline{\gamma}_1(\hat{Y})\right| .$$

$\overline{\Gamma}(\hat{Y})$ is not observed in practice when only a single training set d is available. If n is small, it is recommended to estimate $\overline{\Gamma}$ through re-sampling methods such as bootstrapping. [90]

3.2.2 Bias-Variance-Noise Decompositions of Discrimination

An algorithm that learns models $\hat{Y}_D$ from datasets D is given, and the covariates X and size of the training data n are fixed. We assume that $\hat{Y}_D$ is a deterministic function



(a) For identically distributed protected groups and unaware outcome (see below), bias and noise are equal in expectation. Perceived discrimination is only due to variance.

(b) Heteroskedastic noise, i.e. $\exists x, x' : N(x) \neq N(x')$, may contribute to discrimination even for an optimal model if protected groups are not identically distributed.

(c) One choice of model may be more suited for one protected group, even under negligible noise and variance, resulting in a difference in expected bias, $\overline{B}_0 \neq \overline{B}_1$.

Figure 3-1: Scenarios illustrating how properties of the training set and model choice affect perceived discrimination in a binary classification task, under the assumption that outcomes and predictions are *unaware*, i.e. $p(Y \mid X, A) = p(Y \mid X)$ and $p(\hat{Y} \mid X, A) = p(\hat{Y} \mid X)$. Through bias-variance-noise decompositions (see Section 3.2.2), we can identify which of these dominate in their effect on fairness. We propose procedures for addressing each component in Section 3.3.1, and use them in experiments (see Section 3.4) to mitigate discrimination in income prediction and prediction of ICU mortality.

$\hat{y}_D(x, a)$ given the training set D, e.g., a thresholded scoring function. Following [83], we base our analysis on decompositions of loss functions L evaluated at points (x, a). For decompositions of costs $\gamma_a \in \{\text{ZO}, \text{FPR}, \text{FNR}\}$ we let this be the zero-one loss, $L(y, y') = \mathbb{1}[y \neq y']$, and for $\gamma_a = \text{MSE}$, the squared loss, $L(y, y') = (y - y')^2$. We define the *main prediction* $\tilde{y}(x, a) = \arg\min_{y'} \mathbb{E}_D[L(\hat{Y}_D, y') \mid X = x, A = a]$ as the average prediction over draws of training sets for the squared loss, and the majority vote for the zero-one loss. The *(Bayes) optimal prediction* $y^*(x, a) = \arg\min_{y'} \mathbb{E}_Y[L(Y, y') \mid X = x, A = a]$ achieves the smallest expected error with respect to the random outcome Y.

Definition 2 (Bias, variance and noise)**.** Following [83], we define bias B, variance

V and noise N at a point (x, a) below.

$$B(\hat{Y}, x, a) = L(y^*(x, a), \tilde{y}(x, a)) \qquad N(x, a) = \mathbb{E}_Y[L(y^*(x, a), Y) \mid X = x, A = a]$$
$$V(\hat{Y}, x, a) = \mathbb{E}_D[L(\tilde{y}(x, a), \hat{y}_D(x, a))] \ . \tag{3.2}$$

Here, $y^*, \hat{y}$ and $\tilde{y}$, are all deterministic functions of (x, a), while Y is a random variable.

In words, the bias B is the loss incurred by the main prediction relative to the optimal prediction. The variance V is the average loss incurred by the predictions learned from different datasets relative to the main prediction. The noise N is the remaining loss independent of the learning algorithm, often known as the Bayes error. We use these definitions to decompose $\overline{\Gamma}$ under various definitions of γ_a.

Theorem 1. *With $\overline{\gamma}_a$ the group-specific zero-one loss or class-conditional versions (e.g., FNR, FPR), or the mean squared error, $\overline{\gamma}_a$ and the discrimination level $\overline{\Gamma}$ admit decompositions of the form*

$$\overline{\gamma}_a(\hat{Y}) = \underbrace{\overline{N}_a}_{Noise} + \underbrace{\overline{B}_a(\hat{Y})}_{Bias} + \underbrace{\overline{V}_a(\hat{Y})}_{Variance} \qquad \text{and} \qquad \overline{\Gamma} = \left|(\overline{N}_0 - \overline{N}_1) + (\overline{B}_0 - \overline{B}_1) + (\overline{V}_0 - \overline{V}_1)\right|$$

where we leave out $\hat{Y}$ in the decomposition of $\overline{\Gamma}$ for brevity. With B, V defined as in (3.2)*, we have*

$$\overline{B}_a(\hat{Y}) = \mathbb{E}_X[B(\tilde{y}, X, a) \mid A = a] \quad \text{and} \quad \overline{V}_a(\hat{Y}) = \mathbb{E}_{X,D}[c_v(X)V(\hat{Y}_D, X, a) \mid A = a] \ .$$

For the zero-one loss, $c_v(x, a) = 1$ if $\hat{y}_m(x, a) = y^*(x, a)$, otherwise $c_v(x, a) = -1$. For the squared loss $c_v(x, a) = 1$. The noise term for population losses is

$$\overline{N}_a := \mathbb{E}_X[c_n(X, a)L(y^*(X, a), Y) \mid A = a]$$

and for class-conditional losses w.r.t class $y \in \{0, 1\}$,

$$\overline{N}_a(y) := \mathbb{E}_X[c_n(X, a)L(y^*(X, a), y) \mid A = a, Y = y] \ .$$

For the zero-one loss, and class-conditional variants, $c_n(x, a) = 2\mathbb{E}_D[\mathbb{1}[\hat{y}_D(x, a) = y^*(x, a)]] - 1$ *and for the squared loss,* $c_n(x, a) = 1$.

Proof sketch. Conditioning and exchanging order of expectation, the cases of mean squared error and zero-one losses follow. [83] Class-conditional losses follow from a case-by-case analysis of possible errors. See the supplementary material in Chapter A.5 for a full proof. □

Theorem 1 points to distinct sources of perceived discrimination. Significant differences in bias $\overline{B}_0 - \overline{B}_1$ indicate that the chosen model class is not flexible enough to fit both protected groups well (see Figure 3-1c). This is typical of (misspecified) linear models which approximate non-linear functions well only in small regions of the input space. Regularization or post-hoc correction of models effectively increase the bias of one of the groups, and should be considered only if there is reason to believe that the original bias is already minimal.

Differences in variance, $\overline{V}_0 - \overline{V}_1$, could be caused by differences in sample sizes n_0, n_1 or group-conditional feature variance $\mathrm{Var}(X \mid A)$, combined with a high capacity model. Targeted collection of training samples may help resolve this issue. Our decomposition does not apply to post-hoc randomization methods [136] but we may treat these in the same way as we do random training sets and interpret them as increasing the variance $\overline{V}_a$ of one group to improve fairness.

When noise is significantly different between protected groups, discrimination is partially unrelated to model choice and training set size and may only be reduced by measuring additional variables.

Proposition 1. *If* $\overline{N}_0 \neq \overline{N}_1$*, no model can be 0-discriminatory in expectation without access to additional information or increasing bias or variance w.r.t. to the Bayes optimal classifier.*

Proof. By definition, $\overline{\Gamma} = 0 \implies (\overline{N}_1 - \overline{N}_0) = (\overline{B}_0 - \overline{B}_1) + (\overline{V}_0 - \overline{V}_1)$. As the Bayes optimal classifier has neither bias nor variance, the result follows immediately. □

In line with Proposition 1, most methods for ensuring algorithmic fairness reduce discrimination by trading off a difference in noise for one in bias or variance. However, this trade-off is only motivated if the considered predictive model is close to Bayes optimal *and* no additional predictive variables may be measured. Moreover, if noise is homoskedastic in regression settings, post-hoc randomization is ill-advised, as the difference in Bayes error $\overline{N}_0 - \overline{N}_1$ is zero, and discrimination is caused only by model bias or variance (see the supplementary material for a proof).

Estimating bias, variance and noise Group-specific variance $\overline{V}_a$ may be estimated through sample splitting or bootstrapping. [90] In contrast, the noise $\overline{N}_a$ and bias $\overline{B}_a$ are difficult to estimate when X is high-dimensional or continuous. In fact, no convergence results of noise estimates may be obtained without further assumptions on the data distribution. [14] Under some such assumptions, noise may be approximately estimated using distance-based methods [79], nearest-neighbor methods [110, 71], or classifier ensembles. [289] When comparing the discrimination level of two different models, noise terms cancel, as they are independent of the model. As a result, *differences* in bias may be estimated even when the noise is not known (see the supplementary material).

Testing for significant discrimination When sample sizes are small, perceived discrimination may not be statistically significant. In the supplementary material, we give statistical tests both for the discrimination level $\Gamma(\hat{Y})$ and the difference in discrimination level between two models $\hat{Y}, \hat{Y}'$.

3.3 Methods of Mitigating Algorithmic Discrimination

3.3.1 Reducing Discrimination through Data Collection

In light of the decomposition of Theorem 1, we explore avenues for reducing group differences in bias, variance, and noise without sacrificing predictive accuracy. In practice, predictive accuracy is often artificially limited when data is expensive or impractical to collect. With an investment in training samples or measurement of predictive variables, both accuracy and fairness may be improved.

3.3.2 Increasing Training Set Size

Standard regularization used to avoid overfitting is not guaranteed to improve or preserve fairness. An alternative route is to collect more training samples and reduce the impact of the bias-variance trade-off. When supplementary data is collected from the same distribution as the existing set, covariate shift may be avoided. [249] This is often achievable; labeled data may be expensive, such as when paying experts to label observations, but given the means to acquire additional labels, they would be drawn from the original distribution. To estimate the value of increasing sample size, we predict the discrimination level $\overline{\Gamma}(\hat{Y}_D)$ as D increases in size.

The curve measuring generalization performance of predictive models as a function of training set size n is called a Type II *learning curve*. [82] We call $\overline{\gamma}_a(\hat{Y}, n) := \mathbb{E}[\gamma_a(\hat{Y}_{D_n})]$, as a function of n, the learning curve with respect to protected group a. We define the discrimination learning curve $\overline{\Gamma}(\hat{Y}, n) \coloneqq |\overline{\gamma}_0(\hat{Y}, n) - \overline{\gamma}_1(\hat{Y}, n)|$ (see Figure 3-2 for an example). Empirically, learning curves behave asymptotically as *inverse power-law* curves for diverse algorithms such as deep neural networks, support vector machines, and nearest-neighbor classifiers, even when model capacity is allowed to grow with n. [143, 219] This observation is also supported by theoretical results. [11]

Assumption 1 (Learning curves)**.** *The population prediction loss $\overline{\gamma}(\hat{Y}, n)$, and group-specific losses $\overline{\gamma}_0(\hat{Y}, n), \overline{\gamma}_1(\hat{Y}, n)$, for a fixed learning algorithm $\hat{Y}$, behave asymptoti-*

cally as inverse power-law curves with parameters (α, β, δ)*. That is,* $\exists M, M_0, M_1$ *such that for* $n \geq M, n_a \geq M_a$*,*

$$\overline{\gamma}(\hat{Y}, n) = \alpha n^{-\beta} + \delta \quad \textit{and} \quad \forall a \in \mathcal{A} : \overline{\gamma}_a(\hat{Y}, n_a) = \alpha_a n_a^{-\beta_a} + \delta_a \tag{3.3}$$

Intercepts, δ, δ_a in (3.3) represent the asymptotic bias $\overline{B}(\hat{Y}_{D_\infty})$ and the Bayes error $\overline{N}$, with the former vanishing for consistent estimators. Accurately estimating δ from finite samples is often challenging as the first term tends to dominate the learning curve for practical sample sizes.

In experiments, we find that the inverse power-laws model fit group conditional (γ_a) and class-conditional (FPR, FNR) errors well, and use these to extrapolate $\overline{\Gamma}(\hat{Y}, n)$ based on estimates from subsampled data.

3.3.3 Measuring Additional Variables

When discrimination $\overline{\Gamma}$ is dominated by a difference in noise, $\overline{N}_0 - \overline{N}_1$, fairness may not be improved through model selection alone without sacrificing accuracy (see Proposition 1). Such a scenario is likely when available covariates are not equally predictive of the outcome in both groups. We propose identification of clusters of individuals in which discrimination is high as a means to guide further variable collection—if the variance in outcomes within a cluster is not explained by the available feature set, additional variables may be used to further distinguish its members.

Let a random variable C represent a (possibly stochastic) clustering such that $C = c$ indicates membership in cluster c. Then let $\rho_a(c)$ denote the expected prediction cost for units in cluster c with protected attribute a. As an example, for the zero-one loss we let

$$\rho_a^{\text{ZO}}(c) := \mathbb{E}_X[\mathbb{1}[\hat{Y} \neq Y] \mid A = a, C = c],$$

and define ρ analogously for false positives or false negatives. Clusters c for which $|\rho_0(c) - \rho_1(c)|$ is large identify groups of individuals for which discrimination is worse than average, and can guide targeted collection of additional variables or samples. In

our experiments on income prediction, we consider particularly simple clusterings of data defined by subjects with measurements above or below the average value of a single feature $x(c)$ with $c \in \{1, \ldots, k\}$. In mortality prediction, we cluster patients using topic modeling. As measuring additional variables is expensive, the utility of a candidate set should be estimated before collecting a large sample. [178]

3.4 Experiments

We analyze the fairness properties of standard machine learning algorithms in three tasks: prediction of income based on national census data, prediction of patient mortality based on clinical notes, and prediction of book review ratings based on review text. A synthetic experiment validating group-specific learning curves is left to Appendix 3.4.3. We disentangle sources of discrimination by assessing the level of discrimination for the full data, estimating the value of increasing training set size by fitting Type II learning curves, and using clustering to identify subgroups where discrimination is high. In addition, we estimate the Bayes error through non-parametric techniques.

In our experiments, we omit the sensitive attribute A from our classifiers to allow for closer comparison to previous works, e.g., [136, 314]. In preliminary results, we found that fitting separate classifiers for each group increased the error rates of both groups due to the resulting smaller sample size, as classifiers could not learn from other groups. As our model objective is to maximize accuracy over all data points, our analysis uses a single classifier trained on the entire population.

3.4.1 Income Prediction

Predictions of a person's salary may be used to help determine an individual's market worth, but systematic underestimation of the salary of protected groups could harm their competitiveness on the job market. The Adult dataset in the UCI Machine Learning Repository [191] contains 32,561 observations of yearly income (represented as a binary outcome: over or under \$50,000) and twelve categorical or continuous

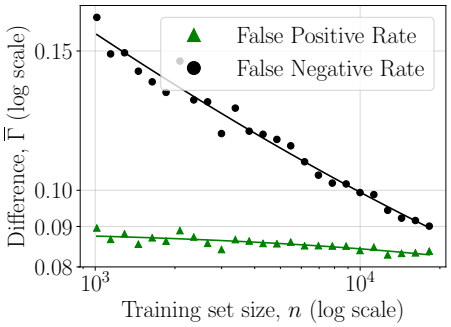


Figure 3-2: Noise level estimation in income prediction with the Adult dataset. Group differences in false positive rates and false negative rates for a random forest classifier decrease with increasing training set size.

Method	E_{low}	E_{up}	group
Mahalanobis [205]	–	0.29	men
	–	0.13	women
Bhattacharyya [27]	0.001	0.040	men
	0.001	0.027	women
Nearest Neighbors [71]	0.10	0.19	men
	0.04	0.07	women

Table 3.1: Discrimination level estimation in income prediction with the Adult dataset. Estimation of Bayes error lower and upper bounds (E_{low} and E_{up}) for zero-one loss of men and women. Intervals for men and women are non-overlapping for Nearest Neighbors.

features including education, age, and marital status. Categorical attributes are dichotomized, resulting in a total of 105 features.

We follow [245] and strive to ensure fairness across genders, which is excluded as a feature from the predictive models. Using an 80/20 train-test split, we learn a random forest predictor, which is well-calibrated for both groups ([35] scores of 0.13 and 0.06 for men and women). We find the difference in zero-one loss $\Gamma^{\text{ZO}}(\hat{Y})$ has a 95%-confidence interval[1] $.085 \pm .069$ with decision thresholds at 0.5. At this threshold, the false negative rates are 0.388 ± 0.026 and 0.448 ± 0.064 for men and women respectively, and the false positive rates 0.111 ± 0.011 and 0.033 ± 0.008. We focus on random forest classifiers, although we found similar results for logistic

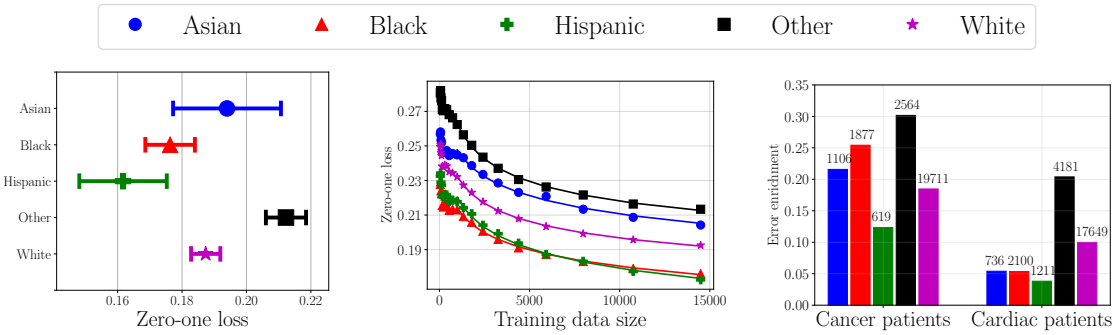
[1]Details for computing statistically significant discrimination can be found in the supplementary material.

regression and decision trees.

We examine the effect of varying training set size n on discrimination. We fit inverse power-law curves to estimates of $\text{FPR}(\hat{Y}, n)$ and $\text{FNR}(\hat{Y}, n)$ using repeated sample splitting where at least 20% of the full data is held out for evaluating generalization error at every value of n. We tune hyperparameters for each training set size for decision tree classifiers and logistic regression but tuned over the entire dataset for random forest. We include full training details in the supplementary material. Metrics are averaged over 50 trials. See Figure 3-2 for the results for random forests. Both FPR and FNR decrease with additional training samples. The discrimination level Γ^{FNR} for false negatives decreases by a striking 40% when increasing the training set size from 1000 to 10,000. This suggests that trading off accuracy for fairness at small sample sizes may be ill-advised. Based on fitted power-law curves, we estimate that for unlimited training data drawn from the same distribution, we would have $\Gamma^{\text{FNR}}(\hat{Y}) \approx 0.04$ and $\Gamma^{\text{FPR}}(\hat{Y}) \approx 0.08$.

In Figure 3.1, we compare estimated upper and lower bounds on noise (E_{low} and E_{up}) for men and women using the Mahalanobis and Bhattacharyya distances [79], and a k-nearest neighbor method [71] with $k = 5$ and 5-fold cross validation. Men have consistently higher noise estimates than women, which is consistent with the differences in zero-one loss found using all models. For nearest neighbors estimates, intervals for men and women are non-overlapping, which suggests that noise may contribute substantially to discrimination.

To guide attempts at reducing discrimination further, we identify clusters of individuals for whom false negative predictions are made at different rates between protected groups, with the method described in Section 3.3.3. We find that for individuals in executive or managerial occupations (12% of the sample), false negatives are more than twice as frequent for women (0.412) as for men (0.157). For individuals in all other occupations, the difference is significantly smaller, 0.543 for women and 0.461 for men, despite the fact that the disparity in outcome base rates in this cluster is large (0.26 for men versus 0.09 for women). A possible reason is that in managerial occupations the available variable set explains a larger portion of the variance



(a) Using Tukey's range test, we present the 95%-significance level for the zero-one loss for each group over 5-fold cross validation.

(b) As training set size increases, zero-one loss over 50 trials decreases over all groups and appears to converge to an asymptote.

(c) Topic modeling reveals subpopulations with high differences in zero-one loss, for example cancer patients and cardiac patients.

Figure 3-3: Mortality prediction from clinical notes using logistic regression. Best viewed in color.

in salary for men than for women. If so, further sub-categorization of managerial occupations could help reduce discrimination in prediction.

3.4.2 Intensive Care Unit Mortality Prediction

Unstructured medical data such as clinical notes can reveal insights for questions like mortality prediction; however, disparities in predictive accuracy may result in discrimination of protected groups. Using the MIMIC-III dataset of all clinical notes from 25,879 adult patients from Beth Israel Deaconess Medical Center [158], we predict hospital mortality of patients in critical care. Fairness is studied with respect to five self-reported ethnic groups of the following proportions: Asian (2.2%), Black (8.8%), Hispanic (3.4%), White (70.8%), and Other (14.8%). Notes were collected in the first 48 hours of an intensive care unit (ICU) stay; discharge notes were excluded. We only included patients that stayed in the ICU for more than 48 hours. We use the tf-idf statistics of the 10,000 most frequent words as features. Training a model on 50% of the data, selecting hyper-parameters on 25%, and testing on 25%, wefind that logistic regression with L1-regularization achieves an AUC of 0.81. The logistic regression is well-calibrated with Brier scores ranging from 0.06-0.11 across the five

groups; we note better calibration is correlated with lower prediction error.

We report cost and discrimination level in terms of generalized zero-one loss. [245] Using an ANOVA test [102] with $p < 0.001$, we reject the null hypothesis that loss is the same among all five groups. To map the 95% confidence intervals, we perform pairwise comparisons of means using Tukey's range test [288] across 5-fold cross-validation. As seen in Figure 3-3a, patients in the Other and Hispanic groups have the highest and lowest generalized zero-one loss, respectively, with relatively few overlapping intervals. Notably, the largest ethnic group (White) does not have the best accuracy, whereas smaller ethnic groups tend towards extremes. While racial groups differ in hospital mortality base rates (Table 1 in the Supplementary material), Hispanic (10.3%) and Black (10.9%) patients have very different error rates despite similar base rates.

To better understand the discrimination induced by our model, we explore the effect of changing training set size. To this end, we repeatedly subsample and split the data, holding out at least 20% of the full data for testing. In Figure 3-3b, weshow loss averaged over 50 trials of training a logistic regression on increasingly larger training sets; estimated inverse power-law curves show good fits. We see that some pairwise differences in loss decrease with additional training data.

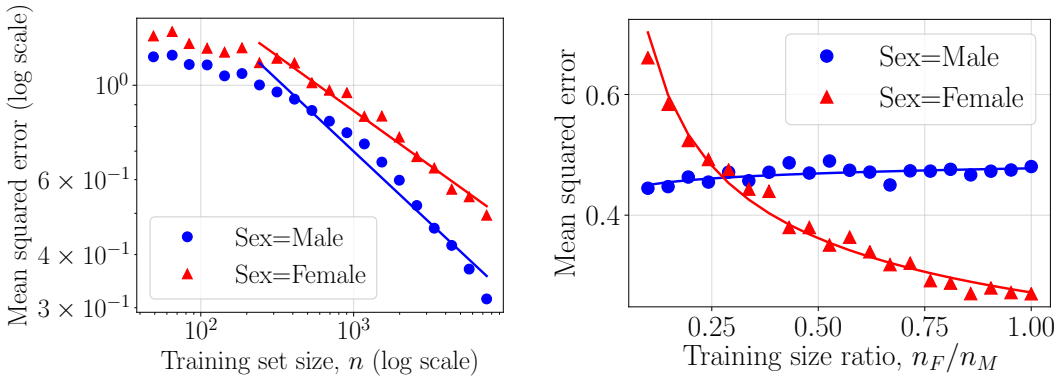
Next, we identify clusters for which the difference in prediction errors between protected groups is large. We learn a topic model with $k = 50$ topics generated using Latent Dirichlet Allocation [30] as implemented by the MALLET toolkit [211] with 1000 iterations. Topics are concatenated into an $n \times k$ matrix Q where q_{ic} designates the proportion of topic $c \in [k]$ in note $i \in [n]$. Following prior work on enrichment of topics in clinical notes [206, 119], we estimate the probability of patient mortality Y given a topic c as $\hat{p}(Y|C = c) := (\sum_{i=1}^{n} y_i q_{ic})/(\sum_{i=1}^{n} q_{ic})$ where y_i is the hospital mortality of patient i. We compare relative error rates given protected group and topic using binary predicted mortality $\hat{y}_i$, actual mortality y_i, and group a_i for patient i with

$$\hat{p}(\hat{Y} \neq Y \mid A = a', C = c) = \frac{\sum_{i=1}^{n} \mathbf{1}(y_i \neq \hat{y}_i)\mathbf{1}(a_i = a')q_{ic}}{\sum_{i=1}^{n} \mathbf{1}(a_i = a')q_{ic}}$$

which follows using substitution and conditioning on A. These error rates were computed using a logistic regression with L1 regularization using an 80/20 train-test split over 50 trials.

While many topics have consistent error rates across groups, some topics (e.g., cardiac patients or cancer patients as shown in Figure 3-3c) have large differences in error rates across groups. We include more detailed topic descriptions in the supplementary material. Once we have identified a subpopulation with particularly high error, for example cancer patients, we can consider collecting more features or collecting more data from the same data distribution. We find that error rates differ between 0.12 and 0.30 across protected groups of cancer patients, and between 0.05 and 0.20 for cardiac patients.

3.4.3 Book Review Ratings



(a) As training set size increases for random forest, MSE decreases but maintains difference between groups. Intercepts from fitted power-laws show no difference in noise.

(b) Holding number of reviews for male authors n_M steady and varying number of reviews for female authors n_F, we can achieve higher MSE for one group than with the full dataset.

Figure 3-4: Goodreads dataset for book rating prediction. Adding training data decreases overall mean squared error (MSE) for both groups while adding training data to only one group has a much bigger impact on reducing $\overline{\Gamma}$. Increasing the number of features reduces MSE but does not reduce $\overline{\Gamma}$.

Sentiment and rating prediction from text reveal quantitative insights from unstructured data; however deficiencies in algorithmic prediction may incorrectly repre-

sent populations. We study prediction of book review ratings from review texts. [126] Using a dataset of 13,244 reviews collected from Goodreads [126] with inferred author sex scraped from Wikipedia, we seek to predict the review rating based on the review text. We use as features the Tf-Idf statistics of the 5000 most frequent words. Our protected attribute is gender of the author of the book, and the target attribute is the rating (1-5) of the review. The data is heavily imbalanced, with 18% reviews about female authors versus 82% reviews about male authors.

We observe statistically significant levels of discrimination with respect to mean squared error (MSE) with linear regression, decision trees and random forests. Using a random forest and training on 80% of the dataset and testing on 20%, we find that our $\Gamma^{\text{MSE}}(\hat{Y})$ has 95%-confidence interval 0.136 ± 0.048 with $\text{MSE}_M = 0.224$ for reviews for male authors and $\text{MSE}_F = 0.358$ for reviews for female authors using a difference in means statistical test. Results were found after hyperparameter turning for each training set size and taking an average over 50 trials. We observe similar patterns with linear regression and decision trees.

To estimate the impact of additional training data, we evaluate the effect of varying training set size n on predictive performance and discrimination. Through repeated sample spitting, we train a random forest on increasing training set sizes, reserving at least 20% of the dataset for testing. In Figure 3-4a, additional training data lowers MSE_F and MSE_M, fitting an inverse power-law. Based on the intercept terms of the extrapolated power-laws ($\delta_M = 0.0011$ for reviews with male authors and $\delta_F = 0.0013$ for reviews with female authors), we may expect that $\overline{\Gamma}$ can be explained more by differences in bias and variance than by noise since our estimated difference in noise $|\delta_F - \delta_M| \approx 0$.

In order to further measure the effect of collecting more samples, we analyze a one-sized increase in training data. Because of the initial skew of author genders in the dataset, we vary the number of reviews for female authors, creating a shift in populations in the training data. We fix the training set size of reviews for male authors at $n_M = 1939$, which represents the size of the full data for female authors N_F, reserving 20% of the dataset as test data. We then vary the training data size for

female authors n_F such that the ratio n_F/n_M varies evenly between 0.1 to 1.0. Using a linear regression in Figure 3-4b, we see that as the ratio n_F/n_M increases, MSE_F decreases far below MSE_M and far below our best reported MSE of the random forest on the full dataset. This suggests that shifting the data ratio and collecting more data for the under-represented group can adapt our model to reduce discrimination.

3.5 Discussion

We identify that existing approaches for reducing discrimination induced by prediction errors may be unethical or impractical to apply in settings where predictive accuracy is critical, such as in healthcare or criminal justice. As an alternative, we propose a procedure for analyzing the different sources contributing to discrimination. Decomposing well-known definitions of cost-based fairness criteria in terms of differences in bias, variance, and noise, we suggest methods for reducing each term through model choice or additional training data collection. Case studies on three real-world datasets confirm that collection of additional samples is often sufficient to improve fairness, and that existing post-hoc methods for reducing discrimination may unnecessarily sacrifice predictive accuracy when other solutions are available.

Looking forward, we can see several avenues for future research. In this work, we argue that identifying clusters or subpopulations with high predictive disparity would allow for more targeted ways to reduce discrimination. We encourage future research to dig deeper into the question of local or context-specific unfairness in general, and into algorithms for addressing it. Additionally, extending our analysis to intersectional fairness [39, 138], e.g., looking at both gender and race or all subdivisions, would provide more nuanced grappling with unfairness. Finally, additional data collection to improve the model may cause unexpected delayed impacts [195] and negative feedback loops [93] as a result of distributional shifts in the data. More broadly, we believe that the study of fairness in non-stationary populations is an interesting direction to pursue.

Chapter 4

Auditing Algorithmic Bias in Predictive Algorithms for Health Insurance

4.1 Introduction

Health insurance companies wield tremendous impact over the health of millions of people. Case management programs in particular seek to assign additional help to policy members who will require additional medical care in the future. Leveraging a constellation of programs and dedicated healthcare professionals, health insurance companies seek to identify and recruit members who would benefit from these additional resources using predictive algorithms. Prior work has shown that these case management risk scores may contain racial bias due to nuances in outcome label selection. [227] Given the wealth of methodological work on algorithmic fairness and the wide-reaching impact of case management programs, it is crucial to examine case management algorithms for potential bias and potential mitigation steps.

In practice, auditing real-world case management algorithms has several technical challenges. First, as with most real-world health data, insurance claims data can be sparse and multivariate as members interact infrequently with the healthcare system

but produce high-dimensional data when they do. Second, case management scores can represent a composite of subscores, relying on many components including health prediction algorithms as well as business prioritization metrics. Additionally, member race is often not collected or included in the dataset. Researchers must then rely on either probabilistic imputation using techniques like Bayesian Improved First Name Surname Geocoding. [5] When individual-level information is not needed, researchers may leverage geographic/census tract or zip code level information about racial composition. Lastly, predicted scores are generally evaluated against a true observed event; however, in case management algorithms, the outcome of interest can sometimes be unclear. Specific targeted programs, including injury management or reduction of adverse pregnancy outcomes, may have more defined outcomes. When needed, noisy proxies are used instead, for example future observed health events.

4.1.1 Contributions

This work was done in collaboration with Stephanie Gervasi, Yuria Utsumi, Sol Rodriguez, Johnathan Kyle Armstrong, Aaron Smith-McLallen, David Sontag, Michael Vennera, and Ravi Chawla.

In partnership with Independence Blue Cross (IBC), a health insurer based in Philadelphia, we examine the case management algorithm, specifically two components of said algorithm, used to prioritize outreach for case management enrollment. We select two specialized algorithms that feed into the risk stratification engine of the overall case management algorithm to examine: 1) likelihood of hospitalization (LOH), which predicts future acute and non-elective hospitalization in the next six months, and 2) a high-risk pregnancy (HRP), which predicts high-risk pregnancy for women of childbearing age.

We select LOH and HRP because of the easily measured outcomes: future hospitalization in next six months and high-risk pregnancy, respectively. In each case, we derive our own predictive models from the IBC claims data as an illustrative model. Although IBC constructs their own algorithmic risk scores, we choose to develop separate predictive models in order to study how the entire model development process

is affected. Because member race is not consistently measured, we use alignment with electronic medical records to extract the patient-provided race group. Leveraging the techniques developed in Chapter 3 and [51], we analyze the LOH and HRP cohorts to examine any algorithmic bias and potential areas of further exploration.

4.1.2 Related Work

Likelihood of Hospitalization

According to Centers for Medicare & Medicaid Services (CMS), hospitalizations represented the largest component of national health care expenditures in 2017 and 2018. [1].While many acute inpatient events such as maternity and trauma admissions are unavoidable, others are preventable through effective primary and specialty care, disease management, availability of interventions at outpatient facilities, or all of the above. In 2017 the Agency for Healthcare Research and Quality (AHRQ) estimated that 3.5 million preventable inpatient hospitalizations accounted for $33.7 billion in hospital costs. [165]

Machine learning models that predict the likelihood of an avoidable inpatient hospitalization (known as likelihood of hospitalization models) can help target interventions, prevent adverse health outcomes, and reduce individual and population health care costs. [148, 291, 76]

However, observing an acute hospitalization event in the data is contingent on access to and use of health care services, both of which are influenced by racial and socioeconomic disparities. [122] Disparities in access and use mean that some subpopulations are underrepresented in the target population and in the data used to predict the outcome of interest. Thus, the resulting model output may reflect those systemic biases, and interventions or policy decisions based on the model outputs risk reinforcing and exacerbating existing inequities.

Similar to disease onset models, one way to address the data disparities is through inclusion of additional data sources that show patterns in primary or preventative care that can prevent unplanned hospitalization. Electronic medical records (EMR) data

can add granularity to clinical events, capturing diagnostic and other health information that may not be recorded on claims. However, integrating EMR and claims data can introduce additional bias [261] stemming from missing or incomplete records for patients who experience barriers to consistent care. Importantly, missing clinical codes can indicate lack of key diagnostics, procedures, or primary care support along a patient's health care journey that might have precluded the need for inpatient hospitalization. Similar symptoms may be treated differently among providers, leading to downstream effects on hospitalization. Social determinants of health data can also improve the performance, and potentially interpretations of likelihood of hospitalization prediction tasks.

High-Risk Pregnancy

Although the large majority of women experience a normal, uncomplicated pregnancy, approximately 15-20% of all pregnancies are considered to be at high-risk. [189] Here, we define a high-risk pregnancy as one where biomedical factors related to the mother's present or previous medical condition could put the mother's or baby's well being at risk. Despite continued advancements in medical care, rates of maternal mortality and morbidity and pre-term birth have been rising in the U.S. [203] In fact, maternal and infant mortality rates in the U.S. are far higher than those in similarly large and wealthy countries. [50]

At the same time, there exist systemic health disparities related to maternal and infant mortality rates. Black and American Indian and Alaska Native Resources women have pregnancy-related mortality rates that are over three and two times higher, respectively, compared to the rate for White women (40.8 and 29.7 compared to 12.7 per 100,000 live births). [238] This disparity persists even in California, where maternal mortality rates are lower than the national average and have been on the decline; yet, the rate is more than three times higher among Black women compared to White women. [17]

Prior research using machine learning to predict pregnancy complications have achieved relatively high performance (0.75-0.85 AUC) with pre-eclampsia and pre-

maturity identified as most predictive features. Notably, however, it is still unclear how to define “at risk" individuals because of the large range of short-term and long-term complications that can arise. [26]

4.2 Data

4.2.1 Likelihood of Hospitalization

We extract the 6-month lookahead LOH outcomes for the months of August 2020, October 2020, and January 2021. Because the LOH algorithm differs slightly based on the type of insurance plan and the frequency of utilization, we focus only on the members with government plans with greater than 4 prescription claims and greater than 5 Medicare claims in the last 12 months. Based on this criteria, we create a cohort of 64,410 members. Based on enrollment in Medicare programs, we are able to extract the self-reported race for over 97% of the patients, as seen in Table 4.1, using enrollment information in Medicare programs. These designations including 5 single race groups, with no option to include more than one race category. Smaller sample size groups are not named explicitly and instead grouped into “Other”, e.g., Asian and Hispanic groups.

Our primary outcome is the confirmed hospitalization in the following 6 months (e.g., August 2020 - February 2021). A beneficiary was confirmed as being hospitalized based on an acute in-patient stay, excluding elective, maternity, and hospice inpatient visits. The cohort has a hospitalization prevalence of 7.8%.

We extract medical data available in the claims data, including medical diagnoses, prescriptions, procedures, and specialty visits. We bucket the features into time intervals of 30 days, 180 days, 365 days, and ever appearing in the clinical history. We include 12 non-temporal features spanning demographics, ages, and patient race extracted for medical claims. The total feature set includes 73,774 features.

4.2.2 High-Risk Pregnancy

We construct a pregnancy cohort in order for our high-risk pregnancy prediction task. An algorithm predicting high-risk pregnancies could help identify members who would benefit from the Baby Blueprints case management program.

Following existing protocol from [209], we infer the most recent pregnancy-related outcome for the full set of IBC patients. The recorded outcomes span 17 years, from July 2005 to February 2021. The median outcome date is May 2017. Pregnancy episodes were included in the final cohort if the patient was female, between 12 and 55 years of age at the time of pregnancy-related outcome and had continuous enrollment during the pregnancy episode. Based on these criteria, we create a cohort of 43,358 members. We extract race based on available electronic medical records from hospitals in the IBC health system exchange, which contains race information for about half of the current commercial members, as seen in Table 4.1. Smaller sample size groups are not named explicitly and instead grouped into "Other", e.g., Asian and Hispanic groups.

Our prediction outcome is the confirmed live birth of an identified pregnancy, meaning a live birth is coded as 1 and all other outcomes are coded as 0. A beneficiary is confirmed as a healthy live birth based on the absence of pregnancy complications. Based on the claims data, we identify several pregnancy outcomes including live birth, stillbirth, ectopic pregnancy, spontaneous abortion, induced abortion, preterm birth, hypertension pre-eclampsia, pre-eclampsia, and neo-natal intensive care unit (NICU) stay. Once a pregnancy-related outcome is identified, we define the pregnancy window. We search over the time window prior to the outcome to identify the first indication of pregnancy, designated the pregnancy start date. After the start date is defined, we search over the full pregnancy window to identify additional outcomes from the claims data. For instance, if the pregnancy resulted in a live birth and also a NICU stay, we replace the outcome with NICU instead of live birth. The healthy live-birth prevalence is 31.5%.

We extract medical data available in the claims data, including medical conditions,

Task	Race	Data Percentage
Likelihood of Hospitalization	White	76.2%
	Black	14.8%
	Other	4.0%
	Unknown	2.7%
High-Risk Pregnancy	White	39.0%
	Black	5.7%
	Other	7.3%
	Unknown	48.0%

Table 4.1: Percentage of patient data by race for Likelihood of Hospitalization and High-Risk Pregnancy tasks.

prescriptions, procedures, and specialty visits. We include medical data from the full history before the recorded pregnancy outcome. We bucket the features into time intervals of 30 days, 180 days, 365 days, and in the clinical history. We include 12 non-temporal features spanning demographics, ages, and patient race extracted for medical claims. The feature set totals 156,712 features.

4.3 Methods

We apply our analysis using methods outlined in Chapter 3. Below, we outline experimental analysis.

4.3.1 Impact of Additional Training Data

To assess the impact of variance on the model, we vary the amount of training data used. We partition the data into 70% training data, 15% validation data for hyper-parameter tuning, and 15% test data. Each data partition is select using random sampling. We vary percentage of the training data and plot the resulting impact on the error, defined as either 1 - accuracy or 1 - positive predictive value (PPV).

Confidence intervals are computed for on a 95%-confidence interval assuming a Gaussian distribution. The Bernoulli success-failure trials of of each independent data sample can be modeled with a binomial distribution. The central limit theorem allows

for the approximation of the accuracy with a Gaussian distribution when the sample size is sufficiently large. Similarly, the PPV confidence intervals can be computed using the same Gaussian assumption. [311]

4.3.2 Bayes Error Estimation

We estimate the Bayes error using the Mahalanobis distance [205] and the Bhattacharyya distance [27], and a k-nearest neighbor method [71] with $k = 5$ and 5-fold cross validation. Because distance metrics can fail on high-dimensional data, we feature select using a highly regularized ($C = 0.001$) L1-logistic regression. We select a logistic regression due to the wide-spread and consistent performance in clinical models. [25] The most predictive features are used for distance computation. This results in 57 features selected for the Likelihood of Hospitalization task and 83 features selected for the High-Risk Pregnancy task.

4.3.3 Subpopulation Identification Using Topic Modeling

We follow the same clustering approach using topic modeling as outlined in 3.4.2. Unlike traditional topic modeling, we create a "bag of events" [127] for each patient including clinical diagnoses, medications, and medical procedures. For each patient, a row of clinical events is encoded into a sparse matrix representation. The traditional topic modeling "note" is now a patient, and each "word" now corresponds to one claim code, e.g., diagnoses, medications, and lab tests.

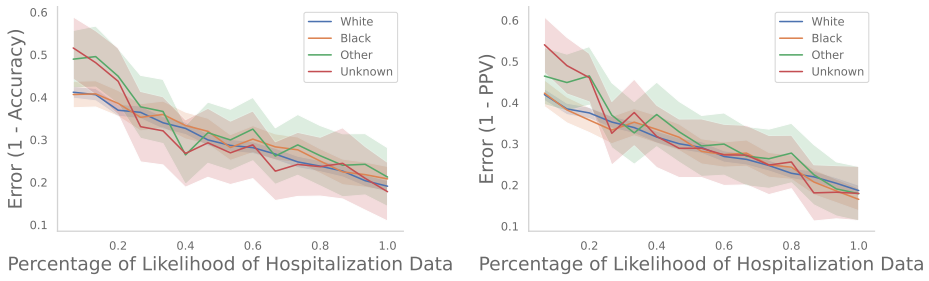
Similar to prior work [251], we encode temporal patterns by concatenating several matrices per patient, with each matrix corresponding to whether the clinical events occurred within a given time window. We use four time windows: the previous 30 days, the previous 180 days, the previous 365 days, and whether the event happened ever in the patient history.

4.4 Results

4.4.1 Impact of Additional Training Data

For the Likelihood of Hospitalization task, the confidence intervals intersect for both error defined by AUC and by accuracy in Figure 4-1. This suggests that there is not great algorithmic bias at both large and small dataset sizes. In fact, adding more data from the same distribution would likely not improve predictive performance.

For the High-Risk Pregnancy task, the confidence intervals clearly intersect for error defined by AUC in Figure 4-2. However, AUC may not capture ranking nuances for fairness analyses. [162] In the error defined by accuracy plots, the Black patients have lower error on average with Unknown and Other race patients having higher error. Unlike the Likelihood of Hospitalization cohort, High-Risk Pregnancy has a large prevalence of Unknown patients, as seen in Table 4.1, which may make prediction on this cohort challenging. The consistently high error for the Unknown race patients across all percentages of the total data could suggest that the noise and statistical bias terms are affecting the differences in error.

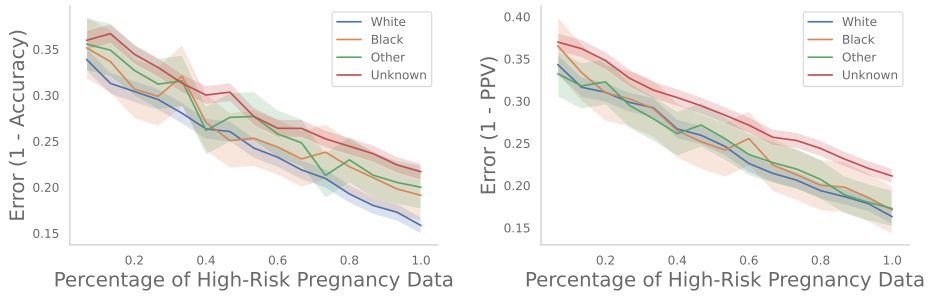


(a) (b)

Figure 4-1: Logistic regression performance for Likelihood of Hospitalization, measured by **a)** 1 - accuracy and **b)** 1 - positive predictive value (PPV) versus the percentage of total training data.

4.4.2 Bayes Error Estimation

We show the Bayes error estimates in Table 4.2 for both Likelihood of Hospitalization and High-Risk Pregnancy. For LOH, all estimation techniques show a particu-



(a) (b)

Figure 4-2: Logistic regression performance for High-Risk Pregnancy, measured by **a)** 1 - accuracy and **b)** 1 - positive predictive value (PPV) versus the percentage of total training data

larly low amount of estimated Bayes error for the Other and Unknown race patients with a slightly higher noise estimates for Black and White patients. Note that the Bhattacharyya distance computation for the LOH task yielded singular matrix errors during computation.

For HRP, the Bayes error estimates show consistently high error estimates for the patients with Unknown race. This finding matches our understanding of how race data is extracted for these patients. Patients without a known race value may also lack rich claims data as well for the HRP prediction task.

4.4.3 Subpopulation Identification Using Topic Modeling

We extract representative claim codes for condition, procedure, specialty visit, and drugs for each of the learned topics, e.g., in Table 4.3 and Table 4.4. Because of the windowed data method, the variable evaluation period may differ, emphasizing the chronic nature of many of these conditions. We provide the full 50 topics for both LOH and HRP in supplementary materials Chapter B.

We observe a wide distribution of topic prevalence. In Figure 4-3, the topics prevalence weights are plotted by topic number. Using topic descriptions from the supplementary materials, we find that hypertension (Topic 28) is one of the most prevalent topics for LOH while prenatal care (Topic 15) is one of the most prevalent topics for HRP.

When evaluating the algorithmic performance across these topics, there is a high range between highest error enrichment values and lowest error enrichment values across patient races, as seen in Figure 4-4. However, we do not want to evaluate solely on differences between patient races because topic prevalence weights may vary. Very small topic weights would correspond to topics that appear very infrequently. We are therefore interested in topics with high topic prevalence weight as well as large differences in error enrichment values across races.

The renal and respiratory failure topic in the LOH task (Topic 35, described in Table 4.3) has the highest difference in error across races (0.071) and a relatively high topic prevalence weight (0.061). Similarly, the fatigue topic in the HRP task (Topic 44, described in Table 4.4) has a high difference in error across races (0.117) and a high topic prevalence weight (0.046).

