Predictive and Prescriptive Analytics for Optimizing Concussion Management Decisions

by

Gian-Gabriel Garcia

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2020

Doctoral Committee:

Associate Professor Mariel S. Lavieri, Chair
Professor Steven P. Broglio
Professor Brian T. Denton
Assistant Professor Ruiwei Jiang

Gian-Gabriel Garcia
garciagg@umich.edu
ORCID iD: 0000-0001-9315-0195

Dedication

To my parents, Francis and Nida Garcia, who instilled in me the value of education and always pushed me to do the best that I could.

Acknowledgments

It gives me great joy to acknowledge the people who have enriched my PhD experience and made my time at the University of Michigan so enjoyable.

First and foremost, I express my deepest gratitude to my advisor, Mariel Lavieri. Your guidance and mentorship have helped me grow tremendously as a researcher, academic, and person. I have learned so much from you about what it means to support your students' success. I hope that my future students gain as much from me as I have gained from you. The achievements that I have had throughout my PhD are a reflection of your dedication as an advisor. I hope to carry your lessons with me through the remainder of my career.

Secondly, I want to thank Steve Broglio. This dissertation would not have been possible without your collaboration and support. Thank you for welcoming me into the world of concussion research and continually providing opportunities for me to make a positive impact on the field. I look forward to continuing research on concussion throughout the rest of my career.

I want to acknowledge Ruiwei Jiang and Brian Denton, who gracefully agreed to serve on my dissertation committee. Your feedback and collaboration on this research has been invaluable. I am inspired by the rigor you bring with you in each and every one of your projects. I also want to thank you for the advice and support you have provided for me in many ways outside of research, including various INFORMS-related activities and my time on the job market. I am also grateful for Joshua Stein, whose support and collaboration have been essential in my development as an academic.

IOE has been my intellectual home for the past five years. I am especially grateful for the support and generosity shown to me by various faculty in the department, including Amy Cohn, Marina Epelman, Joi Mondisa, Larry Seiford, and Mark Van Oyen. Thank you for your guidance, feedback, and support. I also want to thank the IOE staff, who, at various points throughout my PhD, have made my life much easier with respect to technology, room

scheduling, service commitments, teaching, and more. To Tina, Wanda, Rebekah, Chris, Rod, Mint, Valerie, Cathy, Eyevind, and Tom — thank you for your warmth and kindness. I am grateful for the custodial staff at the University of Michigan, especially Alejandro — gracias por ayudarme con mi español. I also want to thank all those whom I have met through CHEPS, which has been an extremely valuable resource for me during my PhD (I am sure that all but two of my posters have been printed by CHEPS). Thanks Gene Kim and Liz Fisher!

I have learned a great deal from my various collaborators. Aside from those I have already mentioned, I want to acknowledge Chris Andrews, Xiang Liu, Gregg Schell, Caroline Schumb, Erik Koffijberg, Jing Yang, Liz Lobazo, Spencer Liebel, Lauren Czerniak, Mike McCrea, and Tom McAllister. I also want to thank Andrea Almeida, Matt Lorincz, Jodi Harland, Nicole Johnson, the CARE Consortium PIs, and the CARE Consortium data team, whose efforts and insights have made this dissertation research possible.

The friends that I have met during the past few years have made my PhD experience much more enjoyable than I could have imagined. Wesley Marrero, Romulo Goes, Alejandro Vigo, Qi Luo, Laura Motta, Ted Nowak, Justin Haney, Thomas Chen, Richard Mwakasege-Minaya, Naz Mwakasege-Minaya, Darwin Guevarra, Robbie Lee, Andrea Belgrade, Tom Logan, Lauren Steimle, Donald Richardson, Yadrianna Acosta-Sojo, Wilmer Henao, Adam VanDeusen, Caroline Johnson, Anna White, Lauren Biernacki, Sammi Meister, Pranjal Singh, James Day, Ece Sanci, Karmel Shehadeh, Niusha Navidi, Emily Tucker, Victor Wu, Brandon Pitts, and Kayse Maass — thanks for helping me fill the time between sleep and research with concerts, fountain lunch, workouts, basketball, tennis, golf, barbecues, hot takes, Game of Thrones, West World, celebrations, eating competitions, ice cream, doggy dates, coffee runs, drinks, and all around good times.

I never would have started this PhD without the support of various people from the University of Pittsburgh. Thanks Lisa Maillart and Bryan Norman for supporting me as an undergraduate researcher and continuing to support me as I progressed through my PhD. I am especially indebted to the Pitt Excel program, especially Sim Saunders, Yvette Moore, and Allaine Allen, whose encouragement, support, and counsel have pushed me to strive for excellence and shaped me to be the person I am today. Big shout out to the Pitt Excel PhDs, especially Rodney Kizito, Chris Cameron, and Brittany Givens Rassoolkhani — I really appreciate your support since the day we all applied for PhD programs. I am

glad that we could lean on each other over the past few years. Thank you to my dearest friends from Pitt who have continued to provide support from afar, especially Zach Smith, Ali Roperti, Alvaro Cardoza, Matt Macar, Ryan Kennedy, Mike Doucette, Tony and Dan Mercader, and Jarrett Eakins.

My family and extended family have been a constant source of support and motivation throughout my life. Thanks Mom, Dad, Kuya Brian, Ad, Lin, Rui Rui, and Jill for your unconditional love, support, and encouragement. Thanks Tita Gigi, Tita Mimi, Tito Wilson, Tito Poch, Danica, and Mei-mei for always being there for me and supporting my education. Tito Martin, Tita Lisa, Jack, and Ana — I am glad we have been able to reconnect over the past few years. I am so grateful for your continued support and always look forward to your hospitality when I visit. I am especially grateful for my partner, Steph Fajardo. Thanks for joining me on this journey and showing me your love and support on every step along the way. Finally, I want to thank our dogs, Coco and Nana, whose daily kalokohan make me smile and laugh.