Dataset	Noise Estimation	Lower	Upper	Subgroup
Likelihood of Hospitalization	Mahalanobis	–	0.137	White
		–	0.145	Black
		–	0.105	Other
		–	0.110	Unknown
Likelihood of Hospitalization	Bhattacharyya	–	–	White
		–	–	Black
		–	–	Other
		–	–	Unknown
Likelihood of Hospitalization	Nearest Neighbors	0.044	0.083	White
		0.046	0.088	Black
		0.034	0.065	Other
		0.032	0.063	Unknown
High-Risk Pregnancy	Mahalanobis	–	0.402	White
		–	0.343	Black
		–	0.416	Other
		–	0.426	Unknown
High-Risk Pregnancy	Bhattacharyya	0.093	0.291	White
		0.032	0.176	Black
		0.041	0.199	Other
		0.233	0.422	Unknown
High-Risk Pregnancy	Nearest Neighbors	0.229	0.353	White
		0.166	0.277	Black
		0.229	0.353	Other
		0.255	0.380	Unknown

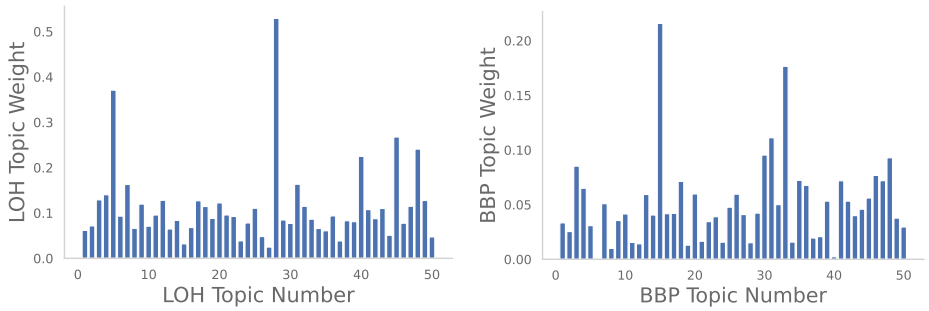
Table 4.2: Bayes error noise estimates, including lower and upper bounds, for Likelihood of Hospitalization and High-Risk Pregnancy tasks. Mahalanobis distance Bayes error estimates only contain an upper bound. Other entries with dash lines indicate singular matrix errors from matrix inverse operations.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Past 365 days	Subsequent hospital care, per day, for the evaluation and management of a patient	12424	41695
	Past 365 days	Initial hospital care, per day, for the evaluation and management of a patient	3186	11080
	Past 365 days	Initial hospital care, per day, for the evaluation and management of a patient	2469	9435
	Past 365 days	Subsequent hospital care, per day, for the evaluation and management of a patient	5517	18183
	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only	5246	24560
Specialty	Past 365 days	Emergency Medicine	4967	25070
	Past 180 days	Emergency Medicine	2365	11250
	Past 365 days	Diagnostic Radiology	6691	38475
	Past 180 days	Diagnostic Radiology	3187	17224
	Past 365 days	Internal Medicine	14403	89915
Condition	Past 365 days	Patient dependence on care provider	3392	13704
	Past 365 days	Acute renal failure syndrome	5329	17082
	Past 365 days	Abnormal findings on diagnostic imaging of lung	1974	9888
	Past 180 days	Patient dependence on care provider	1809	6645
	Past 365 days	Dyspnea	7916	41875
Drug	Past 180 days	pantoprazole 40 MG Delayed Release Oral Tablet	1859	10661
	Past 365 days	pantoprazole 40 MG Delayed Release Oral Tablet	3259	19744
	Past 365 days	metoprolol tartrate 25 MG Oral Tablet	1683	11148
	Past 365 days	oxycodone hydrochloride 5 MG Oral Tablet	495	3336
	Past 180 days	oxycodone hydrochloride 5 MG Oral Tablet	285	1834

Table 4.3: Most representative labs, procedures, conditions, specialty visits, and medications for Topic 35 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections	2935	7772
	Entire History	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections; 1 vaccine single or combination vaccine/toxoid	3630	10205
	Past 365 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections; 1 vaccine single or combination vaccine/toxoid	2036	5108
	Past 180 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections; 1 vaccine single or combination vaccine/toxoid	1239	2923
	Entire History	Cytopathology, cervical or vaginal any reporting system	4236	12158
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Entire History	Internal Medicine	1923	6415
	Entire History	Emergency Medicine	1503	6683
Condition	Entire History	Fatigue	509	2318
	Past 730 days	Fatigue	420	1883
	Entire History	Blood chemistry abnormal	90	593
	Entire History	Vitamin D deficiency	703	2643
	Past 365 days	Fatigue	296	1346
Drug	Entire History	0.5 ML influenza A virus A 0.	105	325
	Past 730 days	0.5 ML influenza A virus A 0.	105	319
	Past 730 days	0.5 ML influenza A virus A	91	305
	Entire History	0.5 ML influenza A virus A	91	310
	Entire History	0.5 ML influenza A virus A	54	199

Table 4.4: Most representative labs, procedures, conditions, specialty visits, and medications for topic 44 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.



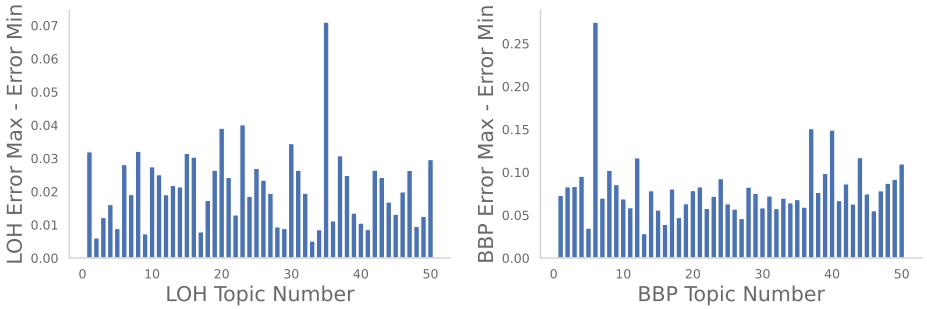
(a) (b)

Figure 4-3: Topic weights for **a)** Likelihood of Hospitalization task and **b)** High-Risk Pregnancy task. Prevalent topics include Topic 28 (hypertension) for LOH and Topic 15 for HRP (prenatal care). See supplementary materials Chapter B for full descriptions of the learned topics with representative conditions, medications, specialty visits, and procedures.

4.5 Discussion

We demonstrate techniques for decomposing discrimination on claims data from Independence Blue Cross, specifically the two tasks of likelihood of hospitalization and high-risk pregnancy. Analyzing the impact of bias, variance, and noise, we suggest areas for future investigation towards equitable algorithms.

Looking forward, we can see several avenues for future research. In this work, the presence or identification of racial identity relies on electronic health records, which may be incomplete. Techniques to increase observation or inference of racial information must ensure that individual identity is respected while also being available for research purposes. Additionally, questions of intersectional fairness complicate techniques that rely on a few clearly defined groups. [268] Finally, actionable insights towards improving the health of individuals must also examine any differences in impact. More broadly, the long-term impact in non-stationary populations is an interesting direction to pursue.



(a) (b)

Figure 4-4: Range between maximum and minimum error topic values across race groups for **a)** Likelihood of Hospitalization task and **b)** High-Risk Pregnancy task.

Topic Num.	Topic Prev.	Diff. between max and min error	White Error	Black Error	Other Error	Unknown Error
35	0.061	0.071	0.127	0.113	0.091	0.162
23	0.039	0.040	0.048	0.043	0.022	0.008
20	0.122	0.039	0.100	0.081	0.081	0.061
30	0.077	0.034	0.108	0.104	0.073	0.106
8	0.066	0.032	0.062	0.048	0.030	0.051
1	0.061	0.032	0.091	0.076	0.059	0.066
15	0.032	0.031	0.120	0.088	0.096	0.099
37	0.038	0.031	0.077	0.068	0.046	0.077
16	0.068	0.030	0.069	0.061	0.073	0.043
50	0.047	0.030	0.101	0.091	0.093	0.121
6	0.093	0.028	0.072	0.065	0.080	0.052
10	0.071	0.027	0.061	0.056	0.075	0.047
25	0.110	0.027	0.057	0.050	0.036	0.063
42	0.087	0.026	0.079	0.081	0.083	0.105
31	0.163	0.026	0.047	0.052	0.041	0.067

Table 4.5: Likelihood of Hospitalization task topic prevalence, difference between maximum and minimum error values across races, and race-specific errors for top 15 topics sorted by difference between maximum and minimume error values across races. Recall that the the dataset has 76.2% White patients, 14.8% Black patients, 4.0% Other patients, and 2.7% Unknown patients (see Table 4.1).

Topic Num.	Topic Prev.	Diff. between max and min error	White Error	Black Error	Other Error	Unknown Error
6	0.001	0.275	0.086	0.233	0.216	0.361
37	0.019	0.151	0.199	0.188	0.230	0.339
40	0.002	0.149	0.112	0.098	0.092	0.241
44	0.046	0.117	0.139	0.186	0.249	0.256
12	0.014	0.117	0.150	0.178	0.266	0.173
50	0.030	0.110	0.096	0.160	0.185	0.205
8	0.010	0.102	0.141	0.183	0.167	0.243
39	0.053	0.099	0.111	0.175	0.183	0.209
4	0.065	0.095	0.145	0.188	0.216	0.241
24	0.016	0.092	0.131	0.184	0.179	0.224
49	0.038	0.092	0.143	0.171	0.232	0.235
48	0.093	0.087	0.200	0.227	0.249	0.287
42	0.053	0.086	0.171	0.193	0.236	0.258
9	0.035	0.086	0.090	0.114	0.152	0.176
3	0.085	0.083	0.207	0.207	0.253	0.291

Table 4.6: High-Risk Pregnancy task topic prevalence, difference between maximum and minimum error values across races, and race-specific errors for top 15 topics sorted by difference between maximum and minimume error values across races. Recall that the the dataset has 39.0% White patients, 5.7% Black patients, 7.3% Other patients, and 48.0% Unknown patients (see Table 4.1).

Chapter 5

Clustering Interval-Censored Multivariate Time-Series Data

5.1 Introduction

Cluster analysis of time-series data is a task of interest across a variety of scientific disciplines including biology [193], meteorology [44], and astrophysics. [252] Automating the discovery of latent patterns in real-world data can be challenging due to noise. We focus on mitigating errata in pattern discovery from interval censoring. [216]

Interval censoring arises when time-series data are only observed within a known interval. Both *left-censorship*, which occurs when a phenomenon is observed at some time but it is unknown when it began, and *right-censorship*, which occurs an event removes a time-series from observation after some point, can lead common techniques for clustering to erroneous conclusions about the underlying patterns. To address this, practitioners must often manually align data to a meaningful start time in order to find non-trivial groups via unsupervised clustering—a process whose difficulty can range from expensive and time-consuming in some problems to infeasible in others. In this work, we develop a machine learning algorithm that clusters time-series data while simultaneously correcting for interval censoring. In doing so, we automate the time-consuming process of manual data alignment and use our method to reveal structure that would otherwise not be found by straightforward application of clustering

analysis.

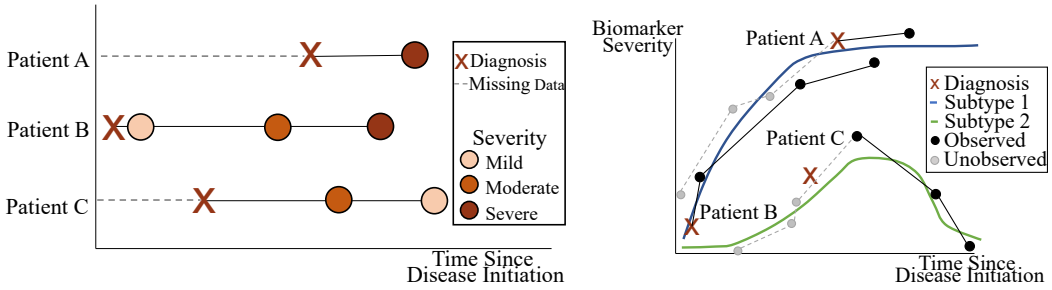


Figure 5-1: (a) Patient data can be interval-censored, meaning longitudinal data can be missing based on entry to the dataset, e.g., first diagnosis, and the data may lack a common outcome against which to align. Patients may enter the dataset at any stage of the disease. (b) Interval censoring can make clustering patient time-series data challenging because data may be aligned incorrectly, e.g., first diagnosis. We seek to understand disease heterogeneity by inferring subtypes after correcting for misalignment.

For a simplified illustration of the problem in the context of disease phenotyping, see Figure 5-1(a) which depicts the common reality of observational health data whereas Figure 5-1(b) depicts the idealized latent substructure we would like to identify. Existing subtyping models applied to clinical data assume (potentially erroneously) that patients are aligned at entry into the dataset or study. The problem with such an assumption is that the disparity between true disease stage and observed observation time can result in unsupervised learning algorithms uncovering the wrong, or perhaps less interesting, structure. For example, a naive clustering algorithm might simply return clusters corresponding to the disease stage at entry

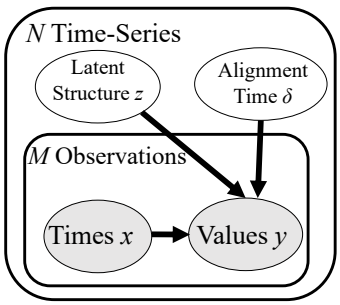


Figure 5-2: Graphical model of SubLign

into the study, which may simply recapitulate some of the biases mentioned earlier.

5.1.1 Contributions

To that end, this chapter makes the following contributions:

1. We introduce and formalize the problem of cluster recovery from interval-censored time-series data.

2. We introduce a practical algorithm, SubLign, based on variational learning of a deep generative model, which:

    - Makes no assumption on the distribution of delayed entry alignment values besides extrema.
    - Operates on multivariate time-series with varying lengths and missing values, both characteristics often found in real-world datasets.
    - Is unsupervised, meaning that neither subtype labels nor alignment values are provided during training.

3. We prove an identifiability result showing that, in a noiseless setting, both the degree of delayed entry from left-censorship and the subtype identity are recoverable.

We show robust quantitative results on synthetic data where, over multiple runs, our method outperforms baselines for subtyping and patient alignment.

5.1.2 Related Work

Learning alignment and clustering has been studied in fields across computer vision, signal processing, and health. Approaches often make assumptions including few discrete time steps [313, 153]; a single piecewise linear function [313] or Gaussian mixture model [153]; significantly more samples per object than number of objects [210, 112]; very small windows of potential misalignment [197, 193]; or known lag time. [190] Methods that directly measure similarity between time-series, e.g., dynamic time

Algorithm 1 SubLign

Input: Observation times $X \in \mathbb{R}^{N \times M}$, biomarkers $Y \in \mathbb{R}^{N \times M \times D}$
Output: τ_k for each subtype and $\hat{\delta}_i$ for each patient
Step 1: Learning
repeat
  Encode time-series: $h_i = \text{RNN}([X_i, Y_i])\ \forall i \in \{1, \ldots, N\}$
  Compute variational distribution $q(Z_i|X_i, Y_i) = \mathcal{N}(\mu(h_i; \phi_2), \Sigma(h_i; \phi_3)) \forall i \in \{1, \ldots, N\}$
  for patient $i = 1$ **to** N **do**
    Run grid-search to find $\hat{\delta}_i = \arg\max_{q(\delta_i)} \mathcal{L}(Y_i|X_i; \gamma, \phi, q(\delta_i))$ (Eq. 5.3)
  end for
  Update γ, ϕ via stochastic gradient ascent on $\mathcal{L}(Y|X; \gamma, \phi, \hat{\delta})$
until convergence
Step 2: Inference and Clustering
Infer $\mathcal{Z} = \{z_i | z_i = \mu(h_i; \phi_2)\}$ for X_i, Y_i
Find K clusters using k-means on $\mathcal{Z}$ and compute cluster centers μ_k
Infer parameters of subtype trajectories $\tau_k = g(\mu_k)$

warping [74] or methods that aggregate multiple imputation methods [95] can also be used for clustering time-series data. Our method aims to cluster interval-censored multivariate time series without these constraints.

5.2 SubLign: Subtype & Align

There are two stages to SubLign. First, we learn a generative model of the observed data which disentangles variation in the observed data due to delayed entry from variation related to subtype identity. Second, we infer subtype representations and (optionally) cluster the representations to obtain the explicit subtype identity for each time-series. Figure 5-1(c) describes the graphical model, and Algorithm 1 depicts the pseudocode for this procedure.

5.2.1 Generative Model

Consider the following setup. We observe N multivariate time-series (one for each patient), each of length up to M: $[(x_{1,1}, y_{1,1}), \ldots, (x_{1,M}, y_{1,M})], \ldots [(x_{N,1}, y_{N,1}), \ldots, (x_{N,M}, y_{N,M})]$. $y_{i,m} \in \mathbb{R}^D$ is a vector of observations for time-series i at time-stamp

$x_{i,m} \in \mathbb{R}^+$. We denote collections of observations as $Y_i = \{y_{i,1}, \ldots, y_{i,M}\}$ and time-stamps as $X_i = \{x_{i,1}, \ldots, x_{i,M}\}$ for patient i.

Figure 5-1(c) depicts the graphical model corresponding to the latent-variable generative model of continuous-time multivariate data:

$$\begin{aligned}
&\forall i = \{1, \ldots, N\},\ \forall m \in \{1, \ldots, M\}, \\
&\forall d \in D_{i,m},\ \delta_i \sim \mathrm{Cat}(\mathcal{D}), z_i \sim \mathcal{N}(\mathbf{0}, \mathbb{I}),\ \Theta = g(z_i; \gamma), \\
&\overline{y}_{i,m}[d] = f(\kappa(x_{i,m} + \delta_i; \Theta[d])),\ y_{i,m} \sim \mathcal{N}(\overline{y}_{i,m}[d], \mathbf{I})
\end{aligned} \tag{5.1}$$

We drop indices denoting patient and dimension where unnecessary. $D_{i,m}$ denotes the set (and $|D_{i,m}|$ denotes the number) of observed biomarkers for patient i at their m-th observation. To accommodate missing data, not all biomarkers are required to be measured at every observation. Each delayed entry value $\delta_i \in \mathbb{R}^+$ has a maximum alignment deviation value δ^+ over all time-series. We discretize the closed interval $[0, \delta^+]$ as $\mathcal{D} = [0, \epsilon, 2\epsilon, \ldots, \delta^+]$ with hyperparameter ϵ and use a categorical distribution over $\mathcal{D}$ with uniform probabilities over each element as our prior over δ_i. Function $g : \mathbb{R}^{N_z} \to \mathbb{R}^{D \times (P+1)}$ has parameters γ and maps from the latent variable to $\Theta \in \mathbb{R}^{D \times (P+1)}$, a matrix of parameters for D polynomials, each of degree P. f is a known link function that describes how observed values relate to observed time-points.

Parameterization We discuss the specific parameterizations of Equation 5.1. In the context of our motivating application of disease phenotyping, these functions represent common characteristics in the progression of patient biomarkers.

Link function and polynomials: For f and P, we study the following choices: *Sigmoid*: $P = 1$ and $f(x) = \frac{1}{1+\exp(-x)}$ and *Quadratic*: $P = 2$ and $f(x) = x$. The sigmoid function can represent bounded and monotonically increasing clinical variables. The quadratic function represents cases where disease severity, as measured by biomarkers, decreases (likely in response to therapy) and then increases (once therapy fails), or vice versa. Other choices for P, f are permissible as long as they are differentiable

with respect to the model parameters. We allow for the possibility that individual biomarkers have different parameterizations.

Modeling polynomial parameters: We parameterize $g(z_i; \gamma)$ using a two layer neural network with ReLU activation functions with parameters γ. To be concrete, if $D = 1, P = 2$, and f is the sigmoid function, then the outputs of g are $[\beta_0(z), \beta_1(z)]$ and $y = \frac{1}{1+\exp^{-(\beta_0(z)x+\beta_1(z))}}$. Similarly, if $D = 1, P = 1$, and f is the quadratic function then the outputs of g are $[a(z), b(z), c(z)]$ and $y = a(z)x^2 + b(z)x + c(z)$.

5.2.2 Inference

Step 1: Learning

We learn the parameters γ of the model in Equation 5.1 via maximum likelihood estimation. Since the model is a non-linear latent variable model, we maximize a variational lower bound on the conditional likelihood of data given the time-stamps corresponding to observations.

$$\log \prod_{m=1}^{M} \prod_{d \in D_{i,m}} p(y_{i,m}[d] | x_{i,m}; \gamma) = \log p(Y_i | X_i; \gamma) \tag{5.2}$$

$$\geq \mathcal{L}(Y_i | X_i; \gamma, \phi)$$

$$= \mathbb{E}_{q(Z_i|X_i,Y_i;\phi)} \Big[\log \sum_{\delta_i} p(Y_i | X_i, \delta_i, Z_i; \gamma) p(\delta_i) +$$

$$+ \log \frac{p(Z_i)}{q(Z_i | X_i, Y_i; \phi)} \Big] \geq \mathcal{L}(Y_i | X_i; \gamma, \phi, q(\delta_i))$$

$$= \mathbb{E}_{q(Z_i|X_i,Y_i;\phi)} \Big[\mathbb{E}_{q(\delta_i)} \Big[\log p(Y_i | X_i, \delta_i, Z_i; \gamma)$$

$$+ \log \frac{p(\delta_i)}{q(\delta_i)} \Big] + \log \frac{p(Z_i)}{q(Z_i | X_i, Y_i; \phi)} \Big] \tag{5.3}$$

The first lower bound uses a variational distribution for Z parameterized via an inference network [174, 254] with parameters ϕ. The second lower bound is a variational distribution over δ. The function $q(\delta_i)$ parameterizes the space of one-hot distributions in $\mathcal{D}$, i.e., a categorical distribution over discrete choices of δ_i.

Our learning algorithm alternates between two steps. We first maximize the lower

bound using subgradient ascent. To do so, we solve: $\hat{\delta}_i = \arg\max_{q(\delta_i)} \mathcal{L}(Y_i|X_i;\gamma,\phi,q(\delta_i))$. For our choice of variational distribution, this maximization can be performed via a grid search. We then derive gradients $\nabla_{\gamma,\phi}\mathcal{L}(Y_i|X_i;\gamma,\phi,\hat{\delta}_i)$ to update the generative model and inference network via stochastic gradient ascent.

Step 2: Obtaining Learned Subtypes

After learning the model, we may re-use the inference network to predict the latent variable $z_i = \mu(h_i;\phi_2)$ for each patient in the training set. Combining z_i across all time-series gives us the set $\mathcal{Z}$. When reasonable, we refer to $\mu(h_i;\phi_2)$ as μ_i.

Although latent variable z_i encodes latent structure from each time-series, we may be interested in explicit subtypes for a given value of K. To obtain discrete subtypes, we can run clustering algorithms on $\mathcal{Z}$ to obtain K cluster centers $\{\mu_1,\ldots,\mu_K\}$. Because we use a Gaussian prior for our biomarker values, measuring distances in the space of the latent variable can be done with the Euclidean norm, making the k-means a reasonable choice of clustering algorithm.

We compute $\{\tau_1,\ldots,\tau_K\}$ where $\tau_k = g(\mu_k)$ as the progression-patterns corresponding to each of the discrete subtypes of the disease. For example, if $f \circ \kappa$ is linear, then we obtain K different slopes and biases, each of which describes how the time-series behaves in that subtype. In practice, K may be chosen based on domain knowledge; alternatively, qualitative results can be assessed for each version of K, e.g., by plotting the corresponding f functions.

5.2.3 Remarks

Role of the latent variable: The latent variable z plays an important role in quantifying how each biomarker behaves. Each time-series's latent variable is used to predict the parameters of D polynomial functions and $f \circ \kappa$ maps from observation times onto the observed biomarkers. Time-series whose representation space z are *close* hail from the same subtype, and consequently manifest similar patterns in their biomarkers. This variation in z results in variation in the parameters Θ and therefore in variation

among the data as a function of the time-points.

Note this latent representation can be used for other settings beyond disease phenotyping, for example supervised prediction tasks. For example, when $P = 1$ and f is the identity (i.e., a linear function), some time-series might increase (positive slope) versus others that decrease (negative slope). The latent representation z captures subtypes by learning to predict the slope of the function that models variation among time-series.

As an illustration, in Figure 5-1(b) in the blue phenotype, f is the sigmoid function. Depending on the latent space, we could imagine a one-dimensional z where $z < 0$ represents the curve (and subtype) in blue and $z \geq 0$ to represent the curve in red. The value of δ_i indicates the degree of delayed entry associated with each time-series. The delayed entry from interval censoring is corrected by applying the scalar δ_i element-wise to X_i and then transforming it by f.

Choice of link function: Our current choice of link function f is motivated to mimic degenerative disorders wherein patient biomarkers gradually increase over time (denoting worse outcomes). There is precedence in prior work to restrict function forms that characterize how biomarkers behave to be monotonic. [241] However, we emphasize that the model is not restricted to only sigmoid or quadratic functions for the link function f used in the synthetic and clinical experiments. Our work permits alternative choices of f—assuming the choice is smooth and differentiable with respect to model parameters—and allows individual biomarkers to have different parameterizations.

Scalability: The runtime of SubLign is impacted by the grid search over model parameter δ_i. The corresponding lines 7-9 in Algorithm 1 have complexity $\mathcal{O}(NSF)$ where F is the complexity associated with a single forward pass of the inference network and the generative model, N is the number of examples, and $S = \delta^+/\epsilon$ is the number of time steps. The model practitioners may therefore balance computational resources with S. In our experiments, we found comparable performance for S as low as 5.

Real-world clinical data: SubLign is motivated by, and designed to capture, vari-

ation in clinical biomarkers while taking into account the challenges of clinical data. Observational healthcare data are often irregularly spaced, and contains missingness. The use of a continuous time model allows us to naturally handle the former issue since we only maximize the likelihood of data corresponding to time-points where they are observed. When a single biomarker is missing while others are observed, it may be marginalized out (by ignoring the corresponding loss term).

Accommodating different kinds of censorship: Equation 5.1 naturally characterizes delayed entry arising from left-censorship. SubLign can also accommodate right-censorship by reversing the sequence of time-series and applying Algorithm 1 (resulting in our ability to infer the degree of right censorship). When both left-censorship and right-censorship are present, it corrects for left-censorship explicitly (using δ) while right-censorship is implicitly accounted for since we only maximize the likelihood of data up to the point that we observe time-series.

5.3 Identifiability Under a Noiseless Model

While SubLign presents a viable, practical model for clustering and aligning censored time-series data, it is worth reflecting upon whether we can ever identify subtype and alignment from observational data. In what follows, we present theoretical conditions that show that there exists conditions under which the problem we study is identifiable.

5.3.1 Generative Process

We assume distinct time stamps for the M observations in X_i. The generative process we assume for Y_i, conditional on X_i, is:

$$\begin{aligned}
&\forall i = \{1, \ldots, N\},\ s_i \sim \text{Cat}(K), \\
&\forall m \in \{1, \ldots, M\},\ d \in D_{i,m}, \\
&y_{i,m}[d] = f(\kappa(x_{i,m} + \delta_i; \theta^P[s_i, d]))
\end{aligned} \tag{5.4}$$

where $s_i \in \{1, \ldots, K\}$ is the subtype for time-series i, $D_{i,m}$ denotes the set of all observations at time-step m for time-series i where $\forall i, m, |D_{i,m}| \leq D$, and δ_i is the delayed entry value. The link function $f : \mathbb{R} \to \mathbb{R}$ has no parameters whereas $\kappa : \mathbb{R} \to \mathbb{R}$ is an unknown polynomial function of degree $P \in \mathbb{Z}^+$. We denote the parameters of κ, for each subtype and dimension (e.g., biomarker), as $\theta^P \in \mathbb{R}^{K \times D \times (P+1)}$. We denote $\theta^P[s_i, d]$ as selecting the (s_i, d)-th vector of size $P + 1$ from the tensor θ^P. Similarly, $\theta^P[s_i]$ selects the s_i-th matrix of size $X \times (P + 1)$. We define θ_p as the p-th coefficient of any polynomial function parameter set θ.

By construction, we have that $s_i = s_{i'} \iff \theta^P[s_i] = \theta^P[s_{i'}]$, i.e., for two subtypes, the values of each time-series are described either by $y = f(\kappa(x; \theta^P[s_i]))$ or $y = f(\kappa(x; \theta^P[s_{i'}]))$.

We begin with a set of assumptions for identifiability.

Assumption 1. f is invertible, and $\kappa(x, \theta) = \theta_0 + \sum_{p=1}^{P} \theta_p x^p$ describes a family of polynomial functions in x with parameter θ and degree $P > 0$. The parameters of each subtype are unique.

Assumption 2. $M \geq P + 1$, i.e., for each patient time-series, there exists at least one of the D features where we observe at least $P + 1$ values.

Assumption 3. For each subtype s_k, there exists a time-series i whose alignment $\delta_i = 0$, meaning no delayed entry.

Theorem 1. Under assumptions 1, 2, 3 for the model in Equation 5.4, we can identify the time-delays $\delta_1, \ldots, \delta_N$. We can identify the polynomial coefficients θ^P up-to a permutation of its rows and columns and the identity of $s_1, \ldots, s_N$ up-to a permutation over K choices.

Proof sketch. Consider the case where we have a single biomarker for each patient. The proof is constructive; first we transform the data using the inverse of f resulting in a set of data drawn from polynomial equations. The polynomial coefficients may be estimated from the observed data; we can then find the roots of these polynomials and pick the smallest root. These roots exactly quantify the degree of delayed entry; i.e.,

they tell us how much each polynomial has been shifted by. We can correct each time-series for this shift, re-estimate polynomial coefficients from the shifted-polynomials and cluster them to reveal the underlying subtype identity for each time-series (and consequently each patient).

Algorithm 2 Procedure for the identification of model parameters

Input: Observation times $X \in \mathbb{R}^{N \times M}$, biomarkers $Y \in \mathbb{R}^{N \times M \times D}$, polynomial degree P, invertible function f
Output: θ^P, $\delta_1, \ldots, \delta_N$, $s_1, \ldots, s_N$ for each patient
Step 1: Transform the observed biomarkers; $Q = f^{-1}(Y)$
Step 2: Obtain time-shifts using a single biomarker;
a) For each patient i, estimate the parameters $\dot{\theta^1}_i$ of $\kappa(x; \dot{\theta^1}_i)$ using a single biomarker $((x_{i,1}, q_{i,1}), \ldots, (x_{i,M}, q_{i,M})$ via polynomial regression,
b) Compute up to P roots of polynomial $\kappa(x, \dot{\theta^1}_i)$ for each patient i as $R_i = \{r_1, \ldots, r_P\}$ and set $\xi_i = \min \text{Real}(R_i)$ where Real denotes the real part of (potentially complex) roots.
c) Estimate $\tilde{\theta^1}_i$ for polynomials in a *canonical position* using $((x_{i,1} - \xi_i, q_{i,1}), \ldots, (x_{i,M} - \xi_i, q_{i,M})$ via polynomial regression,
d) Cluster $\tilde{\theta^j}_i$ across patients via K-means clustering to yield cluster identities $s_1, \ldots, s_N$
e) $\forall k, \;\; \eta_k = \min\{\xi_i \,|\, i \text{ s.t. } s_i = k\}$ and $\forall i, \;\; \delta_i = \xi_i - \eta_{s_i}$
Step 3: Estimate true polynomial coefficients using shifted observation times;
for biomarker $j = 1$ **to** J **do**
    For each patient, estimate the parameters θ_i^j of $\kappa(x; \hat{\theta^j}_i)$ using $((x_{1,1} - \delta_i, q_{1,1}[j]), \ldots, (x_{1,M} - \delta_i, q_{1,M}[j])$ via polynomial regression,
end for
Return $\theta^P = [\theta^1 | \ldots | \theta^J], \{\delta_1, \ldots, \delta_N\}, \{s_1, \ldots, s_N\}$

Proof. The proof is constructive; i.e., we give an algorithm for the identification of the parameters of the model in Equation 5.4. The algorithm for identification is presented in Algorithm 2 and proceeds in three steps.

Step 1: The first step transforms the observed biomarkers by applying the inverse of function f, which exists by Assumption 1. This leaves us with data as:

$$f^{-1}(y_{i,m}) = \kappa(x_{i,m} + \delta_i; \theta^P_{s_i}) \;\; \forall i \in N, m \in M$$

i.e., for all bio-markers, across all patients, we have data arising from different poly-

nomial functions.

Step 2: Without loss of generality, the second step uses the first biomarker to identify the values of δ_i for each patient.

a) First, we estimate the polynomial coefficients for each patient separately; we are guaranteed exact recovery of the coefficients by Assumption 2.

b) Next we find the roots for each polynomial. If they are complex, consider their real part, and define ξ_i to be the smallest root of the polynomial. At least one (real or complex) root is guaranteed to exist by the Fundamental Theorem of Algebra for every non-constant polynomial (Assumption 1). Note that the choice of using the smallest root is arbitrary; what matters is that a consistent choice of root is selected for each patient's polynomials.

c) The goal of this step to learn a new polynomial for each patient which is translated to ensure that the root selected in step b) lies at $x = 0$.

To do so, we first shift the observational time-steps by ξ_i, and we re-estimate the coefficients of each *shifted* polynomial.

We make use of the fact that if ξ_i is the smallest complex root of a polynomial $\kappa(x)$ then the polynomial $\kappa(x + \xi_i)$ has its smallest complex root at 0. We can recover the parameters of this polynomial exactly by shifting our observations and re-estimating the coefficients.

This operation recovers the coefficients of every patient's polynomial in its *canonical position* i.e., a translated polynomial whose smallest root (or its real component) is at $x = 0$.

This step can be viewed as a de-biasing step which allows us to re-estimate $\tilde{\theta}$ without while ignoring the effect that left-censorship has on parameter estimates.

d) We cluster the coefficients estimated in step c). By construction, we know that $s_i = s_{i'} \iff \theta_i = \theta_{i'}$ which guarantees that clustering recovers the true-underlying subtype for each patient (up to a permutation over K choices).

e) Finally we stratify patients by their subtype, and we define δ_i as the difference between their smallest root and the smallest value of ξ_i among all other patients within that subtype.

   By Assumption 3, we know that for each subtype, there exists a patient for whom $\delta_i = 0$, this reference patient will also be the one whose polynomial has the smallest root. We note here that without Assumption 3, we would still have identification of δ_i up to a constant.

   Therefore, by shifting each patient's smallest root by their reference patient's smallest root, we can recover the original time-shifts.

Step 3: Given the values of $\delta_1, \ldots, \delta_N$ from Step 2, we can now estimate the true values of the polynomial coefficients exactly in the noiseless setting via polynomial regression. □

Remarks: Theorem 1 describes conditions under which delayed entry and the polynomial parameters of cluster biomarker progression are identifiable. This encouraging result demonstrates scenarios where the parameters of the model in Equation 5.4 can provably be identified.

On the assumptions for identification: It is possible to relax Assumption 3 to only require the existence of a single time-series from each subtype; this modification only allows identifiability of $\delta_1, \ldots, \delta_N$ up to a translation within each subtype. The above result relies on the existence of at least one biomarker for which there are sufficiently many observations – this is a reasonable assumption in the context of clinical data since there is often a *canonical* biomarker tracked over time for each disease.

On the strategy for identification: We conjecture our analysis for identification of subtype and alignment is of independent interest for identifying causal effects in survival analysis where an important challenge is how to handle confounding that jointly affects both survival time and censorship. Related work in this field [267, 61] has focused on restricting the class of models used to characterize the survival function. Our work presents distinct parameteric assumptions towards this goal.

5.4 Experiments

5.4.1 Datasets

Synthetic

We generate two classes of synthetic datasets from the *sigmoid* and *quadratic* parameterizations in the 'SubLign: Subtype & Align' section. For the sigmoid dataset, we generate data from $K = 2$ subtypes and $\forall i, m, |D_{i,m}| = 3$ biomarker dimensions. For the quadratic dataset, we generate data with $K = 2$ and $\forall i, m, |D_{i,m}| = 1$. See appendix for the data generation process for the six quadratic datasets.

In both synthetic settings, we sample $N = 1000$ patients with $M = 4$ observations, variance $\sigma^2 = 0.25$, and max disease stage $T^+ = 10$. For each patient, we sample subtype $s \sim \text{Bern}(0.5)$. The true disease stage is drawn $t_m \sim \text{Unif}(0, T^+)$ for observation $m \in \{1, \dots, M\}$. The biomarker values are drawn $y_m \sim N(\lambda_m, \sigma^2)$ where $\lambda_m = \sum_{k \in \{1,\dots,K\}} 1(s_i = k) f_k(t_m)$. For the sigmoid dataset, the first subtype generating function across three dimensions is $f_1(t) = [\sigma(-4+t), \sigma(-1+t), \sigma(-8+8t)]$ and the second subtype generating function is $f_2(t) = [\sigma(-1+t), \sigma(-8+8t), \sigma(-25+3.5t)]$. The observed disease time x_m is shifted such that the first patient observation is at time 0. Therefore $x_m = t_m - \zeta$ where $\zeta = \min_{j \in \{1,\dots,M\}} t_j$ is the earliest true disease time for the patient.

5.4.2 Hyperparameters and Baselines

We find optimal hyperparameters via grid search. For both synthetic and clinical experiments, we search over hyperparameters including dimensions of the latent space z (2, 5, 10), the number of hidden units in the RNN (50, 100, 200), the number of hidden units in the multi-layer perceptron (50, 100, 200), the learning rate (0.001, 0.01, 0.1, 1.), regularization parameter (0., 0.1, 1.), and regularization type (L1, L2). We select the hyperparameter configuration with the best validation loss, as measured by Equation 5.3. Our models are implemented in Python 3.7 using PyTorch [234] and are learned via Adam [173] on a single NVIDIA k80 GPU for 1000 epochs. We

set alignment extrema $\delta^+ = 10$ based on the maximum of the synthetic dataset and the maxima of the HF and PD datasets. We search over 50 time steps with $\epsilon = 0.1$. We initialize the clustering with k-means++ [16].

For all models, we run for 1000 epochs and use the model with the best training loss over the 1000 epochs for evaluation. For the sigmoid dataset, the optimal hyper-parameters are latent space of dimension 5, 100 hidden units in the RNN, 50 hidden units in the multi-layer perceptron, learning rate of 0.01, and no regularization.

We compare to seven different baselines. Our greedy baseline, denoted as KMeans+Loss, first clusters the observed values using k-means clustering. Then, using the inferred labels s, we simultaneously learn θ_k for each subtype and δ_i for each patient by minimizing:

$$\arg\min_{\theta,\delta} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{d=1}^{D} \sum_{k=1}^{K} 1[s_i = k][y_{i,j,d} - f(x_{i,j,d} + \delta_i; \theta_k)]^2 \tag{5.5}$$

using Broyden–Fletcher–Goldfarb–Shanno. This naive clustering based approach (in the space of the original data) attempts to correct for shifts in time. We also compare to:

1. SubNoLign: a modified SubLign with no alignment value. This model is comparable to [318], which learns a deep patient representation while controlling for model architecture
2. SuStaIn [313]: a subtype and stage inference algorithm for disease progression,
3. BayLong: a Bayesian model of longitudinal clinical data [153],
4. PAGA [307]: a state-of-the-art single-cell trajectory pseudo-time method,
5. Clustering with dynamic time warping (DTW) using kernel methods [74] and soft-DTW [75],
6. SPARTan: A tensor factorization based approach for phenotyping from time-series data [237]

We detail baseline implementations below.

SuStaIn

SuStaIn [313] is a disease progression algorithm that recovers subtype and stage from cross-sectional data. We transform our longitudinal data by dropping patient affiliation across visits. We transform the data by subtracting the mean for each feature and dividing by the standard deviation for each feature. We assume the Z-scored values have a max of 5. We run for 1,000,000 epochs for the Markov Chain Monte Carlo sampling and 1,000 epochs for optimization. We use an open source implementation by the authors: `https://github.com/ucl-pond/pySuStaIn`

Bayesian Approach

The Bayesian approach [153] assumes longitudinal data, but there must be a small number of measured time points. We assume that there are 10 observed time points where observed data can begin as well as a window of 10 time points before the observed window where a patient's values can be aligned to. Because biomarker values are scaled between 0 and 10, we assume that values change between time points based on a Gaussian with $\sigma = 2$ and that subtype means for each time point are drawn from a Gaussian with $\sigma = 5$. We draw 4000 samples and use the maximum a posteriori estimate to determine stage and subtype for test patients. Because we could not find an open-source option, we implemented the algorithm ourselves based on the description in the paper.

Dynamic Time Warping

Dynamic time warping (DTW) defines similarity between time series that can be combined with clustering techniques. DTW methods include using soft-DTW [75] and kernel [74] before using K-means with the chosen similarity metric. We use open source implementations of DTW algorithms to generate our baseline comparisons: `https://pypi.org/project/dtw-python/`

PAGA Partition-based graph abstraction, or PAGA, [307] assumes cross-sectional data, so we create separate visits for each patient visit. For algorithm parameters,

we set resolution to 0.05, number of neighbors to 15, and connectivity cutoff of 0.05. We use an open source implementation by the authors:
`https://github.com/dynverse/ti_paga/blob/master/run.py`

Tensor factorization Sparse tensor factorization has been used for disease phenotyping. The decomposition of large and sparse datasets using canonical polyadic decomposition can create an interpretable output for phenotyping. We use the Matlab open source implementation of SPARTan: `https://github.com/kperros/SPARTan` We found these baseline results to yield poor clustering performance despite aggressive hyperparameter tuning. We surmise this is because transforming our data from continuous to discrete time resulted in a very large and extremely sparse matrix factorizing which is a tricky optimization problem.

5.4.3 Evaluation

We evaluate models on 5 trials, each with a different randomized data split and random seed. For each trial, we learn on a training set (60%), find the best performance across all hyperparameters on the validation set (20%), and report the performance metrics on the held-out test set (20%). The same data folds are used across all models for each trial.

We report the performance on the test set over three metrics.

1. Adjusted Rand index (ARI) measures whether pairs of samples are correctly assigned in the same or different subtypes. [152]

2. The Swaps metric reports the number of swaps needed to sort the predicted disease times into the true disease stages, expressed as percentage of total possible swaps. For true sorted alignment values $a_1, \ldots, a_N$ for N patients, we define the swaps metric $\mathcal{S}$ of proposed alignment values $b_1, \ldots, b_N$ as the number of swaps needed to sort the predicted disease times into the true disease stages, expressed as percentage of total possible swaps.

Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	**0.94 ± 0.02**	**0.09 ± 0.00**	**0.85 ± 0.04**
SubNoLign	0.81 ± 0.21	–	–
KMeans+Loss	0.67 ± 0.04	0.21 ± 0.03	0.49 ± 0.01
SuStaIn [313]	0.66 ± 0.02	0.16 ± 0.00	0.30 ± 0.02
BayLong [153]	0.19 ± 0.18	0.48 ± 0.00	0.01 ± 0.02
PAGA [307]	0.32 ± 0.05	0.52 ± 0.07	0.04 ± 0.20
Soft-DTW [74]	0.06 ± 0.01	–	–
Kernel-DTW [81]	0.06 ± 0.07	–	–
SPARTan [237]	0.22 ± 0.18	–	–

Table 5.1: Means and standard deviations over 5 trials for synthetic sigmoid dataset with 1000 patients, 3 dimensions, and 4 observations per patient.

$$\mathcal{S} = \frac{\sum_{i,j;i<j} 1(a_i < b_i, a_j > b_j)}{N(N-1)/2}$$

3. The Pearson correlation coefficient expresses correlation between the predicted and true disease stage.

ARI measures the clustering performance while the Swaps metric and the Pearson correlation coefficient quantify how well the learning algorithm infers the alignment values.

5.4.4 Statistical Significance

We evaluate held-out performance over 5 trials. Each trial consists of randomized 60/20/20 training/validation/test data folds and a different random seed. In order to compare models across 5 trials, we report the means and standard deviations from the 5 trials.

5.5 Results

5.5.1 Recovering Subtypes with Interval Censoring

SubLign is able to recover subtypes despite interval censoring, outperforming all baselines. For sigmoid synthetic data (Figure 5.1, ARI column), SubLign can recover subtypes (mean ARI of 0.94) better than the KMeans+Loss baseline (0.67) which assumes a greedy approach. Not correcting for alignment time decreases the quality of inferred subtypes as can be seen in SubNoLign (0.81).

Some baselines appear to suffer because SubLign leverages the longitudinal nature of patient data compared to the cross-sectional assumptions of PAGA and SuStaIn. Other baselines have strong priors, i.e., [153], which may explain its poor performance. Dynamic time warping methods appear to perform poorly for datasets with few observations or high missingness rates. Lastly, the tensor factorization method [237] fails, likely because transforming data from continuous to discrete time results in a very large and extremely sparse matrix factorization that is a tricky optimization problem.

We include additional results in the appendix, including visualizations of the SubLign subtypes compared to baselines, model misspecification analysis, and experiments varying the level of missingness.

5.5.2 Recovering Known Alignment Values

For the synthetic sigmoid data, SubLign outperforms baselines in inferring alignment values (Figure 5.1(a), Swaps and Pearson columns). SubLign recovers alignment values better according to the Swaps metric (mean value of 0.09) and the Pearson metric (mean value of 0.85) compared to the next best baselines of KMeans+Loss and SuStaIn. Note that many baselines only recover subtypes and do not learn patient alignment values.

5.6 Discussion

We study the task of clustering interval-censored time-series data. We present our method, SubLign, to learn latent representations of disease progression that correct for temporal misalignment in real-world observations and consider conditions for identifiability of subtype and alignment values. Empirically, our method outperforms seven baselines, and analysis of subtypes reveals clinically plausible findings. Better modeling of disease heterogeneity through alignment can help clinicians and scientists to better understand and predict how chronic diseases with many subtypes may progress. We hope that our model—in learning a continuous latent space to model heterogeneity—may be applied to other domains where subtypes and temporal alignment are entangled, for example gene expression analysis [20] or cancer pathways. [309]

Our model introduces directions for future work. Practically, SubLign assumes that δ and z are marginally independent. Intuitively, this means that, across all subtypes of a disease, the time at which the patient enters the study cohort is independent of any other factor. There are certainly cases where this assumption may easily be violated, and therefore it remains an important area for further improvement e.g., [167].

One area of interest is how to incorporate conditioning δ on features that are predictive to its value (e.g., clinical history) or other complexities, such as differences in treatment effect. [18] Our results for theoretical and noiseless identification do not naturally extend to the noisy setting since root finding without further assumptions can be sensitive to noise in the coefficients of the polynomials. Alternative strategies for identification that generalize our result are fertile ground for future work.

More broadly, this work contributes to developing clinical models that are robust to real-world factors [120], biases that might affect patient interactions with the healthcare system [56], and the practicalities of clinical care. [246]

Chapter 6

Chronic Disease Progression Subtyping

6.1 Introductions

Interval censoring in healthcare data presents a significant challenge for disease phenotyping. Left-censorship in clinical datasets can occur when patients have *delayed entry*, meaning data are unavailable before a diagnosis or first hospital visit. Factors including geographic proximity to a hospital [49], financial access to care [215] or mistrust of the healthcare system [33] can affect when a patient seeks medical help and consequently the beginning of data availability with respect to the underlying progression of their disease. Other datasets align patients by death time, but right-censorship restricts sample size to patients who have died. Since many factors affect mortality, other causes of death may confound results. For heart failure, a chronic disease that progresses over many years with a large range of onset ages and survival outcomes, interval censoring can confound attempts to analyse disease heterogeneity using observational data.

Many diseases are biologically heterogeneous despite a common diagnosis—for example autism [84], heart failure [269], diabetes [290], and Parkinson's disease. [97] The variation in biomarkers (e.g., glucose or creatinine) across patients can stem from different patient subtypes that manifest in distinct disease trajectories. Scientists seek to

understand this disease heterogeneity by identifying groups of people whose biomarkers behave similarly. For example, cardiologists use a measurement called ejection fraction as a heuristic to separate heart failure patients into two categories [231], with at least one of the two subtypes believed to be heterogeneous. [269] Similarly, neurologists studying Parkinson's disease (PD) have raised similar phenotyping. [207] To better understand patient heterogeneity, clinicians may turn to longitudinal, observational, and often irregularly-measured patient data for disease phenotype discovery. Better and more accurate disease phenotyping can lead to improved understanding of the disease, patient prognosis, and targeting of clinical trials.