Finally, I am fortunate to have received financial support for my dissertation research. I want to acknowledge the Rackham Merit Fellowship and its community, especially Emma Flores-Scott. I am extremely grateful for the support of Mrs. Merill Bonder and the Seth Bonder Foundation. This dissertation is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256260.

Table of Contents

Dedication ii

Acknowledgments iii

List of Figures xi

List of Tables xiv

Abstract xvii

Chapter 1. Introduction 1
- 1.1 Concussion Management . . . 2
  - 1.1.1 CARE Consortium Data . . . 3
- 1.2 Organization and Contributions . . . 4
  - 1.2.1 Part 1: Estimating the Risk of Acute Concussion . . . 6
  - 1.2.2 Part 2: Optimization and Analysis of Diagnosis Decisions . . . 7
  - 1.2.3 Part 3: Optimization and Analysis of RTP Decisions . . . 9

Chapter 2. Quantifying the Value of Multidimensional Assessment Models for Acute Concussion 11
- 2.1 Introduction . . . 11
- 2.2 Methodology . . . 13
  - 2.2.1 Study Measures . . . 13
  - 2.2.2 Data Analysis . . . 14
  - 2.2.3 Quantifying Differences Between Acute Concussions and Normal Performances . . . 14
  - 2.2.4 Performance Measures . . . 16
- 2.3 Results . . . 16
  - 2.3.1 Univariate Logistic Regression Analysis . . . 18
  - 2.3.2 Multivariate Logistic Regression Analysis . . . 20
- 2.4 Discussion . . . 26

2.5 Conclusion 30

Chapter 3. Optimizing Components of the Sport Concussion Assessment Tool for Acute Concussion Assessment 31
3.1 Introduction 31
3.2 Methodology 32
3.2.1 Study Design 32
3.2.2 Test Methods 32
3.2.3 Statistical Analysis 34
3.3 Results 36
3.3.1 Model Descriptions 36
3.3.2 Model Performance 40
3.4 Discussion 46
3.4.1 Limitations 48
3.5 Conclusion 49
3.A Analysis via 10-fold Cross-Validation 50
3.B Removal of SAC Delayed Recall 55

Chapter 4. Data-driven Diagnosis Decision Thresholds for Risk Estimation Models 58
4.1 Introduction 58
4.2 Relevant Literature 61
4.2.1 Operations Research in Disease Screening and Diagnosis Decisions . 62
4.2.2 Determining Diagnosis Decision Boundaries 63
4.3 Modeling Approach 64
4.3.1 Problem Setting and Notation 65
4.3.2 Stochastic Programming Formulation 66
4.3.3 Approximating the Two-threshold Problem 67
4.3.4 Structural Properties 68
4.3.5 Utility-based Frameworks 73
4.3.6 Extensions to Multi-class Diagnosis 77
4.4 Data-driven Solution Methods 84
4.4.1 Quantile Estimation 85
4.4.2 Data-driven Distributionally Robust Optimization 85
4.4.3 Model Calibration 87
4.5 Simulation Analysis 91
4.5.1 Computational Time for TTP*-Q and TTP*-DR 91
4.5.2 Feasibility and Optimality of TTP*-Q and TTP*-DR 91
4.5.3 Comparing One- and Two-threshold Classification Schemes 95
4.5.4 Analysis of Ordinal Threshold Problem 97
4.6 Case Study: Acute Concussion Assessment 100

4.6.1 Concussion Assessment Data . . . 100
4.6.2 Risk Estimation Model . . . 101
4.6.3 Example Solution and Post-hoc Analysis . . . 102
4.6.4 Analyzing The Efficiency of Two-threshold Solutions . . . 103
4.6.5 Comparing TTP* Performance To Existing Methods . . . 107
4.7 Conclusion . . . 111
4.7.1 Managerial Insights . . . 111
4.7.2 Limitations and Future Work . . . 114
4.A Quantile Estimation Representations of TTP . . . 116
4.B Reformulating TTP*-DR Constraints . . . 117
4.C Simulation Analysis of Section 4.5 With Trapezoidal Distribution . . . 120

Chapter 5. Data-driven Approach to Unlikely, Possible, Probable, and Definite Concussion 123
5.1 Introduction . . . 123
5.2 Materials and Methods . . . 124
5.2.1 Study Population and Design . . . 124
5.2.2 Sample Selection . . . 124
5.2.3 Study Variables . . . 125
5.2.4 Data Analysis . . . 125
5.2.5 Model Calibration . . . 126
5.2.6 Model Validation . . . 128
5.2.7 Model Evaluation . . . 128
5.3 Results . . . 130
5.3.1 Multivariate Logistic Regression . . . 130
5.3.2 Classifying Unlikely, Possible, Probable, and Definite Concussion . . . 132
5.3.3 Distribution of Acute Concussions and Normal Performances . . . 132
5.3.4 Interclass Differences . . . 135
5.3.5 Intraclass Differences . . . 135
5.4 Discussion . . . 139

Chapter 6. Cluster Analysis of Possible and Probable Concussions 144
6.1 Introduction . . . 144
6.2 Materials and Methods . . . 145
6.2.1 Study Population and Design . . . 145
6.2.2 Sample Selection . . . 145
6.2.3 Study Variables . . . 145
6.2.4 Data Analysis . . . 146
6.3 Results . . . 149
6.3.1 Characteristics of Study Data . . . 149