6.1.1 Contributions

On two real-world observational clinical datasets—Parkinson's disease and heart failure—we discover disease subtypes. We correct for potential patient delays in entry to the dataset using SubLign, an algorithm outlined in Chapter 5. For Parkinson's disease, we uncover subtypes between healthy controls and patients with Parkinson's disease which outperforms a suite of baselines and matches known clinical findings. For heart failure, we demonstrate in a semi-synthetic setting that we are able to extract the degree of patient delay in entry to a reliable degree. We also find heart failure subtypes that align with recent clinical literature from the last five years.

6.1.2 Related Work

Clinicians and scientists learn disease subtypes to better understand heterogeneity in disease progression in a process known as disease phenotyping. Existing approaches often rely on the assumption that the observed measurements are aligned — and therefore not censored. Researchers then apply clustering techniques like hierarchical clustering of time series [84], affinity clustering [202], or matrix factorization. [290, 237] Other models define disease subtypes as stages of disease progression. [9] For this chapter, building on methods outlined in Chapter 5, we define disease subtypes as distinct from disease stage and jointly learn both.

6.2 Subtyping Parkinson's Disease Patients

6.2.1 Data

We use publicly-available data from the Parkinson's Progression Markers Initiative (PPMI), an observational clinical study, totalling $N_t = 423$ PD patients and $N_c = 196$ healthy controls where $N = N_t + N_c$. We extract four biomarker measurements of autonomic, motor, non-motor, and cognitive ability from $N = 619$ total participants with $M = 17$ maximum observations per patient. Our baseline data include 25 features including demographic information and patient history used to validate subtypes. For PD patients, the first recorded visit is within 2 years of the patient's PD diagnosis. Measurements are scaled between 0 and 1 with larger values corresponding to more abnormal values. We use the sigmoid parameterization of the SubLign model for both datasets because HF and PD are chronic and incurable diseases.

6.2.2 Experiments

For PD, we report the held-out clustering performance for healthy control patients and patients with PD. We use disease status (PD patient or healthy control) as labels and $K = 2$ subtypes.

Hyperparameter Selection

For the Parkinson's disease dataset, we searched on a slightly smaller set of hyperparameters for SubLign and found optimal hyperparameters of $\beta = 0.01$, no regularization, 10 latent dimensions, 10 hidden units for the multi-layer perceptron, 200 units for the recurrent neural network, and learning rate of 0.1.

6.2.3 Results

Parkinson's disease. Biomarkers used to track PD are self-reported, which can be biased, subjective, and noisy. From these biomarkers, SubLign discovers subtypes

that match known clinical findings on two cohorts: 619 combined PD and healthy control patients, and 423 PD patients.

For PD and healthy control patients, we run SubLign with $K = 2$ to uncover characteristics of the two known groups. SubLign subtype A clearly corresponds to healthy controls whereas subtype B designates PD patients. Statistically significant baseline features include all components of the University of Pennsylvania Smell Identification Test (UPSIT), which is a measure of smell dysfunction and highly linked to PD [131], and having a full sibling or biological dad with PD, which aligns with research suggesting PD may be hereditary. [175]

For PD patients only, we set $K = 3$ in SubLign to explore and discover potential disease heterogeneity. Statistically significant baseline features include race and gender, which parallel recent clinical findings about heterogeneity in PD manifestation [277, 87] and indicate new potential areas for future work. See appendix for tables of statistically significant baseline features stratified by the discovered subtypes for HF and PD.

SubLign recovers known subtypes in the PD dataset with statistically significantly higher ARI over baselines (see Figure 6.1(b)). When ARI performance intervals overlap, we use a t-test on pairwise differences over trials to compute statistical significance.

6.3 Subtyping Heart Failure Patients

6.3.1 Data

We use electronic health records from Beth Israel Deaconess Medical Center, a large health system in the Boston, Massachusetts in the United States. We identify patients who enter the emergency department with a diagnosis of HF and extract echocardiogram values—measurements from an ultrasound of the heart—from the full patient history. We include echocardiogram features that are present in more than 60% of echo studies. Our dataset includes $N = 1534$ patients and $|D_{i,m}| \leq 12; \forall i, m$ features

with $M = 38$ maximum observations per patient over a potential span of 10 years in the dataset. We extract 27 baseline features including race, sex, and comorbidities (e.g., renal failure) to validate subtypes only and not for use in the model. The dataset values are linearly scaled such that values are between 0 and 1 with larger values denoting more abnormality.

6.3.2 Methods

Applying SubLign onto real-world clinical data has several technical challenges. First, clinical data can be noisy, incomplete, or missing based on patient interactions with the healthcare system. Additionally, patient trajectories may not be aligned. As described in Section 5, patients may enter a dataset in different times due to factors including geographic proximity to a hospital or medical mistrust. Lastly, the research task is unsupervised, meaning evaluation against ground-truth labels for either clusters or alignment values is not possible.

Statistical significance.

We evaluate held-out performance over 5 trials. Each trial consists of randomized 60/20/20 training/validation/test data folds and a different random seed. In order to compare models across 5 trials, we report the means and standard deviations from the 5 trials.

6.3.3 Experiments

Missing values

SubLign allows for missing biomarker dimensions and missing patient visits to accommodate the sparsity of clinical data. For missing visits, we adapt the recognition network to handle variable sequence lengths. We mask out missing observations so they have no contribution to the learning stage, except for the recognition network input. We linearly interpolate missing values for each patient only for recognition

network input. For baselines that cannot handle missing data, we linearly interpolate missing values for each patient.

Semi-Synthetic Experiment

Because real world data often lack ground truth labels for subtype or alignment, we create two semi-synthetic experiments with clinical datasets. For HF, we evaluate SubLign's ability to infer relative disease stage by introducing additional censoring into the test sets. Specifically, we train SubLign using 80% data (train and validation data) as usual. We then modify the remaining data (20%) by removing the first year of patient observations, creating distorted test set (X', Y'), and by removing the last year of patient observations, creating (X'', Y''). The same amounts of observations are removed from each set to control for length of observations. We infer alignment values using the trained SubLign model: δ' from (X', Y') and δ'' from (X'', Y''). By construction, $\delta' > \delta''$. We report the percentage of patients for which SubLign is able to recover this relationship.

Hyperparameter Selection

For the heart failure dataset, we searched on a slightly smaller set of hyperparameters for SubLign and found optimal hyperparameters of $\beta = 0.001$, no regularization, 10 latent dimensions, 20 hidden units for the multi-layer perceptron, 50 units for the recurrent neural network, and learning rate of 0.01.

Baselines

We compare to seven different baselines. Our greedy baseline, denoted as KMeans+Loss, first clusters the observed values using k-means clustering. Then, using the inferred labels s, we simultaneously learn θ_k for each subtype and δ_i for each patient by minimizing:

$$\arg\min_{\theta,\delta} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{d=1}^{D} \sum_{k=1}^{K} 1[s_i = k][y_{i,j,d} - f(x_{i,j,d} + \delta_i; \theta_k)]^2 \tag{6.1}$$

using Broyden–Fletcher–Goldfarb–Shanno. This naive clustering based approach (in the space of the original data) attempts to correct for shifts in time. We also compare to:

1. SubNoLign: a modified SubLign with no alignment value. This model is comparable to [318], which learns a deep patient representation while controlling for model architecture
2. SuStaIn [313]: a subtype and stage inference algorithm for disease progression,
3. BayLong: a Bayesian model of longitudinal clinical data [153],
4. PAGA [307]: a state-of-the-art single-cell trajectory pseudo-time method,
5. Clustering with dynamic time warping (DTW) using kernel methods [74] and soft-DTW [75],
6. SPARTan: A tensor factorization based approach for phenotyping from time-series data [237]

Evaluation

We evaluate models on 5 trials, each with a different randomized data split and random seed. For each trial, we learn on a training set (60%), find the best performance across all hyperparameters on the validation set (20%), and report the performance metrics on the held-out test set (20%). The same data folds are used across all models for each trial. We report the performance on the test set over three metrics: ARI, Swaps metric, and the Pearson correlation coefficient. See Chapter 5 for more information about baselines and evaluation techniques.

6.3.4 Results

Semi-Synthetic HF Experiment

Although the real-world clinical datasets do not contain true alignment values, we use the previously described HF setup with an artificially censored test set. We find that SubLign predicts known alignment relationships in an altered dataset. When

Model	ARI
SubLign	**0.58 ± 0.12**
SubNoLign	0.42 ± 0.14
KMeans+Loss	0.05 ± 0.04
SuStaIn [313]	0.12 ± 0.11
BayLong [153]	0.04 ± 0.17
PAGA [307]	0.02 ± 0.02
Soft-DTW [74]	0.46 ± 0.43
Kernel-DTW [81]	0.21 ± 0.36
SPARTan [237]	0.15 ± 0.10

Table 6.1: Means and standard deviations over 5 trials for 619 patients in the PPMI dataset including 423 Parkinson's disease patients and 196 healthy controls.

evaluated on the manipulated test data, SubLign recovers the constructed relationship of $\delta' > \delta''$ with a higher performance ($71\% \pm 2\%$) over five trials compared to K-Means+Loss ($57.8 \pm 4\%$) and SuStaIn ($53.8 \pm 3\%$). See appendix for full results.

Clinical Insights from Correcting for Misalignment

We validate SubLign subtypes learned from the HF and PD datasets using known clinical findings. In Figure 6-3, we show the statistically significant baseline features from our SubLign results on heart failure patients. In Figure 6-1, we show the statistically significant baseline features for our SubLign results on both Parkinson's disease patients and healthy controls. In Figure 6-2, we show the baseline features found when using SubLign on only Parkinson's disease patients. Supplementary material Chapter C includes a table of baseline features, which are not included as input to SubLign, with statistically significant differences in subtypes: 7 features (out of 26) for PD and 11 features (out of 27) for HF.

Heart failure. Cardiologists classify patients into two groups based on ejection fraction: HF with reduced ejection fraction (systolic HF) and HF with preserved ejection fraction (diastolic HF). However, in our HF dataset, over 30% of patients correspond to neither group based on clinical diagnosis. When evaluating SubLign, we set $K = 3$, one more than the number of known groups, to explore a new subtype

and to potentially better understand heterogeneity in the existing ejection fraction classifications. [269]

Without ground truth subtype labels, we observe that SubLign finds systolic HF and diastolic HF as statistically significant baseline features. Of the three subtypes, subtype C corresponds to systolic HF, and subtype A and B correspond to diastolic HF, mirroring known clinical heterogeneity in diastolic HF. [269] Of the two diastolic HF subtypes, subtype A has a higher proportion of women while subtype B has a higher rate of obese patients, both subgroups with documented heterogeneity in diastolic HF. [87, 275] In contrast, the subtypes found by the KMeans+Loss baseline do not include known systolic HF and diastolic HF as statistically significant features. See appendix for full results.

Feature	A (321)	B (298)
Biological Dad With PD	0.02	0.06
Sibling With PD	0.01	0.05
UPSIT Part 1	7.55	5.49
UPSIT Part 2	7.64	5.69
UPSIT Part 3	6.98	5.23
UPSIT Part 4	7.53	5.62
UPSIT Total	29.73	22.05

Figure 6-1: Subtypes found by SubLign for 619 Parkinson's disease patients and healthy controls. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ with a Benjamini-Hochberg correction are listed.

Feature	A (156)	B (112)	B (155)
Male	0.51	0.68	0.79
White	0.98	0.95	0.92
UPSIT Part 1	6.01	5.07	5.43
UPSIT Part 2	6.65	5.30	5.52
UPSIT Part 3	5.92	4.79	4.90
UPSIT Part 4	6.28	5.20	5.42
UPSIT Total	24.87	20.36	21.26

Figure 6-2: Subtypes found by SubLign for 423 Parkinson's disease patients only. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ with a Benjamini-Hochberg correction are listed.

Feature	A (321)	B (298)
Biological Dad With PD	0.028	0.068
Sibling With PD	0.010	0.058
UPSIT Part 1	7.558	5.493
UPSIT Part 2	7.648	5.695
UPSIT Part 3	6.988	5.238
UPSIT Part 4	7.539	5.624
UPSIT Total	29.73	22.05

Table 6.2: Subtypes found by SubLign from Parkinson's disease patients and healthy controls using sparsely collected biomarkers. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ are listed.

Feature	A (674)	B (444)	C (416)
Age	75.98	74.73	69.43
Female	0.71	0.23	0.43
Anemia	0.23	0.16	0.14
Atherosclerosis	0.28	0.34	0.40
Atrial Fibrillation	0.44	0.55	0.43
Chronic Kidney Disease	0.27	0.34	0.34
Diastolic Heart Failure	0.50	0.36	0.06
Obese	0.56	0.65	0.46
Old Myocardial Infarction	0.12	0.14	0.24
Pulmonary Heart Disease	0.29	0.22	0.19
Systolic HF	0.09	0.27	0.53

Figure 6-3: Subtypes found by SubLign from heart failure patients using echocardiogram biomarkers. Only statistically significant means between subtypes according to an ANOVA test with $p < 0.05$ with a Benjamini-Hochberg correction are listed.

In comparison, we show the KMeans+Loss subtypes on the heart failure dataset (Table 6.3).

6.4 Discussion

There are numerous potential next steps to determine clinically interesting subtypes.

A natural next step would be to apply our methods to bigger and more heterogeneous datasets. For this chapter, we used datasets containing no more than a few thousand patients to study Parkinson's disease and heart failure. Applying these methods on a larger-scale would yield an opportunity to study chronic conditions

Feature	A (240)	B (802)	C (492)
Age	71.567	74.565	73.793
Hyperlipidemia	0.529	0.448	0.541
Chronic Kidney Disease	0.346	0.273	0.370
Esophageal Reflux	0.375	0.259	0.289
Pulmonary Heart Disease	0.367	0.204	0.256
Kidney Disease	0.254	0.200	0.278
Atherosclerosis	0.196	0.131	0.201
Anemia	0.217	0.163	0.213
Obese	0.688	0.500	0.608

Table 6.3: Heart Failure KMeans+Loss subtypes (patient counts in parentheses), described by mean baseline features. Only statistically significant features are listed and do not include systolic and diastolic HF, two known phenotypes of HF.

over a wider time-scale, with a larger set of features, and across a larger patient population. These methods could have a significant impact on better understanding much larger longitudinal health datasets, e.g., the All of Us Research Program. [229]

It is important to consider heterogeneity in treatment selection, which can complicate the subtyping process. Different clinicians may choose more aggressive or conservative approaches across their patients. Without explicit model, these differences may confound the inference process of learning the disease progression across subtypes. One simplistic modification would be to consider each new treatment protocol as a new subtype because of the potential resulting changes in disease progression. However, because small changes in treatment protocol can have large effects, the number of subtypes can grow quickly. More explicit modeling of treatment protocol may greatly improve our subtyping methods.

Lastly, we can broaden the type of health data used in the subtyping process. Currently SubLign uses continuous temporal data as input, usually assumed to be observational longitudinal data. We could consider in addition either static demographic data which is taken at baseline observation or with binary temporal data like diagnostic information which is measured throughout the disease progression. Traditionally, multimodal machine learning [19] allows for the fusion of different types of data, for example concatenating encoded representations, towards the goal of im-

proved predictive model performance. Adapting these techniques towards disease subtyping could be promising ground for future work.

Chapter 7

Early Detection of Intimate Partner Violence Using Radiology Reports

7.1 Introduction

Intimate partner violence (IPV) is defined as physical, sexual, psychological, or economic violence that occurs between former or current intimate partners. While men can also be affected, IPV is a gendered phenomenon largely perpetrated against women by male partners. [111] The Centers for Disease Control report that more than 1 in 3 women, and 1 in 10 men in the U.S. will experience physical violence, sexual violence, psychological violence, and/or stalking by an intimate partner during their lifetime. [29] IPV victims have a greater risk of health problems including higher rates of mental health illnesses, chronic pain, reproductive difficulties, and generally poorer health.[45, 281, 92] According to the United Nations, half of the women who are intentionally killed globally are killed by their intimate partners or family members. [230] It is essential to detect IPV victims early to provide timely intervention.

Healthcare providers have the opportunity to screen patients for IPV, but several barriers at both patient and provider levels limit the effectiveness. IPV victims often seek treatment within healthcare settings; [306] however, despite its high prevalence, IPV is substantially underdiagnosed due to underreporting of violence by the victim

to health care providers. Because IPV victims generally do not present with obvious trauma, even in emergency departments, [78] they do not readily receive IPV-specific resources.

Imaging studies provide an objective measurement of patient status, especially for vulnerable individuals who are not forthcoming. [260] In a prior observational study, researchers identified IPV-related injury patterns including soft-tissue and musculoskeletal injuries from imaging studies of victims who visited the emergency department. They also found that IPV victims receive more radiology studies than a comparable control cohort. [116]

7.1.1 Contributions

In this work, we present algorithms to predict IPV and injury from radiology reports. We predict IPV from a dataset of 24,131 radiology reports from 262 IPV victims who were referred to a violence prevention support program and 794 controls from the same hospital who were age and sex-matched based on a subset of the IPV victims. We demonstrate strong quantitative results with our best model achieves a mean area under the received operator curve (AUC) of 0.852. With a sensitivity of 95% and a specificity of 71%, we are able to predict IPV a median of 1.34 years in advance of entry into the violence prevent support program. To better detect severe forms of IPV, we predict injury from a dataset of radiology reports from only IPV victims with labels from four emergency radiology fellowship-trained radiologists. Our best model achieves a mean AUC of 0.887.

We analyze our models for validity and usability. Because IPV can manifest differently across race, [192] gender, [280] age, [253] and marital status, [265] we present error analysis comparing accuracy, sensitivity, and specificity across these groups using demographic information extracted from the clinical record. As IPV continues to affect vulnerable individuals—especially in times of great crisis [295, 128]—we demonstrate how automated predictive algorithms can be used to identify patients at high risk of IPV and injury.

7.1.2 Related Work

Intimate Partner Violence

Early detection in IPV is critical to facilitate early intervention in the cycle of abuse, thereby preventing worsening health conditions, [45, 281, 92] life threatening injuries, and potentially homicides. [271] The main obstacle to early intervention is underreporting by the patient due to variety of factors including shame, economic dependency, or lack of trust in healthcare providers. [144] Automated screening can help physicians identify high risk individuals—potentially from radiology studies [171], substance abuse disorders [180], or other clinical data—and intervene quickly.

Clinical Prediction

Machine learning methods can assess patients and other individuals for different levels of risk to allocate resources and improve clinical workflows. [121, 120] The strength of machine learning lies in its ability to learn latent patterns from observational data and make robust predictions on new and previously unseen patients. Researchers have shown promising results about the use of machine learning on chronic diseases like diabetes [251], rare conditions like preterm infant illnesses, [266] and public health concerns like child welfare. [63, 36] In particular, supervised learning models excel in structured settings with large datasets and clearly defined labels, e.g., radiology report text and whether the patient ultimate enters a violence prevention program.

Natural Language Processing

Natural language processing (NLP) techniques can extract information from unstructured text. [310] In healthcare settings, researchers have leveraged NLP on clinical text such as nursing notes, discharge summaries, and radiology and pathology reports for disease surveillance [169, 113], cohort creation [59, 6], prediction of adverse events [247, 151, 259], and diagnosis. [239, 31]

A promising new area of natural language processing research is the use of contextual word embeddings. Whereas traditional approaches represent text as a non-

sequential bag of words or a sequence of static word embeddings, more recent approaches construct unique representations for each word (or sub-word) depending on its surrounding context. For instance, the abbreviation “MS" may refer to mitral stenosis or multiple sclerosis depending on the surrounding context. BERT [80], RoBERTa [199], AlBERT [184], and numerous other recent models are pretrained on large amounts of text using language modelling objectives and then fine-tuned on a smaller task-specific dataset. Among other examples, large open-source clinical datasets [159] have enabled researchers to release clinical contextual word embedding models. ClinicalBERT is a publicly available BERT model initialized from BioBERT [187] and further trained on intensive care unit notes [10]

7.2 Data

We predict IPV using a dataset of IPV victims and age-matched control patients. We predict injury using a dataset of only IPV victims, with labels from emergency radiologists.

7.2.1 IPV Patient Selection

The study cohort consisted of victims who were referred to a large academic hospital's violence prevention support program between January 2013 and June 2018. For the early detection of IPV through IPV prediction, we randomly selected 265 women reporting physical abuse. We excluded all victims without any radiological studies from both groups or whose radiology report lacked clinically meaningful information after data cleaning. The final IPV dataset consists of 262 patients total.

For injury prediction, we examine a wider set of patients from two groups of victims referred to a large academic hospital's violence prevention support program between January 2013 and June 2018. For the first group, we randomly selected 940 victims out of 2948 reporting any type of IPV-physical, psychosocial, or sexual. The second group comprised of all 308 IPV victims (including 265 women) reporting physical abuse. We excluded all victims without any radiological studies from both

groups or whose radiology report lacked clinically meaningful information after data cleaning. The final IPV dataset consists of 685 patients total.

7.2.2 Control Group Selection

We age-matched against 265 women with physical abuse and filtered for patients with at least one radiology study that was not canceled. We selected the first 795 of the resulting 1006 patients to build our control cohort. Note that the control cohort was matched against the 265 female IPV victims and does not contain any men.

7.2.3 Injury Labels

The full set of radiological studies and reports of the injury prediction patient cohort were analyzed for the presence of injury for each study. Any radiological findings unrelated to potential physical injury such as pancreatitis, malignancy, subarachnoid hemorrhage due to aneurysm rupture, etc. were not recorded as "injury". All images were reviewed by four emergency radiology fellowship-trained radiologists who were aware of history of IPV but were blinded to the date of identification of IPV and clinical notes. The readers had full access to the radiology reports. The radiologists also recorded any injuries such as soft tissue swelling, rib fracture, etc. which might be overlooked or not mentioned in the original radiology reports. Each report was reviewed separately and labeled with an injury or not. Of the 15,639 radiology reports reviewed, 2.57% of them were found to have an injury.

7.2.4 Data Cleaning

For each radiology report, we remove extraneous information to improve clarity for the predictive models. We remove all header and footer information, punctuation, and line breaks. We change the text to lowercase and create tokens from each word through bag of words or clinicalBERT. [10] Radiology reports that lack meaningful information after this cleaning are removed from the dataset. Patients who do not have any radiology reports after this step are removed from the dataset completely.

7.2.5 Demographic Data

We extract demographic data from IPV victims and controls including age, gender, race, and marital status. To structure free-form responses for some fields, we consolidate each field into several categories. For age, we discretize the field into < 30, 30-50, 51-65, and 66+. The average age of patients in dataset is 43.8 ± 18.5, with IPV victims average age at 40.9 ± 13.3 and control population average age at 46.3 ± 4.7. For race, we consider white, Black, Hispanic, and "other" categories with patients allowed to belong to more than one group. For marital status, we categorize single, married, and other. Note that because our control population was sex and age-matched against a cohort of female IPV victims, our control population contains no men. We do not use demographic information for predictions and use only radiology reports. For summary statistics about the dataset, see Table 7.1.

		IPV Prediction			Injury Prediction		
		Total	IPV	Control	Total	Injury	No Injury
# Patients		1,056	262	794	530	135	395
# Radiology Reports		24,131	5,127	19,004	10,009	172	9,837
Age	< 30	6.8%	14.4%	4.8%	9.8%	10.5%	7.6%
	30-50	32.0%	47.4%	27.8%	53.6%	58.7%	58.6%
	51-65	37.2%	32.1%	38.5%	29.3%	21.5%	25.7%
	66+	23.9%	6.1%	28.7%	7.3%	9.3%	8.1%
Gender	Female	100.0%	100.0%	100.0%	93.4%	90.7%	94.8%
	Male	0.0%	0.0%	0.0%	6.6%	9.3%	5.2%
Race	Black	50.8%	34.6%	55.2%	28.3%	29.1%	23.6%
	Hispanic	12.0%	24.7%	8.6%	22.2%	19.2%	21.9%
	White	10.0%	29.4%	4.8%	38.7%	40.7%	43.6%
	Other	27.6%	11.6%	32.0%	11.6%	11.6%	12.0%
Marital Status	Single	45.0%	56.6%	41.8%	50.8%	57.0%	47.9%
	Married	36.1%	19.4%	40.7%	29.7%	27.3%	36.2%
	Other	18.9%	24.0%	17.5%	19.5%	15.7%	15.9%

Table 7.1: Summary statistics for dataset, with percentages of radiology reports.

7.3 Methods

7.3.1 Experiment Setup

We train our models on 60% of the patients, validate and select hyperparameters based on 20% of the patients, and report test performance on 20% of the patients. To avoid data leakage, we split our data based on patient rather than radiology study. Once a patient is assigned to train, validation, or test dataset, we assign all radiology reports and labels for that patient to the corresponding dataset. We perform analysis on five trials with shuffled splits of the data. All models are compared against the same five dataset splits.

7.3.2 Models

We compare two tasks and five models. We predict IPV and injury based on collected labels. We consider data from extracted demographic data, radiology reports, and a combination of the two. We use logistic regression, random forest, gradient boosted trees, neural network with bag of words representation, and neural network with clinicalBERT [10] representation.

For logistic regression, we search over hyperparameters of regularization constant $C = [0.001, 0.01, 0.1, 1., 2., 5.]$ and regularization type of L1 or L2. For random forest, we search over maximum depth of trees of 10, 50, 100, 500, or no maximum depth. For gradient boosted trees, we search over hyperparameters learning rate of 0.01, 0.1, 0.5, or 1 and maximum depth of 2, 3, and 4. We use the sklearn-learn Python package [235] with otherwise default settings.

We train two neural network models using the AllenNLP library. [114] Both models contain an embedding layer followed by two feed forward layers with rectified linear unit function and linear activations. The first model represents each note as a vector of word frequencies ("Bag of Words") projected down to a lower dimensional vector while the second model leverages clinicalBERT's contextual word embeddings to represent each note.

To facilitate more rapid training on CPUs, we freeze the clinicalBERT embeddings and only train the feed forward layers. The first model was trained for 40 epochs with an early stopping period of 5 epochs, and the second model was trained for 10 epochs due to computational constraints. Gradient norms were rescaled to a max of 5.0, and training examples were batched by note length to minimize excess padding. Hyperparameters were selected according to validation set performance, resulting in a learning rate of 0.001, weight decay of 0.0001 and batch size of 32 for both models.

7.3.3 Evaluation

Prediction and Predictive Features

We report the predictive performance as the area under the receiver operator curve (AUC) on the same train, validation, and test datasets for all models compared. We compute AUC means and standard deviations for the test datasets of the five shuffled splits of the data. We present predictive features by finding words with high feature importance. Because many compared models are non-linear, it is difficult to use interpretability methods to find predictive words. As logistic regression performance is comparable to that of other other non-linear methods (see Table 7.2), we present linear coefficients of the logistic regression across five test sets of the shuffled splits of the data.

Error Analysis

As clinical models are used in increasingly high stakes decisions, it is important that machine learning reduce health disparities [54] rather than amplify existing biases. [250] We audit our best prediction model for IPV and injury by comparing accuracies, sensitivity, and specificity for different subgroups, including age, race, gender, and marital status. [58, 52, 135] We compute means and standard deviations of accuracy, sensitivity, and specificity for each subgroup with overall model sensitivity set to 0.95. Predicted probabilities are computed for test datasets and compared to the true labels for the five shuffled splits of the data.

Report-Program Date Gap

One practical measure of IPV prediction is how much earlier can our model predict IPV compared to the date of patient's entry into a violence prevention program. For each radiology report, we compare the radiology report date with the entry date into the program. We call this difference in dates the *report-program date gap*. A radiology report with a large report-program date gap is one that occurs long before program entry whereas a low report-program date gap occurs shortly before program entry. A model that can make predictions with a high maximum report-program date gap per IPV victim would allow us to allocate resources and support to high risk individuals more efficiently. For each IPV victim, we compute the largest report-program date gap for which the model predicts IPV above a chosen threshold.

We select the prediction threshold to satisfy sensitivity constraints. A trivial way to maximize the metric would be to predict IPV for every patient in the dataset, which would yield redundant results. However, IPV prediction requires high sensitivity (true positive rate) since the clinical healthcare system can accommodate many false positives—e.g., offering a conversation with a social worker—whereas false negatives can be more dire—e.g., not providing an IPV victim with additional resources for help. Accordingly, we fix our sensitivity level to be at least 95% and compute the corresponding model threshold. We report the median earliest report-program date gap for all IPV victims for whom the model predicts correctly.

7.4 Results

7.4.1 IPV and Injury Prediction Performance and Predictive Features

We are able to predict IPV (best mean AUC of 0.852, random forest classifier) and injury (best mean AUC of 0.887, random forest classifier). For more results, see Table 7.2. We find that words that are most predictive for IPV and injury match clinical literature in IPV injury patterns from radiology reports. In Table 7.3, we

Model	IPV	Injury
Logistic Regression	0.841 ± 0.033	0.866 ± 0.016
Random Forest	**0.852 ± 0.022**	**0.887 ± 0.019**
Gradient Boosted Trees	0.842 ± 0.027	0.858 ± 0.030
Neural Network (Bag of Words)	0.849 ± 0.026	0.879 ± 0.010
Neural Network (clinicalBERT [10])	0.843 ± 0.022	0.852 ± 0.021

Table 7.2: Model AUC means and standard deviations over five data splits for IPV and injury prediction using radiology reports. Bold rows indicate best performance for task.

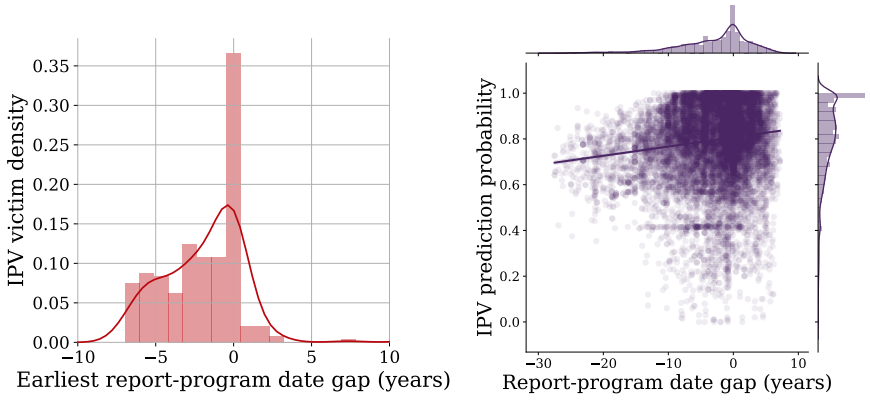


Figure 7-1: **Left:** Earliest possible report-program date gap per patient. **Right:** Scatterplot and marginal histograms of report-program date gap (x-axis) and IPV prediction probability (y-axis) for all radiology reports of IPV victims for random forest classifier. See Section 7.3.3 for definition of report-program date gap.

show words with highest feature importance from logistic regression for both tasks. Findings include soft-tissue abnormalities such as swelling and hematomas and musculoskeletal injuries such as fractures. These findings reflect prior research on IPV injury patterns. [116]

7.4.2 Error Analysis

We find differences in performance in subgroups of age, gender, race in our error analysis (see Table 7.4). We focus on sensitivity because in cases of IPV and injury, it is much more important to detect all true positives. In particular, older patients (51-65, 66+) have lower sensitivity for both IPV and injury prediction. Other groups have low sensitivity for either IPV or injury prediction, but not both. For example,

Task	Predictive words
IPV	ordering, final, trauma, hematoma, technique, swelling, cell, fracture, type, fractures, lymphoma, electronically, male, pancreatitis, reason, gms, implants, unresponsive, assault, none, cancer, pregnancy, mca
Injury	hematoma, fracture, fractures, swelling, trauma, subchorionic, foreign, ankle, third, hand, nondisplaced, fall, stab, phalanx, finger, deformity, skullbase, fifth, wound, laceration, sob, digit, measuring

Table 7.3: Predictive words for IPV and injury averaged across five trials based on linear coefficients of logistic regression. Underline indicates words consistent with clinical literature.

Black patients have lower sensitivity for injury prediction. White patients have low sensitivity for IPV prediction. It appears that patients who are not single or married (e.g., widowed, separated) have lower sensitivity for injury whereas married patients have lower sensitivity for IPV prediction.

7.4.3 Report-Program Date Gap

We can detect IPV from radiology reports much earlier than a patient's entry into a violence prevention program. We compute the *report-program date gap*, or the gap between radiology report date and entry date into the program. The dataset contains radiology reports many years before program entry (see Figure 7-1, left), and our best model for IPV prediction (see Table 7.2) with sensitivity threshold to 95% as explained in Section 7.3.3 yields a median report-program date gap of 1.34 years. For a visual representation of report-program date gap compared to predicted probabilities for IPV victims, see Figure 7-1 (right).

Most radiology reports occur close to the program entry date (x-axis marginal histogram), and the model predicts that nearly all the IPV victims in the dataset are IPV victims (y-axis marginal histogram). The increasing trendline in the scatter plot indicates that as the date of program entry nears, the model increases the IPV predicted probability.

		IPV Prediction			Injury Prediction		
		Accuracy	TPR	TNR	Accuracy	TPR	TNR
Age	< 30	$83 \pm 4\%$	$97 \pm 1\%$	$53 \pm 11\%$	$62 \pm 11\%$	$93 \pm 9\%$	$61 \pm 11\%$
	30-50	$87 \pm 1\%$	$96 \pm 1\%$	$49 \pm 5\%$	$54 \pm 12\%$	$94 \pm 1\%$	$52 \pm 13\%$
	51-65	$\mathbf{71 \pm 5\%}$	$92 \pm 2\%$	$49 \pm 2\%$	$41 \pm 18\%$	$\mathbf{89 \pm 3\%}$	$40 \pm 19\%$
	66+	$\mathbf{60 \pm 5\%}$	$\mathbf{84 \pm 2\%}$	$45 \pm 9\%$	$\mathbf{33 \pm 16\%}$	$98 \pm 4\%$	$31 \pm 17\%$
Gender	Female	$77 \pm 1\%$	$94 \pm 1\%$	$48 \pm 4\%$	$50 \pm 15\%$	$93 \pm 1\%$	$48 \pm 15\%$
	Male	—	—	—	$\mathbf{31 \pm 21\%}$	$96 \pm 4\%$	$28 \pm 21\%$
Race	Black	$\mathbf{72 \pm 2\%}$	$95 \pm 0\%$	$41 \pm 7\%$	$47 \pm 14\%$	$\mathbf{88 \pm 3\%}$	$46 \pm 14\%$
	Hispanic	$91 \pm 2\%$	$97 \pm 0\%$	$51 \pm 11\%$	$58 \pm 13\%$	$96 \pm 3\%$	$57 \pm 13\%$
	White	$84 \pm 1\%$	$\mathbf{90 \pm 2\%}$	$43 \pm 5\%$	$41 \pm 18\%$	$95 \pm 3\%$	$39 \pm 18\%$
	Other	$\mathbf{68 \pm 3\%}$	$98 \pm 0\%$	$55 \pm 5\%$	$58 \pm 13\%$	$95 \pm 6\%$	$57 \pm 13\%$
Marital Status	Single	$81 \pm 2\%$	$95 \pm 0\%$	$45 \pm 7\%$	$49 \pm 13\%$	$95 \pm 1\%$	$48 \pm 14\%$
	Married	$\mathbf{70 \pm 1\%}$	$\mathbf{92 \pm 2\%}$	$49 \pm 3\%$	$49 \pm 18\%$	$92 \pm 7\%$	$48 \pm 19\%$
	Other	$83 \pm 2\%$	$95 \pm 2\%$	$49 \pm 9\%$	$46 \pm 16\%$	$\mathbf{88 \pm 3\%}$	$45 \pm 17\%$

Table 7.4: Error analysis for IPV and injury predictions from random forest classifier. Means and standard deviations of accuracy, sensitivity (TPR), and specificity (TNR) computed over 5 data splits with overall model sensitivity set to 0.95. Bold indicates subgroups with particularly low metrics.

7.5 Discussion

We present a range of findings on the use of prediction algorithms to address IPV in the clinical setting through the analysis of radiology reports. Our results demonstrate several main takeaways. First, we are able accurately predict IPV and injury with AUCs of 0.852 and 0.887, respectively. Second, the linear coefficients of our models confirm known clinical findings about injury patterns for IPV victims. Lastly, while our algorithm demonstrates some bias in the form of differences in accuracy, sensitivity, and specificity with respect to age, gender, race, and marital status, we are able to predict a median report-program date gap of over 1.34 years with sensitivity of 95% and specificity of 71%.

Our work leads naturally to many directions for future research. One limitation of our current work is that we consider one radiology report at a time for IPV and injury prediction and exclude clinical history. Because IPV victims seek greater medical care from clinical settings like the emergency department, [306, 78] patient data including previous visits, clinical notes, and diagnoses could yield more accurate predictions and therefore earlier detection. [171] Additionally, predictive algorithms can help identify the best intervention for an IPV victim. Currently screening programs for IPV vary in execution and effect, [228] and once screened, IPV victims face many obstacles before leaving an abusive relationship. [172] Deeper understanding of targeted interventions could provide a crucial contribution to patient advocacy.

Deployment of a predictive model for IPV and injury detection faces several practical challenges. As with many machine learning algorithms in clinical settings, question of generalization across hospitals [120] and across subgroups [58] raise concerns about robustness and fairness. Moreover, better understanding of physician reliance on, distrust of, and confusion towards predictive models in clinical settings is an active area of research. [286]

We have shown in our analysis that automated detection through machine learning can predict IPV and injury from radiology reports. We look forward to future work towards the deployment of an IPV early detection model in a clinical setting.

Chapter 8

Conclusion

In this dissertation, we explored several approaches to using machine learning towards equitable healthcare. As we noted in Chapter 1, the data and technical challenges of machine learning approaches towards equitable healthcare are immense. While much work remains, we hope that such methods will contribute to improved decision making for patient care for all in the future.

We covered work that spans many facets of the machine learning development pipeline, from the post-deployment discrimination considerations described in Chapter 3 to better scoping of an understudied and under-reported problem in Chapter 7. In tackling these problems, our focus was on learning machine learning approaches that generalize across applications and address the challenges of health data, including missing, noisy, and potentially biased data. In general, our research methodology centered around the identification of sources of bias and inequity that can then inform actionable steps for model developments and clinicians.

We believe that the creation of machine learning methods to distill large amounts of heterogeneous health data into equitable clinical support will advance scientific understanding, improved clinical protocols, and improved human health.

Here are a few exciting opportunities for work in this vein to continue.

Overreliance on Group-Based Membership In this dissertation and in much of the contemporary work, the natural starting point for fairness considerations relies

on group memberships, e.g., self-reported patient race or gender or socioeconomic status. However, there are two main problems with these approaches. First, the overreliance on these group-based memberships mean that equity assessments can be brittle when the group membership is unavailable. Although these group memberships are inferred, e.g. through proxy variables like surnames, these approaches can have severe challenges. [161] Second, on a more fundamental level, the medical community has long struggled with the idea of how to use race in clinical settings. [240, 299]

Instead of focusing on large pre-specified groups, we could explore how to characterize heterogeneity in algorithmic performance in a more fine-grained method. As one example, *individual fairness*

It is essential to develop algorithms for the places of greatest need in the healthcare system. This thesis has focused on risk stratification or scientific discovery from longitudinal observational datasets as a necessary first step. As a natural next step, new work in this area should consider adaptive models to assist clinical decision making for changing dynamics in healthcare systems. For example, how can we update models based on changing patient populations? How can we ensure that underrepresented patient populations do not suffer from eroded model performance? How do we modify algorithmically-recommended treatment protocols based on new clinical domain expertise? Understanding areas of greatest statistical uncertainty in clinical systems can direct the development of relevant computational techniques, for example overcoming limited labeled datasets using weakly-supervised or self-supervised methods.

Shifting Power Towards Patient Another area of great promise is the newly available combined and multimodal health datasets. Each modality of health data contains additional signal—for example, the density of electronic health records, the consistency of health app data, and the depth of genomic data. Given a lack of representation in one modality, underserved patient populations may benefit from a wider range of data sources. Additionally, machine learning methods have proven strong in learning latent representations from datasets with different underlying structures.

Health data itself may have underlying structure; it may be cyclical (e.g., menstrual data), nonlinear (e.g., medical interventions), or non-monotonic (e.g., relapsing and remitting disorders). A multifaceted understanding of health across a patient's entire life can allow for more targeted interventions.

Practical Tools to Address Algorithmic Bias for Policy Makers A focus on real-world applications grounds all technical work towards algorithmic equity. When working on problems that affect humans in high-stakes settings, it is essential for model practitioners to deepen collaborations with policy markers to craft new and update outdated regulation. Beyond healthcare, fields including insurance, areas including employment, criminal employment, and education require specialized approaches towards algorithmic bias. How should policy makers audit algorithms for bias? How should legal precedents like disparate impact and disparate treatment adapt to algorithmic decision making? What is a reasonable burden of algorithmic equity on model developers? Clarity on these questions will have widespread impact.

Appendix A

Additional Information for Chapter 3

A.1 Testing for significant discrimination

In general, neither Γ nor $\overline{\Gamma}$ can be computed exactly, as the expectations $\gamma_a = \mathbb{E}_p[L(Y,\hat{Y}) \mid A = a]$ and $\overline{\gamma}$, for $a \in \mathcal{A}$ are known only approximately through a set of samples $S = \{(x_i, a_i, y_i)\}_{i=1}^m \sim p^m$ drawn from the (possibly class-conditional) population p. The Monte Carlo estimate,

$$\gamma_a^S(\hat{Y}) = \frac{1}{m_a} \sum_{i=1}^{m} L(y_i, \hat{y}_i) \mathbb{1}[a_i = a] \ ,$$

with $m_a = \sum_{i=1}^m \mathbb{1}[a_i = a]$, may be used to form an estimate $\Gamma^S(\hat{Y}) = |\gamma_0^S(\hat{Y}) - \gamma_1^S(\hat{Y})|$. By the central limit theorem, for sufficiently large m, $\gamma_a^S(\hat{Y}) \sim \mathcal{N}(\mu_a, \sigma_a^2/m_a)$ and $(\gamma_0^S - \gamma_1^S) \sim \mathcal{N}(\mu_0 - \mu_1, \sigma_0^2/m_0 + \sigma_1^2/m_1)$. As a result, the significance of $\Gamma^S(\hat{Y})$ can be tested with a two-tailed z-test or using the test of [308]. If sample sizes are small and the target binary, more appropriate tests are available [37]. In addition, we will often want to compare the discrimination levels $\Gamma(\hat{Y}), \Gamma(\hat{Y}')$ of predictors $\hat{Y}, \hat{Y}'$, resulting from different learning algorithms, models, or sets of observed variables. The random variable $|\Gamma^S(\hat{Y}) - \Gamma^S(\hat{Y}')|$ is not Normal distributed, but is an absolute difference of folded-normal variables. However, for any $\alpha \in \{-1, 1\}$, $Z_\alpha := \alpha(\gamma_0^S(\hat{Y}) - \gamma_1^S(\hat{Y})) - (\gamma_0^S(\hat{Y}') - \gamma_1^S(\hat{Y}'))$ is Normal distributed. Further, by enumerating the signs of $(\gamma_0^S(\hat{Y}) - \gamma_1^S(\hat{Y}))$ and $(\gamma_0^S(\hat{Y}') - \gamma_1^S(\hat{Y}'))$, we can show that $|\Gamma^S - \Gamma^{S'}| = \min_{\alpha \in \{-1,1\}} |Z_\alpha|$.

As a result, to reject the null hypothesis $H_0 : \Gamma = \Gamma'$, we require that the observed values of both Z_{-1} and Z_1 are unlikely under H_0 at given significance.

A.2 Additional experimental details

A.2.1 Datasets

- Adult Income Dataset [191]. The dataset has 32,561 instances. The target variable indicates whether or not income is larger than 50K dollars, and the sensitive feature is Gender. Each data object is described by 14 attributes which include 8 categorical and 6 numerical attributes. We quantize the categorical attributes into binary features and keep the continuous attributes, which results in 105 features for prediction. We note the label imbalance as 30% of male adults have income over 50K whereas only 10% of female adults have income over 50K. Additionally 24% of all adults have salary over 50K, and the dataset has 33% women and 67% men.

- Goodreads reviews [126], only included in the supplemental materials. The dataset was collected from Oct 12, 2017 to Oct 21, 2017 and has 13,244 reviews. The target variable is the rating of the review, and the sensitive feature is the gender of the author. Genders were gathered by querying Wikipedia and using pronoun inference, and the dataset is a subset of the original Goodreads dataset because it only includes reviews about the top 100 most popular authors. Each datum consists of the review text, vectorized using Tf-Idf. The review scores occurred with counts 578, 2606, 4544, 5516 for scores 1,3,4, and 5 respectively. Books by women authors and men authors had average scores of 4.088 and 4.092 respectively.

- MIMIC-III dataset [158]. The dataset includes 25,879 adult patients admitted to the intensive care unit of the Beth Israel Deaconess Medical Center in downtown Boston. Clinical notes from the first 48 hours are used to predict hospital mortality after 48 hours. Of all adult patients, 13.8% patients died in

the hospital. We are interested in the difference in performance between the five self-reported ethnic groups and following data sizes and hospital mortality rates.

Race	# patients	% total	Hospital Mortality
Asian	583	2.3	14.2
Black	2,327	9.0	10.9
Hispanic	832	3.2	10.3
Other	3,761	14.5	18.4
White	18,377	71.0	13.4

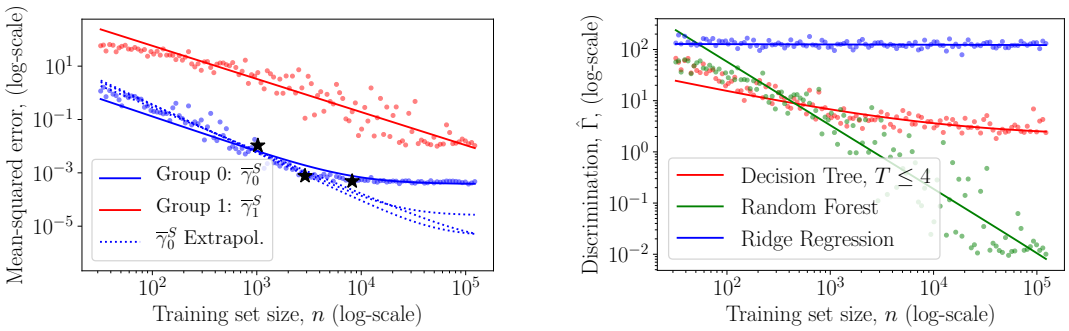
Table A.1: Summary statistics of clinical notes dataset

A.2.2 Synthetic experiments

To illustrate the effect of training set size and model choice, and the validity of the power-law learning curve assumption, we conduct a small synthetic experiment in which $p(A = 1) = 0.3$ and $X \sim \mathcal{N}(\mu_A, \sigma_A^2)$ with $\mu_0 = 0, \mu_1 = 1, \sigma_0 = 1, \sigma_1 = 2$. The outcome is a quadratic function with heteroskedastic noise, $Y = 2X^2 - 2X + .1 + \epsilon X^2$, with $\epsilon \sim \mathcal{N}(0, 1)$. We fit decision tree, random forest and ridge regressors of the outcome Y to X using default parameters in the implementation in scikit-learn [235], but limiting the decision tree to depth $T \leq 4$. The size of the training set is varied exponentially between 2^5 and 2^{17} samples, and at each size, trees are fit 200 times. In Figure A-1, we show the resulting learning curves $\overline{\gamma}_0(\hat{Y}, n)$ and $\overline{\gamma}_1(\hat{Y}, n)$ as well as fits of Pow3 curves to them. Shown in dotted lines are extrapolations of learning curves from different sample sizes, illustrating the difficulty of estimating the intercepts δ_a and the Bayes error with high accuracy.