6.3.2 Characteristics of Possible and Probable Concussions 149
6.3.3 Clustering Variables 151
6.3.4 Analysis of Significant Differences 152
6.3.5 Analysis of Clusters with the Lowest Gini Index 155
6.4 Discussion 158

Chapter 7. Estimating the Value of Incorporating Patient Behavior in Return-to-play Decisions 163
7.1 Introduction 163
7.2 Literature Review 167
7.3 Modeling Framework 169
7.3.1 Problem Description 169
7.3.2 Patient's Health Dynamics 170
7.3.3 Doctor's Treatment Cessation Problem 172
7.3.4 Patient's Symptom-reporting Problem 175
7.4 Analytical Results 176
7.4.1 Analysis of the BLM-POMDP 178
7.4.2 Analysis of the BAM-POMDP 180
7.4.3 Analysis of Patient Strategies 189
7.5 Solution Methodology 194
7.6 RTP From Sports-Related Concussion 195
7.6.1 Background on Sports-Related Concussion 195
7.6.2 Modeling RTP From Concussion 196
7.6.3 Model Parameters and Data Sources 197
7.6.4 Benchmark Policies 198
7.6.5 Simulation Framework 198
7.6.6 Analysis of BAM-POMDP RTP Policies 198
7.6.7 Effect of Symptom-reporting Behavior on RTP Policy Performance 201
7.6.8 Estimating the Value of Incorporating Patient Behavior 204
7.7 Conclusion, Limitations, and Future Directions 206
7.A Modeling RTP From Sports-Related Concussion 209
7.A.1 Concussion Recovery Dynamics 209
7.A.2 Post-RTP Markov Process 211
7.A.3 Rewards 212
7.B Derivation of Model Inputs 214
7.B.1 Parameterizing the HMM 214
7.B.2 Post-RTP Injury Rates 216
7.B.3 Rewards 218
7.C Importance of Initial Beliefs for BLM-POMDP 219

Chapter 8. Conclusion and Future Work **221**
8.1 An Integrated Approach to Personalized Concussion Management 222
8.2 Data-driven Decision-making in the Management of Military Concussions . 224
8.3 Design and Analysis of Shared Decision-Making Frameworks 225
8.4 Conclusion 226

Bibliography **227**

List of Figures

1.1 Organization of dissertation, connections between chapters, and connections between chapters and the concussion management protocol 5

2.1 Receiver operating characteristic curves for selected multivariate models based on validation against testing sets at <6h and 24–48 hours. AUC, area under the receiver operating characteristic curve; SCAT, Sport Concussion Assessment Tool 23

3.1 Receiver operating characteristic curves for Opt-k models and Summary Scores models at <6h and 24-48h. AUC, area under the receiver operating characteristic curve 44

3.2 Receiver operating characteristic curves for Opt-RS-k models and Summary Scores models at <6h and 24-48h. AUC, area under the receiver operating characteristic curve 45

4.1 Median $\log_{10}$(computational time in seconds) for solving TTP*-Q and TTP*-DR. Shaded area represents the 5^{th} and 95^{th} percentiles. 92

4.2 Distribution of optimality gap and maximum constraint violation under varying quality of risk estimation model (AUROC) and sample size (N^+, N^-). 94

4.3 Comparison of overall accuracy between one-threshold (1T*) and two-threshold (TTP*-Q) classification schemes under varying quality of risk estimation model (AUROC) and post-defer classification accuracy (k). AUROC, area under the receiver operating characteristic curve. 98

4.4 Comparison of accuracy and mean squared error in ordinal classification for the Ordinal Threshold Problem (OTP) and Equidistant Thresholds (ET) 100

4.1 Optimal upper threshold (u^*) and lower threshold (l^*) when solving TTP*-Q with $\gamma^{fp} = 0.028$ and $\gamma^{fn} = 0.020$. Risk estimates $> u^*$ are classified as true-positives while risk estimates $< l^*$ are classified as true-negatives. Diagnosis decisions are deferred for risk estimates between u^* and l^*. True positives (+) and true-negatives (x) from validation data are shown with estimated kernel densities. 103

4.2 (a) $q = 0.2$. (b) $q = 0.4$. (c) $q = 0.6$. (d) $q = 0.8$. Rate of correct classifications per deferred decision (e_1) vs. Rate of correct classifications per misclassification (e_3) for TTP*-Q under various parameter combination and proportion of true-positives, q. Pareto optimal (circles) and dominated (crosses) parameter combinations are shown. Labels are shown only for parameter combinations which are Pareto optimal for at least two values of q. These labels correspond to parameter combinations in Table 4.3. . . 105
4.3 (a) False-positive vs. false-negative rate satisfying minimum levels of sensitivity. (b) Sensitivity vs. false-positive rate satisfying maximum levels of false-negative rate. Comparison of Pareto optimal parameter combinations for TTP*-Q, TTP*-DR, optimal single-threshold solutions (1T*), and normative value comparison (NC). 109
4.4 (a) $q = 0.2$. (b) $q = 0.4$. (c) $q = 0.6$. (d) $q = 0.8$. Probability of correct classification vs. probability of misclassification for varying proportions of true-positives, q, and Pareto optimal parameter combinations of TTP*-Q , TTP*-DR, optimal single-threshold solutions (1T*), and normative value comparison (NC). 110
4.C.1 Distribution of optimality gap and maximum constraint violation under varying quality of risk estimation model (AUROC) and sample size (N^+, N^-). AUROC, area under the receiver operating characteristic curve. 121

5.1 Illustration of methodological framework for developing data-driven models which were used to classify athletes as Unlikely, Possible, Probable, or Definite concussion based on certainty of acute concussion. CART, Classification Tree; ADASYN, Adaptive Synthetic Sampling 127
5.1 Classification trees for determining Possible and Probable concussions at <6h and 24-48h post-injury 133

6.1 Illustration of data analysis procedure 147
6.1 Frequency of clustering variables among cluster sets in the lowest 200 Gini Indices by sex at (a) <6h post-injury and (b) 24-48h post-injury 154
6.2 Frequency of significant differences in study variables among cluster sets in the lowest 200 Gini Indices by sex at (a) <6h post-injury and (b) 24-48h post-injury 156

7.1 Illustration of BLM-POMDP Treatment Cessation Decision Process . . . 170
7.1 The BAM-POMDP's value function is not piecewise linear and convex in π like the POMDP's. 183

7.2 (a) Theorem 7.3 implies that $\bar{\Gamma}_1 \subseteq \Gamma_1 \subseteq \Gamma_L$ with $\mu^*(\pi) = 1$ in all shaded regions and $\mu^*(\pi) = 0$ in all non-shaded regions. (b) Illustration of variable-resolution grid regions in Section 7.5 based on Theorem 7.3. Allocation of grid points should be prioritized in $\mathcal{R}_3$, followed by $\mathcal{R}_1$ and $\mathcal{R}_2$. 186
7.1 BAM-POMDP policies for honest athletes. Black indicates that it is optimal to RTP and gray indicates that it is optimal to wait. For belief states not shown, it is optimal to wait. 199
7.3 Differences in BAM-POMDP policies relative to honest behavior (i.e., $b_0 = 0$). RTP, return-to-play. 200
7.4 Illustration of BLM-POMDP and POMDP health state belief evolution for sample athletes. Differences in performance between the BLM-POMDP and POMDP are primarily due to differences in health belief updates rather than differences in RTP policy. RTP, return-to-play; $\bar{s}_t$, reported symptom; o_t, objective assessment 201
7.5 Total discounted health utilities for each RTP policy and behavior type $b_0 \in \mathcal{B}$. Markers indicate mean values and shaded areas indicate 95% confidence intervals of the mean. 202
7.7 Violin plots illustrating distribution of relative RTP delay. Median RTP delay is shown by a circle for BLM-POMDPs and a square for POMDPs in each violin plot, while lower and upper bars indicate 5^{th} and 95^{th} percentiles, respectively. Probability of Premature RTP is shown below BLM-POMDP violin plots and above POMDP violin plots. ** No athletes RTP for POMDP when $b_0 = -1$. RTP, return-to-play. 203
7.8 Lower ($\mathbb{V}_{LB}$) and upper ($\mathbb{V}_{UB}$) bounds on the Value of Incorporating Patient Behavior (VoIPB) over the POMDP. 205
7.A.1 Illustration of Pre-RTP States, Post-RTP States, and State Transitions. . 210
7.B.1 Distribution of SAC total scores and SCAT total symptom severity scores for testing data and HMM predictions for all $\theta \in \Theta$. SAC = Standard Assessment of Concussion; SCAT = Sport Concussion Assessment Tool; D_{KL} = Kullback-Leibler divergence 217
7.C.1 Total discounted health utilities for BLM-POMDP with different initial behavior beliefs. ϕ_0^1, ϕ_0^2, and ϕ_0^3 correspond to initial beliefs of honest, under-reporting, and over-reporting, respectively. Markers indicate mean values and shaded areas indicate 95% confidence interval of the mean. . . 220

List of Tables

2.1 Characteristics of the study data with reference to selected variables by timepoint . . . 17
2.2 Results of the univariate logistic regression for association between risk factors and concussion for <6h and 24-48h hours . . . 19
2.3 Testing and training set estimates for performance measures of univariate models at <6h and 24-48h hours . . . 20
2.4 Factors in multivariate logistic regression (full model) associated with acute concussion at <6h and 24-48h . . . 21
2.5 Testing set and training set estimates for performance measures of multivariate models at <6h and 24-48h . . . 22
2.6 Association in limited models between risk factors and concussion for <6 hours and 24-48 hours . . . 24
2.7 Raw score analog of factors in multivariate logistic regression (raw score model) associated with concussion diagnosis at <6 hours and 24-48 hours 25
2.8 Objective factors in multivariate logistic regression (objective model) associated with concussion diagnosis at <6 hours and 24-48 hours . . . 25
2.9 Comparison of standard assessments for males and females at <6 hours and 24-48 hours . . . 29

3.1 Characteristics of study data by timepoint . . . 37
3.2 Change scores for the SCAT symptom checklist, SAC, and BESS by timepoint 38
3.3 Model variables, coefficient values, and performance measures for Opt-k models at <6h . . . 39
3.4 Model variables, coefficient values, and performance measures for Opt-k models at 24-48h . . . 40
3.5 Model variables, coefficient values, and performance measures for Summary Scores models at <6h and 24-48h . . . 41
3.6 Model variables, coefficient values, and performance measures for Opt-RS-k models at <6h . . . 42
3.7 Model variables, coefficient values, and performance measures for Opt-RS-k models at 24-48h . . . 43