A.2.3 Clinical notes

Here we include additional details about topic modeling. Topics were sampled using Markov Chain Monte Carlo after 2,500 iterations. We present the topics with highest and lowest variance in error rates among groups in Table A.2. Error rates were computed using a logistic regression with L1 regularization over 10,000 TF-IDF features



(a) Learning curves, $\overline{\gamma}_0, \overline{\gamma}_1$ for random forest

(b) Discrimination, $\overline{\Gamma} = |\overline{\gamma}_0 - \overline{\gamma}_1|$ for various models

Figure A-1: Inverse power-laws (Pow3) fit to generalization error as a function of training set size on synthetic data. Dotted lines are extrapolations from sample sizes indicated by black stars. This illustrates the difficulty of estimating the Bayes error through extrapolation, here at $\overline{N}_0 = 3 \cdot 10^{-4}$ and $\overline{N}_1 = 7 \cdot 10^{-3}$ respectively.

using 80/20 training and testing data split over 50 trials. Based on the most representative words for each topic, we can infer topic descriptions, for example cancer patients for topic 48 and cardiac patients for topic 45.

We identified patients with notes corresponding to topic 48, corresponding to cancer, as a subpopulation with large differences in errors between groups. By varying the training size while saving 20% of the data for testing, we estimate that more data would not be beneficial for decreasing error (see Figure A-2c). The mean over 50 trials is reported with hyperparameters chosen for each training size. Instead, we recommend collecting more features (e.g. structured data from lab results, more detailed patient history) as a way of improving error for this subpopulation.

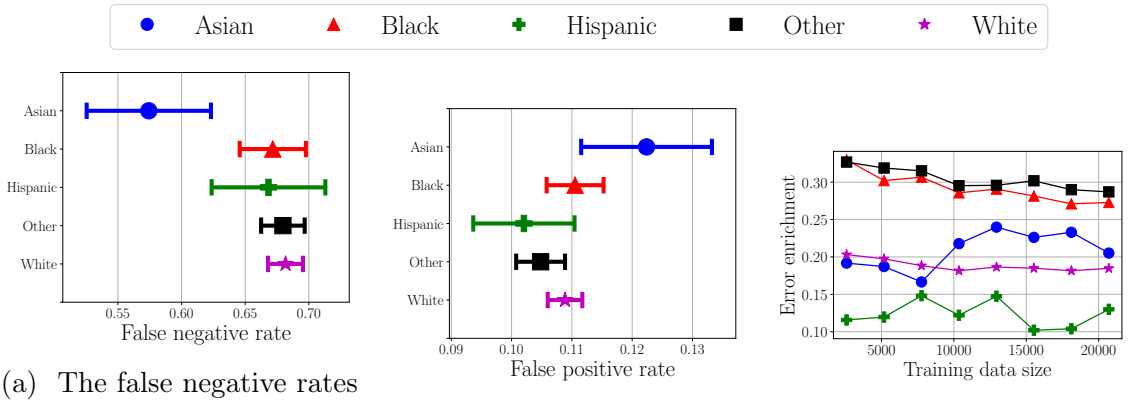
Furthermore, we compute the 95% confidence intervals for false positive and false negative rates for a logistic regression with L1 regularization in Figure A-2a and Figure A-2b.

A.3 Exploring model choice

If a difference in bias is the dominating source of discrimination between groups, changing the class of models under consideration could have a large impact on discrimination.Consider for example Figure 1c in which the true outcome has higher

Topic	Top words	Asian	Black	Hispanic	Other	White
31	no(t pain present normal edema tube history pulse absent left respiratory monitor	5.9	8.4	17.6	30.8	11.1
17	hospital lymphoma continue s/p unit bmt thrombocytopenia line rash	34.3	13.6	34.9	30.2	26.0
43	bowel abdominal abd abdomen surgery s/p small pain obstruction fluid ngt	16.6	11.8	5.7	26.8	13.2
45	artery carotid aneurysm left identifier numeric vertebral internal clip	5.4	5.3	3.8	20.4	10.0
48	mass cancer metastatic lung tumor patient cell left malignant breast hospital	21.6	25.4	12.3	30.2	18.5
1	neo gtt pain resp neuro wean clear plan insulin good	3.3	1.8	1.6	3.6	2.7
2	assessment insulin mg/dl plan pain meq/l mmhg chest cabg action	0.3	0.6	0.9	3.6	2.2
0	chest reason tube clip left artery s/p pneumothorax cabg pulmonary	3.2	5.5	2.5	5.6	4.0
25	c/o pain clear denies oriented sats plan alert stable monitor	7.3	3.9	5.9	8.2	6.5
47	pacer pacemaker icd s/p paced rhythm ccu amiodarone cardiac	8.2	9.1	8.3	13.8	10.1

Table A.2: Top and bottom 5 topics (of 50) based on variance in error rates of groups. Error rates by group and topic $p(\hat{Y} \neq Y|K, A)$ are reported in percentages.



(a) The false negative rates for logistic regression with L1 regularization do not differ across five ethnic groups, shown by the overlapping 95%-confidence intervals intervals, except for Asian patients.

(b) The false positive rates also does not differ much across groups with many overlapping intervals. Note that Asian patients have high false positive rate but low false negative rates.

(c) Adding training data size on error enrichment for cancer (topic 48) does not necessarily reduce error for all groups. This may suggest we should focus on collecting more features instead.

Figure A-2: Additional clinical notes experiments highlight the differences in false positive and false negative rates. We also examine the effect of training size on cancer patients in the dataset.

complexity in regions where one protected group is more densely distributed than the other. Increasing model capacity in such cases, or exploring other model classes of similar capacity, may reduce as long as the bias-variance trade-off is beneficial. Bias is not identifiable in general, as this requires estimation or bounding of noise components N_a, or an assumption that they are equal, $\overline{N}_0 = \overline{N}_1$, or negligible, $\overline{N}_a \approx 0$. However, as noise is in-dependent of model choice, a difference in bias of different models is identifiable even if the noise is not known, provided that the variance is estimated. With $\Delta\overline{B} = \overline{B}_0 - \overline{B}_1$, and $\Delta\overline{V} = \overline{V}_0 - \overline{V}_1$, and $\hat{Y}, \hat{Y}'$, two predictors for comparison, we may test the hypothesis $H_0 : \Delta\overline{B}(\hat{Y}) + \Delta\overline{V}(\hat{Y}) = \Delta\overline{B}(\hat{Y}') + \Delta\overline{V}(\hat{Y}')$.

A.4 Regression with homoskedastic noise

By definition of $\overline{N}$, we can state the following result.

Proposition 2. *Homoskedastic noise, i.e.* $\forall x \in \mathcal{X}, a \in \mathcal{A} : N(x, a) = N$, *does not contribute to discrimination level* $\overline{\Gamma}$ *under the squared loss* $L(y, y') = (y - y')^2$.

Proof. Under the squared loss, $\forall a : \overline{N}_a = \mathbb{E}_X[N(X,a)] = N$, as $c_n(x,a) = 1$. □

In contrast, for the zero-one loss and class-specific variants, the expected noise terms $\overline{N}_a$ do not cancel, as they depend on the factor $c_n(x,a)$.

A.5 Bias-variance decomposition. Proof of Theorem 1.

Lemma A1 (Squared loss and zero-one loss)**.** *The following claim holds for both:*
a) $L(y,y') = [y \neq y']$ *the zero-one loss with* $c_1(x,a) = 2\mathbb{E}[\mathbb{1}[\hat{Y}_D(x,a) = \hat{y}_*(x,a)]] - 1$ *and* $c_2(x,a) = \{1,$ *if* $\hat{y}^*(x,a) = \hat{y}^m(x,a); -1$ *otherwise*$\}$,
b) a) $L(y,y') = (y-y')^2$ *the squared loss with* $c_1(x,a) = c_2(x,a) = 1$.

$$\mathbb{E}[L(Y,\hat{Y}_D) \mid X = x, A = a] = c_1(x,a)\mathbb{E}[L(y,\hat{Y}^*) \mid x,a] + L(\hat{y}^m(x,a),\hat{y}^*(x,a)) + c_2\mathbb{E}[L(\hat{y}^m(x,a),\hat{Y}_D) \mid x,a] \ .$$

Proof. See [83]. □

Lemma A2 (Class-specific zero-one loss)**.** *With* $L(y,y') = [y \neq y']$ *the zero-one loss, it holds with* $c_1(x,a) = 2\mathbb{E}[\mathbb{1}[\hat{Y}_D(x,a) = \hat{y}_*(x,a)]] - 1$ *and* $c_2(x,a) = \{1,$ *if* $\hat{y}^*(x,a) = \hat{y}^m(x,a); -1$ *otherwise*$\}$

$$\forall y \in \{0,1\} : \mathbb{E}[L(y,\hat{Y}_D) \mid X = x, A = a] = c_1(x,a)L(y,\hat{Y}^*) + L(\hat{y}^m(x,a),\hat{y}^*(x,a)) + c_2\mathbb{E}[L(\hat{y}^m(x,a),\hat{Y}_D) \mid x,a] \ .$$

Proof. We begin by showing that $L(y,\hat{Y}_D(x,a)) = L(\hat{y}^*(x,a),\hat{Y}_D(x,a)) + c_0(x,a)L(y,\hat{y}^*(x,a))$

with $c_0(x,a) = \{+1, \text{ if } \hat{y}^*(x,a) = \hat{Y}_D(x,a); -1, \text{otherwise}\}$.

$$L(y,\hat{Y}_D) - L(\hat{y}^*(x,a),\hat{Y}_D(x,a)) + c_0(x,a)L(y,\hat{y}^*(x,a))$$
$$= \begin{cases} 0, & \text{if } \hat{Y}_D(x,a) = \hat{y}^*(x,a) = 0 \\ -1 - c_0(x,a), & \text{if } \hat{Y}_D(x,a) = 0, \hat{y}^*(x,a) = 1 \\ 0, & \text{if } \hat{Y}_D(x,a) = 1, \hat{y}^*(x,a) = 0 \\ 1 - c_0(x,a), & \text{if } \hat{Y}_D(x,a) = \hat{y}^*(x,a) = 1 \end{cases}$$

As the above should be zero for all options, this implies that $c_0 = 2 * \mathbb{1}[\hat{Y}_D(x,a) = \hat{y}^*(x,a)] - 1$.

We now show that,

$$\mathbb{E}[L(\hat{y}^*(x,a), Y_d) \mid x,a] = L(\hat{y}^*(x,a),\hat{y}^m(x,a)) + c_2(x,a)\mathbb{E}[L(\hat{y}^m(x,a),\hat{Y}) \mid x,a] \ .$$

We have that if $\hat{y}^m(x,a) \neq \hat{y}^*(x,a)$,

$$\begin{aligned} \mathbb{E}[L(\hat{y}^*(x,a),\hat{Y}_D) \mid x,a] &= p(\hat{y}^*(x,a) \neq \hat{Y}_D \mid x,a) = 1 - p(\hat{y}^*(x,a) = \hat{Y}_D \mid x,a) \\ &= 1 - p(\hat{y}^m(x,a) = \hat{Y}_D \mid x,a) = 1 - \mathbb{E}[L(\hat{y}^m(x,a),\hat{Y}_D) \mid x,a] \\ &= L(\hat{y}^*(x,a),\hat{y}^m(x,a)) - \mathbb{E}[L(\hat{y}^m(x,a),\hat{Y}_D) \mid x,a] \\ &= L(\hat{y}^*(x,a),\hat{y}^m(x,a)) + c_2(x,a)\mathbb{E}[L(\hat{y}^m(x,a),\hat{Y}_D) \mid x,a] \ . \end{aligned}$$

A similar calculation for the case where $\hat{y}^m(x,a) = \hat{y}^*(x,a)$ yields the claim.

Finally, We have that

$$\begin{aligned} \mathbb{E}[L(y,\hat{Y}_D)] &= \mathbb{E}[L(\hat{y}^*(x,a),\hat{Y}_D) + c_0(x,a)L(y,\hat{y}^*(x,a)) \mid x,a] \\ &= \mathbb{E}[L(\hat{y}^*(x,a),\hat{Y}_D) \mid x,a] + \mathbb{E}[c_0(x,a) \mid x,a]L(y,\hat{y}^*(x,a)) \\ &= L(\hat{y}^*(x,a),\hat{y}^m(x,a)) + c_2(x,a)\mathbb{E}[L(\hat{y}^m(x,a),\hat{Y}_D) \mid x,a] \\ &+ \mathbb{E}[c_0(x,a) \mid x,a]L(y,\hat{y}^*(x,a)) \end{aligned}$$

which gives us our result. □

Since datasets are drawn independently of the protected attribute A,

$$\begin{aligned}\overline{\gamma}_a(\hat{Y}) &= \mathbb{E}_D[\mathbb{E}_{X,Y}[L(Y,\hat{Y}_D) \mid D, A=a] \mid A=a] \\ &= \mathbb{E}_X[\mathbb{E}_{D,Y}[L(Y,\hat{Y}_D) \mid X, A=a] \mid A=a] \\ &= \mathbb{E}_X[B(\hat{Y},X,a) + c_2(X,a)V(\hat{Y},X,a) + c_1(X,a)N(X,a) \mid A=a] \ ,\end{aligned}$$

and an analogous results hold for class-specific losses, Theorem 1 follows from lemmas A1–A2.

A.6 Difference between power law curves

Let $f(x) = ax^{-b} + c$ and $g(x) = dx^{-e} + h$. Then $d(x) = f(x) - g(x)$ has at most 2 local minima. We see this by re-writing $d(x)$

$$d(x) = ax^{-b} + \tilde{c} - dx^{-e}$$

and so

$$d'(x) = (-b)ax^{-b-1} + dex^{-e-1}$$

Setting the derivative to zero,

$$(-b)ax^{-b-1} + dex^{-e-1} = 0$$

$$x^{b-e} = \frac{ba}{de}$$

which has a unique positive root

$$x = (\frac{ba}{de})^{\frac{1}{b-e}} \ .$$

Since $f(x)$ has a single critical point (for $x > 0$), $f(x)$ can switch signs at most twice. The curves $f(x) = \frac{100}{x^2} + 1$ and $g(x) = \frac{50}{x}$ intersect twice on $x \in [0, \infty]$. If $b = e$, $d(x)$

has a single zero,

$$d(x) = (a - d)x^{-b} + \tilde{c} = 0$$

yields

$$x = (\frac{\tilde{c}}{d - a})^{\frac{1}{-b}} \ .$$

Appendix B

Additional Information for Chapter 4

B.1 Topic Model for Likelihood of Hospitalization

See Table B.1 to B.50 for the most representative conditions, procedures, specialty visits, and drugs for the 50 topics produced in the Likelihood of Hospitalization topic modeling.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Cerebral infarction	9675	64458
	Entire History	Transient cerebral ischemia	4677	34335
	Entire History	Carotid artery obstruction	7107	45836
	Past 365 days	Cerebral infarction	3275	17256
	Entire History	Cerebrovascular disease	2654	16717
Procedure	Entire History	Magnetic resonance eg, proton imaging, brain including brain stem; without contrast material	1510	11117
	Entire History	Duplex scan of extracranial arteries; complete bilateral study	3581	25265
	Entire History	Magnetic resonance angiography, head; without contrast materials	663	5043
	Entire History	Computed tomography, head or brain; without contrast material	7252	42614
	Entire History	Computed tomographic angiography, neck, with contrast materials, including noncontrast images, if performed, and image postprocessing	446	3247
Specialty	Entire History	Neurology	5376	43892
	Past 365 days	Neurology	1209	8835
	Entire History	Neuroradiology	3148	21318
	Entire History	Vascular Surgery	4717	25301
	Entire History	Diagnostic Radiology	27485	187103
Drug	Entire History	clopidogrel 75 MG Oral Tablet	16430	107051
	Entire History	atorvastatin 40 MG Oral Tablet	14881	124415
	Past 365 days	clopidogrel 75 MG Oral Tablet	2367	14994
	Past 180 days	clopidogrel 75 MG Oral Tablet	1297	8066
	Past 365 days	atorvastatin 40 MG Oral Tablet	2890	22660

Table B.1: Most representative labs, procedures, conditions, specialty visits, and medications for topic 1 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Injury of head	2416	15465
	Past 365 days	Injury of head	530	3363
	Entire History	Patient dependence on care provider	10198	43939
	Entire History	Injury of neck	383	2188
	Entire History	Laceration of head	360	3037
Procedure	Entire History	Computed tomography, head or brain; without contrast material	7252	42614
	Entire History	Computed tomography, cervical spine; without contrast material	1683	9448
	Entire History	Emergency department visit for the evaluation and management of a patient	11691	77101
	Past 365 days	Emergency department visit for the evaluation and management of a patient	2304	14474
	Past 365 days	Computed tomography, head or brain; without contrast material	1778	8667
Specialty	Past 365 days	Emergency Medicine	4967	25070
	Entire History	Emergency Medicine	20356	113387
	Past 180 days	Emergency Medicine	2365	11250
	Entire History	Neuroradiology	3148	21318
	Entire History	Surgical Critical Care	712	4087
Drug	Entire History	cephalexin 500 MG Oral Capsule	3951	28466
	Entire History	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.016 MG/ML / Bordetella pertussis pertactin vaccine, inactivated 0.005 MG/ML / Bordetella pertussis toxoid vaccine, inactivated 0.016 MG/ML / diphtheria toxoid vaccine, inactiv...	123	1715
	Past 365 days	cephalexin 500 MG Oral Capsule	667	4507
	Entire History	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.016 MG/ML / Bordetella pertussis pertactin vaccine, inactivated 0.005 UNT/ML / Bordetella pertussis toxoid vaccine, inactivated 0.016 MG/ML / diphtheria toxoid vaccine, inacti...	69	584
	Past 180 days	cephalexin 500 MG Oral Capsule	371	2304

Table B.2: Most representative labs, procedures, conditions, specialty visits, and medications for topic 2 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Actinic keratosis	11051	135967
	Entire History	Neoplasm of uncertain behavior of skin	6971	85098
	Entire History	Senile hyperkeratosis	4721	65927
	Past 365 days	Actinic keratosis	1516	19691
	Past 365 days	Senile hyperkeratosis	715	11082
Specialty	Entire History	Dermatology	11950	151270
	Past 365 days	Dermatology	2119	27485
	Past 180 days	Dermatology	868	11595
	Past 30 days	Dermatology	192	2570
	Entire History	Dermatopathology	1144	14086
Procedure	Entire History	Destruction eg, laser surgery, electrosurgery, cryosurgery, chemosurgery, surgical curettement, premalignant lesions	4518	56334
	Entire History	Destruction eg, laser surgery, electrosurgery, cryosurgery, chemosurgery, surgical curettement, premalignant lesions eg, actinic keratoses; second through 14 lesions, each List separately in addition to code for first lesion	3234	38793
	Entire History	Biopsy of skin, subcutaneous tissue and/or mucous membrane including simple closure, unless otherwise listed; single lesion	2785	34700
	Entire History	Tangential biopsy of skin eg, shave, scoop, saucerize, curette; single lesion	848	10632
	Past 365 days	Level IV Surgical pathology, gross and microscopic examination Abortion	2094	22421
Drug	Entire History	mupirocin 0.02 MG/MG Topical Ointment	1714	14639
	Entire History	fluorouracil 50 MG/ML Topical Cream	259	2828
	Entire History	cephalexin 500 MG Oral Capsule	3951	28466
	Entire History	0.5 ML influenza A virus A	1454	16514
	Past 365 days	0.5 ML influenza A virus A	1454	16513

Table B.3: Most representative labs, procedures, conditions, specialty visits, and medications for topic 3 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Anesthesia for procedures on eye; lens surgery	2840	31686
	Entire History	Extracapsular cataract removal with insertion of intraocular lens prosthesis 1 stage procedure, manual or mechanical technique eg, irrigation and aspiration or phacoemulsification; without endoscopic cyclophotocoagulation	3683	43408
	Entire History	Ophthalmic biometry by partial coherence interferometry with intraocular lens power calculation	1521	17996
	Entire History	Ophthalmological services; comprehensive, established patient, 1 or more visits	13904	147733
	Entire History	Ophthalmological services; intermediate, established patient	9993	98335
Condition	Entire History	Cataract	3034	32227
	Entire History	Nuclear senile cataract	16225	199309
	Entire History	Secondary cataract	3703	38099
	Entire History	Bilateral cataracts	1596	15327
	Entire History	Senile cataract	998	12868
Drug	Entire History	prednisolone acetate 10 MG/ML Ophthalmic Suspension	2461	28139
	Entire History	ketorolac tromethamine 5 MG/ML Ophthalmic Solution	915	8936
	Entire History	ofloxacin 3 MG/ML Ophthalmic Solution	869	9234
	Entire History	polymyxin B 10000 UNT/ML / trimethoprim 1 MG/ML Ophthalmic Solution	493	5326
	Entire History	difluprednate 0.5 MG/ML Ophthalmic Suspension [Durezol]	389	4245
Specialty	Entire History	Ophthalmology	37394	394685
	Entire History	Anesthesiology	11414	99859
	Past 365 days	Ophthalmology	5151	59577
	Past 365 days	Anesthesiology	2232	18738
	Entire History	No matching concept	491045	3843783

Table B.4: Most representative labs, procedures, conditions, specialty visits, and medications for topic 4 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Emergency department visit for the evaluation and management of a patient	18712	97802
	Entire History	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	11322	70817
	Entire History	Emergency department visit for the evaluation and management of a patient	11691	77101
	Entire History	Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only	23068	121633
	Entire History	Radiologic examination, chest, 2 views, frontal and lateral	8319	54989
Specialty	Entire History	Emergency Medicine	20356	113387
	Entire History	Internal Medicine	60351	431318
	Entire History	Diagnostic Radiology	27485	187103
	Entire History	Radiology	9555	67151
	Entire History	Interventional Cardiology	9925	57143
Condition	Entire History	Chest pain	26088	191977
	Entire History	Dizziness and giddiness	12515	105276
	Entire History	Abdominal pain	8970	67345
	Entire History	Dyspnea	37891	217024
	Entire History	Benign essential hypertension	27610	247456
Drug	Entire History	2 ML ondansetron 2 MG/ML Injection	2315	17067
	Entire History	1000 ML sodium chloride 9 MG/ML Injection	2066	13064
	Entire History	6 azithromycin 250 MG Oral Tablet Pack	5289	49956
	Entire History	tramadol hydrochloride 50 MG Oral Tablet	10979	79865
	Entire History	ciprofloxacin 500 MG Oral Tablet	2814	26196

Table B.5: Most representative labs, procedures, conditions, specialty visits, and medications for topic 5 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Entire History	Urology	11932	99276
	Past 365 days	Urology	2167	17915
	Past 180 days	Urology	1009	8218
	Entire History	Female Pelvic Medicine and Reconstructive Surgery	469	5173
	Past 30 days	Urology	195	1759
Procedure	Entire History	Measurement of post-voiding residual urine and/or bladder capacity by ultrasound, non-imaging	4964	45161
	Entire History	Cystourethroscopy separate procedure	1670	13757
	Entire History	Ultrasound, retroperitoneal eg, renal, aorta, nodes, real time with image documentation; complete	1913	14008
	Entire History	Computed tomography, abdomen and pelvis; without contrast material in one or both body regions, followed by contrast materials and further sections in one or both body regions	738	6245
	Entire History	Computed tomography, abdomen and pelvis; without contrast material	2593	15395
Condition	Entire History	Urinary tract infectious disease	28336	179556
	Entire History	Hematuria syndrome	4746	37138
	Entire History	Finding of frequency of urination	6535	64424
	Past 365 days	Urinary tract infectious disease	5531	29130
	Entire History	Microscopic hematuria	2372	25744
Drug	Entire History	ciprofloxacin 500 MG Oral Tablet	2814	26196
	Entire History	nitrofurantoin, macrocrystals 25 MG / nitrofurantoin, monohydrate 75 MG Oral Capsule	1121	9473
	Entire History	sulfamethoxazole 800 MG / trimethoprim 160 MG Oral Tablet	3114	25009
	Entire History	ciprofloxacin 250 MG Oral Tablet	1420	11299
	Past 365 days	nitrofurantoin, macrocrystals 25 MG / nitrofurantoin, monohydrate 75 MG Oral Capsule	292	2276

Table B.6: Most representative labs, procedures, conditions, specialty visits, and medications for topic 6 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	4529	36711
	Entire History	Echocardiography, transthoracic, real-time with image documentation 2D, includes M-mode recording, when performed, complete, with spectral Doppler echocardiography, and with color flow Doppler echocardiography	13647	94396
	Past 365 days	Echocardiography, transthoracic, real-time with image documentation 2D, includes M-mode recording, when performed, complete, with spectral Doppler echocardiography, and with color flow Doppler echocardiography	2536	15943
	Entire History	Myocardial perfusion imaging, tomographic SPECT including attenuation correction, qualitative or quantitative wall motion, ejection fraction by first pass or gated technique, additional quantification, when performed; multiple studies, at rest and/or	3623	29071
	Past 180 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	1642	13241
Condition	Entire History	Non-rheumatic mitral valve stenosis with regurgitation	9351	67245
	Entire History	Dyspnea	37891	217024
	Entire History	Palpitations	6398	73029
	Entire History	Electrocardiogram abnormal	9932	79898
	Past 365 days	Dyspnea	7916	41875
Drug	Entire History	no match	1561	12940
	Entire History	5 ML regadenoson 0.08 MG/ML Prefilled Syringe	822	6257
	Entire History	24 HR metoprolol succinate 25 MG Extended Release Oral Tablet	11914	100033
	Past 365 days	24 HR metoprolol succinate 25 MG Extended Release Oral Tablet	2037	16942
	Past 365 days	5 ML regadenoson 0.08 MG/ML Prefilled Syringe	174	1489
Specialty	Past 180 days	No matching concept	54309	371773
	Entire History	Interventional Cardiology	9925	57143
	Past 30 days	No matching concept	9979	71669
	Past 365 days	No matching concept	112625	811174
	Entire History	No matching concept	491045	3843783

Table B.7: Most representative labs, procedures, conditions, specialty visits, and medications for topic 7 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Chiropractic manipulative treatment CMT; spinal, 3-4 regions	4884	76582
	Entire History	Chiropractic manipulative treatment CMT; spinal, 1-2 regions	799	13085
	Entire History	Radiologic examination, spine, cervical; 4 or 5 views	481	5401
	Past 365 days	Chiropractic manipulative treatment CMT; spinal, 3-4 regions	711	9979
	Entire History	Magnetic resonance eg, proton imaging, spinal canal and contents, cervical; without contrast material	531	5197
Condition	Entire History	Somatic dysfunction of lumbar region	5419	74174
	Entire History	Neck pain	7323	102690
	Entire History	Cervical somatic dysfunction	3625	54311
	Entire History	Somatic dysfunction of thoracic region	3204	46260
	Entire History	Low back pain	25443	256965
Specialty	Past 30 days	No matching concept	9979	71669
	Past 180 days	No matching concept	54309	371773
	Past 365 days	No matching concept	112625	811174
	Entire History	No matching concept	491045	3843783
	Entire History	Optometry	5353	62272
Drug	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	3795	36927
	Entire History	cyclobenzaprine hydrochloride 10 MG Oral Tablet	1402	14365
	Entire History	amoxicillin 500 MG Oral Capsule	4366	49635
	Entire History	meloxicam 15 MG Oral Tablet	2507	32953
	Past 365 days	21 methylprednisolone 4 MG Oral Tablet Pack	458	4850

Table B.8: Most representative labs, procedures, conditions, specialty visits, and medications for topic 8 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Type 2 diabetes mellitus	55892	341272
	Entire History	Type 2 diabetes mellitus without complication	115331	903917
	Past 365 days	Type 2 diabetes mellitus without complication	15782	126227
	Past 365 days	Type 2 diabetes mellitus	15602	93670
	Past 180 days	Type 2 diabetes mellitus without complication	6883	54153
Drug	Entire History	metformin hydrochloride 1000 MG Oral Tablet	7199	65537
	Past 365 days	metformin hydrochloride 1000 MG Oral Tablet	800	9056
	Past 180 days	metformin hydrochloride 1000 MG Oral Tablet	436	4887
	Entire History	sitagliptin 100 MG Oral Tablet [Januvia]	3142	22592
	Entire History	3 ML insulin glargine 100 UNT/ML Pen Injector [Lantus]	3405	20491
Procedure	Past 180 days	Collection of venous blood by venipuncture	7906	61965
	Past 365 days	Collection of venous blood by venipuncture	17457	137379
	Entire History	Ophthalmological services; comprehensive, established patient, 1 or more visits	13904	147733
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	20683	182276
	Entire History	Collection of venous blood by venipuncture	90495	754265
Specialty	Entire History	Endocrinology	7227	51085
	Past 180 days	No matching concept	54309	371773
	Entire History	Ophthalmology	37394	394685
	Entire History	No matching concept	491045	3843783
	Past 365 days	No matching concept	112625	811174

Table B.9: Most representative labs, procedures, conditions, specialty visits, and medications for topic 9 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Obstructive sleep apnea syndrome	26098	220857
	Past 365 days	Obstructive sleep apnea syndrome	4928	45495
	Entire History	Sleep apnea	3112	23356
	Past 180 days	Obstructive sleep apnea syndrome	2307	21544
	Entire History	Obesity	12121	119225
Specialty	Entire History	Sleep Medicine	3575	23269
	Past 365 days	Sleep Medicine	834	5251
	Past 180 days	Sleep Medicine	419	2533
	Entire History	Pulmonary Disease	2589	14061
	Entire History	Psychiatry	4181	41283
Procedure	Entire History	Home sleep test hst with type iii portable monitor, unattended; minimum of 4 channels: 2 respiratory movement/airflow, 1 ecg/heart rate and 1 oxygen saturation	381	4011
	Entire History	Polysomnography; age 6 years or older, sleep staging with 4 or more additional parameters of sleep, attended by a technologist	288	2804
	Entire History	Polysomnography; age 6 years or older, sleep staging with 4 or more additional parameters of sleep, with initiation of continuous positive airway pressure therapy or bilevel ventilation, attended by a technologist	334	2629
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	14173	98050
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient	2560	21467
Drug	Entire History	verapamil hydrochloride 240 MG Extended Release Oral Tablet	1264	10170
	Entire History	zolpidem tartrate 10 MG Oral Tablet	3468	33192
	Past 365 days	verapamil hydrochloride 240 MG Extended Release Oral Tablet	86	849
	Entire History	topiramate 25 MG Oral Tablet	244	3602
	Entire History	zolpidem tartrate 5 MG Oral Tablet	1801	17370

Table B.10: Most representative labs, procedures, conditions, specialty visits, and medications for topic 10 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Chronic kidney disease stage 3	45634	256984
	Past 365 days	Chronic kidney disease stage 3	10704	63160
	Entire History	Chronic kidney disease due to hypertension	14247	78309
	Past 180 days	Chronic kidney disease stage 3	4657	28300
	Entire History	Chronic kidney disease	14259	63704
Specialty	Entire History	Nephrology	11217	41428
	Past 365 days	Nephrology	3044	10577
	Past 180 days	Nephrology	1554	5154
	Past 30 days	Nephrology	194	771
	Past 180 days	No matching concept	54309	371773
Drug	Entire History	allopurinol 100 MG Oral Tablet	5515	31846
	Past 365 days	allopurinol 100 MG Oral Tablet	865	4913
	Past 180 days	allopurinol 100 MG Oral Tablet	482	2688
	Entire History	allopurinol 300 MG Oral Tablet	3726	27601
	Entire History	amlodipine 10 MG Oral Tablet	13835	124188
Procedure	Entire History	Ultrasound, retroperitoneal eg, renal, aorta, nodes, real time with image documentation; complete	1913	14008
	Entire History	Ultrasound, retroperitoneal eg, renal, aorta, nodes, real time with image documentation; limited	1366	9344
	Entire History	Computed tomography, abdomen and pelvis; without contrast material	2593	15395
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	14173	98050
	Past 180 days	Collection of venous blood by venipuncture	7906	61965

Table B.11: Most representative labs, procedures, conditions, specialty visits, and medications for topic 11 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Entire History	Gastroenterology	11340	92146
	Entire History	Anesthesiology	11414	99859
	Past 365 days	Gastroenterology	2208	16696
	Past 365 days	Anesthesiology	2232	18738
	Past 365 days	No matching concept	112625	811174
Procedure	Entire History	Colonoscopy, flexible; with biopsy, single or multiple	1827	23746
	Entire History	Colonoscopy, flexible; with removal of tumors, polyps, or other lesions by snare technique	1546	19603
	Entire History	Anesthesia for lower intestinal endoscopic procedures, endoscope introduced distal to duodenum	2396	29417
	Entire History	Level IV Surgical pathology, gross and microscopic examination Abortion	13388	147528
	Entire History	Anesthesia for lower intestinal endoscopic procedures, endoscope introduced distal to duodenum; not otherwise specified	743	8163
Condition	Entire History	Diverticulosis of large intestine without diverticulitis	4663	48400
	Entire History	Polyp of colon	1919	26116
	Entire History	Hemorrhoids	2193	25407
	Entire History	Benign neoplasm of colon	2489	32518
	Entire History	Benign neoplasm of transverse colon	1049	11924
Drug	Entire History	2 480 ML magnesium sulfate 0.0277 MEQ/ML / potassium sulfate 0.0374 MEQ/ML / sodium sulfate 0.257 MEQ/ML Oral Solution Pack [Suprep Bowel Prep Kit]	352	4746
	Entire History	20 ML propofol 10 MG/ML Injection	1319	10742
	Past 365 days	0.5 ML influenza A virus A	1454	16513
	Entire History	0.5 ML influenza A virus A	1454	16514
	Entire History	0.5 ML Streptococcus pneumoniae serotype 1 capsular antigen diphtheria CRM197 protein conjugate vaccine 0.0044 MG/ML / Streptococcus pneumoniae serotype 14 capsular antigen diphtheria CRM197 protein conjugate vaccine 0.0044 MG/ML / Streptococcus pneumo...	1352	15626

Table B.12: Most representative labs, procedures, conditions, specialty visits, and medications for topic 12 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Congestive heart failure	31063	99781
	Entire History	Heart failure	21748	63138
	Past 365 days	Congestive heart failure	8914	26160
	Entire History	Hypertensive heart failure	8071	27210
	Past 365 days	Heart failure	6234	15912
Drug	Entire History	furosemide 40 MG Oral Tablet	13499	59492
	Past 365 days	furosemide 40 MG Oral Tablet	2516	9939
	Entire History	furosemide 20 MG Oral Tablet	12695	64959
	Past 180 days	furosemide 40 MG Oral Tablet	1411	5538
	Entire History	spironolactone 25 MG Oral Tablet	4015	24055
Procedure	Entire History	Echocardiography, transthoracic, real-time with image documentation 2D, includes M-mode recording, when performed, complete, with spectral Doppler echocardiography, and with color flow Doppler echocardiography	13647	94396
	Past 365 days	Echocardiography, transthoracic, real-time with image documentation 2D, includes M-mode recording, when performed, complete, with spectral Doppler echocardiography, and with color flow Doppler echocardiography	2536	15943
	Entire History	Subsequent hospital care, per day, for the evaluation and management of a patient	19410	72503
	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	4529	36711
	Entire History	Interrogation device evaluations remote, up to 90 days; single, dual, or multiple lead implantable defibrillator system with interim analysis, reviews and reports by a physician or other qualified health care professional	1325	7459
Specialty	Entire History	Clinical Cardiac Electrophysiology	9954	58559
	Past 365 days	Clinical Cardiac Electrophysiology	2539	14483
	Entire History	Advanced Heart Failure and Transplant Cardiology	1825	9825
	Entire History	Interventional Cardiology	9925	57143
	Past 180 days	Clinical Cardiac Electrophysiology	1220	7288

Table B.13: Most representative labs, procedures, conditions, specialty visits, and medications for topic 13 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Atrial fibrillation	66638	361726
	Entire History	Paroxysmal atrial fibrillation	35288	203474
	Past 365 days	Paroxysmal atrial fibrillation	9039	51763
	Past 365 days	Atrial fibrillation	10675	49249
	Past 180 days	Paroxysmal atrial fibrillation	4213	23266
Procedure	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	4529	36711
	Entire History	Echocardiography, transthoracic, real-time with image documentation 2D, includes M-mode recording, when performed, complete, with spectral Doppler echocardiography, and with color flow Doppler echocardiography	13647	94396
	Entire History	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	31136	253643
	Entire History	Programming device evaluation in person with iterative adjustment of the implantable device to test the function of the device and select optimal permanent programmed values with analysis, review and report by a physician or other qualified health care	1789	10804
	Past 180 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	1642	13241
Specialty	Entire History	Clinical Cardiac Electrophysiology	9954	58559
	Past 365 days	Clinical Cardiac Electrophysiology	2539	14483
	Past 180 days	Clinical Cardiac Electrophysiology	1220	7288
	Entire History	Interventional Cardiology	9925	57143
	Past 30 days	No matching concept	9979	71669
Drug	Entire History	apixaban 5 MG Oral Tablet [Eliquis]	5466	33209
	Past 365 days	apixaban 5 MG Oral Tablet [Eliquis]	1886	12306
	Past 180 days	apixaban 5 MG Oral Tablet [Eliquis]	1116	7111
	Entire History	rivaroxaban 20 MG Oral Tablet [Xarelto]	3295	28430
	Entire History	warfarin sodium 5 MG Oral Tablet	5453	37662

Table B.14: Most representative labs, procedures, conditions, specialty visits, and medications for topic 14 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Positron emission tomography PET with concurrently acquired computed tomography CT for attenuation correction and anatomical localization imaging; skull base to mid-thigh	970	5115
	Entire History	Continuing medical physics consultation, including assessment of treatment parameters, quality assurance of dose delivery, and review of patient treatment documentation in support of the radiation oncologist, reported per week of therapy	1075	11788
	Entire History	Basic radiation dosimetry calculation, central axis depth dose calculation, TDF, NSD, gap calculation, off axis factor, tissue inhomogeneity factors, calculation of non-ionizing radiation surface and depth dose, as required during course of treatment, onl	604	5590
	Entire History	Therapeutic radiology treatment planning; complex	262	2381
	Entire History	Chemotherapy administration, intravenous infusion technique; up to 1 hour, single or initial substance/drug	3691	18869
Specialty	Entire History	Radiation Oncology	3731	38392
	Entire History	Medical Oncology	13900	77931
	Past 365 days	Medical Oncology	4325	18897
	Past 180 days	Medical Oncology	2341	9428
	Past 365 days	Radiation Oncology	968	7127
Drug	Entire History	ondansetron 8 MG Oral Tablet	710	3865
	Entire History	dexamethasone phosphate 10 MG/ML Injectable Solution	1153	5131
	Entire History	2 ML fentanyl 0.05 MG/ML Injection	2683	20558
	Entire History	no match	1561	12940
	Entire History	dexamethasone 4 MG Oral Tablet	386	2599
Condition	Entire History	Abnormal findings on diagnostic imaging of lung	7018	37848
	Entire History	Solitary nodule of lung	4634	32768
	Entire History	Primary malignant neoplasm of respiratory tract	4414	24327
	Entire History	Localized enlarged lymph nodes	850	5741
	Entire History	Primary malignant neoplasm	813	4622

Table B.15: Most representative labs, procedures, conditions, specialty visits, and medications for topic 15 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Entire History	Neurology	5376	43892
	Past 365 days	Neurology	1209	8835
	Entire History	Psychiatry	4181	41283
	Past 180 days	Neurology	591	4069
	Past 365 days	Psychiatry	1171	9244
Condition	Entire History	Amnesia	2422	22664
	Entire History	Seizure	2551	15323
	Entire History	Tremor	1130	9252
	Entire History	Epilepsy	1860	13373
	Entire History	Minimal cognitive impairment	1416	9228
Procedure	Entire History	Magnetic resonance eg, proton imaging, brain including brain stem; without contrast material	1510	11117
	Entire History	Magnetic resonance eg, proton imaging, brain including brain stem; without contrast material, followed by contrast materials and further sequences	918	9056
	Entire History	Collection of venous blood by venipuncture	90495	754265
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	14173	98050
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	124201	1073514
Drug	Entire History	carbidopa 25 MG / levodopa 100 MG Oral Tablet	1711	9238
	Past 365 days	carbidopa 25 MG / levodopa 100 MG Oral Tablet	308	1811
	Past 180 days	carbidopa 25 MG / levodopa 100 MG Oral Tablet	180	973
	Entire History	gabapentin 100 MG Oral Capsule	4907	35506
	Entire History	lamotrigine 100 MG Oral Tablet	271	4945

Table B.16: Most representative labs, procedures, conditions, specialty visits, and medications for topic 16 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Entire History	Urology	11932	99276
	Past 365 days	Urology	2167	17915
	Past 180 days	Urology	1009	8218
	Entire History	No matching concept	491045	3843783
	Past 365 days	No matching concept	112625	811174
Condition	Entire History	Benign prostatic hypertrophy with outflow obstruction	11654	115370
	Entire History	Benign prostatic hyperplasia	8929	89672
	Entire History	Raised prostate specific antigen	8291	94581
	Entire History	Nocturia	6084	63260
	Past 365 days	Benign prostatic hypertrophy with outflow obstruction	1972	20789
Procedure	Entire History	Measurement of post-voiding residual urine and/or bladder capacity by ultrasound, non-imaging	4964	45161
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	124201	1073514
	Past 365 days	Measurement of post-voiding residual urine and/or bladder capacity by ultrasound, non-imaging	622	6523
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	102234	994395
	Entire History	Collection of venous blood by venipuncture	90495	754265
Drug	Entire History	tamsulosin hydrochloride 0.4 MG Oral Capsule	13505	102723
	Past 365 days	tamsulosin hydrochloride 0.4 MG Oral Capsule	2166	17328
	Past 180 days	tamsulosin hydrochloride 0.4 MG Oral Capsule	1182	9416
	Entire History	finasteride 5 MG Oral Tablet	7046	48462
	Past 365 days	finasteride 5 MG Oral Tablet	1024	7526

Table B.17: Most representative labs, procedures, conditions, specialty visits, and medications for topic 17 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Past 365 days	Ophthalmology	5151	59577
	Entire History	Ophthalmology	37394	394685
	Past 180 days	Ophthalmology	2163	25313
	Entire History	Optometry	5353	62272
	Past 365 days	Optometry	836	9967
Condition	Entire History	Tear film insufficiency	7065	74335
	Past 365 days	Tear film insufficiency	1012	12451
	Entire History	Blepharitis	1330	13494
	Entire History	Vitreous degeneration	10865	116057
	Entire History	Blepharitis of left eyelid	783	7881
Procedure	Entire History	Ophthalmological services; intermediate, established patient	9993	98335
	Past 365 days	Ophthalmological services; comprehensive, established patient, 1 or more visits	2014	23158
	Entire History	Ophthalmological services; comprehensive, established patient, 1 or more visits	13904	147733
	Entire History	Determination of refractive state	3835	42176
	Past 365 days	Ophthalmological services; intermediate, established patient	1498	15825
Drug	Entire History	erythromycin 0.005 MG/MG Ophthalmic Ointment	928	8812
	Entire History	dexamethasone 1 MG/ML / tobramycin 3 MG/ML Ophthalmic Suspension	468	4314
	Entire History	dexamethasone 0.001 MG/MG / neomycin 0.0035 MG/MG / polymyxin B 10 UNT/MG Ophthalmic Ointment	252	3140
	Entire History	cyclosporine 0.5 MG/ML Ophthalmic Suspension [Restasis]	1012	7990
	Entire History	prednisolone acetate 10 MG/ML Ophthalmic Suspension	2461	28139

Table B.18: Most representative labs, procedures, conditions, specialty visits, and medications for topic 18 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Radiologic examination, foot; complete, minimum of 3 views	2424	21935
	Entire History	Radiologic examination, ankle; complete, minimum of 3 views	1396	11268
	Past 365 days	Radiologic examination, foot; complete, minimum of 3 views	415	3717
	Entire History	Arthrocentesis, aspiration and/or injection, intermediate joint or bursa eg, temporomandibular, acromioclavicular, wrist, elbow or ankle, olecranon bursa; without ultrasound guidance	533	6176
	Past 365 days	Radiologic examination, ankle; complete, minimum of 3 views	207	1961
Specialty	Entire History	Podiatry	21829	162483
	Past 365 days	Podiatry	4912	36842
	Entire History	Orthopedic Surgery	12317	122002
	Past 180 days	Podiatry	2198	16509
	Entire History	Diagnostic Radiology	27485	187103
Condition	Entire History	Arthralgia of the ankle and/or foot	3196	36734
	Entire History	Pain in right foot	2364	24907
	Entire History	Pain in left foot	2365	23405
	Entire History	Localized, primary osteoarthritis of the ankle and/or foot	1386	14473
	Entire History	Plantar fascial fibromatosis	1290	20124
Drug	Entire History	diclofenac sodium 0.01 MG/MG Topical Gel	1088	9258
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	5820	43831
	Entire History	dexamethasone phosphate 4 MG/ML Injectable Solution	777	5342
	Entire History	cephalexin 500 MG Oral Capsule	3951	28466
	Entire History	ibuprofen 600 MG Oral Tablet	1403	14479

Table B.19: Most representative labs, procedures, conditions, specialty visits, and medications for topic 19 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Chronic obstructive lung disease	54838	215275
	Past 365 days	Chronic obstructive lung disease	10895	43890
	Past 180 days	Chronic obstructive lung disease	4917	20125
	Entire History	Dyspnea	37891	217024
	Entire History	Uncomplicated asthma	4664	36291
Procedure	Entire History	Diffusing capacity eg, carbon monoxide, membrane List separately in addition to code for primary procedure	2276	14426
	Entire History	Bronchodilation responsiveness, spirometry as in 94010, pre- and post-bronchodilator administration	2105	13560
	Entire History	Spirometry, including graphic record, total and timed vital capacity, expiratory flow rate measurements, with or without maximal voluntary ventilation	2312	17337
	Entire History	Computed tomography, thorax; without contrast material	3657	22943
	Entire History	Plethysmography for determination of lung volumes and, when performed, airway resistance	1269	8326
Drug	Entire History	NDA021457 200 ACTUAT albuterol 0.09 MG/ACTUAT Metered Dose Inhaler [ProAir]	4563	30328
	Entire History	prednisone 10 MG Oral Tablet	4810	29049
	Entire History	albuterol 0.83 MG/ML Inhalation Solution	2942	13700
	Entire History	NDA021457 200 ACTUAT albuterol 0.09 MG/ACTUAT Metered Dose Inhaler	1134	7574
	Entire History	NDA020983 200 ACTUAT albuterol 0.09 MG/ACTUAT Metered Dose Inhaler [Ventolin]	2106	14254
Specialty	Entire History	Pulmonary Disease	2589	14061
	Entire History	Sleep Medicine	3575	23269
	Past 30 days	No matching concept	9979	71669
	Entire History	Diagnostic Radiology	27485	187103
	Past 365 days	No matching concept	112625	811174