3.8 Variable Inflation Factors for Opt-k Models at <6h and 24-48h 46
3.A.1 Frequency of variables in Opt-k models at <6h using 10-fold cross-validation 51
3.A.2 Frequency of variables in Opt-k models at 24-48h using 10-fold cross-validation 52
3.A.3 Frequency of variables in Opt-RS-k models at <6h using 10-fold cross-validation 53
3.A.4 Frequency of variables in Opt-RS-k models at 24-48h using 10-fold cross-validation 54
3.A.5 Performance of Opt-k and Opt-RS-k models based on 10-fold cross-validation 55
3.B.1 Modified Opt-8, Opt-12, and Opt-16 Models without SAC Delayed Recall at <6h post-injury 56
3.B.2 Modified Opt-8, Opt-12, and Opt-16 Models without SAC Delayed Recall at 24-48h post-injury 57

4.1 Model notation 65
4.1 Description of training and validation data 101
4.2 Comparison of athletes who were correctly diagnosed and deferred under example TTP*-Q solution 104
4.3 Selected Pareto Optimal parameter combinations for efficiency analysis . 106

5.2 Multivariate logistic regression coefficients at <6h and 24-48h post-injury 130
5.1 Data characteristics of training and validation data with respect to each timepoint 131
5.3 Distribution of acute concussions and normal performances among risk classifications at <6h and 24-48h post-injury 134
5.4 Comparison of study variables for acute concussions classified as Unlikely, Possible, Probable, and Definite concussion at <6h and 24-48h post-injury 136
5.5 Comparison of study variables for normal performances classified as Unlikely, Possible, Probable, and Definite concussion at <6h and 24-48h post-injury 137

6.1 Symptom groups 146
6.1 Characteristics of normal performances by timepoint 150
6.2 Characteristics of acute concussions by timepoint 151
6.3 Characteristics of Possible and Probable concussions at <6h 152
6.4 Characteristics of Possible and Probable concussions at 24-48h 153
6.5 Two-cluster cluster sets with the lowest Gini index for males and females at <6h 157

6.6 Two-cluster cluster sets with the lowest Gini index for males and females at 24-48h 158

7.1 Optimality Gap UB (%) for BLM-POMDP policies based on simulation estimates 201
7.2 BLM-POMDP reduction in Probability of Premature RTP (PoPRTP) and gain in Total Health-adjusted Athletic Exposures (THAEs) over benchmark policies 206
7.B.1 Injury rate parameters and sources 218
7.B.2 Health utility values for the doctor's reward function for each $\theta \in \Theta$ . . . 219
7.B.3 Health utility values for the patient's reward function for each $\theta \in \Theta$ and $b \in \mathcal{B}$. 219

Abstract

As the volume and granularity of health data continue to increase, clinical decision-makers are faced with two key questions: (Q1) How can large clinical datasets be used to gain a patient-specific understanding of disease risk and disease progression? (Q2) How can a data-driven understanding of patient-specific disease risk and disease progression be combined with multiple stakeholders' perspectives to optimize medical decision-making? These challenges are especially pertinent to managing patients with concussion. Concussion, an emergent public health issue, affects millions of people in the United States each year. Characterized by wide-ranging symptoms and impairment in neurocognitive function, researchers believe that improving concussion management can mitigate the long-term consequences associated with the injury. In this dissertation, we answer (Q1) and (Q2) by analyzing three key aspects of concussion management: acute concussion assessment, diagnosis decisions, and return-to-play (RTP) decisions. Throughout this dissertation, we develop, parameterize, and validate our models using data from the Concussion Assessment, Research, and Education (CARE) Consortium — a large dataset on concussion among collegiate athletes from 29 universities and military service academies across the United States.

In our analysis of acute concussion assessment, we design predictive models to assess the relationship between acute concussion and clinical assessments, individual risk modifiers (e.g., age, sex, number of previous concussions), and time of injury characteristics (e.g., loss of consciousness). This research provides valuable contributions in (1) quantifying the value of a multi-dimensional approach to acute concussion assessment and (2) identifying specific components of the Sport Concussion Assessment Tool which best identify acute concussion.

To analyze concussion diagnosis decisions, we formulate and solve the Two-Threshold Problem (TTP): a data-driven stochastic programming approach to determine optimal diagnosis decision thresholds with risk estimation models. Using the personalized risk estimation models from the first part of this dissertation as an input, we apply the TTP to acute concus-

sion diagnosis and identify its implications for clinicians. The contributions of this research include (1) the development of a novel data-driven framework for optimizing diagnosis decisions, (2) an algorithmic approach to classifying the certainty in acute concussion diagnosis decisions (i.e., Unlikely, Possible, Probable, and Definite concussion), and (3) the characterization of athletes who are most difficult to diagnose, i.e., Possible and Probable concussions.

The final part of this dissertation analyzes the timing of RTP from concussion. We first formulate and solve a novel Behavior-Learning Multi-agent POMDP (BLM-POMDP): a multi-agent, stochastic dynamic programming model which incorporates the patient's and doctor's perspectives while accounting for uncertainty in the patient's health and symptom-reporting behavior. We then apply the BLM-POMDP to CARE Consortium data to estimate the value of incorporating patient behavior in RTP decisions. The contributions of this work include (1) the formulation and characterization of a novel dynamic programming framework which naturally models patient-doctor interactions in sequential treatment planning regimes and (2) the development and analysis of an optimal RTP policy which can be tailored to each athlete and outperforms current practice.

In summary, this dissertation combines data analytics and operations research to address major challenges in concussion management. Our modeling frameworks span the range of predictive models for risk estimation to data-driven sequential decision-making under uncertainty. While this research was motivated by and applied to concussion management decisions, this research can be adapted to a broader range of application areas where data, prediction, and decisions play a crucial role.

Chapter 1

Introduction

Over the past decade, the *Big Data movement* — characterized by increasing volume, velocity, and variety of data — has revolutionized and transformed numerous industries including retail, manufacturing, sports, and healthcare (Mcafee and Brynjolfsson, 2012). Within healthcare, Big Data has presented the opportunity to develop new insights across several facets of health systems, including healthcare operations, pricing of health services, drug development, and medical decision-making (Bates et al., 2014). Yet, solely increasing the amount of data and the availability of data is not enough to improve healthcare — new data analytics techniques must be developed to transform these data into insights and these insights into value (Lavelle et al., 2011). Two key questions which must be addressed to transform these data into actionable and valuable information are:

1. How can large clinical datasets be used to gain a patient-specific understanding of disease risk and disease progression?

2. How can a data-driven understanding of patient-specific disease risk and disease progression be combined with multiple stakeholders' perspectives to optimize medical decision-making?

While these challenges are universally important to many disease areas, they are especially pertinent to the clinical assessment and management of concussion.

1.1 Concussion Management

Concussion, the most common type of traumatic brain injury (TBI), affects millions of people in the United States each year and has become a recent focal point in public health (CDC, 2016; Langlois, Rutland-Brown, and Wald, 2006; McCrory et al., 2017). In the short-term, concussion is associated with wide-ranging symptoms including headache, amnesia, sensitivity to light and noise, loss of neurocognitive function, and impaired balance (Broglio et al., 2014; Giza et al., 2013; Harmon et al., 2013; McCrory et al., 2017). In the long-term, concussion is associated with persistent post-concussion symptoms, Alzheimer's disease, dementia, depression, and neurodegenerative diseases such as chronic traumatic encephalopathy (Daneshvar et al., 2011; De Beaumont et al., 2007; Kerr et al., 2018a; McCrory et al., 2013).

Concussion management plays a critical role in mitigating these short- and long-term consequences, especially in sports, which accounts for roughly 30% of concussions among youth (Voss et al., 2015). While the specifics of the concussion management protocol continue to evolve, two key components of the process include (1) concussion assessment and diagnosis and (2) determining the timing of return-to-play (RTP) from concussion.

The accurate assessment and diagnosis of concussion is critical to initiating the concussion management protocol. Delayed diagnosis can increase risks for secondary injuries, lengthen recovery time, and exacerbate concussion symptoms (Asken et al., 2018; Asken et al., 2016). However, achieving this combination of accuracy and timeliness in diagnosis is a major challenge for physicians and clinicians. First and foremost, a perfect diagnostic marker for concussion does not yet exist. Therefore, clinicians must rely on available clinical measures and assessments to determine whether athletes do or do not have concussion. To this end, while several clinical assessments and assessment batteries have been developed in the few years (McCrory et al., 2017), the exact combination of assessments which best detects concussion for each type of athlete (e.g., by sex and concussion history) is unclear.

Once athletes have been diagnosed with concussion, they typically progress through a "graded RTP protocol" in which they gradually increase their activity levels until deemed fit to RTP (Kutcher and Giza, 2014). Over 25 distinct approaches to concussion assessment and RTP decisions have been published since 2001. However, many of these guidelines are based on expert consensus rather than scientific evidence (McCrea et al., 2005). Often

relying on physician experience, these guidelines make it difficult to accurately determine when a concussion has fully resolved. Further, athletes may under-report symptoms to achieve earlier RTP (Williamson and Goodman, 2006). To this end, premature RTP is associated with increased risks for secondary injury, post-concussion syndrome, and second impact syndrome (Broglio et al., 2014; Harmon et al., 2013; Makdissi et al., 2010; McGrath, 2010) while late RTP can result in lost opportunities or decreased quality of life (Kutcher and Giza, 2014; Schneider et al., 2013). Hence, determining the optimal timing for RTP is a major challenge in concussion management.

By combining operations research and data analytics, this dissertation develops new methodological frameworks in *predictive analytics* (i.e., data-driven prediction and estimation of unknown information) and *prescriptive analytics* (i.e., data-driven decision-making in the face of known and unknown information) to address these challenges within concussion assessment and RTP decisions.

1.1.1 CARE Consortium Data

Throughout this dissertation, we develop our models and perform our analysis using data from the National Collegiate Athletic Association-Department of Defense (NCAA-DoD) Concussion Assessment, Research, and Education (CARE) Consortium (Broglio et al., 2017). The CARE Consortium was established to study the natural history and neurobiological recovery of concussion to improve the overall management of concussion for military personnel and student-athletes (i.e., participants). From evidence-based guidelines, the CARE Consortium defines concussion as "a change in brain function following a force to the head, which may accompanied by temporary loss of consciousness, but is identified in awake individuals with measures of neurologic and cognitive function" (Carney et al., 2014).

CARE data were collected from varsity athletes at 30 NCAA universities and military service academies who underwent preseason baseline evaluations during the 2014-2019 academic years. These player-seasons consisted of male (57.08%) and female (42.90%) participants from 27 sports, including football (19.2%), cross country/track (11.4%), and soccer (9.6%). Only non-varsity military service academy cadets were included in these data. These data contained 38,379 player-seasons from 29,712 participants who completed a pre-season baseline evaluation and 2,971 concussions across 2,628 participants. Participants diagnosed

with concussion by the local institution's medical staff obtained assessments within 6 hours post-injury (<6h), 24-48 hours post-injury (24-48h), when cleared to begin the return-to-play (RTP) protocol (asymptomatic), when cleared for unrestricted RTP, and 6 months after the time of unrestricted RTP. RTP protocol decisions were made at the discretion of the local medical staff at each institution based on the CARE Consortium study protocol (Broglio et al., 2017). Some participants did not complete post-injury assessments at each timepoint, leading to missingness in the data. All participants provided written consent that was approved by their local institutional review board and the United States Army Human Research Protection Office.