Table B.20: Most representative labs, procedures, conditions, specialty visits, and medications for topic 20 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Scanning computerized ophthalmic diagnostic imaging, posterior segment, with interpretation and report, unilateral or bilateral; retina	10548	94642
	Past 365 days	Scanning computerized ophthalmic diagnostic imaging, posterior segment, with interpretation and report, unilateral or bilateral; retina	1902	18436
	Past 365 days	Ophthalmological services; comprehensive, established patient, 1 or more visits	2014	23158
	Past 180 days	Scanning computerized ophthalmic diagnostic imaging, posterior segment, with interpretation and report, unilateral or bilateral; retina	840	8253
	Entire History	Fundus photography with interpretation and report	3933	43688
Specialty	Past 365 days	Ophthalmology	5151	59577
	Past 180 days	Ophthalmology	2163	25313
	Entire History	Ophthalmology	37394	394685
	Past 30 days	Ophthalmology	490	5671
	Past 365 days	Optometry	836	9967
Condition	Entire History	Nonexudative age-related macular degeneration	10210	92371
	Entire History	Vitreous degeneration	10865	116057
	Past 365 days	Vitreous degeneration	1561	18464
	Past 365 days	Nonexudative age-related macular degeneration	1736	15631
	Entire History	Epiretinal membrane	5702	60967
Drug	Entire History	0.05 ML aflibercept 40 MG/ML Injection [Eylea]	1751	13532
	Entire History	4 ML bevacizumab 25 MG/ML Injection [Avastin]	638	7372
	Entire History	0.05 ML ranibizumab 10 MG/ML Prefilled Syringe [Lucentis]	849	8392
	Past 365 days	0.05 ML aflibercept 40 MG/ML Injection [Eylea]	393	3314
	Past 365 days	0.05 ML ranibizumab 10 MG/ML Prefilled Syringe [Lucentis]	318	3205

Table B.21: Most representative labs, procedures, conditions, specialty visits, and medications for topic 21 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Past 365 days	Osteoarthritis of knee	4630	58400
	Entire History	Osteoarthritis of knee	31776	355147
	Entire History	Pain in right knee	5799	75649
	Entire History	Pain in left knee	5167	70866
	Past 180 days	Osteoarthritis of knee	1971	23362
Procedure	Past 365 days	Arthrocentesis, aspiration and/or injection, major joint or bursa eg, shoulder, hip, knee, subacromial bursa; without ultrasound guidance	1192	12460
	Entire History	Arthrocentesis, aspiration and/or injection, major joint or bursa eg, shoulder, hip, knee, subacromial bursa; without ultrasound guidance	9807	90529
	Entire History	Radiologic examination, knee; complete, 4 or more views	2259	22196
	Past 180 days	Arthrocentesis, aspiration and/or injection, major joint or bursa eg, shoulder, hip, knee, subacromial bursa; without ultrasound guidance	512	5300
	Entire History	Radiologic examination, knee; 3 views	1959	21100
Specialty	Past 365 days	Orthopedic Surgery	1971	20910
	Entire History	Orthopedic Surgery	12317	122002
	Past 180 days	Orthopedic Surgery	848	8943
	Entire History	Sports Medicine	3071	36242
	Past 365 days	Sports Medicine	619	6970
Drug	Entire History	triamcinolone acetonide 40 MG/ML Injectable Suspension [Kenalog]	2020	20368
	Past 365 days	triamcinolone acetonide 40 MG/ML Injectable Suspension [Kenalog]	433	4221
	Past 180 days	triamcinolone acetonide 40 MG/ML Injectable Suspension [Kenalog]	186	1843
	Entire History	2 ML sodium hyaluronate 15 MG/ML Prefilled Syringe [Orthovisc]	1034	14153
	Entire History	methylprednisolone acetate 40 MG/ML Injectable Suspension [Depo-Medrol]	1675	14526

Table B.22: Most representative labs, procedures, conditions, specialty visits, and medications for topic 22 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Diagnostic mammography, including computer-aided detection CAD when performed; bilateral	961	13082
	Entire History	Diagnostic digital breast tomosynthesis, unilateral or bilateral list separately in addition to 77065 or 77066	589	9526
	Entire History	Diagnostic mammography, including computer-aided detection CAD when performed; unilateral	1087	16562
	Entire History	Ultrasound, breast, unilateral, real time with image documentation, including axilla when performed; limited	731	10684
	Entire History	Computer-aided detection computer algorithm analysis of digital image data for lesion detection with further review for interpretation, with or without digitization of film radiographic images; diagnostic mammography List separately in addition to code	757	11167
Condition	Entire History	Abnormal findings on diagnostic imaging of breast	1524	23549
	Entire History	Primary malignant neoplasm of female breast	8401	78690
	Entire History	Breast lump	903	10769
	Past 365 days	Primary malignant neoplasm of female breast	1362	9869
	Past 365 days	Abnormal findings on diagnostic imaging of breast	203	3170
Specialty	Entire History	Medical Oncology	13900	77931
	Past 365 days	Medical Oncology	4325	18897
	Entire History	Radiation Oncology	3731	38392
	Past 180 days	Medical Oncology	2341	9428
	Entire History	Diagnostic Radiology	27485	187103
Drug	Entire History	anastrozole 1 MG Oral Tablet	972	12285
	Entire History	2 ML fentanyl 0.05 MG/ML Injection	2683	20558
	Past 365 days	anastrozole 1 MG Oral Tablet	112	1816
	Entire History	letrozole 2.5 MG Oral Tablet	610	6912
	Past 180 days	anastrozole 1 MG Oral Tablet	65	995

Table B.23: Most representative labs, procedures, conditions, specialty visits, and medications for topic 23 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Past 365 days	Gastroenterology	2208	16696
	Entire History	Gastroenterology	11340	92146
	Past 180 days	Gastroenterology	1041	7209
	Past 365 days	Diagnostic Radiology	6691	38475
	Past 180 days	Diagnostic Radiology	3187	17224
Condition	Past 365 days	Abdominal pain	1467	9305
	Entire History	Abdominal pain	8970	67345
	Entire History	Disease of liver	2655	20634
	Entire History	Steatosis of liver	1729	20327
	Past 180 days	Abdominal pain	645	4189
Procedure	Entire History	Ultrasound, abdominal, real time with image documentation; complete	1328	11947
	Entire History	Computed tomography, abdomen and pelvis; with contrast materials	4128	27291
	Past 365 days	Computed tomography, abdomen and pelvis; with contrast materials	954	5404
	Entire History	Ultrasound, abdominal, real time with image documentation; limited eg, single organ, quadrant, follow-up	958	7375
	Past 365 days	Ultrasound, abdominal, real time with image documentation; complete	230	1949
Drug	Entire History	metronidazole 500 MG Oral Tablet	724	6332
	Past 365 days	metronidazole 500 MG Oral Tablet	86	815
	Entire History	dicyclomine hydrochloride 10 MG Oral Capsule	827	8585
	Entire History	Iohexol 755 MG/ML Injectable Solution [Omnipaque]	476	3925
	Entire History	no match	1561	12940

Table B.24: Most representative labs, procedures, conditions, specialty visits, and medications for topic 24 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Scanning computerized ophthalmic diagnostic imaging, posterior segment, with interpretation and report, unilateral or bilateral; optic nerve	3702	45569
	Entire History	Visual field examination, unilateral or bilateral, with interpretation and report; extended examination eg, Goldmann visual fields with at least 3 isopters plotted and static determination within the central 30, or quantitative, automated threshold perim	4118	50066
	Entire History	Gonioscopy separate procedure	1579	20416
	Past 365 days	Scanning computerized ophthalmic diagnostic imaging, posterior segment, with interpretation and report, unilateral or bilateral; optic nerve	476	6681
	Past 365 days	Visual field examination, unilateral or bilateral, with interpretation and report; extended examination eg, Goldmann visual fields with at least 3 isopters plotted and static determination within the central 30, or quantitative, automated threshold perim	523	6870
Condition	Entire History	Open-angle glaucoma borderline	5095	65835
	Entire History	Primary open angle glaucoma	14200	150586
	Past 365 days	Primary open angle glaucoma	2007	22925
	Past 365 days	Open-angle glaucoma borderline	942	13479
	Entire History	Glaucoma	3112	31727
Specialty	Past 365 days	Ophthalmology	5151	59577
	Past 180 days	Ophthalmology	2163	25313
	Entire History	Ophthalmology	37394	394685
	Past 365 days	Optometry	836	9967
	Entire History	Optometry	5353	62272
Drug	Entire History	latanoprost 0.05 MG/ML Ophthalmic Solution	8894	88327
	Past 365 days	latanoprost 0.05 MG/ML Ophthalmic Solution	1299	13442
	Past 180 days	latanoprost 0.05 MG/ML Ophthalmic Solution	707	7339
	Entire History	12 HR timolol 5 MG/ML Ophthalmic Solution	1584	17828
	Entire History	dorzolamide 20 MG/ML / timolol 5 MG/ML Ophthalmic Solution	1795	19676

Table B.25: Most representative labs, procedures, conditions, specialty visits, and medications for topic 25 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Anesthesia for intraperitoneal procedures in upper abdomen including laparoscopy; not otherwise specified	646	4829
	Entire History	Computed tomography, abdomen and pelvis; with contrast materials	4128	27291
	Entire History	Level III Surgical pathology, gross and microscopic examination Abortion, induced Abscess Aneurysm	1069	10831
	Entire History	Subsequent hospital care, per day, for the evaluation and management of a patient	50818	194961
	Entire History	Initial hospital care, per day, for the evaluation and management of a patient	11692	47164
Condition	Entire History	Abdominal pain	8970	67345
	Entire History	Acquired absence of organ	6727	15642
	Entire History	Disorder of digestive system	840	5920
	Entire History	Intestinal obstruction	4076	13445
	Entire History	Gallstone	1278	9042
Specialty	Entire History	Pathology	2759	22435
	Entire History	Anesthesiology	11414	99859
	Entire History	Gastroenterology	11340	92146
	Entire History	Emergency Medicine	20356	113387
	Entire History	Infectious Disease	4972	20544
Drug	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	5820	43831
	Entire History	metronidazole 500 MG Oral Tablet	724	6332
	Entire History	ciprofloxacin 500 MG Oral Tablet	2814	26196
	Entire History	oxycodone hydrochloride 5 MG Oral Tablet	2221	17459
	Entire History	neomycin sulfate 500 MG Oral Tablet	54	386

Table B.26: Most representative labs, procedures, conditions, specialty visits, and medications for topic 26 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Past 365 days	No matching concept	112625	811174
	Entire History	No matching concept	491045	3843783
	Past 180 days	No matching concept	54309	371773
	Entire History	Allergy / Immunology	1390	23156
	Past 365 days	Allergy / Immunology	201	4219
Procedure	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	102234	994395
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	124201	1073514
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	15845	157150
	Entire History	Immunization administration	32	3990
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	20683	182276
Condition	Entire History	Acute pharyngitis	1140	17098
	Entire History	Acute upper respiratory infection	3360	37397
	Entire History	Allergic rhinitis	5305	74756
	Entire History	Viral disease	757	9541
	Entire History	Mild intermittent asthma	1768	15867
Drug	Entire History	amoxicillin 80 MG/ML Oral Suspension	36	1434
	Entire History	NDA021457 200 ACTUAT albuterol 0.09 MG/ACTUAT Metered Dose Inhaler [ProAir]	4563	30328
	Entire History	fluticasone propionate 0.05 MG/ACTUAT Metered Dose Nasal Spray	7434	85406
	Entire History	albuterol 0.83 MG/ML Inhalation Solution	2942	13700
	Entire History	0.5 ML Neisseria meningitidis serogroup A capsular polysaccharide diphtheria toxoid protein conjugate vaccine 0.104 MG/ML / Neisseria meningitidis serogroup C capsular polysaccharide diphtheria toxoid protein conjugate vaccine 0.104 MG/ML / Neisseria m...	7	393

Table B.27: Most representative labs, procedures, conditions, specialty visits, and medications for topic 27 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Past 365 days	Essential hypertension	44851	374474
	Entire History	Essential hypertension	242191	2039327
	Past 180 days	Essential hypertension	19628	162539
	Entire History	Hyperlipidemia	109212	1061584
	Past 365 days	Hyperlipidemia	16119	160363
Specialty	Entire History	No matching concept	491045	3843783
	Past 365 days	No matching concept	112625	811174
	Past 180 days	No matching concept	54309	371773
	Entire History	Internal Medicine	60351	431318
	Past 365 days	Internal Medicine	14403	89915
Procedure	Entire History	Collection of venous blood by venipuncture	90495	754265
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	102234	994395
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	124201	1073514
	Past 365 days	Collection of venous blood by venipuncture	17457	137379
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	20683	182276
Drug	Entire History	hydrochlorothiazide 25 MG Oral Tablet	11459	127334
	Past 365 days	0.5 ML influenza A virus A	1454	16513
	Entire History	0.5 ML influenza A virus A	1454	16514
	Entire History	amlodipine 10 MG Oral Tablet	13835	124188
	Entire History	0.5 ML Streptococcus pneumoniae serotype 1 capsular antigen diphtheria CRM197 protein conjugate vaccine 0.0044 MG/ML	1352	15626

Table B.28: Most representative labs, procedures, conditions, specialty visits, and medications for topic 28 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Past 365 days	Low back pain	4842	46622
	Entire History	Lumbar radiculopathy	9518	87723
	Entire History	Low back pain	25443	256965
	Past 365 days	Lumbar radiculopathy	2312	20982
	Entire History	Spinal stenosis of lumbar region	12366	101865
Procedure	Entire History	Magnetic resonance eg, proton imaging, spinal canal and contents, lumbar; without contrast material	1492	14075
	Entire History	Radiologic examination, spine, lumbosacral; 2 or 3 views	1670	14926
	Entire History	Radiologic examination, spine, lumbosacral; minimum of 4 views	1232	12055
	Past 365 days	Magnetic resonance eg, proton imaging, spinal canal and contents, lumbar; without contrast material	249	2042
	Entire History	Injections, anesthetic agent and/or steroid, transforaminal epidural, with imaging guidance fluoroscopy or CT; lumbar or sacral, single level	1533	14679
Specialty	Past 30 days	No matching concept	9979	71669
	Past 180 days	No matching concept	54309	371773
	Past 365 days	No matching concept	112625	811174
	Past 365 days	Orthopedic Surgery	1971	20910
	Entire History	No matching concept	491045	3843783
Drug	Entire History	tramadol hydrochloride 50 MG Oral Tablet	10979	79865
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	3795	36927
	Past 365 days	tramadol hydrochloride 50 MG Oral Tablet	1540	11877
	Entire History	gabapentin 100 MG Oral Capsule	4907	35506
	Past 180 days	tramadol hydrochloride 50 MG Oral Tablet	877	6353

Table B.29: Most representative labs, procedures, conditions, specialty visits, and medications for topic 29 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Subsequent nursing facility care, per day, for the evaluation and management of a patient	5069	19621
	Entire History	Subsequent nursing facility care, per day, for the evaluation and management of a patient	5011	21151
	Entire History	Initial nursing facility care, per day, for the evaluation and management of a patient	1484	5852
	Entire History	Initial nursing facility care, per day, for the evaluation and management of a patient	1232	5567
	Entire History	Direct skilled nursing services of a registered nurse rn in the home health or hospice setting, each 15 minutes	43230	161698
Condition	Entire History	Muscle weakness	11997	85140
	Entire History	Difficulty walking	13010	92780
	Entire History	Patient dependence on care provider	10198	43939
	Entire History	Asthenia	8960	52546
	Entire History	Altered mental status	6731	19726
Specialty	Entire History	Geriatric Medicine	13566	112957
	Entire History	Infectious Disease	4972	20544
	Entire History	Psychiatry	4181	41283
	Entire History	Radiology	9555	67151
	Entire History	Podiatry	21829	162483
Drug	Entire History	pantoprazole 40 MG Delayed Release Oral Tablet	18146	125273
	Entire History	furosemide 20 MG Oral Tablet	12695	64959
	Entire History	metoprolol tartrate 25 MG Oral Tablet	11019	77113
	Entire History	ciprofloxacin 250 MG Oral Tablet	1420	11299
	Entire History	nitrofurantoin, macrocrystals 25 MG / nitrofurantoin, monohydrate 75 MG Oral Capsule	1121	9473

Table B.30: Most representative labs, procedures, conditions, specialty visits, and medications for topic 30 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Osteoarthritis of knee	31776	355147
	Entire History	Knee pain	5369	51151
	Entire History	Pain in right knee	5799	75649
	Entire History	Localized, primary osteoarthritis	4840	54269
	Entire History	Pain in left knee	5167	70866
Procedure	Entire History	Arthrocentesis, aspiration and/or injection, major joint or bursa eg, shoulder, hip, knee, subacromial bursa; without ultrasound guidance	9807	90529
	Entire History	Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	23929	295239
	Entire History	Manual therapy techniques eg, mobilization/ manipulation, manual lymphatic drainage, manual traction, 1 or more regions, each 15 minutes	11933	171792
	Entire History	Radiologic examination, knee; complete, 4 or more views	2259	22196
	Entire History	Radiologic examination, knee; 3 views	1959	21100
Specialty	Entire History	Orthopedic Surgery	12317	122002
	Entire History	Sports Medicine	3071	36242
	Entire History	Anesthesiology	11414	99859
	Entire History	Diagnostic Radiology	27485	187103
	Entire History	Pathology	2759	22435
Drug	Entire History	tramadol hydrochloride 50 MG Oral Tablet	10979	79865
	Entire History	triamcinolone acetonide 40 MG/ML Injectable Suspension [Kenalog]	2020	20368
	Entire History	oxycodone hydrochloride 5 MG Oral Tablet	2221	17459
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	5820	43831
	Entire History	amoxicillin 500 MG Oral Capsule	4366	49635

Table B.31: Most representative labs, procedures, conditions, specialty visits, and medications for topic 31 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Onychomycosis due to dermatophyte	17805	128199
	Past 365 days	Onychomycosis due to dermatophyte	3687	27428
	Past 180 days	Onychomycosis due to dermatophyte	1569	12057
	Entire History	Pain in toe	6509	56661
	Entire History	Callosity	5478	35814
Specialty	Entire History	Podiatry	21829	162483
	Past 365 days	Podiatry	4912	36842
	Past 180 days	Podiatry	2198	16509
	Past 30 days	Podiatry	479	3354
	Past 365 days	No matching concept	112625	811174
Procedure	Entire History	Debridement of nails by any methods; 6 or more	9704	69381
	Past 365 days	Debridement of nails by any methods; 6 or more	1970	14789
	Entire History	Paring or cutting of benign hyperkeratotic lesion eg, corn or callus; 2 to 4 lesions	3494	24521
	Past 180 days	Debridement of nails by any methods; 6 or more	879	6663
	Entire History	Avulsion of nail plate, partial or complete, simple; single	1496	12904
Drug	Entire History	cephalexin 500 MG Oral Capsule	3951	28466
	Entire History	ketoconazole 20 MG/ML Topical Cream	1111	10425
	Entire History	furosemide 20 MG Oral Tablet	12695	64959
	Entire History	0.5 ML influenza A virus A	1490	17052
	Entire History	ciclopirox 80 MG/ML Topical Solution	169	2318

Table B.32: Most representative labs, procedures, conditions, specialty visits, and medications for topic 32 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Radiologic examination, hand; minimum of 3 views	1102	11821
	Entire History	Radiologic examination, wrist; complete, minimum of 3 views	901	8732
	Entire History	Needle electromyography, each extremity, with related paraspinal areas, when performed, done with nerve conduction, amplitude and latency/velocity study; complete, five or more muscles studied, innervated by three or more nerves or four or more spinal lev	848	8468
	Entire History	Injections; single tendon sheath, or ligament, aponeurosis eg, plantar fascia	994	12204
	Entire History	Anesthesia for all procedures on nerves, muscles, tendons, fascia, and bursae of forearm, wrist, and hand	372	4152
Condition	Entire History	Wrist joint pain	2000	19014
	Entire History	Carpal tunnel syndrome	3576	40943
	Entire History	Localized, primary osteoarthritis of the hand	1530	15541
	Entire History	Pain in right hand	639	8106
	Entire History	Hand pain	742	8674
Specialty	Entire History	Orthopedic Surgery	12317	122002
	Entire History	Occupational Therapy	630	9676
	Entire History	Anesthesiology	11414	99859
	Entire History	Sports Medicine	3071	36242
	Past 365 days	Orthopedic Surgery	1971	20910
Drug	Entire History	betamethasone 3 MG/ML / betamethasone acetate 3 MG/ML Injectable Suspension	429	4467
	Entire History	triamcinolone acetonide 40 MG/ML Injectable Suspension [Kenalog]	2020	20368
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	5820	43831
	Entire History	acetaminophen 300 MG / codeine phosphate 30 MG Oral Tablet	3206	23218
	Entire History	cephalexin 500 MG Oral Capsule	3951	28466

Table B.33: Most representative labs, procedures, conditions, specialty visits, and medications for topic 33 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Entire History	Vascular Surgery	4717	25301
	Past 365 days	Vascular Surgery	1080	5536
	Past 180 days	Vascular Surgery	524	2370
	Entire History	Infectious Disease	4972	20544
	Entire History	Plastic surgery	1436	10974
Condition	Entire History	Peripheral vascular disease	23375	124767
	Entire History	Peripheral venous insufficiency	8229	45265
	Entire History	Atherosclerosis of artery of lower limb	6965	37181
	Past 365 days	Peripheral vascular disease	5157	26488
	Entire History	Localized edema	9123	51662
Procedure	Entire History	Duplex scan of extremity veins including responses to compression and other maneuvers; complete bilateral study	2762	14389
	Entire History	Complete bilateral noninvasive physiologic studies of upper or lower extremity arteries, 3 or more levels eg, for lower extremity: ankle/brachial indices at distal posterior tibial and anterior tibial/dorsalis pedis arteries plus segmental blood pressure	1018	5436
	Entire History	Duplex scan of extremity veins including responses to compression and other maneuvers; unilateral or limited study	2599	17565
	Entire History	Duplex scan of lower extremity arteries or arterial bypass grafts; complete bilateral study	784	4127
	Entire History	Debridement, subcutaneous tissue includes epidermis and dermis, if performed; first 20 sq cm or less	2114	8994
Drug	Entire History	cephalexin 500 MG Oral Capsule	3951	28466
	Entire History	sulfamethoxazole 800 MG / trimethoprim 160 MG Oral Tablet	3114	25009
	Past 365 days	cephalexin 500 MG Oral Capsule	667	4507
	Entire History	mupirocin 0.02 MG/MG Topical Ointment	1714	14639
	Past 180 days	cephalexin 500 MG Oral Capsule	371	2304

Table B.34: Most representative labs, procedures, conditions, specialty visits, and medications for topic 34 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Past 365 days	Subsequent hospital care, per day, for the evaluation and management of a patient	12424	41695
	Past 365 days	Initial hospital care, per day, for the evaluation and management of a patient	3186	11080
	Past 365 days	Initial hospital care, per day, for the evaluation and management of a patient	2469	9435
	Past 365 days	Subsequent hospital care, per day, for the evaluation and management of a patient	5517	18183
	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only	5246	24560
Specialty	Past 365 days	Emergency Medicine	4967	25070
	Past 180 days	Emergency Medicine	2365	11250
	Past 365 days	Diagnostic Radiology	6691	38475
	Past 180 days	Diagnostic Radiology	3187	17224
	Past 365 days	Internal Medicine	14403	89915
Condition	Past 365 days	Patient dependence on care provider	3392	13704
	Past 365 days	Acute renal failure syndrome	5329	17082
	Past 365 days	Abnormal findings on diagnostic imaging of lung	1974	9888
	Past 180 days	Patient dependence on care provider	1809	6645
	Past 365 days	Dyspnea	7916	41875
Drug	Past 180 days	pantoprazole 40 MG Delayed Release Oral Tablet	1859	10661
	Past 365 days	pantoprazole 40 MG Delayed Release Oral Tablet	3259	19744
	Past 365 days	metoprolol tartrate 25 MG Oral Tablet	1683	11148
	Past 365 days	oxycodone hydrochloride 5 MG Oral Tablet	495	3336
	Past 180 days	oxycodone hydrochloride 5 MG Oral Tablet	285	1834

Table B.35: Most representative labs, procedures, conditions, specialty visits, and medications for topic 35 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Magnetic resonance eg, proton imaging, spinal canal and contents, lumbar; without contrast material	1492	14075
	Entire History	Needle electromyography, each extremity, with related paraspinal areas, when performed, done with nerve conduction, amplitude and latency/velocity study; complete, five or more muscles studied, innervated by three or more nerves or four or more spinal lev	848	8468
	Entire History	Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	23929	295239
	Entire History	Radiologic examination, spine, lumbosacral; 2 or 3 views	1670	14926
	Entire History	Radiologic examination, spine, lumbosacral; minimum of 4 views	1232	12055
Condition	Entire History	Spinal stenosis of lumbar region	12366	101865
	Entire History	Lumbar radiculopathy	9518	87723
	Entire History	Low back pain	25443	256965
	Entire History	Degeneration of lumbar intervertebral disc	5562	54257
	Entire History	Peripheral neuritis	4355	41620
Specialty	Entire History	Orthopedic Surgery	12317	122002
	Entire History	Neurology	5376	43892
	Entire History	Anesthesiology	11414	99859
	Entire History	Neuroradiology	3148	21318
	Entire History	Sports Medicine	3071	36242
Drug	Entire History	tramadol hydrochloride 50 MG Oral Tablet	10979	79865
	Entire History	gabapentin 100 MG Oral Capsule	4907	35506
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	3795	36927
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	5820	43831
	Entire History	acetaminophen 325 MG / hydrocodone bitartrate 5 MG Oral Tablet	2957	23037

Table B.36: Most representative labs, procedures, conditions, specialty visits, and medications for topic 36 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Radiologic examination, hip, unilateral, with pelvis when performed; 2-3 views	1799	14288
	Past 365 days	Radiologic examination, hip, unilateral, with pelvis when performed; 2-3 views	457	3660
	Entire History	Radiologic examination, hip, unilateral, with pelvis when performed; 1 view	354	2664
	Entire History	Radiologic examination, pelvis; 1 or 2 views	775	5692
	Entire History	Radiologic examination, femur; minimum 2 views	469	2697
Condition	Entire History	Hip pain	6741	75425
	Past 365 days	Hip pain	1566	18373
	Entire History	Osteoarthritis of hip	5280	50035
	Entire History	Closed fracture of neck of femur	2281	13880
	Past 365 days	Osteoarthritis of hip	1069	10674
Specialty	Entire History	Orthopedic Surgery	12317	122002
	Past 365 days	Orthopedic Surgery	1971	20910
	Past 180 days	Orthopedic Surgery	848	8943
	Entire History	Anesthesiology	11414	99859
	Entire History	Internal Medicine	60351	431318
Drug	Entire History	tramadol hydrochloride 50 MG Oral Tablet	10979	79865
	Entire History	oxycodone hydrochloride 5 MG Oral Tablet	2221	17459
	Entire History	alendronic acid 70 MG Oral Tablet	3694	49060
	Past 365 days	tramadol hydrochloride 50 MG Oral Tablet	1540	11877
	Past 365 days	oxycodone hydrochloride 5 MG Oral Tablet	495	3336

Table B.37: Most representative labs, procedures, conditions, specialty visits, and medications for topic 37 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
	Entire History	2 ML fentanyl 0.05 MG/ML Injection	2683	20558
	Entire History	2 ML ondansetron 2 MG/ML Injection	2315	17067
	Past 365 days	2 ML fentanyl 0.05 MG/ML Injection	661	4736
	Entire History	cefazolin 1000 MG Injection	1028	7613
	Entire History	20 ML propofol 10 MG/ML Injection [Diprivan]	746	7108
Specialty	Past 365 days	Anesthesiology	2232	18738
	Entire History	Anesthesiology	11414	99859
	Past 180 days	Anesthesiology	1011	8094
	Entire History	Pathology	2759	22435
	Past 365 days	Pathology	502	3323
Procedure	Entire History	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	11322	70817
	Entire History	Blood typing, serologic; Rh D	1937	10514
	Entire History	Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only	23068	121633
	Entire History	Blood typing, serologic; ABO	2016	10791
	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only	5246	24560
Condition	Entire History	Postoperative state	3590	31273
	Past 365 days	Postoperative state	956	7543
	Entire History	Inguinal hernia	1058	11088
	Entire History	Persistent pain following procedure	1248	12017
	Past 180 days	Postoperative state	406	3046

Table B.38: Most representative labs, procedures, conditions, specialty visits, and medications for topic 38 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	23929	295239
	Past 365 days	Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	5255	61086
	Entire History	Manual therapy techniques eg, mobilization/ manipulation, manual lymphatic drainage, manual traction, 1 or more regions, each 15 minutes	11933	171792
	Entire History	Radiologic examination, shoulder; complete, minimum of 2 views	2411	21987
	Entire History	Therapeutic activities, direct one-on-one patient contact use of dynamic activities to improve functional performance, each 15 minutes	11887	121940
Condition	Entire History	Shoulder joint pain	10350	133283
	Past 365 days	Shoulder joint pain	1831	25989
	Entire History	Localized, primary osteoarthritis of the shoulder region	4171	38462
	Entire History	Nontraumatic rotator cuff tear	2459	31509
	Entire History	Impingement syndrome of shoulder region	1485	22131
Specialty	Entire History	Orthopedic Surgery	12317	122002
	Past 365 days	Orthopedic Surgery	1971	20910
	Entire History	Sports Medicine	3071	36242
	Past 30 days	No matching concept	9979	71669
	Past 180 days	Orthopedic Surgery	848	8943
Drug	Entire History	triamcinolone acetonide 40 MG/ML Injectable Suspension [Kenalog]	2020	20368
	Entire History	oxycodone hydrochloride 5 MG Oral Tablet	2221	17459
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	5820	43831
	Entire History	tramadol hydrochloride 50 MG Oral Tablet	10979	79865
	Entire History	methylprednisolone acetate 40 MG/ML Injectable Suspension [Depo-Medrol]	1675	14526

Table B.39: Most representative labs, procedures, conditions, specialty visits, and medications for topic 39 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
	Entire History	fluticasone propionate 0.05 MG/ACTUAT Metered Dose Nasal Spray	7434	85406
	Entire History	6 azithromycin 250 MG Oral Tablet Pack	5289	49956
	Entire History	amoxicillin 875 MG / clavulanate 125 MG Oral Tablet	2142	20855
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	3795	36927
	Past 365 days	fluticasone propionate 0.05 MG/ACTUAT Metered Dose Nasal Spray	894	11565
Condition	Entire History	Allergic rhinitis	5305	74756
	Entire History	Cough	11673	90141
	Entire History	Acute sinusitis	1511	18865
	Entire History	Chronic sinusitis	1796	21689
	Entire History	Acute upper respiratory infection	3360	37397
Specialty	Entire History	Otolaryngology	5044	55011
	Entire History	No matching concept	491045	3843783
	Past 365 days	No matching concept	112625	811174
	Past 180 days	No matching concept	54309	371773
	Entire History	Emergency Medicine	20356	113387
Procedure	Entire History	Nasal endoscopy, diagnostic, unilateral or bilateral separate procedure	1044	11714
	Entire History	Radiologic examination, chest; 2 views	7509	42340
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	15845	157150
	Past 180 days	Office or other outpatient visit for the evaluation and management of an established patient	7527	73624
	Entire History	Computed tomography, maxillofacial area; without contrast material	594	5379

Table B.40: Most representative labs, procedures, conditions, specialty visits, and medications for topic 40 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Entire History	Hypothyroidism	46453	429322
	Past 365 days	Hypothyroidism	9519	86538
	Entire History	Acquired hypothyroidism	14023	137462
	Past 180 days	Hypothyroidism	4144	36202
	Past 30 days	Hypothyroidism	795	7817
Drug	Entire History	levothyroxine sodium 0.075 MG Oral Tablet	5701	58532
	Entire History	levothyroxine sodium 0.1 MG Oral Tablet	4828	45194
	Entire History	levothyroxine sodium 0.025 MG Oral Tablet	4343	37326
	Past 365 days	levothyroxine sodium 0.075 MG Oral Tablet	778	8039
	Past 180 days	levothyroxine sodium 0.075 MG Oral Tablet	422	4326
Procedure	Entire History	Collection of venous blood by venipuncture	90495	754265
	Past 365 days	Collection of venous blood by venipuncture	17457	137379
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	124201	1073514
	Past 180 days	Collection of venous blood by venipuncture	7906	61965
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	102234	994395
Specialty	Entire History	No matching concept	491045	3843783
	Past 365 days	No matching concept	112625	811174
	Past 180 days	No matching concept	54309	371773
	Entire History	Endocrinology	7227	51085
	Past 30 days	No matching concept	9979	71669

Table B.41: Most representative labs, procedures, conditions, specialty visits, and medications for topic 41 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
	Past 365 days	Emergency Medicine	4967	25070
	Past 180 days	Emergency Medicine	2365	11250
	Entire History	Emergency Medicine	20356	113387
	Past 365 days	Internal Medicine	14403	89915
	Past 365 days	Diagnostic Radiology	6691	38475
Procedure	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	2292	13037
	Past 365 days	Emergency department visit for the evaluation and management of a patient	3988	18945
	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; interpretation and report only	5246	24560
	Past 365 days	Emergency department visit for the evaluation and management of a patient	2304	14474
	Past 180 days	Electrocardiogram, routine ECG with at least 12 leads; tracing only, without interpretation and report	924	5303
Condition	Past 365 days	Chest pain	4273	28076
	Past 365 days	Patient dependence on care provider	3392	13704
	Entire History	Chest pain	26088	191977
	Entire History	Patient dependence on care provider	10198	43939
	Past 365 days	Dizziness and giddiness	2145	17224
Drug	Past 365 days	1000 ML sodium chloride 9 MG/ML Injection	763	4090
	Entire History	1000 ML sodium chloride 9 MG/ML Injection	2066	13064
	Past 365 days	heparin sodium, porcine 5000 UNT/ML Injectable Solution	744	2550
	Entire History	heparin sodium, porcine 5000 UNT/ML Injectable Solution	2323	9796
	Past 365 days	2 ML ondansetron 2 MG/ML Injection	594	3764

Table B.42: Most representative labs, procedures, conditions, specialty visits, and medications for topic 42 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Condition	Past 365 days	Atherosclerosis of coronary artery without angina pectoris	15160	83943
	Entire History	Atherosclerosis of coronary artery without angina pectoris	67027	402617
	Past 180 days	Atherosclerosis of coronary artery without angina pectoris	6603	34933
	Entire History	Coronary atherosclerosis	8360	56127
	Entire History	Coronary arteriosclerosis	11285	72081
Drug	Entire History	clopidogrel 75 MG Oral Tablet	16430	107051
	Entire History	atorvastatin 40 MG Oral Tablet	14881	124415
	Past 365 days	clopidogrel 75 MG Oral Tablet	2367	14994
	Past 180 days	clopidogrel 75 MG Oral Tablet	1297	8066
	Entire History	nitroglycerin 0.4 MG Sublingual Tablet	705	4799
Procedure	Entire History	Myocardial perfusion imaging, tomographic SPECT	3623	29071
	Past 365 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	4529	36711
	Entire History	Catheter placement in coronary arterys for coronary angiography	681	5950
	Entire History	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	31136	253643
	Entire History	Echocardiography, transthoracic, real-time with image documentation 2D, includes M-mode recording, when performed, complete, with spectral Doppler echocardiography, and with color flow Doppler echocardiography	13647	94396
Specialty	Entire History	Interventional Cardiology	9925	57143
	Past 365 days	Interventional Cardiology	2530	13377
	Past 180 days	Interventional Cardiology	1261	6491
	Past 180 days	No matching concept	54309	371773
	Entire History	No matching concept	491045	3843783

Table B.43: Most representative labs, procedures, conditions, specialty visits, and medications for topic 43 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
	Entire History	Medical Oncology	13900	77931
	Past 365 days	Medical Oncology	4325	18897
	Past 180 days	Medical Oncology	2341	9428
	Past 30 days	Medical Oncology	472	1741
	Past 30 days	No matching concept	9979	71669
Condition	Entire History	Anemia	45831	279530
	Entire History	Iron deficiency anemia	21893	119903
	Past 365 days	Anemia	10068	50692
	Past 365 days	Iron deficiency anemia	5034	23853
	Entire History	Vitamin B deficiency	7851	70086
Procedure	Entire History	Therapeutic, prophylactic, or diagnostic injection specify substance or drug; subcutaneous or intramuscular	7846	49525
	Past 365 days	Therapeutic, prophylactic, or diagnostic injection specify substance or drug; subcutaneous or intramuscular	2020	10138
	Past 180 days	Collection of venous blood by venipuncture	7906	61965
	Entire History	Intravenous infusion, for therapy, prophylaxis, or diagnosis specify substance or drug; initial, up to 1 hour	2687	13385
	Past 180 days	Therapeutic, prophylactic, or diagnostic injection specify substance or drug; subcutaneous or intramuscular	922	4617
Drug	Entire History	vitamin B12 1 MG/ML Injectable Solution	1536	10024
	Past 365 days	vitamin B12 1 MG/ML Injectable Solution	311	2155
	Entire History	15 ML ferric carboxymaltose 50 MG/ML Injection [Injectafer]	240	1206
	Past 180 days	vitamin B12 1 MG/ML Injectable Solution	132	939
	Entire History	hydroxyurea 500 MG Oral Capsule	470	5066

Table B.44: Most representative labs, procedures, conditions, specialty visits, and medications for topic 44 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Entire History	Dermatology	11950	151270
	Past 365 days	Dermatology	2119	27485
	Entire History	Ophthalmology	37394	394685
	Past 365 days	No matching concept	112625	811174
	Entire History	No matching concept	491045	3843783
Condition	Entire History	Senile hyperkeratosis	4721	65927
	Entire History	Neoplasm of uncertain behavior of skin	6971	85098
	Entire History	Inflamed seborrheic keratosis	2962	43084
	Entire History	Actinic keratosis	11051	135967
	Entire History	Inflammatory dermatosis	2738	29866
Procedure	Entire History	Biopsy of skin, subcutaneous tissue and/or mucous membrane including simple closure, unless otherwise listed; single lesion	2785	34700
	Entire History	Level IV Surgical pathology, gross and microscopic examination Abortion	13388	147528
	Entire History	Destruction eg, laser surgery, electrosurgery, cryosurgery, chemosurgery, surgical curettement, premalignant lesions	4518	56334
	Entire History	Destruction eg, laser surgery, electrosurgery, cryosurgery, chemosurgery, surgical curettement, of benign lesions other than skin tags or cutaneous vascular proliferative lesions; up to 14 lesions	1473	22718
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient	2135	23172
Drug	Entire History	triamcinolone acetonide 1 MG/ML Topical Cream	1623	13634
	Entire History	6 azithromycin 250 MG Oral Tablet Pack	5289	49956
	Entire History	ketoconazole 20 MG/ML Topical Cream	1111	10425
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	3795	36927
	Entire History	0.5 ML influenza A virus A	1454	16514

Table B.45: Most representative labs, procedures, conditions, specialty visits, and medications for topic 45 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Specialty	Entire History	Rheumatology	4955	47434
	Past 365 days	Rheumatology	848	8898
	Past 180 days	Rheumatology	401	4324
	Entire History	Endocrinology	7227	51085
	Past 180 days	No matching concept	54309	371773
Procedure	Entire History	Ultrasound, soft tissues of head and neck eg, thyroid, parathyroid, parotid, real time with image documentation	874	12835
	Entire History	Dual-energy X-ray absorptiometry DXA, bone density study, 1 or more sites; axial skeleton eg, hips, pelvis, spine	2591	40159
	Past 365 days	Ultrasound, soft tissues of head and neck eg, thyroid, parathyroid, parotid, real time with image documentation	140	1990
	Past 180 days	Collection of venous blood by venipuncture	7906	61965
	Entire History	Screening mammography, bilateral 2-view study of each breast, including computer-aided detection CAD when performed	7292	129306
Condition	Entire History	Osteoporosis	14405	164047
	Entire History	Non-toxic uninodular goiter	2517	31352
	Entire History	Non-toxic multinodular goiter	2008	28223
	Past 365 days	Osteoporosis	2378	28241
	Entire History	Rheumatoid arthritis	6759	50732
Drug	Entire History	prednisone 5 MG Oral Tablet	3982	20653
	Entire History	1 ML denosumab 60 MG/ML Prefilled Syringe [Prolia]	572	6052
	Entire History	hydroxychloroquine sulfate 200 MG Oral Tablet	1977	15009
	Entire History	methotrexate 2.5 MG Oral Tablet	2126	19447
	Past 365 days	1 ML denosumab 60 MG/ML Prefilled Syringe [Prolia]	138	1478

Table B.46: Most representative labs, procedures, conditions, specialty visits, and medications for topic 46 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Esophagogastroduodenoscopy, flexible, transoral; with biopsy, single or multiple	2392	24384
	Entire History	Anesthesia for upper gastrointestinal endoscopic procedures, endoscope introduced proximal to duodenum	1602	12675
	Entire History	Anesthesia for upper gastrointestinal endoscopic procedures, endoscope introduced proximal to duodenum; not otherwise specified	965	7258
	Entire History	Level IV Surgical pathology, gross and microscopic examination Abortion	13388	147528
	Entire History	Esophagogastroduodenoscopy, flexible, transoral; diagnostic, including collection of specimens by brushing or washing, when performed separate procedure	515	3742
Condition	Entire History	Gastroesophageal reflux disease without esophagitis	21479	204765
	Entire History	Gastritis	1969	16917
	Entire History	Diaphragmatic hernia	2962	25587
	Past 365 days	Gastroesophageal reflux disease without esophagitis	3969	38874
	Entire History	Disorder of upper gastrointestinal tract	1475	14172
Specialty	Entire History	Gastroenterology	11340	92146
	Past 365 days	Gastroenterology	2208	16696
	Entire History	Anesthesiology	11414	99859
	Past 180 days	Gastroenterology	1041	7209
	Entire History	Pathology	2759	22435
Drug	Entire History	pantoprazole 40 MG Delayed Release Oral Tablet	18146	125273
	Entire History	omeprazole 40 MG Delayed Release Oral Capsule	9776	86403
	Past 365 days	pantoprazole 40 MG Delayed Release Oral Tablet	3259	19744
	Past 180 days	pantoprazole 40 MG Delayed Release Oral Tablet	1859	10661
	Entire History	omeprazole 20 MG Delayed Release Oral Capsule	13051	119603

Table B.47: Most representative labs, procedures, conditions, specialty visits, and medications for topic 47 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Screening mammography, bilateral 2-view study of each breast	7292	129306
	Entire History	Screening digital breast tomosynthesis, bilateral List separately in addition to code for primary procedure	3276	63579
	Entire History	Computer-aided detection computer algorithm analysis of digital image data for lesion detection with further review for interpretation	3798	62313
	Entire History	Dual-energy X-ray absorptiometry DXA, bone density study, 1 or more sites; axial skeleton eg, hips, pelvis, spine	2591	40159
	Past 365 days	Screening mammography, bilateral 2-view study of each breast, including computer-aided detection CAD when performed	681	15388
Specialty	Entire History	Obstetrics / Gynecology	2624	46649
	Entire History	Diagnostic Radiology	27485	187103
	Entire History	No matching concept	491045	3843783
	Past 365 days	No matching concept	112625	811174
	Past 365 days	Obstetrics / Gynecology	347	6528
Condition	Entire History	Disorder of bone	4866	75644
	Entire History	Osteoporosis	14405	164047
	Entire History	Menopause present	865	14584
	Entire History	Abnormal findings on diagnostic imaging of breast	1524	23549
	Entire History	Disorder of bone and articular cartilage	1234	20116
Drug	Entire History	alendronic acid 70 MG Oral Tablet	3694	49060
	Entire History	0.5 ML Streptococcus pneumoniae serotype 1 capsular antigen diphtheria CRM197 protein conjugate vaccine 0.0044 MG/ML	1352	15626
	Past 365 days	0.5 ML influenza A virus A		
	Entire History	0.5 ML influenza A virus A		
	Entire History	fluconazole 150 MG Oral Tablet	742	7518

Table B.48: Most representative labs, procedures, conditions, specialty visits, and medications for topic 48 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Comprehensive audiometry threshold evaluation and speech recognition 92553 and 92556 combined	1787	18841
	Entire History	Tympanometry impedance testing	943	10434
	Entire History	Tympanometry and reflex threshold measurements	721	7544
	Entire History	Removal impacted cerumen requiring instrumentation, unilateral	2071	18573
	Entire History	Removal of impacted cerumen one or both ears by physician on same date of service as audiologic function testing	403	4235
Specialty	Entire History	Otolaryngology	5044	55011
	Past 365 days	Otolaryngology	848	9759
	Past 180 days	Otolaryngology	347	4192
	Entire History	Audiology	348	3359
	Entire History	Ophthalmology	37394	394685
Condition	Entire History	Sensorineural hearing loss, bilateral	4613	45548
	Entire History	Impacted cerumen	5616	52915
	Past 365 days	Impacted cerumen	832	8099
	Entire History	Sensorineural hearing loss	1369	14430
	Past 365 days	Sensorineural hearing loss, bilateral	797	8338
Drug	Entire History	fluticasone propionate 0.05 MG/ACTUAT Metered Dose Nasal Spray	7434	85406
	Entire History	meclizine hydrochloride 25 MG Oral Tablet	1329	10141
	Entire History	hydrocortisone 10 MG/ML / neomycin 3.5 MG/ML / polymyxin B 10000 UNT/ML Otic Suspension	149	1570
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	3795	36927
	Entire History	meclizine hydrochloride 12.5 MG Oral Tablet	931	8217

Table B.49: Most representative labs, procedures, conditions, specialty visits, and medications for topic 49 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

Variable type	Variable evaluation period	Variable description	Num. hosp.	Num. no hosp.
Procedure	Entire History	Arterial catheterization or cannulation for sampling, monitoring or transfusion separate procedure; percutaneous	623	3755
	Entire History	Subsequent hospital care, per day, for the evaluation and management of a patient	19410	72503
	Entire History	Subsequent hospital care, per day, for the evaluation and management of a patient	50818	194961
	Entire History	Critical care, evaluation and management of the critically ill or critically injured patient; first 30-74 minutes	5140	21515
	Entire History	Computed tomographic angiography, chest noncoronary, with contrast materials, including noncontrast images, if performed, and image post-processing	1124	6381
Condition	Entire History	Atelectasis	3645	17761
	Entire History	Pleural effusion	7252	28151
	Entire History	Embolism from thrombosis of vein of lower extremity	4641	25445
	Entire History	Transplanted heart valve present	6074	31525
	Entire History	Postoperative state	3590	31273
Specialty	Entire History	Thoracic Surgery	929	5692
	Entire History	Vascular Surgery	4717	25301
	Entire History	Anesthesiology	11414	99859
	Entire History	Interventional Cardiology	9925	57143
	Entire History	Radiology	9555	67151
Drug	Entire History	warfarin sodium 5 MG Oral Tablet	5453	37662
	Entire History	apixaban 5 MG Oral Tablet [Eliquis]	5466	33209
	Entire History	furosemide 20 MG Oral Tablet	12695	64959
	Past 365 days	apixaban 5 MG Oral Tablet [Eliquis]	1886	12306
	Entire History	furosemide 40 MG Oral Tablet	13499	59492

Table B.50: Most representative labs, procedures, conditions, specialty visits, and medications for topic 50 for the Likelihood of Hospitalization task, including number of patients with and without hospitalization for each feature.