Throughout this dissertation, we updated the CARE Consortium data as new data became available. Hence, each chapter may have used a different version of the CARE Consortium Data. For example, the analysis in Chapter 3 was conducted more recently than that of Chapter 2, explaining the differences in sample sizes.

1.2 Organization and Contributions

Figure 1.1 illustrates the organization of this dissertation, the descriptions of each chapter, and the connections between chapters. This dissertation is divided into three parts. The first part of this dissertation (Chapters 2 and 3) develops predictive models to estimate the risk of acute concussion based on a combination of individual risk factors, injury characteristics, and clinical measures. The second part of this dissertation (Chapters 4, 5, and 6) develops and analyzes a data-driven stochastic programming framework which optimizes diagnosis decision thresholds based on risk estimation models (like those developed in part 1). These decision thresholds guide clinicians in determining which patients should be diagnosed as positive or negative, or whether diagnosis decisions should be deferred due to elevated inaccuracy inherent to the risk estimation model. In the third part of this dissertation (Chapter 7), we formulate and analyze a novel multi-agent stochastic dynamic programming framework which optimizes the timing of treatment decisions for potentially strategic patients. We now briefly describe each chapter and its contributions.

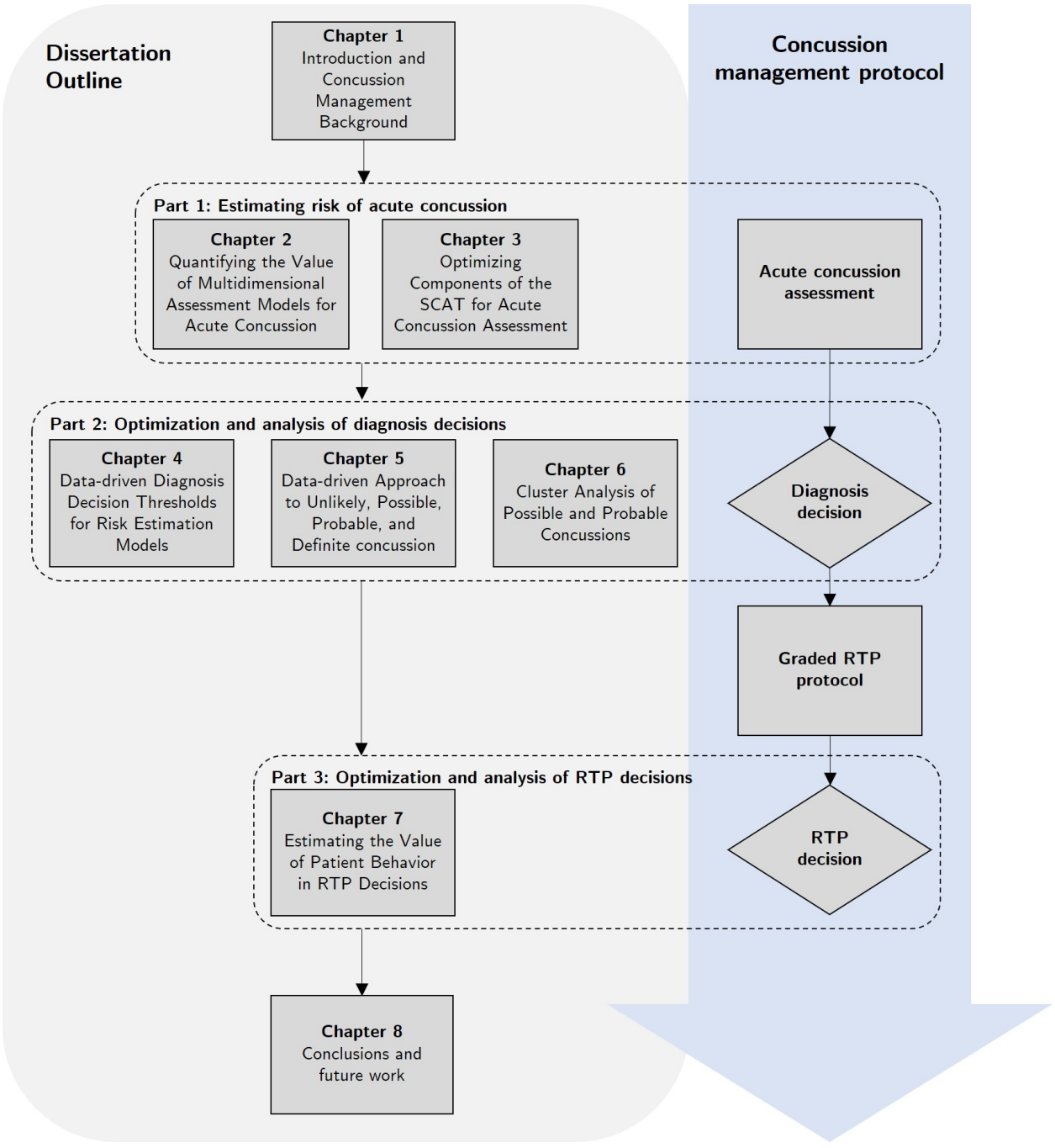


Figure 1.1: Organization of dissertation, connections between chapters, and connections between chapters and the concussion management protocol

1.2.1 Part 1: Estimating the Risk of Acute Concussion

In Chapter 2, we develop univariate and multivariate logistic regression models to quantify and analyze the association between acute concussion and several important clinical variables, including individual risk factors, injury characteristics, and standard clinical assessments. The main contributions of Chapter 2 are as follows.