B.2 Topic Model for High-Risk Pregnancy

See Table B.51 to Table B.100 for the most representative conditions, procedures, specialty visits, and drugs for the 50 topics produced in the High-Risk Pregnancy topic modeling.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Level IV Surgical pathology, gross and microscopic examination Abortion	2125	7200
	Past 730 days	Level IV Surgical pathology, gross and microscopic examination Abortion	1650	5401
	Entire History	Colposcopy of the cervix including upper/adjacent vagina; with biopsys of the cervix and endocervical curettage	289	934
	Past 365 days	Level IV Surgical pathology, gross and microscopic examination Abortion	1238	3952
	Past 730 days	Colposcopy of the cervix including upper/adjacent vagina; with biopsys of the cervix and endocervical curettage	182	519
Condition	Entire History	Atypical squamous cells of undetermined significance on cervical Papanicolaou smear	373	1341
	Past 730 days	Atypical squamous cells of undetermined significance on cervical Papanicolaou smear	251	813
	Past 365 days	Atypical squamous cells of undetermined significance on cervical Papanicolaou smear	189	524
	Entire History	Cervical intraepithelial neoplasia grade 1	125	478
	Entire History	Human papilloma virus infection	124	439
Specialty	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Entire History	Pathology	526	2282
Drug	Entire History	6 azithromycin 250 MG Oral Tablet Pack	1636	5194
	Past 730 days	6 azithromycin 250 MG Oral Tablet Pack	1242	3657
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	761	2855
	Past 730 days	21 methylprednisolone 4 MG Oral Tablet Pack	552	2005
	Entire History	fluconazole 150 MG Oral Tablet	964	3242

Table B.51: Most representative labs, procedures, conditions, specialty visits, and medications for topic 1 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Joint pain	417	1480
	Past 730 days	Joint pain	286	825
	Entire History	Malaise	908	2786
	Entire History	Lyme disease	71	236
	Past 365 days	Joint pain	197	509
Procedure	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	3949	12284
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	3151	9889
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient	799	3332
	Past 180 days	Office or other outpatient visit for the evaluation and management of an established patient	2194	7137
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	4595	14418
Specialty	Entire History	Rheumatology	101	475
	Past 730 days	Rheumatology	74	370
	Past 365 days	Rheumatology	53	286
	Past 180 days	Rheumatology	37	212
	Entire History	Neurology	174	894
Drug	Entire History	doxycycline hyclate 100 MG Oral Capsule	448	1860
	Past 730 days	doxycycline hyclate 100 MG Oral Capsule	330	1320
	Entire History	meloxicam 15 MG Oral Tablet	94	419
	Entire History	doxycycline hyclate 100 MG Oral Tablet	266	1164
	Past 730 days	prednisone 5 MG Oral Tablet	36	188

Table B.52: Most representative labs, procedures, conditions, specialty visits, and medications for topic 2 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Specialty	Entire History	Dermatology	1518	4934
	Past 730 days	Dermatology	1224	3862
	Past 365 days	Dermatology	895	2855
	Past 180 days	Dermatology	537	1734
	Entire History	Dermatopathology	244	805
Condition	Entire History	Neoplasm of uncertain behavior of skin	805	2394
	Entire History	Acne	869	2435
	Past 730 days	Neoplasm of uncertain behavior of skin	568	1615
	Past 730 days	Acne	596	1594
	Entire History	Benign neoplasm of skin of trunk	703	1749
Procedure	Entire History	Level IV Surgical pathology, gross and microscopic examination Abortion	2125	7200
	Entire History	Biopsy of skin, subcutaneous tissue and/or mucous membrane including simple closure, unless otherwise listed; single lesion	682	1926
	Past 730 days	Level IV Surgical pathology, gross and microscopic examination Abortion	1650	5401
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A detailed history; A detailed examination; Medical decision making of low complexity. Counseling and/or coordination of care with	3770	11108
	Past 730 days	Biopsy of skin, subcutaneous tissue and/or mucous membrane including simple closure, unless otherwise listed; single lesion	456	1192
Drug	Entire History	ketoconazole 20 MG/ML Medicated Shampoo	97	320
	Entire History	clindamycin 10 MG/ML Topical Lotion	72	279
	Past 730 days	ketoconazole 20 MG/ML Medicated Shampoo	67	217
	Entire History	tretinoin 0.25 MG/ML Topical Cream	75	244
	Past 730 days	clindamycin 10 MG/ML Topical Lotion	54	192

Table B.53: Most representative labs, procedures, conditions, specialty visits, and medications for topic 3 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Specialty	Entire History	Orthopedic Surgery	356	1481
	Past 730 days	Orthopedic Surgery	244	944
	Entire History	Podiatry	383	1248
	Past 730 days	Podiatry	266	896
	Past 365 days	Orthopedic Surgery	170	612
Procedure	Entire History	Radiologic examination, foot; complete, minimum of 3 views	318	1173
	Entire History	Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	755	2579
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A detailed history; A detailed examination; Medical decision making of low complexity. Counseling and/or coordination of care with	3770	11108
	Past 730 days	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A detailed history; A detailed examination; Medical decision making of low complexity. Counseling and/or coordination of care with	2919	8043
	Past 730 days	Radiologic examination, foot; complete, minimum of 3 views	209	726
Condition	Entire History	Pain in limb	581	1712
	Entire History	Joint pain	417	1480
	Entire History	Arthralgia of the ankle and/or foot	249	924
	Entire History	Pain in lower limb	261	915
	Past 730 days	Pain in limb	377	804
Drug	Entire History	naproxen 500 MG Oral Tablet	337	1257
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	901	3379
	Past 730 days	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	605	2043
	Past 730 days	naproxen 500 MG Oral Tablet	226	795
	Entire History	meloxicam 15 MG Oral Tablet	94	419

Table B.54: Most representative labs, procedures, conditions, specialty visits, and medications for topic 4 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Annual gynecological examination, established patient	1477	4520
	Past 730 days	Periodic comprehensive preventive medicine reevaluation and management of an individual	4311	10308
	Past 730 days	Annual gynecological examination, established patient	1155	3398
	Entire History	Periodic comprehensive preventive medicine reevaluation and management of an individual	4932	12638
	Past 730 days	Cytopathology, cervical or vaginal any reporting system	3582	9726
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Past 30 days	Obstetrics/Gynecology	811	2670
Condition	Entire History	Routine antenatal care	2291	4396
	Past 730 days	Routine antenatal care	1621	2185
	Past 730 days	Malaise	589	1489
	Entire History	Malaise	908	2786
	Entire History	Pregnancy test negative	159	520
Drug	Past 730 days	Multivitamin preparation Oral Tablet	615	1270
	Past 365 days	Multivitamin preparation Oral Tablet	504	956
	Entire History	Multivitamin preparation Oral Tablet	755	1779
	Entire History	Multivitamin preparation Oral Capsule	340	776
	Past 180 days	Multivitamin preparation Oral Tablet	417	800

Table B.55: Most representative labs, procedures, conditions, specialty visits, and medications for topic 5 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Skeletal muscle relaxants; 3 or more	19	70
	Entire History	Drugs or substances, definitive, qualitative or quantitative, not otherwise specified; 7 or more	18	53
	Entire History	Antiepileptics, not otherwise specified; 7 or more	18	52
	Entire History	Analgesics, non-opioid; 6 or more	17	54
	Past 730 days	Drugs or substances, definitive, qualitative or quantitative, not otherwise specified; 7 or more	12	36

Table B.56: Most representative labs, procedures, conditions, specialty visits, and medications for topic 6 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Routine antenatal care	2291	4396
	Entire History	Single live birth	2012	5167
	Entire History	Delivery normal	1585	3954
	Entire History	Postpartum care	775	1793
	Entire History	Supervision of high risk pregnancy	1340	4949
Procedure	Entire History	Neuraxial labor analgesia/anesthesia for planned vaginal delivery this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor	1615	4075
	Entire History	Routine obstetric care including antepartum care, vaginal delivery with or without episiotomy, and/or forceps and postpartum care	1587	3834
	Entire History	Ultrasound, pregnant uterus, after first trimester > or = 14 weeks 0 days, transabdominal approach	1271	3645
	Entire History	Fetal non-stress test	1067	3160
	Entire History	Ultrasound, pregnant uterus plus detailed fetal anatomic examination, transabdominal approach	1163	3384
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Entire History	Anesthesiology	1556	5727
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 730 days	Anesthesiology	1133	4126
Drug	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	901	3379
	Entire History	ibuprofen 600 MG Oral Tablet	1015	3482
	Entire History	Multivitamin preparation Oral Tablet	755	1779
	Past 730 days	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	605	2043
	Past 730 days	ibuprofen 600 MG Oral Tablet	703	2308

Table B.57: Most representative labs, procedures, conditions, specialty visits, and medications for topic 7 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Fetal chromosomal aneuploidy , must include analysis of chromosomes 13, 18, and 21	318	1421
	Past 730 days	Fetal chromosomal aneuploidy , must include analysis of chromosomes 13, 18, and 21	238	1061
	Past 365 days	Fetal chromosomal aneuploidy , must include analysis of chromosomes 13, 18, and 21	147	687
	Past 180 days	Fetal chromosomal aneuploidy , must include analysis of chromosomes 13, 18, and 21	130	613
	Past 30 days	Fetal chromosomal aneuploidy , must include analysis of chromosomes 13, 18, and 21	52	163
Condition	Entire History	Multigravida of advanced maternal age	249	1251
	Past 730 days	Multigravida of advanced maternal age	185	986
	Entire History	High risk pregnancy	776	3791
	Entire History	Elderly primigravida	108	648
	Past 730 days	Supervision of high risk pregnancy	879	3228
Specialty	Past 180 days	Obstetrics/Gynecology	2658	8715
	Entire History	Dermatology	1518	4934
	Past 730 days	Psychology	156	560
	Entire History	Geriatric Medicine	382	1440
	Entire History	Sleep Medicine	68	330
Drug	Entire History	0.5 ML influenza A virus vaccine, A-Texas-50-2012 H3N2-like virus 0.03 MG/ML	41	56
	Entire History	zolpidem tartrate 12.5 MG Extended Release Oral Tablet	9	30
	Entire History	glycopyrrolate 2 MG Oral Tablet	6	13
	Entire History	metronidazole 0.01 MG/MG Topical Gel	15	76
	Past 730 days	hydrocortisone 25 MG/ML Topical Cream [Proctozone HC]	8	44

Table B.58: Most representative labs, procedures, conditions, specialty visits, and medications for topic 8 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Essential hypertension	128	1674
	Past 730 days	Essential hypertension	96	1423
	Past 365 days	Essential hypertension	65	1133
	Past 180 days	Essential hypertension	38	831
	Entire History	Benign essential hypertension	45	553
Procedure	Entire History	Subsequent hospital care, per day, for the evaluation and management of a patient	158	977
	Past 730 days	Subsequent hospital care, per day, for the evaluation and management of a patient	104	655
	Entire History	Initial hospital care, per day, for the evaluation and management of a patient	105	744
	Entire History	Echocardiography, transthoracic, real-time with image documentation 2D, includes M-mode recording, when performed, complete, with spectral Doppler echocardiography, and with color flow Doppler echocardiography	275	1358
	Past 365 days	Subsequent hospital care, per day, for the evaluation and management of a patient	59	425
Drug	Past 730 days	labetalol hydrochloride 100 MG Oral Tablet	8	216
	Entire History	labetalol hydrochloride 100 MG Oral Tablet	10	257
	Past 365 days	labetalol hydrochloride 100 MG Oral Tablet	5	177
	Past 180 days	labetalol hydrochloride 100 MG Oral Tablet	3	151
	Entire History	amlodipine 5 MG Oral Tablet	2	134
Specialty	Past 730 days	Radiology	571	2583
	Entire History	Infectious Disease	83	377
	Entire History	Endocrinology	238	1062
	Entire History	Nephrology	20	195
	Past 730 days	Infectious Disease	58	275

Table B.59: Most representative labs, procedures, conditions, specialty visits, and medications for topic 9 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Single live birth	2012	5167
	Past 730 days	Single live birth	1328	2952
	Entire History	Delivery normal	1585	3954
	Entire History	High risk pregnancy	776	3791
	Entire History	Supervision of high risk pregnancy	1340	4949
Procedure	Entire History	Neuraxial labor analgesia/anesthesia for planned vaginal delivery this includes any repeat subarachnoid needle placement and drug injection and/or any necessary replacement of an epidural catheter during labor	1615	4075
	Entire History	Routine obstetric care including antepartum care, vaginal delivery with or without episiotomy, and/or forceps and postpartum care	1587	3834
	Entire History	Ultrasound, pregnant uterus, re-evaluation of fetal size by measuring standard growth parameters and amniotic fluid volume, re-evaluation of organ systems suspected or confirmed to be abnormal on a prev	999	3500
	Entire History	Ultrasound, pregnant uterus plus detailed fetal anatomic examination, transabdominal approach	1163	3384
	Entire History	Ultrasound, pregnant uterus, first trimester $<$ 14 weeks 0 days, transabdominal approach	1724	6255
Specialty	Entire History	Anesthesiology	1556	5727
	Past 730 days	Anesthesiology	1133	4126
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 365 days	Obstetrics/Gynecology	3865	12456
Drug	Entire History	ibuprofen 600 MG Oral Tablet	1015	3482
	Past 730 days	ibuprofen 600 MG Oral Tablet	703	2308
	Entire History	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.01 MG/ML	467	1049
	Past 730 days	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.01 MG/ML	276	626
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	901	3379

Table B.60: Most representative labs, procedures, conditions, specialty visits, and medications for topic 10 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient, which requires at least 2 of these 3 key components: A problem focused history; A problem focused examination; Straightforward medical decision making. Counselin	1834	5640
	Past 730 days	Molecular diagnostics; interpretation and report	35	43
	Past 730 days	Molecular diagnostics; amplification, target, multiplex, each additional nucleic acid sequence beyond 2 List separately in addition to code for primary procedure	35	43
	Past 730 days	Molecular diagnostics; isolation or extraction of highly purified nucleic acid, each nucleic acid type ie, DNA or RNA	35	43
	Past 730 days	Molecular diagnostics; amplification, target, multiplex, first 2 nucleic acid sequences	33	44
Drug	Past 730 days	naproxen 375 MG Oral Tablet	34	76
	Past 730 days	ondansetron 8 MG Disintegrating Oral Tablet	83	256
	Entire History	ondansetron 8 MG Disintegrating Oral Tablet	118	435
	Past 730 days	2 ML sodium chloride 9 MG/ML Cartridge	2	9
	Entire History	naproxen 375 MG Oral Tablet	49	136
Condition	Past 365 days	Torticollis	13	37
	Past 365 days	High risk sexual behavior	41	73
	Past 180 days	Intra-abdominal and pelvic swelling, mass and lump	22	70
	Past 730 days	Disorder of menstruation	260	504
Specialty	Past 365 days	Medical Toxicology	6	23

Table B.61: Most representative labs, procedures, conditions, specialty visits, and medications for topic 11 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 30 days	Collection of venous blood by venipuncture	1430	4393
	Past 180 days	Collection of venous blood by venipuncture	2931	9312
	Past 365 days	Collection of venous blood by venipuncture	3801	11835
	Entire History	Collection of venous blood by venipuncture	5296	16203
	Past 730 days	Collection of venous blood by venipuncture	4704	14342
Specialty	Past 180 days	Internal Medicine	841	2734
	Past 30 days	Internal Medicine	157	666
	Past 365 days	Internal Medicine	1283	4087
	Past 730 days	Internal Medicine	1606	5306
	Past 30 days	Diagnostic Radiology	23	227
Condition	Past 30 days	Vitamin D deficiency	49	186
	Entire History	Obesity	352	2061
	Past 730 days	Essential hypertension	96	1423
	Past 365 days	Obesity	205	1101
	Past 730 days	Obesity	283	1588

Table B.62: Most representative labs, procedures, conditions, specialty visits, and medications for topic 12 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Collection of venous blood by venipuncture	5296	16203
	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Entire History	Periodic comprehensive preventive medicine reevaluation and management of an individual	4932	12638
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	7136	19444
	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	6546	17544
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Entire History	Internal Medicine	1923	6415
	Past 730 days	Internal Medicine	1606	5306
	Past 365 days	Obstetrics/Gynecology	3865	12456
Condition	Entire History	Vitamin D deficiency	703	2643
	Entire History	Malaise	908	2786
	Entire History	Fatigue	509	2318
	Entire History	Hyperlipidemia	328	1538
	Entire History	Acute upper respiratory infection	1513	4944
Drug	Entire History	6 azithromycin 250 MG Oral Tablet Pack	1636	5194
	Entire History	amoxicillin 500 MG Oral Capsule	1111	3755
	Entire History	ibuprofen 800 MG Oral Tablet	400	1585
	Entire History	ibuprofen 600 MG Oral Tablet	1015	3482
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	761	2855

Table B.63: Most representative labs, procedures, conditions, specialty visits, and medications for topic 13 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Specialty	Entire History	Ophthalmology	582	1887
	Entire History	Optometry	482	1574
	Past 730 days	Ophthalmology	435	1290
	Past 730 days	Optometry	358	1048
	Past 365 days	Ophthalmology	312	901
Procedure	Entire History	Ophthalmological services: comprehensive, established patient, 1 or more visits	290	909
	Entire History	Determination of refractive state	271	825
	Entire History	Ophthalmological services: medical examination and evaluation with initiation of diagnostic and treatment program; comprehensive, new patient, 1 or more visits	229	725
	Entire History	Ophthalmological services: intermediate, established patient	214	694
	Past 730 days	Ophthalmological services: comprehensive, established patient, 1 or more visits	229	653
Condition	Entire History	Tear film insufficiency	242	702
	Past 730 days	Tear film insufficiency	179	472
	Entire History	Visual disturbance	159	608
	Entire History	Myopia	115	361
	Past 365 days	Tear film insufficiency	128	298
Drug	Entire History	dexamethasone 1 MG/ML / tobramycin 3 MG/ML Ophthalmic Suspension	115	356
	Entire History	prednisolone acetate 10 MG/ML Ophthalmic Suspension	91	256
	Entire History	erythromycin 0.005 MG/MG Ophthalmic Ointment	113	407
	Past 730 days	dexamethasone 1 MG/ML / tobramycin 3 MG/ML Ophthalmic Suspension	78	186
	Past 730 days	erythromycin 0.005 MG/MG Ophthalmic Ointment	71	264

Table B.64: Most representative labs, procedures, conditions, specialty visits, and medications for topic 14 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Periodic comprehensive preventive medicine reevaluation and management of an individual	4932	12638
	Past 730 days	Periodic comprehensive preventive medicine reevaluation and management of an individual	4311	10308
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	7136	19444
	Entire History	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections	3630	10205
	Past 730 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections	2935	7772
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Entire History	Internal Medicine	1923	6415
Drug	Entire History	Multivitamin preparation Oral Tablet	755	1779
	Past 730 days	Multivitamin preparation Oral Tablet	615	1270
	Past 365 days	Multivitamin preparation Oral Tablet	504	956
	Past 180 days	Multivitamin preparation Oral Tablet	417	800
	Past 30 days	Multivitamin preparation Oral Tablet	269	489
Condition	Entire History	Routine antenatal care	2291	4396
	Past 180 days	Routine antenatal care	961	978
	Past 730 days	Routine antenatal care	1621	2185
	Past 365 days	Routine antenatal care	1131	1256
	Past 30 days	Routine antenatal care	753	645

Table B.65: Most representative labs, procedures, conditions, specialty visits, and medications for topic 15 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Entire History	Irregular periods	1120	3811
	Entire History	Polycystic ovary syndrome	154	1005
	Entire History	Female infertility	921	3626
	Past 730 days	Irregular periods	892	2833
	Past 730 days	Polycystic ovary syndrome	131	880
Procedure	Entire History	Ultrasound, transvaginal	1835	6863
	Past 730 days	Ultrasound, transvaginal	1510	5498
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient	799	3332
	Past 365 days	Ultrasound, transvaginal	1237	4363
	Past 730 days	Office or other outpatient visit for the evaluation and management of a new patient	582	2368
Drug	Entire History	clomiphene citrate 50 MG Oral Tablet	416	1719
	Entire History	medroxyprogesterone acetate 10 MG Oral Tablet	172	887
	Past 730 days	medroxyprogesterone acetate 10 MG Oral Tablet	144	688
	Past 730 days	clomiphene citrate 50 MG Oral Tablet	344	1373
	Past 365 days	medroxyprogesterone acetate 10 MG Oral Tablet	109	544
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Entire History	Endocrinology	238	1062

Table B.66: Most representative labs, procedures, conditions, specialty visits, and medications for topic 16 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Breast lump	413	1188
	Past 730 days	Breast lump	285	717
	Entire History	Pain of breast	273	899
	Past 730 days	Pain of breast	204	578
	Past 365 days	Breast lump	201	435
Procedure	Entire History	Screening mammography, bilateral 2-view study of each breast, including computer-aided detection CAD when performed	196	1117
	Entire History	Ultrasound, breast, unilateral, real time with image documentation, including axilla when performed; limited	224	1027
	Past 730 days	Screening mammography, bilateral 2-view study of each breast, including computer-aided detection CAD when performed	174	961
	Past 730 days	Ultrasound, breast, unilateral, real time with image documentation, including axilla when performed; limited	166	790
	Entire History	Diagnostic mammography, including computer-aided detection CAD when performed; bilateral	185	747
Specialty	Entire History	Diagnostic Radiology	966	4641
	Past 730 days	Diagnostic Radiology	703	3497
	Past 365 days	Diagnostic Radiology	492	2428
	Past 730 days	Radiology	571	2583
	Entire History	Radiology	806	3490
Drug	Entire History	ascorbic acid 30 MG	112	183
	Past 730 days	ascorbic acid 30 MG	99	110
	Past 365 days	ascorbic acid 30 MG	81	77
	Past 180 days	ascorbic acid 30 MG	67	67
	Past 30 days	ascorbic acid 30 MG	52	44

Table B.67: Most representative labs, procedures, conditions, specialty visits, and medications for topic 17 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Urinary tract infectious disease	1321	4626
	Past 730 days	Urinary tract infectious disease	943	2980
	Past 365 days	Urinary tract infectious disease	655	1975
	Past 180 days	Urinary tract infectious disease	428	1235
	Entire History	Dysuria	494	1901
Drug	Entire History	nitrofurantoin, macrocrystals 25 MG / nitrofurantoin, monohydrate 75 MG Oral Capsule	671	2405
	Past 730 days	nitrofurantoin, macrocrystals 25 MG / nitrofurantoin, monohydrate 75 MG Oral Capsule	493	1643
	Past 365 days	nitrofurantoin, macrocrystals 25 MG / nitrofurantoin, monohydrate 75 MG Oral Capsule	358	1081
	Past 180 days	nitrofurantoin, macrocrystals 25 MG / nitrofurantoin, monohydrate 75 MG Oral Capsule	236	690
	Entire History	sulfamethoxazole 800 MG / trimethoprim 160 MG Oral Tablet	670	2352
Procedure	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	6546	17544
	Entire History	Cytopathology, cervical or vaginal any reporting system	4236	12158
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	5574	14992
	Past 730 days	Cytopathology, cervical or vaginal any reporting system	3582	9726
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	7136	19444
Specialty	Entire History	Urology	114	595
	Past 730 days	Urology	82	400
	Past 365 days	Urology	56	283
	Past 365 days	Internal Medicine	1283	4087
	Past 730 days	Internal Medicine	1606	5306

Table B.68: Most representative labs, procedures, conditions, specialty visits, and medications for topic 18 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Entire History	Rheumatology	101	475
	Past 730 days	Rheumatology	74	370
	Past 365 days	Rheumatology	53	286
	Past 180 days	Rheumatology	37	212
	Entire History	Allergy/Immunology	234	806
Drug	Entire History	hydroxychloroquine sulfate 200 MG Oral Tablet	22	123
	Past 730 days	hydroxychloroquine sulfate 200 MG Oral Tablet	19	110
	Past 365 days	hydroxychloroquine sulfate 200 MG Oral Tablet	18	95
	Past 180 days	hydroxychloroquine sulfate 200 MG Oral Tablet	15	80
	Entire History	prednisone 10 MG Oral Tablet	333	1368
Condition	Entire History	Systemic lupus erythematosus	24	104
	Past 730 days	Systemic lupus erythematosus	13	85
	Entire History	Allergic rhinitis	962	2984
	Past 365 days	Systemic lupus erythematosus	11	70
	Entire History	Urticaria	137	470
Procedure	Entire History	Office consultation for a new or established patient of care with other physic	1073	4031
	Past 365 days	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A comprehensive history; A comprehensive examination; Medical decision making of moderate complexity. Counseling and/or coordinatio	1024	3114
	Past 730 days	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A comprehensive history; A comprehensive examination; Medical decision making of moderate complexity. Counseling and/or coordinatio	1434	4541
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A comprehensive history; A comprehensive examination; Medical decision making of moderate complexity. Counseling and/or coordinatio	1934	6528

Table B.69: Most representative labs, procedures, conditions, specialty visits, and medications for topic 19 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Entire History	Chiropractic manipulative treatment CMT; spinal, 3-4 regions	743	2063
	Entire History	Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility	755	2579
	Entire History	Chiropractic manipulative treatment CMT; extraspinal, 1 or more regions	529	1571
	Past 730 days	Chiropractic manipulative treatment CMT; spinal, 3-4 regions	577	1444
	Entire History	Manual therapy techniques eg, mobilization/ manipulation, manual lymphatic drainage, manual traction, 1 or more regions, each 15 minutes	681	2145
Condition	Entire History	Low back pain	879	3125
	Past 730 days	Low back pain	670	2232
	Entire History	Neck pain	676	2335
	Past 730 days	Neck pain	511	1629
	Past 365 days	Low back pain	484	1505
Drug	Entire History	cyclobenzaprine hydrochloride 10 MG Oral Tablet	287	1175
	Past 730 days	cyclobenzaprine hydrochloride 10 MG Oral Tablet	201	767
	Entire History	cyclobenzaprine hydrochloride 5 MG Oral Tablet	171	605
	Past 365 days	cyclobenzaprine hydrochloride 10 MG Oral Tablet	152	500
	Entire History	21 methylprednisolone 4 MG Oral Tablet Pack	761	2855
Specialty	Entire History	Orthopedic Surgery	356	1481
	Past 730 days	Orthopedic Surgery	244	944
	Entire History	Sports Medicine	265	1019
	Entire History	Radiology	806	3490
	Past 730 days	Sports Medicine	183	701

Table B.70: Most representative labs, procedures, conditions, specialty visits, and medications for topic 20 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Recurrent miscarriage	134	718
	Past 730 days	Recurrent miscarriage	106	592
	Past 365 days	Recurrent miscarriage	81	488
	Past 180 days	Recurrent miscarriage	58	359
	Entire History	Missed miscarriage	695	2450
Procedure	Entire History	Treatment of missed abortion, completed surgically; first trimester	405	1308
	Entire History	Ultrasound, pregnant uterus, real time with image documentation, transvaginal	1774	5919
	Entire History	Anesthesia for incomplete or missed abortion procedures	325	1148
	Past 730 days	Ultrasound, pregnant uterus, real time with image documentation, transvaginal	1199	4010
	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	738	2925
Specialty	Entire History	Medical Oncology	97	524
	Past 730 days	Medical Oncology	82	418
	Past 365 days	Medical Oncology	53	315
	Past 180 days	Medical Oncology	37	231
	Past 730 days	Psychology	156	560
Drug	Entire History	0.4 ML enoxaparin sodium 100 MG/ML Prefilled Syringe	33	232
	Past 730 days	0.4 ML enoxaparin sodium 100 MG/ML Prefilled Syringe	24	193
	Past 365 days	0.4 ML enoxaparin sodium 100 MG/ML Prefilled Syringe	20	153
	Past 180 days	0.4 ML enoxaparin sodium 100 MG/ML Prefilled Syringe	15	126
	Past 30 days	0.4 ML enoxaparin sodium 100 MG/ML Prefilled Syringe	10	77

Table B.71: Most representative labs, procedures, conditions, specialty visits, and medications for topic 21 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Past 365 days	Collection of venous blood by venipuncture	3801	11835
	Entire History	Collection of venous blood by venipuncture	5296	16203
	Past 730 days	Periodic comprehensive preventive medicine reevaluation and management of an individual	4311	10308
	Past 365 days	Periodic comprehensive preventive medicine reevaluation and management of an individual	3322	7472
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 365 days	Internal Medicine	1283	4087
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 730 days	Internal Medicine	1606	5306
	Entire History	Internal Medicine	1923	6415
Condition	Past 730 days	Vitamin D deficiency	570	2132
	Entire History	Vitamin D deficiency	703	2643
	Past 365 days	Vitamin D deficiency	458	1587
	Past 365 days	Fatigue	296	1346
	Entire History	Hyperlipidemia	328	1538
Drug	Entire History	naproxen 500 MG Oral Tablet	337	1257
	Entire History	ergocalciferol 1.25 MG Oral Capsule	162	689
	Entire History	6 azithromycin 250 MG Oral Tablet Pack	1636	5194
	Past 730 days	ergocalciferol 1.25 MG Oral Capsule	122	493
	Entire History	ciprofloxacin 500 MG Oral Tablet	488	1799

Table B.72: Most representative labs, procedures, conditions, specialty visits, and medications for topic 22 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Past 365 days	Multivitamin preparation Oral Tablet	504	956
	Past 730 days	Multivitamin preparation Oral Tablet	615	1270
	Past 180 days	Multivitamin preparation Oral Tablet	417	800
	Entire History	Multivitamin preparation Oral Tablet	755	1779
	Past 30 days	Multivitamin preparation Oral Tablet	269	489
Specialty	Entire History	Diagnostic Radiology	966	4641
	Past 730 days	Diagnostic Radiology	703	3497
	Past 180 days	Obstetrics/Gynecology	2658	8715
Condition	Entire History	Cystic fibrosis	5	108
	Past 730 days	Cystic fibrosis	5	91
	Past 365 days	Cystic fibrosis	5	84
	Past 180 days	Cystic fibrosis	4	74
	Entire History	Cystic fibrosis without meconium ileus	30	48
Procedure	Entire History	Office or other outpatient visit for the evaluation and management of a new patient	799	3332
	Past 730 days	Office or other outpatient visit for the evaluation and management of a new patient	582	2368
	Past 180 days	Office or other outpatient visit for the evaluation and management of a new patient	259	988
	Past 365 days	Office or other outpatient visit for the evaluation and management of a new patient	435	1688
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A detailed history; A detailed examination; Medical decision making of low complexity. Counseling and/or coordination of care with	3770	11108

Table B.73: Most representative labs, procedures, conditions, specialty visits, and medications for topic 23 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 730 days	Anesthesiology	1133	4126
	Entire History	Anesthesiology	1556	5727
	Entire History	Diagnostic Radiology	966	4641
Procedure	Entire History	Ultrasound, pregnant uterus, after first trimester > or = 14 weeks 0 days, transabdominal approach	1271	3645
	Past 730 days	Ultrasound, pregnant uterus, after first trimester > or = 14 weeks 0 days, transabdominal approach	685	1776
	Entire History	Collection of venous blood by venipuncture	5296	16203
	Entire History	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections	3630	10205
	Entire History	Ultrasound, pregnant uterus, real time with image documentation, first trimester fetal nuchal translucency measurement, transabdominal or transvaginal approach	1017	4027
Condition	Past 730 days	Unplanned pregnancy	725	1885
	Entire History	Unplanned pregnancy	1234	3588
	Entire History	Delivery normal	1585	3954
	Entire History	Amenorrhea	1278	4050
	Past 730 days	Amenorrhea	801	2557
Drug	Entire History	Multivitamin preparation Oral Tablet	755	1779
	Past 730 days	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.01 MG/ML	155	453
	Past 730 days	Multivitamin preparation Oral Tablet	615	1270
	Entire History	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.01 MG/ML	279	891
	Past 730 days	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.01 MG/ML	276	626

Table B.74: Most representative labs, procedures, conditions, specialty visits, and medications for topic 24 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Specialty	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 30 days	Obstetrics/Gynecology	811	2670
Condition	Past 730 days	Urinary tract infectious disease	943	2980
	Entire History	Urinary tract infectious disease	1321	4626
	Past 365 days	Urinary tract infectious disease	655	1975
	Past 180 days	Urinary tract infectious disease	428	1235
	Past 730 days	Finding of frequency of urination	150	665
Procedure	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	5574	14992
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	7136	19444
	Past 180 days	Office or other outpatient visit for the evaluation and management of an established patient	4121	11505
	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	6546	17544
	Past 730 days	Annual gynecological examination, established patient	1155	3398
Drug	Entire History	21 ethinyl estradiol 0.03 MG / norethindrone acetate 1.5 MG Oral Tablet / 7 ferrous fumarate 75 MG Oral Tablet Pack [Junel Fe 1.5/30 28 Day]	96	303
	Entire History	0.1 ML influenza A virus A	19	63
	Past 180 days	0.1 ML influenza A virus A	7	24
	Past 365 days	0.1 ML influenza A virus A	7	33
	Past 730 days	21 ethinyl estradiol 0.03 MG / norethindrone acetate 1.5 MG Oral Tablet / 7 ferrous fumarate 75 MG Oral Tablet Pack [Junel Fe 1.5/30 28 Day]	65	176

Table B.75: Most representative labs, procedures, conditions, specialty visits, and medications for topic 25 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Acquired hypothyroidism	473	1443
	Entire History	Hypothyroidism	324	1627
	Past 730 days	Hypothyroidism	296	1496
	Past 730 days	Acquired hypothyroidism	390	962
	Past 365 days	Hypothyroidism	254	1238
Procedure	Past 365 days	Collection of venous blood by venipuncture	3801	11835
	Past 180 days	Collection of venous blood by venipuncture	2931	9312
	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	3151	9889
	Entire History	Collection of venous blood by venipuncture	5296	16203
Specialty	Entire History	Endocrinology	238	1062
	Past 730 days	Endocrinology	192	856
	Past 365 days	Endocrinology	158	665
	Past 180 days	Endocrinology	118	475
	Past 30 days	Endocrinology	30	145
Drug	Entire History	levothyroxine sodium 0.075 MG Oral Tablet	131	441
	Past 730 days	levothyroxine sodium 0.075 MG Oral Tablet	115	381
	Entire History	levothyroxine sodium 0.025 MG Oral Tablet	134	527
	Past 365 days	levothyroxine sodium 0.075 MG Oral Tablet	98	322
	Past 730 days	levothyroxine sodium 0.025 MG Oral Tablet	115	446

Table B.76: Most representative labs, procedures, conditions, specialty visits, and medications for topic 26 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 180 days	Collection of venous blood by venipuncture	2931	9312
	Past 365 days	Collection of venous blood by venipuncture	3801	11835
	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Entire History	Collection of venous blood by venipuncture	5296	16203
	Past 180 days	Periodic comprehensive preventive medicine reevaluation and management of an individual	1857	4024
Specialty	Past 180 days	Internal Medicine	841	2734
	Past 730 days	Internal Medicine	1606	5306
	Past 365 days	Internal Medicine	1283	4087
	Entire History	Internal Medicine	1923	6415
	Past 180 days	Geriatric Medicine	139	505
Condition	Past 180 days	Vitamin D deficiency	276	1028
	Past 365 days	Vitamin D deficiency	458	1587
	Past 180 days	Fatigue	176	798
	Past 180 days	Hyperlipidemia	102	521
	Entire History	Hyperlipidemia	328	1538
Drug	Entire History	ergocalciferol 1.25 MG Oral Capsule	162	689
	Past 180 days	ergocalciferol 1.25 MG Oral Capsule	65	252
	Past 730 days	ergocalciferol 1.25 MG Oral Capsule	122	493
	Past 365 days	ergocalciferol 1.25 MG Oral Capsule	101	348
	Past 30 days	ergocalciferol 1.25 MG Oral Capsule	11	75

Table B.77: Most representative labs, procedures, conditions, specialty visits, and medications for topic 27 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Past 365 days	Collection of venous blood by venipuncture	3801	11835
	Entire History	Collection of venous blood by venipuncture	5296	16203
	Past 180 days	Collection of venous blood by venipuncture	2931	9312
	Entire History	Blood typing, serologic; ABO	2156	7088
Specialty	Past 730 days	Diagnostic Radiology	703	3497
	Entire History	Emergency Medicine	1503	6683
	Entire History	Diagnostic Radiology	966	4641
	Past 730 days	Emergency Medicine	1158	5288
	Entire History	Obstetrics/Gynecology	5005	16041
Drug	Entire History	acetaminophen 500 MG Oral Tablet	21	123
	Entire History	0.5 ML influenza A virus A	117	418
Condition	Entire History	Genitourinary tract hemorrhage	256	1302

Table B.78: Most representative labs, procedures, conditions, specialty visits, and medications for topic 28 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Entire History	Obesity	352	2061
	Past 730 days	Obesity	283	1588
	Past 365 days	Obesity	205	1101
	Entire History	Swelling / lump finding	179	535
	Entire History	Non-toxic uninodular goiter	150	505
Procedure	Entire History	Ultrasound, soft tissues of head and neck eg, thyroid, parathyroid, parotid, real time with image documentation	279	1048
	Entire History	Medical nutrition therapy; initial assessment and intervention, individual, face-to-face with the patient, each 15 minutes	291	1038
	Past 730 days	Medical nutrition therapy; initial assessment and intervention, individual, face-to-face with the patient, each 15 minutes	193	627
	Past 730 days	Ultrasound, soft tissues of head and neck eg, thyroid, parathyroid, parotid, real time with image documentation	220	732
	Entire History	Medical nutrition therapy; re-assessment and intervention, individual, face-to-face with the patient, each 15 minutes	169	579
Specialty	Entire History	Otolaryngology	332	1478
	Past 730 days	Otolaryngology	246	1055
	Past 365 days	Otolaryngology	171	726
	Entire History	Radiology	806	3490
	Entire History	Diagnostic Radiology	966	4641
Drug	Entire History	omeprazole 40 MG Delayed Release Oral Capsule	150	662
	Entire History	ergocalciferol 1.25 MG Oral Capsule	162	689
	Past 730 days	omeprazole 40 MG Delayed Release Oral Capsule	108	454
	Past 730 days	0.5 ML influenza A virus A	49	189
	Entire History	0.5 ML influenza A virus A	62	239

Table B.79: Most representative labs, procedures, conditions, specialty visits, and medications for topic 29 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Emergency department visit for the evaluation and management of a patient	1851	6928
	Past 730 days	Emergency department visit for the evaluation and management of a patient	1363	5010
	Past 365 days	Emergency department visit for the evaluation and management of a patient	958	3465
	Entire History	Emergency department visit for the evaluation and management of a patient	1027	4487
	Past 730 days	Emergency department visit for the evaluation and management of a patient	748	3172
Specialty	Past 730 days	Emergency Medicine	1158	5288
	Entire History	Emergency Medicine	1503	6683
	Past 365 days	Emergency Medicine	815	3789
	Past 180 days	Emergency Medicine	460	2343
	Past 730 days	Diagnostic Radiology	703	3497
Condition	Entire History	Abdominal pain	1193	4334
	Past 730 days	Abdominal pain	817	2738
	Entire History	Chest pain	625	2777
	Past 730 days	Chest pain	432	1883
	Past 365 days	Abdominal pain	539	1741
Drug	Entire History	2 ML ondansetron 2 MG/ML Injection	617	2531
	Past 730 days	2 ML ondansetron 2 MG/ML Injection	457	1808
	Entire History	ondansetron 4 MG Disintegrating Oral Tablet	355	1368
	Past 730 days	ondansetron 4 MG Disintegrating Oral Tablet	258	929
	Entire History	1000 ML sodium chloride 9 MG/ML Injection	234	1091

Table B.80: Most representative labs, procedures, conditions, specialty visits, and medications for topic 30 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Ultrasound, transvaginal	1510	5498
	Entire History	Ultrasound, transvaginal	1835	6863
	Entire History	Ultrasound, pelvic nonobstetric, real time with image documentation; complete	1141	4301
	Past 730 days	Ultrasound, pelvic nonobstetric, real time with image documentation; complete	867	3064
	Past 365 days	Ultrasound, transvaginal	1237	4363
Condition	Past 730 days	Irregular periods	892	2833
	Entire History	Irregular periods	1120	3811
	Past 365 days	Irregular periods	724	2167
	Past 180 days	Irregular periods	538	1527
	Entire History	Cyst of ovary	520	1856
Specialty	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 30 days	Obstetrics/Gynecology	811	2670
Drug	Entire History	progesterone 200 MG Oral Capsule	243	1203
	Past 730 days	progesterone 200 MG Oral Capsule	207	1027
	Entire History	medroxyprogesterone acetate 10 MG Oral Tablet	172	887
	Entire History	Multivitamin preparation Oral Tablet	755	1779
	Past 365 days	progesterone 200 MG Oral Capsule	171	843

Table B.81: Most representative labs, procedures, conditions, specialty visits, and medications for topic 31 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Vitamin D deficiency	703	2643
	Past 730 days	Vitamin D deficiency	570	2132
	Entire History	Anemia	541	2189
	Past 365 days	Vitamin D deficiency	458	1587
	Entire History	Iron deficiency anemia	211	892
Drug	Entire History	ergocalciferol 1.25 MG Oral Capsule	162	689
	Past 730 days	ergocalciferol 1.25 MG Oral Capsule	122	493
	Past 365 days	ergocalciferol 1.25 MG Oral Capsule	101	348
	Past 180 days	ergocalciferol 1.25 MG Oral Capsule	65	252
	Entire History	vitamin B12 1 MG/ML Injectable Solution	21	95
Procedure	Entire History	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	650	2794
	Past 730 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	474	1944
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A comprehensive history; A comprehensive examination; Medical decision making of moderate complexity. Counseling and/or coordinatio	1934	6528
	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	3949	12284
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	999	3911
Specialty	Entire History	Medical Oncology	97	524
	Past 730 days	Medical Oncology	82	418
	Entire History	Internal Medicine	1923	6415
	Past 365 days	Medical Oncology	53	315
	Past 730 days	Internal Medicine	1606	5306

Table B.82: Most representative labs, procedures, conditions, specialty visits, and medications for topic 32 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Acute pharyngitis	1473	4852
	Past 730 days	Acute pharyngitis	1073	3267
	Entire History	Acute upper respiratory infection	1513	4944
	Past 730 days	Acute upper respiratory infection	1113	3296
	Entire History	Acute sinusitis	1040	3378
Drug	Entire History	6 azithromycin 250 MG Oral Tablet Pack	1636	5194
	Past 730 days	6 azithromycin 250 MG Oral Tablet Pack	1242	3657
	Past 365 days	6 azithromycin 250 MG Oral Tablet Pack	935	2462
	Entire History	amoxicillin 875 MG / clavulanate 125 MG Oral Tablet	969	3276
	Entire History	amoxicillin 500 MG Oral Capsule	1111	3755
Procedure	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	7136	19444
	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	6546	17544
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	5574	14992
	Past 180 days	Office or other outpatient visit for the evaluation and management of an established patient	4121	11505
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	4595	14418
Specialty	Entire History	Internal Medicine	1923	6415
	Past 730 days	Internal Medicine	1606	5306
	Past 365 days	Internal Medicine	1283	4087
	Past 180 days	Internal Medicine	841	2734
	Entire History	Emergency Medicine	1503	6683

Table B.83: Most representative labs, procedures, conditions, specialty visits, and medications for topic 33 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 365 days	Blood typing, serologic; ABO	1256	4052
	Past 180 days	Blood typing, serologic; ABO	941	2913
	Past 180 days	Blood typing, serologic; Rh D	933	2916
	Past 365 days	Antibody screen, RBC, each serum technique	1019	3355
	Past 365 days	Blood typing, serologic; Rh D	1245	4065
Drug	Entire History	purified protein derivative of tuberculin 50 UNT/ML Injectable Solution [Tubersol]	144	404
	Past 730 days	purified protein derivative of tuberculin 50 UNT/ML Injectable Solution [Tubersol]	81	270
	Entire History	ondansetron 4 MG Oral Tablet	258	1016
	Past 730 days	0.5 ML measles virus vaccine live, Enders' attenuated Edmonston strain 2000 UNT/ML / mumps virus vaccine live, Jeryl Lynn strain 25000 UNT/ML / rubella virus vaccine live Wistar RA 27-3 strain 2000 UNT/ML Injection [M-M-R II]	21	94
	Past 365 days	purified protein derivative of tuberculin 50 UNT/ML Injectable Solution [Tubersol]	52	153
Condition	Past 730 days	Nonspecific tuberculin test reaction	48	169
	Past 365 days	Primary physiologic amenorrhea	51	225
	Entire History	Primary physiologic amenorrhea	83	340
	Past 365 days	Nonspecific tuberculin test reaction	37	110
	Past 180 days	Primary physiologic amenorrhea	42	184
Specialty	Past 365 days	Psychiatry	112	576
	Entire History	Psychiatry	190	961

Table B.84: Most representative labs, procedures, conditions, specialty visits, and medications for topic 34 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	6546	17544
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	7136	19444
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	5574	14992
	Entire History	Office or other outpatient visit for the evaluation and management of a new patient, which requires these 3 key components: A detailed history; A detailed examination; Medical decision making of low complexity. Counseling and/or coordination of care with	3770	11108
	Entire History	Comprehensive audiometry threshold evaluation and speech recognition 92553 and 92556 combined	157	627
Condition	Entire History	Inflammatory dermatosis	640	1938
	Entire History	Impacted cerumen	222	739
	Past 730 days	Inflammatory dermatosis	445	1095
	Past 730 days	Impacted cerumen	154	488
	Entire History	Eruption	431	1573
Drug	Entire History	cephalexin 500 MG Oral Capsule	602	2099
	Entire History	sulfamethoxazole 800 MG / trimethoprim 160 MG Oral Tablet	670	2352
	Past 730 days	cephalexin 500 MG Oral Capsule	424	1352
	Past 730 days	sulfamethoxazole 800 MG / trimethoprim 160 MG Oral Tablet	450	1520
	Entire History	mupirocin 0.02 MG/MG Topical Ointment	256	863
Specialty	Entire History	Otolaryngology	332	1478
	Entire History	Podiatry	383	1248
	Past 730 days	Otolaryngology	246	1055
	Past 730 days	Podiatry	266	896
	Past 365 days	Otolaryngology	171	726

Table B.85: Most representative labs, procedures, conditions, specialty visits, and medications for topic 35 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Acute vaginitis	372	1682
	Entire History	Vaginitis	739	2040
	Past 730 days	Acute vaginitis	291	1346
	Entire History	Candida infection of genital region	379	1438
	Entire History	Noninflammatory disorder of the vagina	462	1743
Drug	Entire History	fluconazole 150 MG Oral Tablet	964	3242
	Past 730 days	fluconazole 150 MG Oral Tablet	736	2295
	Entire History	metronidazole 500 MG Oral Tablet	426	1761
	Past 730 days	metronidazole 500 MG Oral Tablet	318	1217
	Past 365 days	fluconazole 150 MG Oral Tablet	521	1622
Specialty	Past 365 days	Obstetrics/Gynecology	3865	12456
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Past 30 days	Obstetrics/Gynecology	811	2670
Procedure	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	5574	14992
	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	6546	17544
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	7136	19444
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	4595	14418
	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	3949	12284

Table B.86: Most representative labs, procedures, conditions, specialty visits, and medications for topic 36 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Entire History	Collection of venous blood by venipuncture	5296	16203
	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Past 730 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections	2935	7772
	Entire History	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections	3630	10205
	Past 365 days	Collection of venous blood by venipuncture	3801	11835
Specialty	Entire History	Internal Medicine	1923	6415
	Past 365 days	Internal Medicine	1283	4087
	Past 730 days	Internal Medicine	1606	5306
	Past 180 days	Internal Medicine	841	2734
	Past 730 days	Geriatric Medicine	317	1133
Drug	Entire History	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.016 MG/ML	271	607
	Entire History	0.5 ML Bordetella pertussis filamentous hemagglutinin vaccine, inactivated 0.01 MG/ML	467	1049
	Entire History	fentanyl 0.05 MG/ML Injection	187	303
	Entire History	0.5 ML influenza A virus A	58	101
	Entire History	0.5 ML influenza A virus A 0.0	87	100

Table B.87: Most representative labs, procedures, conditions, specialty visits, and medications for topic 37 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Orthopantogram eg, panoramic x-ray	190	579
	Past 730 days	Orthopantogram eg, panoramic x-ray	104	298
	Entire History	Anesthesia for intraperitoneal procedures in upper abdomen including laparoscopy; not otherwise specified	67	275
	Entire History	DEEP SEDATION/GENERAL ANESTHESIA-FIRST 30 MINUTES	152	376
	Entire History	Surgical removal of impacted tooth	120	359
Condition	Entire History	Disturbance of tooth eruption or exfoliation	214	506
	Entire History	Diarrhea	432	1562
	Entire History	Kidney stone	113	475
	Past 730 days	Kidney stone	94	322
	Past 730 days	Disturbance of tooth eruption or exfoliation	96	176
Specialty	Entire History	Urology	114	595
	Past 730 days	Urology	82	400
	Past 365 days	Urology	56	283
	Entire History	Emergency Medicine	1503	6683
	Past 730 days	Emergency Medicine	1158	5288
Drug	Past 730 days	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	605	2043
	Entire History	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	901	3379
	Past 365 days	acetaminophen 325 MG / oxycodone hydrochloride 5 MG Oral Tablet	340	1086
	Entire History	tamsulosin hydrochloride 0.4 MG Oral Capsule	14	118
	Past 730 days	amoxicillin 500 MG Oral Capsule	826	2520

Table B.88: Most representative labs, procedures, conditions, specialty visits, and medications for topic 38 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Female infertility	921	3626
	Past 730 days	Female infertility	815	3139
	Entire History	Disorder of endocrine ovary	674	2937
	Past 365 days	Female infertility	730	2757
	Past 730 days	Disorder of endocrine ovary	587	2555
Procedure	Entire History	Catheterization and introduction of saline or contrast material for saline infusion sonohysterography SIS or hysterosalpingography	635	2631
	Entire History	Ultrasound, pelvic nonobstetric, real time with image documentation; limited or follow-up eg, for follicles	561	2341
	Past 730 days	Catheterization and introduction of saline or contrast material for saline infusion sonohysterography SIS or hysterosalpingography	529	2127
	Entire History	Ultrasound, transvaginal	1835	6863
	Past 730 days	Ultrasound, transvaginal	1510	5498
Drug	Entire History	clomiphene citrate 50 MG Oral Tablet	416	1719
	Past 730 days	clomiphene citrate 50 MG Oral Tablet	344	1373
	Past 365 days	clomiphene citrate 50 MG Oral Tablet	282	1110
	Entire History	estradiol 2 MG Oral Tablet	172	931
	Past 730 days	estradiol 2 MG Oral Tablet	157	850

Table B.89: Most representative labs, procedures, conditions, specialty visits, and medications for topic 39 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Past 365 days	Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum	39	371
	Past 730 days	Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum	95	628
	Past 730 days	Cyst of ovary	367	1273
	Entire History	Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum	123	927
	Past 180 days	Maternal obesity complicating pregnancy, childbirth and the puerperium, antepartum	24	270
Procedure	Entire History	Office or other outpatient visit for the evaluation and management of an established patient, that may not require the presence of a physician or other qualified health care professional. Usually, the presenting problems are minimal. Typically, 5 minute	904	3280
	Entire History	Fluoroscopy of Left Subclavian Artery using Low Osmolar Contrast	0	3
	Past 730 days	Fluoroscopy of Left Subclavian Artery using Low Osmolar Contrast	0	3
	Past 30 days	Ultrasound, pregnant uterus, first trimester $<$ 14 weeks 0 days, transabdominal approach	7	190
	Past 730 days	Pretreatment of RBCs for use in RBC antibody detection, identification, and/or compatibility testing; incubation with enzymes, each	0	4
Drug	Past 730 days	0.5 ML influenza A virus A/California/7/2009 H1N1 antigen 0.03 MG/ML	39	111
	Entire History	6 azithromycin 250 MG Oral Tablet Pack	1636	5194
	Entire History	doxycycline monohydrate 150 MG Oral Capsule	0	4
	Entire History	morphine sulfate 15 MG Extended Release Oral Tablet	3	21
	Past 730 days	doxycycline monohydrate 150 MG Oral Capsule	0	3

Table B.90: Most representative labs, procedures, conditions, specialty visits, and medications for topic 40 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Blood typing, serologic; Rh D	1688	5466
	Past 730 days	Blood typing, serologic; ABO	1693	5445
	Entire History	Blood typing, serologic; ABO	2156	7088
	Entire History	Blood typing, serologic; Rh D	2148	7096
	Past 365 days	Blood typing, serologic; ABO	1256	4052
Condition	Entire History	Missed miscarriage	695	2450
	Entire History	Threatened miscarriage	743	2716
	Past 730 days	Threatened miscarriage	581	1984
	Past 730 days	Missed miscarriage	578	1955
	Entire History	Miscarriage without complication	595	1973
Specialty	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Past 730 days	Emergency Medicine	1158	5288
Drug	Entire History	misoprostol 0.2 MG Oral Tablet	126	574
	Past 730 days	misoprostol 0.2 MG Oral Tablet	108	427
	Past 365 days	misoprostol 0.2 MG Oral Tablet	80	307
	Past 180 days	misoprostol 0.2 MG Oral Tablet	50	151
	Entire History	progesterone 200 MG Oral Capsule	243	1203

Table B.91: Most representative labs, procedures, conditions, specialty visits, and medications for topic 41 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Specialty	Entire History	Gastroenterology	423	1899
	Past 730 days	Gastroenterology	312	1407
	Past 365 days	Gastroenterology	224	989
	Past 180 days	Gastroenterology	142	657
	Entire History	Anesthesiology	1556	5727
Condition	Entire History	Abdominal pain	1193	4334
	Entire History	Epigastric pain	294	1152
	Past 730 days	Abdominal pain	817	2738
	Entire History	Gastroesophageal reflux disease	333	1137
	Entire History	Gastroesophageal reflux disease without esophagitis	217	1243
Procedure	Entire History	Esophagogastroduodenoscopy, flexible, transoral; with biopsy, single or multiple	189	940
	Entire History	Level IV Surgical pathology, gross and microscopic examination Abortion	2125	7200
	Entire History	Ultrasound, abdominal, real time with image documentation; complete	269	1163
	Entire History	Anesthesia for upper gastrointestinal endoscopic procedures, endoscope introduced proximal to duodenum	156	704
	Entire History	Anesthesia for lower intestinal endoscopic procedures, endoscope introduced distal to duodenum	160	670
Drug	Entire History	pantoprazole 40 MG Delayed Release Oral Tablet	135	659
	Entire History	omeprazole 40 MG Delayed Release Oral Capsule	150	662
	Past 730 days	pantoprazole 40 MG Delayed Release Oral Tablet	98	479
	Entire History	omeprazole 20 MG Delayed Release Oral Capsule	154	594
	Past 730 days	omeprazole 40 MG Delayed Release Oral Capsule	108	454

Table B.92: Most representative labs, procedures, conditions, specialty visits, and medications for topic 42 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 30 days	Collection of venous blood by venipuncture	1430	4393
	Past 180 days	Collection of venous blood by venipuncture	2931	9312
	Past 365 days	Collection of venous blood by venipuncture	3801	11835
	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Entire History	Collection of venous blood by venipuncture	5296	16203
Condition	Entire History	Routine antenatal care	2291	4396
	Past 30 days	Routine antenatal care	753	645
	Past 365 days	Routine antenatal care	1131	1256
	Past 180 days	Routine antenatal care	961	978
	Past 730 days	Routine antenatal care	1621	2185
Specialty	Past 30 days	Obstetrics/Gynecology	811	2670
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 30 days	Psychology	39	131
	Entire History	Neurology	174	894
Drug	Past 180 days	doxylamine succinate 10 MG	100	230
	Past 30 days	doxylamine succinate 10 MG	83	167
	Entire History	doxylamine succinate 10 MG	152	490
	Past 365 days	doxylamine succinate 10 MG	103	250
	Past 730 days	doxylamine succinate 10 MG	124	336

Table B.93: Most representative labs, procedures, conditions, specialty visits, and medications for topic 43 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections	2935	7772
	Entire History	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections; 1 vaccine single or combination vaccine/toxoid	3630	10205
	Past 365 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections; 1 vaccine single or combination vaccine/toxoid	2036	5108
	Past 180 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections; 1 vaccine single or combination vaccine/toxoid	1239	2923
	Entire History	Cytopathology, cervical or vaginal any reporting system	4236	12158
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Entire History	Internal Medicine	1923	6415
	Entire History	Emergency Medicine	1503	6683
Condition	Entire History	Fatigue	509	2318
	Past 730 days	Fatigue	420	1883
	Entire History	Blood chemistry abnormal	90	593
	Entire History	Vitamin D deficiency	703	2643
	Past 365 days	Fatigue	296	1346
Drug	Entire History	0.5 ML influenza A virus A 0.	105	325
	Past 730 days	0.5 ML influenza A virus A 0.	105	319
	Past 730 days	0.5 ML influenza A virus A	91	305
	Entire History	0.5 ML influenza A virus A	91	310
	Entire History	0.5 ML influenza A virus A	54	199

Table B.94: Most representative labs, procedures, conditions, specialty visits, and medications for topic 44 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 365 days	Collection of venous blood by venipuncture	3801	11835
	Past 730 days	Collection of venous blood by venipuncture	4704	14342
	Past 180 days	Collection of venous blood by venipuncture	2931	9312
	Entire History	Collection of venous blood by venipuncture	5296	16203
	Past 730 days	Immunization administration includes percutaneous, intradermal, subcutaneous, or intramuscular injections; 1 vaccine single or combination vaccine/toxoid	2935	7772
Specialty	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 365 days	Diagnostic Radiology	492	2428
	Past 730 days	Diagnostic Radiology	703	3497
	Past 730 days	Internal Medicine	1606	5306
	Past 180 days	Diagnostic Radiology	268	1455
Condition	Past 365 days	Routine antenatal care	1131	1256
	Past 180 days	Routine antenatal care	961	978
	Past 730 days	Routine antenatal care	1621	2185
	Past 365 days	Problem related to lifestyle	1	22
	Past 180 days	Problem related to lifestyle	1	13
Drug	Entire History	6 azithromycin 250 MG Oral Tablet Pack	1636	5194
	Past 730 days	doxylamine succinate 10 MG	124	336
	Past 365 days	doxylamine succinate 10 MG	103	250
	Entire History	doxylamine succinate 10 MG	152	490
	Past 180 days	doxylamine succinate 10 MG	100	230

Table B.95: Most representative labs, procedures, conditions, specialty visits, and medications for topic 45 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Entire History	Emergency department visit for the evaluation and management of a patient	1851	6928
	Entire History	Periodic comprehensive preventive medicine reevaluation and management of an individual	4932	12638
	Entire History	Cytopathology, cervical or vaginal any reporting system	4236	12158
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	4595	14418
	Entire History	Collection of venous blood by venipuncture	5296	16203
Specialty	Entire History	Obstetrics/Gynecology	5005	16041
	Entire History	Emergency Medicine	1503	6683
	Entire History	Diagnostic Radiology	966	4641
	Entire History	Anesthesiology	1556	5727
	Entire History	Dermatology	1518	4934
Condition	Entire History	Urinary tract infectious disease	1321	4626
	Entire History	Acute pharyngitis	1473	4852
	Entire History	Abdominal pain	1193	4334
	Entire History	Acute upper respiratory infection	1513	4944
	Entire History	Amenorrhea	1278	4050
Drug	Entire History	6 azithromycin 250 MG Oral Tablet Pack	1636	5194
	Entire History	ibuprofen 600 MG Oral Tablet	1015	3482
	Entire History	amoxicillin 500 MG Oral Capsule	1111	3755
	Entire History	amoxicillin 875 MG / clavulanate 125 MG Oral Tablet	969	3276
	Entire History	fluconazole 150 MG Oral Tablet	964	3242

Table B.96: Most representative labs, procedures, conditions, specialty visits, and medications for topic 46 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Condition	Entire History	Allergic rhinitis	962	2984
	Past 730 days	Allergic rhinitis	703	1887
	Entire History	Allergic rhinitis due to pollen	371	1109
	Past 365 days	Allergic rhinitis	500	1289
	Entire History	Asthma	372	1203
Procedure	Past 730 days	Office or other outpatient visit for the evaluation and management of an established patient	3949	12284
	Entire History	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	650	2794
	Entire History	Office or other outpatient visit for the evaluation and management of an established patient	4595	14418
	Past 365 days	Office or other outpatient visit for the evaluation and management of an established patient	3151	9889
	Past 730 days	Electrocardiogram, routine ECG with at least 12 leads; with interpretation and report	474	1944
Drug	Entire History	NDA021457 200 ACTUAT albuterol 0.09 MG/ACTUAT Metered Dose Inhaler [ProAir]	625	2285
	Entire History	montelukast 10 MG Oral Tablet	215	826
	Past 730 days	NDA021457 200 ACTUAT albuterol 0.09 MG/ACTUAT Metered Dose Inhaler [ProAir]	469	1535
	Entire History	fluticasone propionate 0.05 MG/ACTUAT Metered Dose Nasal Spray	943	3074
	Past 730 days	montelukast 10 MG Oral Tablet	176	589
Specialty	Entire History	Allergy/Immunology	234	806
	Past 730 days	Allergy/Immunology	172	571
	Entire History	Neurology	174	894
	Past 730 days	Neurology	123	644
	Past 365 days	Allergy/Immunology	128	423

Table B.97: Most representative labs, procedures, conditions, specialty visits, and medications for topic 47 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
Procedure	Past 730 days	Cytopathology, cervical or vaginal any reporting system	3582	9726
	Past 365 days	Cytopathology, cervical or vaginal any reporting system	2704	6957
	Entire History	Cytopathology, cervical or vaginal any reporting system	4236	12158
	Past 180 days	Cytopathology, cervical or vaginal any reporting system	1761	4343
	Entire History	Periodic comprehensive preventive medicine reevaluation and management of an individual	4932	12638
Specialty	Past 365 days	Obstetrics/Gynecology	3865	12456
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 180 days	Obstetrics/Gynecology	2658	8715
	Past 30 days	Obstetrics/Gynecology	811	2670
Condition	Entire History	Abnormal microbiological finding in specimen from female genital organ	76	287
	Past 730 days	Abnormal microbiological finding in specimen from female genital organ	63	203
	Past 365 days	Abnormal microbiological finding in specimen from female genital organ	36	127
	Past 180 days	Abnormal microbiological finding in specimen from female genital organ	25	87
	Entire History	Atypical squamous cells of undetermined significance on cervical Papanicolaou smear	373	1341
Drug	Past 730 days	amoxicillin 500 MG Oral Capsule	826	2520
	Entire History	24 ethinyl estradiol 0.01 MG / norethindrone acetate 1 MG Oral Tablet / 2 ethinyl estradiol 0.01 MG Oral Tablet / 2 ferrous fumarate 75 MG Oral Tablet Pack [Lo Loestrin Fe 28 Day]	175	564
	Entire History	21 DAY ethinyl estradiol 0.000625 MG/HR / etonogestrel 0.005 MG/HR Vaginal System [NuvaRing]	280	767
	Entire History	amoxicillin 500 MG Oral Capsule	1111	3755
	Past 730 days	21 DAY ethinyl estradiol 0.000625 MG/HR / etonogestrel 0.005 MG/HR Vaginal System [NuvaRing]	213	523

Table B.98: Most representative labs, procedures, conditions, specialty visits, and medications for topic 48 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Entire History	2 ML fentanyl 0.05 MG/ML Injection	342	1570
	Entire History	2 ML ondansetron 2 MG/ML Injection	617	2531
	Past 730 days	2 ML fentanyl 0.05 MG/ML Injection	263	1123
	Past 730 days	2 ML ondansetron 2 MG/ML Injection	457	1808
	Entire History	1 ML ketorolac tromethamine 30 MG/ML Injection	433	1805
Specialty	Entire History	Anesthesiology	1556	5727
	Past 730 days	Anesthesiology	1133	4126
	Past 365 days	Anesthesiology	534	2190
	Entire History	Pathology	526	2282
	Past 180 days	Anesthesiology	214	999
Procedure	Entire History	Level IV Surgical pathology, gross and microscopic examination Abortion	2125	7200
	Past 730 days	Level IV Surgical pathology, gross and microscopic examination Abortion	1650	5401
	Past 365 days	Level IV Surgical pathology, gross and microscopic examination Abortion	1238	3952
	Entire History	Anesthesia for intraperitoneal procedures in lower abdomen including laparoscopy; not otherwise specified	187	807
	Entire History	Level III Surgical pathology, gross and microscopic examination Abortion, induced Abscess Aneurysm	317	1122
Condition	Entire History	Legal termination of pregnancy without complication	91	346
	Past 730 days	Legal termination of pregnancy without complication	55	174
	Entire History	Endometriosis of pelvic peritoneum	59	195
	Entire History	Endometriosis	72	301
	Past 730 days	Postoperative state	56	314

Table B.99: Most representative labs, procedures, conditions, specialty visits, and medications for topic 49 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Variable type	Variable evaluation period	Variable description	Num. l.b.	Num. no l.b.
	Past 365 days	Supervision of high risk pregnancy	359	1774
	Past 730 days	Supervision of high risk pregnancy	879	3228
	Entire History	Supervision of high risk pregnancy	1340	4949
	Past 365 days	High risk pregnancy	251	1770
	Past 180 days	Supervision of high risk pregnancy	184	1157
Procedure	Past 365 days	Ultrasound, pregnant uterus, first trimester $<$ 14 weeks 0 days, transabdominal approach	551	2429
	Past 730 days	Ultrasound, pregnant uterus, first trimester $<$ 14 weeks 0 days, transabdominal approach	1094	3921
	Past 365 days	Ultrasound, pregnant uterus, real time with image documentation, first trimester fetal nuchal translucency measurement, transabdominal or transvaginal approach	145	1080
	Past 730 days	Ultrasound, pregnant uterus, real time with image documentation, first trimester fetal nuchal translucency measurement, transabdominal or transvaginal approach	528	2135
	Entire History	Ultrasound, pregnant uterus, real time with image documentation, first trimester fetal nuchal translucency measurement, transabdominal or transvaginal approach	1017	4027
Specialty	Past 180 days	Obstetrics/Gynecology	2658	8715
	Past 365 days	Obstetrics/Gynecology	3865	12456
	Entire History	Obstetrics/Gynecology	5005	16041
	Past 730 days	Obstetrics/Gynecology	4671	14872
	Past 730 days	Clinical Cytogenetics	38	329
Drug	Past 365 days	doxylamine succinate 10 MG	103	250
	Past 730 days	doxylamine succinate 10 MG	124	336
	Past 365 days	Multivitamin preparation Oral Tablet	504	956
	Past 180 days	doxylamine succinate 10 MG	100	230
	Past 180 days	Multivitamin preparation Oral Tablet	417	800

Table B.100: Most representative labs, procedures, conditions, specialty visits, and medications for topic 50 for the High-Risk Pregnancy task, including number of patients with and without confirmed live birth for each feature.

Appendix C

Additional Information for Chapter 5

- In Section C.1, we present the proof for conditions of identifiability and variational lower bound derivation.
- In Section C.2, we describe additional experiment setup details for the sigmoid, quadratic, and clinical experiments.
- In Section C.3, we present additional experimental results on discovering subtypes including visualization of subtypes, model misspecification, and missingness methods.
  - Visualizations of subtypes show that SubLign subtypes match the data generating function closer than SubNoLign.
  - Model misspecification results show that SubLign is robust to misspecification from cubic spline data generation.
  - Missingness method experiments show that SubLign outperforms baselines for various missingness rates with comparable performance to some baselines for no missingness.
- In Section C.4, we present quadratic experiment results.
  - An additional six synthetic experiments show that SubLign outperforms baselines.

– In cases where alignment values are not identifiable (e.g., one subtype's data generating function is a flat line), SubLign cannot recover those alignment values.

C.1 Identifiability and Inference

C.1.1 Identifiability

Algorithm 3 Restated procedure from Algorithm 2 for the identification of model parameters

Input: Observation times $X \in \mathbb{R}^{N \times M}$, biomarkers $Y \in \mathbb{R}^{N \times M \times D}$, polynomial degree P, invertible function f
Output: θ^P, $\delta_1, \ldots, \delta_N$, $s_1, \ldots, s_N$ for each patient
Step 1: Transform the observed biomarkers; $Q = f^{-1}(Y)$
Step 2: Obtain time-shifts using a single biomarker;
a) For each patient i, estimate the parameters $\dot{\theta^1}_i$ of $\kappa(x; \dot{\theta^1}_i)$ using a single biomarker $((x_{i,1}, q_{i,1}), \ldots, (x_{i,M}, q_{i,M})$ via polynomial regression,
b) Compute up to P roots of polynomial $\kappa(x, \dot{\theta^1}_i)$ for each patient i as $R_i = \{r_1, \ldots, r_P\}$ and set $\xi_i = \min \text{Real}(R_i)$ where Real denotes the real part of (potentially complex) roots.
c) Estimate $\tilde{\theta^1}_i$ for polynomials in a *canonical position* using $((x_{i,1} - \xi_i, q_{i,1}), \ldots, (x_{i,M} - \xi_i, q_{i,M})$ via polynomial regression,
d) Cluster $\tilde{\theta^j}_i$ across patients via K-means clustering to yield cluster identities $s_1, \ldots, s_N$
e) $\forall k, \;\; \eta_k = \min\{\xi_i \,|\, i \text{ s.t. } s_i = k\}$ and $\forall i, \;\; \delta_i = \xi_i - \eta_{s_i}$
Step 3: Estimate true polynomial coefficients using shifted observation times;
for biomarker $j = 1$ **to** J **do**
    For each patient, estimate the parameters θ_i^j of $\kappa(x; \hat{\theta^j}_i)$ using $((x_{1,1} - \delta_i, q_{1,1}[j]), \ldots, (x_{1,M} - \delta_i, q_{1,M}[j])$ via polynomial regression,
end for
Return $\theta^P = [\theta^1 | \ldots | \theta^J], \{\delta_1, \ldots, \delta_N\}, \{s_1, \ldots, s_N\}$

We restate our assumptions.

Assumption 1. f is invertible, and $\kappa(x, \theta) = \theta_0 + \sum_{p=1}^{P} \theta_p x^p$ describes a family of polynomial functions in x with parameter θ and degree $P > 0$. The parameters of each subtype are unique.

Assumption 2. $M \geq P + 1$, i.e., for each object, across all the D features, we observe at least $P + 1$ values.

Assumption 3. For each subtype s_k, there exists an object i whose alignment $\delta_i = 0$.

We provide the proof for Theorem 1 below:

Proof. The proof is constructive; i.e. we give an algorithm for the identification of the parameters of the model in Equation 5.4. The algorithm for identification is presented in Algorithm 2 and proceeds in three steps.

Step 1: The first step transforms the observed biomarkers by applying the inverse of function f, which exists by Assumption 1. This leaves us with data as:

$$f^{-1}(y_{i,m}) = \kappa(x_{i,m} + \delta_i; \theta^P_{s_i}) \ \ \forall i \in N, m \in M$$

i.e. for all bio-markers, across all patients, we have data arising from different polynomial functions.

Step 2: Without loss of generality, the second step uses the first biomarker to identify the values of δ_i for each patient.

a) First, we estimate the polynomial coefficients for each patient separately; we are guaranteed exact recovery of the coefficients by Assumption 2.

b) Next we find the roots for each polynomial. If they are complex, consider their real part, and define ξ_i to be the smallest root of the polynomial. At least one (real or complex) root is guaranteed to exist by the Fundamental Theorem of Algebra for every non-constant polynomial (Assumption 1). Note that the choice of using the smallest root is arbitrary; what matters is that a consistent choice of root is selected for each patient's polynomials.

c) The goal of this step to learn a new polynomial for each patient which is translated to ensure that the root selected in step b) lies at $x = 0$.

To do so, we first shift the observational time-steps by ξ_i, and we re-estimate the coefficients of each *shifted* polynomial.

We make use of the fact that if ξ_i is the smallest complex root of a polynomial $\kappa(x)$ then the polynomial $\kappa(x + \xi_i)$ has its smallest complex root at 0. We can recover the parameters of this polynomial exactly by shifting our observations and re-estimating the coefficients.

This operation recovers the coefficients of every patient's polynomial in its *canonical position* i.e. a translated polynomial whose the smallest root (or its real component) is at $x = 0$.

This step can be viewed as a de-biasing step which allows us to re-estimate $\tilde{\theta}$ without while ignoring the effect that left-censorship has on parameter estimates.

d) We cluster the coefficients estimated in step c). By construction, we know that $s_i = s_{i'} \iff \theta_i = \theta_{i'}$ which guarantees that clustering recovers the true-underlying subtype for each patient (up to a permutation over K choices).

e) Finally we stratify patients by their subtype, and we define δ_i as the difference between their smallest root and the smallest value of ξ_i among all other patients within that subtype.

By Assumption 3, we know that for each subtype, there exists a patient for whom $\delta_i = 0$, this reference patient will also be the one whose polynomial has the smallest root. We note here that without Assumption 3, we would still have identification of δ_i up to a constant.

Therefore, by shifting each patient's smallest root by their reference patient's smallest root, we can recover the original time-shifts.

Step 3: Given the values of $\delta_1, \ldots, \delta_N$ from Step 2, we can now estimate the true values of the polynomial coefficients exactly in the noiseless setting via polynomial regression. □

C.1.2 Variational Lower Bound

$$
\begin{aligned}
&\log p(Y|X;\gamma) \\
&\quad = \log \int_{Z,\delta} p(Y,Z,\delta|X;\gamma)\, dZ\, d\delta \\
&\quad = \log \int_{Z,\delta} q(Z|X,Y;\phi) \frac{p(Y,Z,\delta|X;\gamma)}{q(Z|X,Y;\phi)}\, dZ\, d\delta \\
&\quad \geq \int_{Z,\delta} q(Z|X,Y;\phi) \log \frac{p(Y,Z,\delta|X;\gamma)}{q(Z|X,Y;\phi)}\, dZ\, d\delta \qquad \text{(C.1)} \\
&\quad = \int_{\delta} q(\delta) \int_{Z} q(Z|X,Y;\phi) \log \frac{p(\delta)p(Y,Z|X,\delta;\gamma)}{q(\delta)q(Z|X,Y;\phi)} \\
&\quad = \int_{\delta} q(\delta) \int_{Z} q(Z|X,Y;\phi) \log \frac{p(Y,Z|X,\delta;\gamma)}{q(Z|X,Y;\phi)} + \int_{\delta} q(\delta) \int_{Z} q(Z|X,Y;\phi) \log \frac{p(\delta)}{q(\delta)} \\
&\quad = \mathbb{E}_{q(Z|X,Y;\phi)} \left[\log \frac{p(Y,Z,\delta|X,\delta;\gamma)}{q(Z|X,Y;\phi)} \right] \qquad \text{(C.2)}
\end{aligned}
$$

C.2 Experiment Setup Details

C.2.1 Optimal hyperparameters for sigmoid and clinical baselines

For all models, we run for 1000 epochs and use the model with the best training loss over the 1000 epochs for evaluation. For the sigmoid dataset, the optimal hyperparameters are latent space of dimension 5, 100 hidden units in the RNN, 50 hidden units in the multi-layer perceptron, learning rate of 0.01, and no regularization.

C.2.2 Missing values

SubLign allows for missing biomarker dimensions and missing patient visits to accommodate the sparsity of clinical data. For missing visits, we adapt the recognition network to handle variable sequence lengths. We mask out missing observations so they have no contribution to the learning stage, except for the recognition network input. For the recognition network input, we linearly interpolate missing values for each patient. For baselines that cannot handle missing data, we also linearly interpolate

missing values for each patient.

C.2.3 Statistical significance

To estimate robustness of our models, we evaluate our held-out performance over 5 trials. Each trial consists of randomized 60/20/20 training/validation/test data folds and a different random seed. In order to compare models across 5 trials, we report the means and standard deviations from the 5 trials. When the reported performance intervals overlap, we compute the statistical significance of the pairwise differences using a t-test and a Benjamini-Hochberg correction of 0.05.

C.3 Additional Experiment Results

- In Section C.3.1, we visualize the SubLign and SubNoLign subtypes for sigmoid data.
- In Section C.3.2, we present results on model misspecification.
- In Section C.3.3, we present the empirical results with varying levels of missingness.

C.3.1 SubLign and SubNoLign subtype visualization

In Figure C-1, we show the visualization for SubLign subtypes compared to SubNoLign for the first dimension of the sigmoid dataset. We find that the visualization of the SubNoLign subtypes are not as close to the data generating function as the SubLign subtypes. All other parameters, data dimensions, and experimental conditions are held constant.

C.3.2 Model Misspecification

Because our model implicitly assumes a functional form (e.g., sigmoid or quadratic), we investigate learning under model misspecification. With synthetic datasets created using splines on 5 randomly generated control points with the same noise rates,

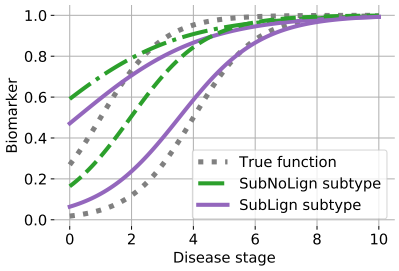


Figure C-1: One of three dimensions of learned SubLign and SubNoLign subtypes from sigmoid synthetic data plotted on top of original data generating functions.

dimensions, and censoring as the original experiments, we create two settings: one where the control points are monotonically increasing (using SubLign with a sigmoid f function) and one with no restrictions (using SubLign with a quadratic f function).

We run SubLign and use a sigmoid f function the monotonically increasing control points (labeled "Incr") and use quadratic f function for splines generated without restrictions (labeled "Any").

In Table C.1, we present results on model misspecification over 2 settings and 5 trials of SubLign learning on data generated from splines with 5 control points over 3 dimensions. We conclude that SubLign is robust against reasonable model misspecification using piecewise polynomial data.

Model	ARI↑	Pearson ↑	Swaps ↓
SubLign-Incr	0.82 ± 0.17	0.83 ± 0.08	0.14 ± 0.04
SubNoLign-Incr	0.77 ± 0.10	–	–
KMeans+Loss-Incr	0.58 ± 0.11	0.43 ± 0.09	0.21 ± 0.06
SubLign-Any	0.46 ± 0.12	0.67 ± 0.39	0.22 ± 0.14
SubNoLign-Any	0.29 ± 0.10	–	–
KMeans+Loss-Any	0.22 ± 0.07	0.23 ± 0.21	0.48 ± 0.11

Table C.1: Model misspecification experiment means and standard deviations using 5 cubic splines datasets.

C.3.3 Missingness Experiments

Here we present experiments varying the amount of missingness in synthetic datasets. Following the Parkinson's disease dataset (PPMI) where 47-60% of the biomarkers

are missing with a maximum of 17 observations for each patient, we modify our synthetic sigmoid dataset and remove biomarkers uniformly randomly with different missingness rates and $M = 17$.

Additionally we experiment with two missingness imputation methods for the baseline models. In addition to the linear interpolation used in the results presented in the main paper, we experiment with two other imputation methods: chained equations (MICE) [304] and a multi-directional recurrent neural network (MRNN) designed for multivariate time-series [312].

We find that SubLign outperforms baselines across higher missingness rates although with no missing observations, some baselines are comparable. In Table C.2, we show results with 50% of the data missing uniformly random. In Table C.3, we show results with 25% of the data missing uniformly random. In Table C.4, we show results with none of the data missing uniformly random.

Table C.2: Experiments on synthetic data with 50% of the data missing. Baselines include SuStaIn [313], BayLong [153], PAGA [307], SPARTan [237], clustering using Soft-DTW [74], and clustering using Kernel-DTW [81]. Imputation methods include MICE [304] and MRNN [312].

Imputation Method	Model	ARI ↑	Swaps ↓	Pearson ↑
–	SubLign	0.813 ± 0.024	0.299 ± 0.019	0.613 ± 0.049
–	SubNoLign	0.789 ± 0.058	–	–
MICE	KMeans+Loss	0.780 ± 0.046	0.327 ± 0.048	0.503 ± 0.018
	SuStaIn	0.459 ± 0.010	0.243 ± 0.004	0.160 ± 0.030
	BayLong	0.028 ± 0.003	0.480 ± 0.002	0.009 ± 0.003
	PAGA	0.003 ± 0.002	0.494± 0.028	0.034 ± 0.001
–	Soft-DTW	0.081 ± 0.004	–	–
–	Kernel-DTW	0.013 ± 0.002	–	–
	SPARTan	0.081 ± 0.013	–	–
MRNN	KMeans+Loss	0.783 ± 0.071	0.321 ± 0.120	0.562 ± 0.042
	SuStaIn	0.450 ± 0.120	0.304 ± 0.120	0.434 ± 0.120
	BayLong	0.028 ± 0.003	0.480 ± 0.002	0.009 ± 0.003
	PAGA	0.004 ± 0.001	0.492± 0.031	0.032 ± 0.003
	Soft-DTW	0.094 ± 0.005	–	–
	Kernel-DTW	0.000 ± 0.003	–	–
	SPARTan	0.091 ± 0.005	–	–

Table C.3: Experiments on synthetic data with 25% of the data missing. Baselines include SuStaIn [313], BayLong [153], PAGA [307], SPARTan [237], clustering using Soft-DTW [74], and clustering using Kernel-DTW [81]. Imputation methods include MICE [304] and MRNN [312].

Imputation Method	Model	ARI ↑	Swaps ↓	Pearson ↑
–	SubLign	0.811 ± 0.406	0.309 ± 0.028	0.554 ± 0.070
–	SubNoLign	0.743 ± 0.040	–	–
MICE	KMeans+Loss	0.741 ± 0.041	0.332 ± 0.048	0.508 ± 0.014
	SuStaIn	0.611 ± 0.406	0.309 ± 0.028	0.554 ± 0.070
	BayLong	0.111 ± 0.010	0.464 ± 0.071	0.011 ± 0.058
	PAGA	0.010 ± 0.003	0.433 ± 0.007	0.048 ± 0.030
	Soft-DTW	0.151 ± 0.031	–	–
	Kernel-DTW	0.002 ± 0.004	–	–
	SPARTan	0.168 ± 0.008	–	–
MRNN	KMeans+Loss	0.653 ± 0.029	0.308 ± 0.358	0.497 ± 0.025
	SuStaIn	0.615 ± 0.112	0.244 ± 0.004	0.577 ± 0.020
	BayLong	0.108 ± 0.016	0.461 ± 0.005	0.010 ± 0.041
	PAGA	0.011 ± 0.002	0.413 ± 0.005	0.031 ± 0.002
	Soft-DTW	0.103 ± 0.011	–	–
	Kernel-DTW	0.006 ± 0.004	–	–
	SPARTan	0.171 ± 0.041	–	–

C.4 Quadratic Data Results

We describe an additional set of experiments using the quadratic functional family. These experiments were designed to better understand where SubLign is able to learn clustering and alignment metrics well.

- In Section C.4.1, we detail the dataset creation.
- In Section C.4.2, we outline the optimal hyperparameters for the quadratic experiments.
- In Section C.4.3, we describe the empirical results for the quadratic experiments.

C.4.1 Setup

For the quadratic dataset, we generate data from 2 subtypes and 1 dimension with generating functions. See Table C.5 for subtype generating functions. Similar to the

Table C.4: Experiments on synthetic data with 0% of the data missing.

Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	0.980 ± 0.000	0.273 ± 0.012	0.714 ± 0.022
SubNoLign	0.809 ± 0.382	–	–
KMeans+Loss	0.980 ± 0.016	0.057 ± 0.112	0.480 ± 0.039
SuStaIn [313]	0.765 ± 0.012	0.144 ± 0.004	0.477 ± 0.020
BayLong [153]	0.201 ± 0.166	0.451 ± 0.050	0.011 ± 0.021
PAGA [307]	0.251 ± 0.031	0.481 ± 0.003	0.015 ± 0.002
Soft-DTW [74]	0.974 ± 0.015	–	–
Kernel-DTW [81]	0.949 ± 0.040	–	–
SPARTan [237]	0.251 ± 0.031	–	–

sigmoid synthetic dataset, for each patient in the datasets, we draw subtype $\kappa \sim Bern(0.5)$. The true disease stage is drawn $t_m \sim U(0, T^+)$ for observation $m \in [M]$. The biomarker values are drawn $y_m \sim N(\mu_m, S)$ where $\mu_m = \sum_{k \in [K]} \mathbb{1}(\kappa = k) f_k(t_m)$. The observed disease time x_m is shifted such that the first patient observation is at time 0. Therefore $x_m = t_m - \tau$ where $\tau = \min_{j \in [M]} t_j$ and is the earliest true disease time for the patient. We sample $N = 1000$ patients with $M = 4$ patient observations times with noise $S = 0.25$ and upper time bound $T^+ = 10$.

We construct our quadratic experiments such that we examine different model classes (i.e. flat, linear, quadratic) as well as examine subtypes that are overlapping or separable.

We include baseline results for the quadratic datasets. Note that SuStaIn [313] assumes monotonically increasing functions and is therefore omitted. We denote degenerate solutions with dashes.

C.4.2 Optimal hyperparameters for quadratic datasets

For the synthetic quadratic dataset corresponding to Figure C-2, we found the optimal hyperparameters for SubLign as no regularization, 5 hidden dimensions for the multi-layer perceptron, 200 latent dimensions, 200 units for the recurrent neural network, and learning rate of 0.001.

For the synthetic quadratic dataset corresponding to Figure C-3, we found the

Figure	Description	Subtype Generating Functions
C-2	Quadratic curve and flat line, separable	$f_1(t) = 0.25t^2 - 2.2t + 5$, $f_2(t) = 2$
C-3	Quadratic curve and flat line, overlapping	$f_1(t) = 0.25t^2 - 2.2t + 5$, $f_2(t) = -2$
C-4	Quadratic curve and sloped line, separable	$f_1(t) = 0.25t^2 - 2.2t + 5$, $f_2(t) = 0.4t$
C-5	Quadratic curve and sloped line, overlapping	$f_1(t) = 0.25t^2 - 2.2t + 5$, $f_2(t) = 0.4t - 5$
C-6	Quadratic curves in opposite directions, separable	$f_1(t) = 0.25t^2 - 2.2t + 3$, $f_2(t) = -0.25t^2 + 2.2 - 5$
C-7	Quadratic curves in opposite directions, overlapping	$f_1(t) = 0.25t^2 - 2.2t + 7$, $f_2(t) = -0.25t^2 + 2.2 - 5$

Table C.5: Quadratic dataset subtype generating functions and corresponding figure numbers

optimal hyperparameters for SubLign as no regularization, 5 hidden dimensions for the multi-layer perceptron, 200 latent dimensions, 200 units for the recurrent neural network, and learning rate of 0.001.

For the synthetic quadratic dataset corresponding to Figure C-4, we found the optimal hyperparameters for SubLign as no regularization, 5 hidden dimensions for the multi-layer perceptron, 200 latent dimensions, 200 units for the recurrent neural network, and learning rate of 0.001.

For the synthetic quadratic dataset corresponding to Figure C-5, we found the optimal hyperparameters for SubLign as no regularization, 5 hidden dimensions for the multi-layer perceptron, 200 latent dimensions, 200 units for the recurrent neural

network, and learning rate of 0.001.

For the synthetic quadratic dataset corresponding to Figure C-6, we found the optimal hyperparameters for SubLign as no regularization, 5 hidden dimensions for the multi-layer perceptron, 200 latent dimensions, 200 units for the recurrent neural network, and learning rate of 0.001.

For the synthetic quadratic dataset corresponding to Figure C-6, the optimal hyperparameters are latent space of dimension 10, 20 hidden units in the RNN, 50 hidden units in the multi-layer perceptron, learning rate of 0.01, and no regularization.

For the synthetic quadratic dataset corresponding to Figure C-7, the optimal hyperparameters are latent space of dimension 5, 100 hidden units in the RNN, 50 hidden units in the multi-layer perceptron, learning rate of 0.01, and no regularization.

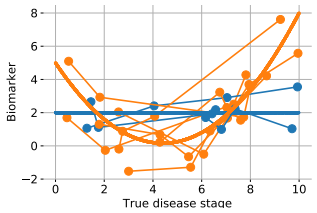
C.4.3 Results

In Figures C-2 to C-7, we present the quantitative results of 6 different quadratic cases as well as a plot of example data and the data generating subtypes. In each, we see that SubLign or SubNoLign outperforms the baselines.

When the subtypes are separable (i.e. Fig C-3, C-5, and C-7), SubLign handily recovers the subtypes. When the subtypes are not separable (i.e. Fig C-2, C-4, and C-6), SubLign still outperforms baselines.

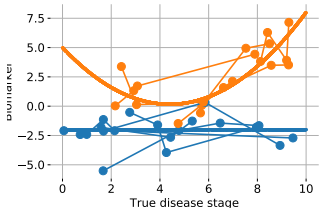
We note that alignment metrics are especially challenging to recover when one subtype is a flat or sloped line as in with Figure C-2 to C-7. Because the alignment metric is entirely unidentifiable, the swaps and Pearson metrics suffer. Note that for the swaps metric, 0.5 corresponds to random guessing, so the lack of identifiability of one of the subtypes would cause a swaps metric of 0.25. When the second subtype has a changing slope, as in Figure C-6, the alignment metrics are more recoverable.

When the model is degenerate and does not return the alignment values, we denote this with an empty cell.



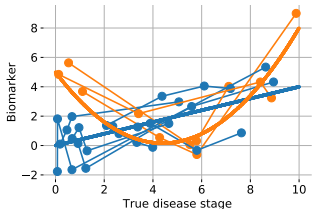
Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	0.277 ± 0.040	0.277 ± 0.014	0.516 ± 0.007
SubNoLign	0.103 ± 0.002	–	–
KMeans+Loss	0.213 ± 0.009	0.498 ± 0.022	0.016 ± 0.051
SuStaIn [313]	0.151	0.203	0.000
Bayesian [153]	0.000 ± 0.000	0.501 ± 0.017	0.018 ± 0.125
PAGA [307]	0.027 ± 0.001	–	–

Figure C-2: Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign outperforms baselines while KMeans+Loss recovers subtypes (ARI metric) better than SubNoLign, but alignment metrics are difficult to recover because of the horizontal subtype



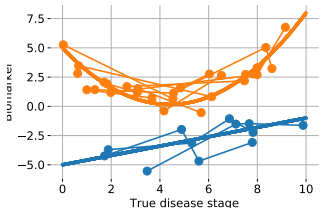
MODEL	**ARI ↑**	**SWAPS ↓**	**PEARSON ↑**
SubLign	0.980 ± 0.000	0.253 ± 0.001	0.527 ± 0.011
SubNoLign	0.980 ± 0.000	–	–
KMeans+Loss	0.883 ± 0.000	0.471 ± 0.011	0.064 ± 0.067
SuStaIn [313]	0.228 ± 0.039	0.182 ± 0.010	0.000 ± 0.000
Bayesian [153]	0.198 ± 0.189	0.446 ± 0.052	0.157 ± 0.286
PAGA [307]	0.227 ± 0.035	–	–

Figure C-3: Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign and SubNoLign have near-perfect clustering accuracy (ARI) while alignment metrics (swaps, Pearson) are difficult recover because of the horizontal subtype.



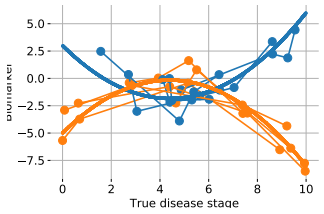
Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	0.122 ± 0.006	0.272 ± 0.001	0.621 ± 0.020
SubNoLign	0.145 ± 0.006	–	–
KMeans+Loss	0.031 ± 0.030	0.302 ± 0.026	0.498 ± 0.010
SuStaIn [313]	0.138 ± 0.019	0.119 ± 0.006	0.000 ± 0.000
Bayesian [153]	0.001 ± 0.002	–	–
PAGA [307]	0.009 ± 0.000	–	–

Figure C-4: Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign outperforms baselines in clustering and alignment metrics although the task is challenging.



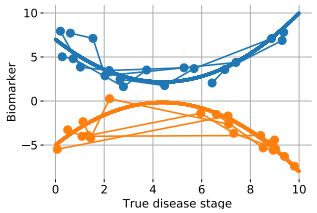
Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	0.968 ± 0.016	0.253 ± 0.015	0.604 ± 0.041
SubNoLign	0.968 ± 0.016	–	–
KMeans+Loss	0.964 ± 0.008	0.486 ± 0.032	0.044 ± 0.101
SuStaIn [313]	0.220 ± 0.011	0.196 ± 0.008	0.000 ± 0.000
Bayesian [153]	0.221 ± 0.214	0.448 ± 0.070	0.165 ± 0.193
PAGA [307]	0.205 ± 0.012	–	–

Figure C-5: Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign, SubNoLign, and KMeans+Loss perform well on clustering.



Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	0.729 ± 0.049	0.153 ± 0.006	0.843 ± 0.016
SubNoLign	0.721 ± 0.035	–	–
KMeans+Loss	0.540 ± 0.034	0.490 ± 0.020	0.043 ± 0.057
SuStaIn [313]	0.198 ± 0.024	0.200 ± 0.008	0.000 ± 0.000
Bayesian [153]	0.003 ± 0.007	–	–
PAGA [307]	0.059 ± 0.006	–	–

Figure C-6: Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign learns subtypes and recovers alignment better than baselines.



Model	ARI ↑	Swaps ↓	Pearson ↑
SubLign	1.000 ± 0.000	0.134 ± 0.012	0.907 ± 0.009
SubNoLign	0.984 ± 0.008	–	–
KMeans+Loss	1.000 ± 0.000	0.508 ± 0.034	0.014 ± 0.095
SuStaIn [313]	0.148 ± 0.032	0.247 ± 0.016	0.000 ± 0.000
Bayesian [153]	0.261 ± 0.349	0.409 ± 0.052	0.303 ± 0.119
PAGA [307]	0.233 ± 0.017	–	–

Figure C-7: Synthetic results over 5 trials. **Top**: Data generating functions for two subtypes (thick lines) and example aligned patients (dots and thin lines). **Bottom**: SubLign learns subtypes and alignment values.

Appendix D

Additional Information for Chapter 6

D.1 Data Privacy and Ethics

The heart failure dataset was collected from a large health system in the United States and shared with us under a research data use agreement. The hospital obtained the relevant consent from individuals to use their data, and the dataset is covered by the hospital's institutional review board (IRB). Since the dataset was de-identified and provided with a data use agreement, our institution's IRB ruled it as exempt. The dataset use agreement was approved by the legal teams from both our academic institution and the hospital.

The Parkinson's Progression Markers Initiative (PPMI) provides open and full access to the study data, which is intended for researchers to study the disease. The PPMI dataset only includes patients who consent to including their data in the study, and the patients are de-identified in the dataset.

For the Parkinson's disease dataset, we searched on a slightly smaller set of hyperparameters for SubLign and found optimal hyperparameters of $\beta = 0.01$, no regularization, 10 latent dimensions, 10 hidden units for the multi-layer perceptron, 200 units for the recurrent neural network, and learning rate of 0.1.

For the heart failure dataset, we searched on a slightly smaller set of hyperparameters for SubLign and found optimal hyperparameters of $\beta = 0.001$, no regularization, 10 latent dimensions, 20 hidden units for the multi-layer perceptron, 50 units for the

recurrent neural network, and learning rate of 0.01.

D.2 Clinical dataset biomarkers and baseline features

For the heart failure dataset, we include the following biomarkers: Aorta - Ascending, Aorta - Valve Level, Aortic Valve - Peak Velocity, Left Atrium - Four Chamber Length, Left Atrium - Long Axis Dimension, Left Ventricle - Diastolic Dimension, Left Ventricle - Ejection Fraction, Left Ventricle - Inferolateral Thickness, Left Ventricle - Septal Wall Thickness, Mitral Valve - E Wave, Mitral Valve - E Wave Deceleration Time, and Right Atrium - Four Chamber Length.

From the Parkinson's Progression Markers Initiative (PPMI) dataset, we include four main biomarkers: 1) MOCA, a cognitive assessment, 2) SCOPA-AUT, an autonomic assessment, 3) NUPDRS1, an assessment of non-motor symptoms, and 4) a maximum taken over NUPDRS3 and NUPDRS2 as an assessment of motor symptoms. We removed patients without extractable biomarker measurements.

The baseline features considered for heart failure are: age, anemia, atherosclerosis, atrial fibrillation, Black, body mass index, chronic kidney disease, diastolic heart failure, esophageal reflux, female, hyperlipidemia, hypertension, hypothyroidism, kidney disease, major depressive disorder, obesity old myocardial infarction, other race, pulmonary heart disease, pneumonia, renal failure, type 2 diabetes, urinary tract infection, and White.

The baseline features considered for Parkinson's disease (PD) are: male, Hispanic/Latino, White, Asian, Black, American Indian, Pacific Islander, not specified race, biological mom with PD, biological dad with PD, full sibling with PD, half sibling with PD, maternal grandparent with PD, paternal grandparent with PD, maternal aunt/uncle with PD, paternal aunt/uncle with PD, kids with PD, years of education, right handed, left handed, University of Pennsylvania Smell Identification Test (UPSIT) part 1, UPSIT part 2, UPSIT part 3, UPSIT part 4, and UPSIT total.

Bibliography

[1] Nhe fact sheet.

[2] Nih offers its first research project grant (r01) on sex and gender.

[3] Proposed standards for race-based and indigenous identity data, July 2020.

[4] Rediet Abebe, Shawndra Hill, Jennifer Wortman Vaughan, Peter M Small, and H Andrew Schwartz. Using search queries to understand health information needs in africa. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 3–14, 2019.

[5] Dzifa Adjaye-Gbewonyo, Robert A Bednarczyk, Robert L Davis, and Saad B Omer. Using the bayesian improved surname geocoding method (bisg) to create a working classification of race and ethnicity in a diverse managed care population: a validation study. *Health services research*, 49(1):268–283, 2014.

[6] Naveed Afzal, Vishnu Priya Mallipeddi, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-Olson. Natural language processing of clinical notes for identification of critical limb ischemia. *International journal of medical informatics*, 111:83–89, 2018.

[7] Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361:k1479, 2018.

[8] Y Collier Ai-ris and Rose L Molina. Maternal mortality in the united states: updates on trends, causes, and solutions. *Neoreviews*, 20(10):e561–e574, 2019.

[9] Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. In *Advances in Neural Information Processing Systems*, pages 11334–11344, 2019.

[10] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019, 2019.

[11] Shun-Ichi Amari. A universal theorem on learning curves. *Neural networks*, 6(2):161–166, 1993.

[12] American Academy of Pediatrics, Committee on Adolescence, American College of Obstetricians and Gynecologists, Committee on Adolescent Health Care. Menstruation in girls and adolescents: using the menstrual cycle as a vital sign. *Pediatrics*, 118(5):2245, 2006.

[13] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

[14] András Antos, Luc Devroye, and Laszlo Gyorfi. Lower bounds for bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):643–645, 1999.

[15] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

[16] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[17] Samantha Artiga, O Pham, K Orgera, and U Ranji. Racial disparities in maternal and infant health: An overview. *Issue Brief. Kaiser Family Foundation. Available online: https://www. kff. org/da8cdf8/j (accessed on 28 December 2020)*, 2020.

[18] Susan Athey. Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 5–6, 2015.

[19] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[20] Ziv Bar-Joseph, Georg Gerber, David K Gifford, Tommi S Jaakkola, and Itamar Simon. A new approach to analyzing gene expression time series data. In *Proceedings of the sixth annual international conference on Computational biology*, pages 39–48, 2002.

[21] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

[22] Victoria L Bartlett, Sanket S Dhruva, Nilay D Shah, Patrick Ryan, and Joseph S Ross. Feasibility of using real-world data to replicate clinical trial evidence. *JAMA network open*, 2(10):e1912869–e1912869, 2019.

[23] Brett K Beaulieu-Jones, William Yuan, Gabriel A Brat, Andrew L Beam, Griffin Weber, Marshall Ruffin, and Isaac S Kohane. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? *NPJ digital medicine*, 4(1):1–6, 2021.

[24] Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. 2017.

[25] David Bellamy, Leo Celi, and Andrew L Beam. Evaluating progress on machine learning for longitudinal electronic healthcare data. *arXiv preprint arXiv:2010.01149*, 2020.

[26] Ayleen Bertini, Rodrigo Salas, Steren Chabert, Luis Sobrevia, and Fabián Pardo. Using machine learning to predict complications in pregnancy: A systematic review. *Frontiers in bioengineering and biotechnology*, 9, 2021.

[27] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

[28] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[29] Michele Black, Kathleen Basile, Matthew Breiding, Sharon Smith, Mikel Walters, Melissa Merrick, Jieru Chen, and Mark Stevens. National intimate partner and sexual violence survey: 2010 summary report. 2011.

[30] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[31] Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26, 2018.

[32] Willie Boag, Harini Suresh, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Racial disparities and mistrust in end-of-life care. *arXiv preprint arXiv:1808.03827*, 2018.

[33] Dwayne T Brandon, Lydia A Isaac, and Thomas A LaVeist. The legacy of tuskegee and trust in medical care: is tuskegee responsible for race differences in mistrust of medical care? *Journal of the National Medical Association*, 97(7):951, 2005.

[34] Paula Braveman. Health disparities and health equity: concepts and measurement. *Annu. Rev. Public Health*, 27:167–194, 2006.

[35] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3, 1950.

[36] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. pages 1–12, 2019.

[37] Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical science*, pages 101–117, 2001.

[38] Raffaele Bugiardini, Beatrice Ricci, Edina Cenko, Zorana Vasiljevic, Sasko Kedev, Goran Davidovic, Marija Zdravkovic, Davor Miličić, Mirza Dilic, Olivia Manfrini, et al. Delayed care and mortality among women and men with myocardial infarction. *Journal of the American Heart Association*, 6(8):e005968, 2017.

[39] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[40] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pages 71–80. IEEE, 2013.

[41] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[42] Edward J Callahan, Shea Hazarian, Mark Yarborough, and John Paul Sánchez. Eliminating lgbtiqq health disparities: the associated roles of electronic health records and institutional culture. *Hastings Center Report*, 44(s4):S48–S52, 2014.

[43] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3995–4004, 2017.

[44] Suzana J Camargo, Andrew W Robertson, Scott J Gaffney, Padhraic Smyth, and Michael Ghil. Cluster analysis of typhoon tracks. part i: General properties. *Journal of Climate*, 20(14):3635–3653, 2007.

[45] Jacquelyn C Campbell. Health consequences of intimate partner violence. *The lancet*, 359(9314):1331–1336, 2002.

[46] John G Canto, Robert J Goldberg, Mary M Hand, Robert O Bonow, George Sopko, Carl J Pepine, and Terry Long. Symptom presentation of women with acute coronary syndromes: myth vs reality. *Archives of internal medicine*, 167(22):2405–2413, 2007.

[47] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[48] Shraddha Chakradhar. Discovery cycle. *Nature Medicine*, 24(8):1082–1086, 2018.

[49] Leighton Chan, L Gary Hart, and David C Goodman. Geographic access to health care for rural medicare beneficiaries. *The Journal of Rural Health*, 22(2):140–146, 2006.

[50] Alice Chen, Emily Oster, and Heidi Williams. Why is infant mortality higher in the united states than in europe? *American Economic Journal: Economic Policy*, 8(2):89–124, 2016.

[51] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 31, 2018.

[52] Irene Y Chen, Monica Agrawal, Steven Horng, and David Sontag. Robustly extracting medical knowledge from ehrs: A case study of learning a health knowledge graph. In *Pac Symp Biocomput*, pages 19–30. World Scientific, 2020.

[53] Irene Y Chen, Emily Alsentzer, Hyesun Park, Richard Thomas, Babina Gosangi, Rahul Gujrathi, and Bharti Khurana. Intimate partner violence and injury prediction from radiology reports. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 55–66. World Scientific, 2020.

[54] Irene Y Chen, Shalmali Joshi, and Marzyeh Ghassemi. Treating health disparities with artificial intelligence. *Nature Medicine*, 26(1):16–17, 2020.

[55] Irene Y Chen, Rahul G Krishnan, and David Sontag. Clustering interval-censored time-series for disease phenotyping. 2022.

[56] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4, 2020.

[57] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4:123–144, 2021.

[58] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

[59] Long Chen, Yu Gu, Xin Ji, Chao Lou, Zhiyong Sun, Haodan Li, Yuan Gao, and Yang Huang. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *Journal of the American Medical Informatics Association*, 26(11):1218–1226, 2019.

[60] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[61] Jaeun Choi and A James O'Malley. Estimating the causal effect of treatment in observational studies with survival time endpoints and unmeasured confounding. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 66(1):159, 2017.

[62] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056*, 2017.

[63] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. pages 134–148, 2018.

[64] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[65] Natalie Coburn, John Fulton, Deborah N Pearlman, Calvin Law, Brenda DiPaolo, and Blake Cady. Treatment variation by insurance status for breast cancer patients. *The breast journal*, 14(2):128–134, 2008.

[66] Benjamin Lê Cook, Thomas G McGuire, and Alan M Zaslavsky. Measuring racial/ethnic disparities in health care: methods and practical issues. *Health services research*, 47(3pt2):1232–1254, 2012.

[67] Benjamin Lê Cook, Samuel H Zuvekas, Nicholas Carson, Geoffrey Ferris Wayne, Andrew Vesper, and Thomas G McGuire. Assessing racial/ethnic disparities in treatment across episodes of mental health care. *Health services research*, 49(1):206–229, 2014.

[68] Andy Coravos, Irene Chen, Ankit Gordhandas, and Ariel Dora Stern. We should treat algorithms like prescription drugs. *Quartz*, 2019.

[69] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[70] Katherine Courtright. Point: Do randomized controlled trials ignore needed patient populations? yes. *Chest*, 149(5):1128–1130, 2016.

[71] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[72] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal modeling for fairness in dynamical systems. *arXiv preprint arXiv:1909.09141*, 2019.

[73] Andreea A Creanga, Cynthia J Berg, Jean Y Ko, Sherry L Farr, Van T Tong, F Carol Bruce, and William M Callaghan. Maternal mortality and morbidity in the united states: where are we now? *Journal of women's health*, 23(1):3–9, 2014.

[74] Marco Cuturi. Fast global alignment kernels. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 929–936, 2011.

[75] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. *arXiv preprint arXiv:1703.01541*, 2017.

[76] Guy David, Aaron Smith-McLallen, and Benjamin Ukert. The effect of predictive analytics-driven interventions on healthcare utilization. *Journal of health economics*, 64:68–79, 2019.

[77] Sharon E Davis, Thomas A Lasko, Guanhua Chen, Edward D Siew, and Michael E Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017.

[78] Stephen R Dearwater, Jeffrey H Coben, Jacquelyn C Campbell, Gregory Nah, Nancy Glass, Elizabeth McLoughlin, and Betty Bekemeier. Prevalence of intimate partner abuse in women treated at community hospital emergency departments. *Jama*, 280(5):433–438, 1998.

[79] Pierre A. Devijver and Josef Kittler. *Pattern recognition: a statistical approach*. Sung Kang, 1982.

[80] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[81] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, 2004.

[82] Tobias Domhan, Jost Tobias Springenberg, and Frank Hutter. Speeding up automatic hyperparameter optimization of deep neural networks by extrapolation of learning curves. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[83] Pedro Domingos. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, pages 231–238. Morgan Kaufmann Stanford, 2000.

[84] Finale Doshi-Velez, Yaorong Ge, and Isaac Kohane. Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*, 133(1), 2014.

[85] Kathleen Dracup, Debra K Moser, Mickey Eisenberg, Hendrika Meischke, Angelo A Alonzo, and Allan Braslow. Causes of delay in seeking treatment for heart attack symptoms. *Social science & medicine*, 40(3):379–392, 1995.

[86] Mark Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84, 2012.

[87] Franz Duca, Caroline Zotter-Tufaro, Andreas A Kammerlander, Stefan Aschauer, Christina Binder, Julia Mascherbauer, and Diana Bonderman. Gender-related differences in heart failure with preserved ejection fraction. *Scientific reports*, 8(1):1–9, 2018.

[88] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[89] Cynthia Dwork and Christina Ilvento. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*, 2018.

[90] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[91] VH Eisenberg, C Weil, G Chodick, and V Shalev. Epidemiology of endometriosis: a large population-based database study from a healthcare provider with 2 million members. *BJOG: An International Journal of Obstetrics & Gynaecology*, 125(1):55–62, 2018.

[92] Mary Ellsberg, Henrica AFM Jansen, Lori Heise, Charlotte H Watts, Claudia Garcia-Moreno, et al. Intimate partner violence and women's physical and mental health in the who multi-country study on women's health and domestic violence: an observational study. *The lancet*, 371(9619):1165–1172, 2008.

[93] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. *CoRR*, abs/1706.09847, 2017.

[94] Faheem Farooq and John J Strouse. Disparities in foundation and federal support and development of new therapeutics for sickle cell disease and cystic fibrosis. *Blood*, 132:4687–4687, 2018.

[95] Lilith Faucheux, Matthieu Resche-Rigon, Emmanuel Curis, Vassili Soumelis, and Sylvie Chevret. Clustering with missing and left-censored data: A simulation study comparing multiple-imputation-based procedures. *Biometrical Journal*, 63(2):372–393, 2021.

[96] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In

Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259–268. ACM, 2015.

[97] Seyed-Mohammad Fereshtehnejad, Yashar Zeighami, Alain Dagher, and Ronald B Postuma. Clinical criteria for subtyping parkinson's disease: biomarkers and longitudinal progression. *Brain*, 140(7):1959–1976, 2017.

[98] Kadija Ferryman. Addressing health disparities in the fda's ai and machine learning regulatory framework. *Journal of the American Medical Informatics Association*, To Appear.

[99] Kadija Ferryman and Mikaela Pitcan. Fairness in precision medicine. *Data & Society*, 2018.

[100] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

[101] Charles K Fisher, Aaron M Smith, and Jonathan R Walsh. Machine learning for comprehensive forecasting of alzheimer's disease progression. *Scientific reports*, 9(1):1–14, 2019.

[102] R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

[103] Peter A Flach. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 194–201, 2003.

[104] Laura E Flink, Robert R Sciacca, Michael L Bier, Juviza Rodriguez, and Elsa-Grace V Giardina. Women at risk for cardiovascular disease lack knowledge of heart attack symptoms. *Clinical cardiology*, 36(3):133–138, 2013.

[105] Center for Devices and Radiological Health. Artificial intelligence and machine learning (ai/ml) medical devices.

[106] Center for Devices and Radiological Health. Artificial intelligence and machine learning in software, Jan 2020.

[107] Christine Fountain and Peter Bearman. Risk as social context: immigration policy and autism in california. In *Sociological Forum*, volume 26, pages 215–240. Wiley Online Library, 2011.

[108] AlexanderM Franks, Alexander D'Amour, and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, pages 1–33, 2019.

[109] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. *arXiv preprint arXiv:1802.04422*, 2018.

[110] Keinosuke Fukunaga and Donald M Hummels. Bayes error estimation using parzen and k-nn procedures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):634–643, 1987.

[111] Emma Fulu, Rachel Jewkes, Tim Roselli, Claudia Garcia-Moreno, et al. Prevalence of and factors associated with male perpetration of intimate partner violence: findings from the un multi-country cross-sectional study on men and violence in asia and the pacific. *The lancet global health*, 1(4):e187–e207, 2013.

[112] Scott J Gaffney and Padhraic Smyth. Joint probabilistic curve clustering and alignment. In *Advances in neural information processing systems*, pages 473–480, 2005.

[113] Shang Gao, Michael T Young, John X Qiu, Hong-Jun Yoon, James B Christian, Paul A Fearn, Georgia D Tourassi, and Arvind Ramanthan. Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association*, 25(3):321–330, 2018.

[114] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. March 2017.

[115] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[116] Elizabeth George, Catherine H Phillips, Nandish Shah, Annie Lewis-O'Connor, Bernard Rosner, Hanni M Stoklosa, and Bharti Khurana. Radiologic findings in intimate partner violence. *Radiology*, 291(1):62–69, 2019.

[117] Michael Geruso and Timothy Layton. Upcoding: Evidence from medicare on squishy risk adjustment. Technical report, National Bureau of Economic Research, 2015.

[118] Stephanie S Gervasi, Irene Y Chen, Aaron Smith-McLallen, David Sontag, Ziad Obermeyer, Michael Vennera, and Ravi Chawla. The potential for bias in machine learning and opportunities for health insurers to address it: Article examines the potential for bias in machine learning and opportunities for health insurers to address it. *Health Affairs*, 41(2):212–218, 2022.

[119] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM*

SIGKDD international conference on Knowledge discovery and data mining, pages 75–84. ACM, 2014.

[120] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health*, 1(4):e157–e159, 2019.

[121] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

[122] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, 2018.

[123] Donna K Ginther, Walter T Schaffer, Joshua Schnell, Beth Masimore, Faye Liu, Laurel L Haak, and Raynard Kington. Race, ethnicity, and nih research awards. *Science*, 333(6045):1015–1019, 2011.

[124] Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H Andrew Schwartz. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1172, 2018.

[125] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[126] Gnanesh. Goodreads book reviews, 2017.

[127] Jen J Gong, Tristan Naumann, Peter Szolovits, and John V Guttag. Predicting clinical outcomes across changing electronic health record systems. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1497–1505, 2017.

[128] Babina Gosangi, Hyesun Park, Richard Thomas, Rahul Gujrathi, Camden P Bay, Ali S Raja, Steven E Seltzer, Marta Chadwick Balcom, Meghan L McDonald, Dennis P Orgill, et al. Exacerbation of physical intimate partner violence during covid-19 lockdown. *Radiology*, page 202866, 2020.

[129] William A Grobman, Yinglei Lai, Mark B Landon, Catherine Y Spong, Kenneth J Leveno, Dwight J Rouse, Michael W Varner, Atef H Moawad, Steve N Caritis, Margaret Harper, et al. Development of a nomogram for prediction of vaginal birth after cesarean delivery. *Obstetrics & Gynecology*, 109(4):806–812, 2007.

[130] Devin Guillory. Combating anti-blackness in the ai community. *arXiv preprint arXiv:2006.16879*, 2020.

[131] A Haehner, S Boesveldt, HW Berendse, A Mackay-Sim, J Fleischmann, PA Silburn, AN Johnston, GD Mellick, B Herting, H Reichmann, et al. Prevalence of smell loss in parkinson's disease–a multicenter study. *Parkinsonism & related disorders*, 15(7):490–494, 2009.

[132] Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459, 2013.

[133] Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731–740, 2016.

[134] Sebastien JP A Haneuse and Susan M Shortreed. *On the use of electronic health records*. Chapman and Hall/CRC New York, NY, 2017.

[135] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. pages 3323–3331, June 2016. Barcelona, Spain.

[136] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[137] Frank E Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[138] Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

[139] Achim Hekler, Jochen S Utikal, Alexander H Enk, Wiebke Solass, Max Schmitt, Joachim Klode, Dirk Schadendorf, Wiebke Sondermann, Cindy Franklin, Felix Bestvater, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, 118:91–96, 2019.

[140] Robert L Helmreich. On error management: lessons from aviation. *Bmj*, 320(7237):781–785, 2000.

[141] Miguel A Hernán and James M Robins. Causal inference, 2010.

[142] Diana Herrera-Perez, Alyson Haslam, Tyler Crain, Jennifer Gill, Catherine Livingston, Victoria Kaestner, Michael Hayes, Dan Morgan, Adam S Cifu, and Vinay Prasad. Meta-research: A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. *Elife*, 8:e45183, 2019.

[143] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[144] Denise Hien and Lesia Ruglass. Interpersonal partner violence and women in the united states: An overview of prevalence rates, psychiatric correlates and consequences and barriers to help seeking. *International journal of law and psychiatry*, 32(1):48, 2009.

[145] Paula J Adams Hillard. Menstruation in Adolescents: What Do We Know? and What Do We Do with the Information? *Journal of Pediatric and Adolescent Gynecology*, 27(6):309–319, 2014.

[146] Esther Hing and Catharine W Burt. Are there patient disparities when electronic health records are adopted? *Journal of health care for the poor and underserved*, 20(2):473–488, 2009.

[147] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.

[148] Jack Holloway, Chris Neely, Xiaojing Yuan, Yuan Zhang, Jason Ouyang, Dawn Cantrell, Janet Chaisson, Tasha Bergeron, Vindell Washington, and Somesh Nigam. Evaluating the performance of a predictive modeling approach to identifying members at high-risk of hospitalization. *Journal of Medical Economics*, 23(3):228–234, 2020.

[149] Woo Suk Hong, Adrian Daniel Haimovich, and R Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7):e0201016, 2018.

[150] Travis A Hoppe, Aviva Litovitz, Kristine A Willis, Rebecca A Meseroll, Matthew J Perkins, B Ian Hutchins, Alison F Davis, Michael S Lauer, Hannah A Valantine, James M Anderson, et al. Topic choice contributes to the lower rate of nih awards to african-american/black scientists. *Science advances*, 5(10):eaaw7238, 2019.

[151] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[152] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[153] Ilkka Huopaniemi, Girish Nadkarni, Rajiv Nadukuru, Vaneet Lotay, Steve Ellis, Omri Gottesman, and Erwin P Bottinger. Disease progression subtype discovery from longitudinal emr data with a majority of missing values and unknown

initial time points. In *AMIA Annual Symposium Proceedings*, volume 2014, page 709. American Medical Informatics Association, 2014.

[154] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):1–9, 2021.

[155] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[156] Sandy James, Jody Herman, Susan Rankin, Mara Keisling, Lisa Mottet, and Ma'ayan Anafi. The report of the 2015 us transgender survey. 2016.

[157] Dean T Jamison et al. *Disease and mortality in sub-Saharan Africa*. World Bank Publications, 2006.

[158] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

[159] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[160] Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xgems: Generating examplars to explain black-box models. *arXiv preprint arXiv:1806.08867*, 2018.

[161] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

[162] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

[163] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874. IEEE, 2010.

[164] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 643–650. IEEE, 2011.

[165] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698, 2011.

[166] Mudit Kapoor, Deepak Agrawal, Shamika Ravi, Ambuj Roy, SV Subramanian, and Randeep Guleria. Missing female patients: an observational analysis of sex ratio among outpatients in a referral tertiary care public hospital in india. *BMJ open*, 9(8):e026850, 2019.

[167] CC Kariyawasan, DA Hughes, MM Jayatillake, and AB Mehta. Multiple myeloma: causes and consequences of delay in diagnosis. *QJM: An International Journal of Medicine*, 100(10):635–640, 2007.

[168] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision making. Technical report, Working paper, 2020.

[169] Kenneth L Kehl, Haitham Elmarakeby, Mizuki Nishino, Eliezer M Van Allen, Eva M Lepisto, Michael J Hassett, Bruce E Johnson, and Deborah Schrag. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA oncology*, 5(10):1421–1429, 2019.

[170] Aaron S Kesselheim and Troyen A Brennan. Overbilling vs. downcoding—the battle between physicians and insurers. *New England Journal of Medicine*, 352(9):855–857, 2005.

[171] Bharti Khurana, Steven E Seltzer, Isaac S Kohane, and Giles W Boland. Making the 'invisible'visible: transforming the detection of intimate partner violence. *BMJ quality & safety*, 29(3):241–244, 2020.

[172] Jinseok Kim and Karen A Gray. Leave or stay? battered women's decision after intimate partner violence. *Journal of Interpersonal Violence*, 23(10):1465–1482, 2008.

[173] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[174] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *The 2nd International Conference on Learning Representations (ICLR)*, 2013.

[175] Christine Klein and Ana Westenberger. Genetics of parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(1):a008888, 2012.

[176] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[177] Elissa V Klinger, Sara V Carlini, Irina Gonzalez, Stella St Hubert, Jeffrey A Linder, Nancy A Rigotti, Emily Z Kontos, Elyse R Park, Lucas X Marinacci, and Jennifer S Haas. Accuracy of race, ethnicity, and language preference in an

electronic health record. *Journal of general internal medicine*, 30(6):719–723, 2015.

[178] Hoyt Koepke and Mikhail Bilenko. Fast prediction of new feature utility. *arXiv preprint arXiv:1206.4680*, 2012.

[179] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *ICML*, 2020.

[180] Fleur L Kraanen, Ellen Vedel, Agnes Scholing, and Paul MG Emmelkamp. Prediction of intimate partner violence by type of substance use disorder. *Journal of substance abuse treatment*, 46(4):532–539, 2014.

[181] Roopal V Kundu and Stavonnie Patterson. Dermatologic conditions in skin of color: part i. special considerations for common skin disorders. *American family physician*, 87(12):850–856, 2013.

[182] M Kunneman, AH Pieterse, AM Stiggelbout, RA Nout, M Kamps, LCHW Lutgens, J Paulissen, OJA Mattheussens, RFPM Kruitwagen, and CL Creutzberg. Treatment preferences and involvement in treatment decision making of patients with endometrial cancer and clinicians. *British journal of cancer*, 111(4):674–679, 2014.

[183] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[184] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[185] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[186] Karen E Lasser, David U Himmelstein, and Steffie Woolhandler. Access to care, health status, and health disparities in the united states and canada: results of a cross-national population-based survey. *American journal of public health*, 96(7):1300–1307, 2006.

[187] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[188] Marie des Neiges Léonard. Census and racial categorization in france: Invisible categories and color-blind politics. *Humanity & society*, 38(1):67–88, 2014.

[189] Rachel Levy-Shiff, Maya Lerman, Dov Har-Even, and Moshe Hod. Maternal adjustment and infant outcome in medically defined high-risk pregnancy. *Developmental psychology*, 38(1):93, 2002.

[190] Gang Li, Baosheng Liu, S Joe Qin, and Donghua Zhou. Dynamic latent variable modeling for statistical process monitoring. In *Proc. IFAC World Congress, Milan, Italy*, pages 12886–12891, 2011.

[191] M. Lichman. UCI machine learning repository, 2013.

[192] Sherry Lipsky, Raul Caetano, Craig A Field, and Gregory L Larkin. The role of intimate partner violence, race, and ethnicity in help-seeking behaviors. *Ethnicity and Health*, 11(1):81–100, 2006.

[193] Jennifer Listgarten, Radford M Neal, Sam T Roweis, Rachel Puckrin, and Sean Cutler. Bayesian detection of infrequent differences in sets of time series with shared structure. In *Advances in neural information processing systems*, pages 905–912, 2007.

[194] Max A Little and Reham Badawy. Causal bootstrapping. *arXiv preprint arXiv:1910.09648*, 2019.

[195] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[196] Siqi Liu, Sidong Liu, Weidong Cai, Sonia Pujol, Ron Kikinis, and Dagan Feng. Early diagnosis of alzheimer's disease with deep learning. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, pages 1015–1018. IEEE, 2014.

[197] Xiaoming Liu, Yan Tong, and Frederick W Wheeler. Simultaneous alignment and clustering for an image ensemble. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1327–1334. IEEE, 2009.

[198] Xiaoxuan Liu, Samantha Cruz Rivera, Livia Faes, Lavinia Ferrante Di Ruffano, Christopher Yau, Pearse A Keane10, Hutan Ashrafian11, Ara Darzi11, Sebastian J Vollmer, and Jonathan Deeks. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat. Med*, 25:1467–1468, 2019.

[199] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[200] Michael Lohaus, Michael Perrot, and Ulrike von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning*, 2020.

[201] M Magaña López, M Bevans, L Wehrlen, L Yang, and GR Wallen. Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities. *Journal of racial and ethnic health disparities*, 4(5):812–818, 2017.

[202] Yuan Luo, Alal Eran, Nathan Palmer, Paul Avillach, Ami Levy-Moonshine, Peter Szolovits, and Isaac S Kohane. A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nature Medicine*, 26(9):1375–1379, 2020.

[203] Marian F MacDorman, Eugene Declercq, Howard Cabral, and Christine Morton. Is the united states maternal mortality rate increasing? disentangling trends from measurement issues short title: Us maternal mortality trends. *Obstetrics and gynecology*, 128(3):447, 2016.

[204] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[205] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[206] Benjamin M Marlin, David C Kale, Robinder G Khemani, and Randall C Wetzel. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 389–398. ACM, 2012.

[207] Connie Marras and Anthony Lang. Parkinson's disease subtypes: lost in translation? *Journal of Neurology, Neurosurgery & Psychiatry*, 84(4):409–415, 2013.

[208] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591, 2019.

[209] Amy Matcho, Patrick Ryan, Daniel Fife, Dina Gifkins, Chris Knoll, and Andrew Friedman. Inferring pregnancy episodes and outcomes within a network of observational databases. *PloS one*, 13(2):e0192033, 2018.

[210] Marwan A Mattar, Allen R Hanson, and Erik G Learned-Miller. Unsupervised joint alignment and clustering using bayesian nonparametrics. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 584–593, 2012.

[211] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. *http://mallet. cs. umass. edu*, 2002.

[212] Matthew B. A. McDermott, Bret Nestor, Evan Kim, Wancong Zhang, Anna Goldenberg, Peter Szolovits, and Marzyeh Ghassemi. A comprehensive evaluation of multi-task learning and multi-task pre-training on ehr time-series data, 2020.

[213] Thomas G McGuire, Anna L Zink, and Sherri Rose. Simplifying and improving the performance of risk adjustment systems. Technical report, National Bureau of Economic Research, 2020.

[214] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[215] Sarah Miller and Laura R Wherry. Health and access to care during the first 2 years of the aca medicaid expansions. *New England Journal of Medicine*, 376(10):947–956, 2017.

[216] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[217] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[218] Ellen Montz, Tim Layton, Alisa B Busch, Randall P Ellis, Sherri Rose, and Thomas G McGuire. Risk-adjustment simulation: plans may have incentives to distort mental health and substance use coverage. *Health Affairs*, 35(6):1022–1028, 2016.

[219] Sayan Mukherjee, Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R Golub, and Jill P Mesirov. Estimating dataset size requirements for classifying dna microarray data. *Journal of computational biology*, 10(2):119–142, 2003.

[220] Kenric M Murayama, Anna M Derossis, Debra A DaRosa, Heather B Sherman, and Jonathan P Fryer. A critical evaluation of the morbidity and mortality conference. *The American journal of surgery*, 183(3):246–250, 2002.

[221] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[222] Anna C Need and David B Goldstein. Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, 25(11):489–494, 2009.

[223] Bret Nestor, Matthew McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature

robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. *Machine Learning for Healthcare*, 2019.

[224] Peter A Noseworthy, Zachi I Attia, LaPrincess C Brewer, Sharonne N Hayes, Xiaoxi Yao, Suraj Kapa, Paul A Friedman, and Francisco Lopez-Jimenez. Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ecg analysis. *Circulation: Arrhythmia and Electrophysiology*, 13(3):e007988, 2020.

[225] Luke Oakden-Rayner. Exploring large-scale public medical image datasets. *Academic Radiology*, 27(1):106–112, 2020.

[226] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

[227] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[228] Lorna O'Doherty, Kelsey Hegarty, Jean Ramsay, Leslie L Davidson, Gene Feder, and Angela Taft. Screening women for intimate partner violence in healthcare settings. *Cochrane database of systematic reviews*, (7), 2015.

[229] All of Us Research Program Investigators. The "all of us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019.

[230] United Nations Office on Drugs and Crime. *Global Study on Homicide: Gender-related Killing of Women and Girls*. UNODC, United Nations Office on Drugs and Crime, 2018.

[231] Theophilus E Owan, David O Hodge, Regina M Herges, Steven J Jacobsen, Veronique L Roger, and Margaret M Redfield. Trends in prevalence and outcome of heart failure with preserved ejection fraction. *New England Journal of Medicine*, 355(3):251–259, 2006.

[232] Ravi B Parikh, Ziad Obermeyer, and Amol S Navathe. Regulation of predictive analytics in medicine. *Science*, 363(6429):810–812, 2019.

[233] Madison Park. Ncaa genetic screening rule sparks discrimination concerns, Aug 2010.

[234] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[235] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[236] Caroline Criado Perez. *Invisible women: Exposing data bias in a world designed for men*. Random House, 2019.

[237] Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. Spartan: Scalable parafac2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 375–384, 2017.

[238] Emily E Petersen, Nicole L Davis, David Goodman, Shanna Cox, Carla Syverson, Kristi Seed, Carrie Shapiro-Mendoza, William M Callaghan, and Wanda Barfield. Racial/ethnic disparities in pregnancy-related deaths—united states, 2007–2016. *Morbidity and Mortality Weekly Report*, 68(35):762, 2019.

[239] Anne-Dominique Pham, Aurélie Névéol, Thomas Lavergne, Daisuke Yasunaga, Olivier Clément, Guy Meyer, Rémy Morello, and Anita Burgun. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. *BMC bioinformatics*, 15(1):1–10, 2014.

[240] Elizabeth G Phimister. Medicine and the racial divide. *New England Journal of Medicine*, 348(12):1081–1082, 2003.

[241] E Pierson, PW Koh, T Hashimoto, D Koller, J Leskovec, N Eriksson, and P Liang. Inferring multidimensional rates of aging from cross-sectional data. *Proceedings of machine learning research*, 89:97–107, 2019.

[242] Emma Pierson. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124*, 2017.

[243] Emma Pierson, Tim Althoff, Daniel Thomas, Paula Hillard, and Jure Leskovec. The menstrual cycle is a primary contributor to cyclic variation in women's mood, behavior, and vital signs. *Working Paper*, 2019.

[244] Leah Pierson and Joseph Millum. Grant reviews and health research priority setting: Do research funders uphold widely endorsed ethical principles? *Working Paper*, 2020.

[245] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5684–5693, 2017.

[246] Tom J Pollard, Irene Chen, Jenna Wiens, Steven Horng, Danny Wong, Marzyeh Ghassemi, Heather Mattie, Emily Lindemer, and Trishan Panch. Turning the crank for machine learning: ease, at what expense? *The Lancet Digital Health*, 1(5):e198–e199, 2019.

[247] Chris Poulin, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley Watts, Laura Flashman, and Thomas McAllister. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733, 2014.

[248] Vinay Prasad, Adam Cifu, and John PA Ioannidis. Reversals of established medical practices: evidence to abandon ship. *Jama*, 307(1):37–38, 2012.

[249] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[250] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[251] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4):277–287, 2015.

[252] Umaa Rebbapragada, Pavlos Protopapas, Carla E Brodley, and Charles Alcock. Finding anomalous periodic time series. *Machine learning*, 74(3):281–313, 2009.

[253] Callie Marie Rennison. *Intimate partner violence and age of victim, 1993-99*. US Department of Justice, Office of Justice Programs, Bureau of Justice ..., 2001.

[254] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, pages II–1278, 2014.

[255] Sherri Rose. A machine learning framework for plan payment risk adjustment. *Health services research*, 51(6):2358–2374, 2016.

[256] Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

[257] Carolyn Rouse. *Uncertain suffering: racial health care disparities and sickle cell disease*. Univ of California Press, 2009.

[258] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):9, 2010.

[259] Anna Rumshisky, Marzyeh Ghassemi, Tristan Naumann, Peter Szolovits, VM Castro, TH McCoy, and RH Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry*, 6(10):e921–e921, 2016.

[260] Anna Russo, Alfonso Reginelli, Maria Pignatiello, Fabrizio Cioce, Giovanni Mazzei, Olimpia Fabozzi, Vincenzo Parlato, Salvatore Cappabianca, and Sabrina Giovine. Imaging of violence against the elderly and the women. 40(1):18–24, 2019.

[261] Eduardo Sabaté and Eduardo Sabaté. *Adherence to long-term therapies: evidence for action*. World Health Organization, 2003.

[262] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ICLR 2020*, 2019.

[263] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. *ICML*, 2020.

[264] Shems Saleh, William Boag, Lauren Erdman, and Tristan Naumann. Clinical collabsheets: 53 questions to guide a clinical collaboration.

[265] Amy Salomon, Shari S Bassuk, and Nicholas Huntington. The relationship between intimate partner violence and the use of addictive substances in poor and homeless single mothers. *Violence Against Women*, 8(7):785–815, 2002.

[266] Suchi Saria, Anand K Rajani, Jeffrey Gould, Daphne Koller, and Anna A Penn. Integration of early physiological responses predicts later illness severity in preterm infants. *Science translational medicine*, 2(48):48ra65–48ra65, 2010.

[267] Shaun Seaman, Oliver Dukes, Ruth Keogh, and Stijn Vansteelandt. Adjusting for time-varying confounders in survival analysis using structural nested cumulative survival time models. *Biometrics*, 76(2):472–483, 2020.

[268] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *arXiv preprint arXiv:2003.00827*, 2020.

[269] Sanjiv J Shah, Daniel H Katz, Senthil Selvaraj, Michael A Burke, Clyde W Yancy, Mihai Gheorghiade, Robert O Bonow, Chiang-Ching Huang, and Rahul C Deo. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation*, 131(3):269–279, 2015.

[270] Scott M Sporer, James N Weinstein, and Kenneth J Koval. The geographic incidence and treatment variation of common fractures of elderly patients. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, 14(4):246–255, 2006.

[271] Heidi Stöckl, Karen Devries, Alexandra Rotstein, Naeemah Abrahams, Jacquelyn Campbell, Charlotte Watts, and Claudia Garcia Moreno. The global prevalence of intimate partner homicide: a systematic review. *The Lancet*, 382(9895):859–865, 2013.

[272] Elizabeth A Stuart, Catherine P Bradshaw, and Philip J Leaf. Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16(3):475–485, 2015.

[273] A Subbaswamy and S Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics (Oxford, England)*, 21(2):345, 2020.

[274] Hannah R Sullivan and Scott J Schweikart. Are current tort liability doctrines adequate for addressing injury caused by ai? *AMA journal of ethics*, 21(2):160–166, 2019.

[275] Marijana Tadic and Cesare Cuspidi. Obesity and heart failure with preserved ejection fraction: a paradox or something else? *Heart Failure Reviews*, 24(3):379–385, 2019.

[276] Suzanne Tamang, Arnold Milstein, Henrik Toft Sørensen, Lars Pedersen, Lester Mackey, Jean-Raymond Betterton, Lucas Janson, and Nigam Shah. Predicting patient 'cost blooms' in denmark: a longitudinal population-based study. *BMJ open*, 7(1):e011580, 2017.

[277] Kate Sophia Mary Taylor, Jonathan Alistair Cook, and Carl Edward Counsell. Heterogeneity in male to female risk for parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(8):905–906, 2007.

[278] Bruce Thompson. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial, 1995.

[279] Chetan Tiwari, Kirsten Beyer, and Gerard Rushton. The impact of data suppression on local mortality rates: the case of cdc wonder. *American journal of public health*, 104(8):1386–1388, 2014.

[280] Patricia Tjaden and Nancy Thoennes. Prevalence and consequences of male-to-female and female-to-male intimate partner violence as measured by the national violence against women survey. *Violence against women*, 6(2):142–161, 2000.

[281] Kristine Tollestrup, David Sklar, Floyd J Frost, Lenora Olson, JoElla Weybright, Joan Sandvig, and Muree Larson. Health indicators and intimate partner violence among women who are members of a managed care organization. *Preventive medicine*, 29(5):431–440, 1999.

[282] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[283] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[284] Justin Travers, Suzanne Marsh, Mathew Williams, Mark Weatherall, Brent Caldwell, Philippa Shirtcliffe, Sarah Aldington, and Richard Beasley. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? *Thorax*, 62(3):219–223, 2007.

[285] Unequal Treatment. Confronting racial and ethnic disparities in health care. *Washington, DC, Institute of Medicine*, 2002.

[286] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, pages 1–6, 2020.

[287] Krystal S Tsosie, Joseph M Yracheta, and Donna Dickenson. Overvaluing individual consent ignores risks to tribal participants. *Nature Reviews Genetics*, 20(9):497–498, 2019.

[288] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[289] Kagan Tumer and Joydeep Ghosh. Estimating the bayes error rate through classifier combining. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, volume 2, pages 695–699. IEEE, 1996.

[290] Miriam S Udler, Jaegil Kim, Marcin von Grotthuss, Sílvia Bonás-Guarch, Joanne B Cole, Joshua Chiou, Christopher D. Anderson on behalf of METASTROKE, the ISGC, Michael Boehnke, Markku Laakso, Gil Atzmon, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS medicine*, 15(9):e1002654, 2018.

[291] Benjamin Ukert, Guy David, Aaron Smith-McLallen, and Ravi Chawla. Do payor-based outreach programs reduce medical cost and utilization? *Health Economics*, 29(6):671–682, 2020.

[292] 23andMe under. 23andme's call for collaborations to study underrepresented populations, Mar 2019.

[293] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382, 2019.

[294] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

[295] N Van Gelder, A Peterman, A Potts, M O'Donnell, K Thompson, N Shah, and S Oertelt-Prigione. Covid-19: Reducing the risk of infection might increase the risk of intimate partner violence. *EClinicalMedicine*, 21, 2020.

[296] D Vidyasagar. Global notes: the 10/90 gap disparities in global health research. *Journal of Perinatology*, 26(1):55–56, 2006.

[297] Kailas Vodrahalli, Roxana Daneshjou, Roberto A Novoa, Albert Chiou, Justin M Ko, and James Zou. Trueimage: a machine learning algorithm to improve the quality of telehealth photos. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 220–231. World Scientific, 2020.

[298] Peter Von Philipsborn, Fridolin Steinbeis, Max E Bender, Sadie Regmi, and Peter Tinnemann. Poverty-related and neglected diseases–an economic and epidemiological analysis of poverty relatedness and neglect in research and development. *Global Health Action*, 8(1):25818, 2015.

[299] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in plain sight — reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine*, 0(0):null, 0.

[300] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.

[301] E Hope Weissler, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F Freitag, James Benoit, Michael C Hughes, Faisal Khan, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1):1–15, 2021.

[302] Brian J Wells, Kevin M Chagin, Amy S Nowacki, and Michael W Kattan. Strategies for handling missing data in electronic health record derived data. *Egems*, 1(3), 2013.

[303] Jevin D West, Jennifer Jacquet, Molly M King, Shelley J Correll, and Carl T Bergstrom. The role of gender in scholarly authorship. *PloS one*, 8(7):e66212, 2013.

[304] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[305] Coady Wing, Kosali Simon, and Ricardo A Bello-Gomez. Designing difference in difference studies: best practices for public health policy research. *Annual review of public health*, 39, 2018.

[306] C Wisner, T Gilmer, L Saltzman, and T Zink. Intimate partner violence against women. *Journal of family practice*, 48:439–443, 1999.

[307] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1):59, 2019.

[308] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *Conference On Learning Theory*, 2017.

[309] Hao Wu, Lin Gao, and Nikola Kasabov. Inference of cancer progression from somatic mutation data. *IFAC-PapersOnLine*, 48(28):234–238, 2015.

[310] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470, 2020.

[311] Gui-Shuang Ying, Maureen G Maguire, Robert J Glynn, and Bernard Rosner. Calculating sensitivity, specificity, and predictive values for correlated eye data. *Investigative ophthalmology & visual science*, 61(11):29–29, 2020.

[312] Jinsung Yoon, William R Zame, and Mihaela van der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5):1477–1490, 2018.

[313] Alexandra L Young, Razvan V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C Firth, David M Cash, David L Thomas, Katrina M Dick, Jorge Cardoso, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with subtype and stage inference. *Nature communications*, 9(1):4273, 2018.

[314] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2017.

[315] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[316] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[317] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[318] Xi Zhang, Jingyuan Chou, Jian Liang, Cao Xiao, Yize Zhao, Harini Sarva, Claire Henchcliffe, and Fei Wang. Data-driven subtyping of parkinson's disease using longitudinal clinical records: a cohort study. *Scientific reports*, 9(1):1–12, 2019.

[319] Anna Zink and Sherri Rose. Fair regression for health care spending. *Biometrics*, 76:973–982, 2020.