- We establish quantitative evidence supporting the clinical utility of a multi-faceted approach to acute concussion assessment.
- We show that measures of symptom presentation are the most important indicator of acute concussion and removing them from an assessment battery can drastically reduce its accuracy in assessing acute concussion.
- We quantify the marginal clinical utility of change scores for acute concussion assessment, showing that acute concussion assessment can still be performed at clinically acceptable levels of accuracy when baseline information is unavailable.
- This research was published in Sports Medicine (Garcia et al., 2018).

In Chapter 3, we extend our previous modeling framework in Chapter 2 by formulating a mixed integer program (MIP) to identify the subset of SCAT symptom checklist, SAC, and BESS components which best identifies acute concussion. The main contributions of Chapter 3 are as follows.

- We developed models (with and without change scores) which take less time to administer the SCAT and can more accurately assess concussion compared to the SCAT's composite scores.
- We identified specific components of the SCAT symptom checklist, SAC, and BESS which are highly associated with acute concussion.
- Our analysis provides additional quantitative support regarding the clinical utility of change scores.
- We provide quantitative evidence that some of the SCAT's components are simply adding noise to the process of assessing acute concussion, suggesting the importance of

identifying the components of the SCAT which are most pertinent for acute concussion assessment.

- This research was accepted for publication in Neurosurgery (Garcia et al., 2020b).

1.2.2 Part 2: Optimization and Analysis of Diagnosis Decisions

In Chapter 4, we formulate the Two-Threshold Problem (TTP) as a data-driven stochastic program to determine decision thresholds for diagnosis decisions with risk estimation models. Without making assumptions on the distribution of risk estimates, TTP provides clinical guidance on whether a patient should receive a positive or negative diagnosis, or whether the risk estimate for that patient is inconclusive. We then apply TTP to acute concussion assessment and show that on a number of performance metrics, TTP outperforms an optimized single threshold approach as well as normative value comparison approaches from clinical practice. The main contributions of Chapter 4 are as follows.

- We introduce a data-driven stochastic optimization framework to determine diagnostic decision thresholds based on the application of a risk estimation model to patients from a fixed population. Compared to previous methods, we avoid the need to estimate outcome-based utilities or make distributional assumptions to account for uncertainty in risk estimates.

- We show that the optimal solution to our proposed model can be characterized by extreme-point solutions of a related linear program. Thus, our model can be solved using quantile estimation — bypassing the need for advanced optimization software.

- We identify additional modeling frameworks, including utility-based and multi-class classification frameworks, for which our analytical results can be applied. Specifically, for our utility-based extensions, we formulate a model for which utilities such as quality-adjusted life-years may be used. In our extensions to multi-class classification, we develop frameworks for both multi-label and ordinal classification.

- Through an analytical study and numerical analysis using both real and simulated data, we determine when two decision thresholds, which allow for a deferred diagnosis

decision, will outperform a single decision threshold, which only allows for binary classification.

- We perform extensive numerical analysis to determine how the modeling parameters should be chosen based on the general characteristics of the population that undergoes diagnostic testing. Our analysis also gives insight to guide the choice of data-driven solution methodology based on sample size and the quality of the underlying risk estimation model.

- We are one of the first groups to apply an optimization framework to develop data-driven diagnostic thresholds for acute concussion based on data from the CARE Consortium — a nationwide collaboration comprising 29 National Collegiate Athletic Association (NCAA) universities and military service academies. By incorporating feedback from concussion experts across the CARE Consortium, we ensure that, in the case of acute concussion assessment, our modeling framework outperforms methods which are commonly used in practice. Furthermore, we provide a valuable framework which quantifies the uncertainty in diagnosis decisions using real data rather than subjective clinical experience. The models developed in this research have the potential to be developed into tools which can supplement clinical decision-making.

- This research was accepted for publication in IISE Transactions (Garcia et al., 2020a).

In Chapter 5, we apply and extend the methodology from Chapter 4 to develop a data-driven framework which classifies acute concussion based on diagnostic certainty, i.e., Unlikely, Possible, Probable, and Definite concussion. We characterize the distribution of acute concussions and normal performances within each risk category as well as analyze key differences within and across athletes placed in each of these risk categories. The main contributions of this chapter are as follows.

- We develop an objective and data-driven framework which stratifies acute concussion assessment by diagnostic certainty. These risk categories lay the foundation for guiding post-injury management decisions.

- We identify key characteristic which can be used to differentiate between acute concussions and normal performances in each risk category.

- We provide additional, quantitative support for the value of a multidimensional battery, the use of change scores in acute concussion assessment, and the potential implications for several demographic factors and time-of-injury characteristics in acute concussion assessment.

- This research was published in the Journal of Neurotrauma (Garcia et al., 2020b).

In Chapter 6, we extend our analysis from Chapter 5 by applying cluster analysis to provide a more granular analysis of Possible and Probable concussions. The main contributions of this chapter are as follows.

- We provide a characterization of athletes whose concussions are most difficult to assess (i.e., Possible and Probable concussions) — a subpopulation which is seldom analyzed in the literature.

- We guide clinical decision-making by highlighting specific symptom groups which best differentiate between acute concussions and normal performances among the Possible and Probable concussions.

- We quantify the utility of objective assessments and change scores among Possible and Probable concussions.

- This research will be submitted to a medical journal.

1.2.3 Part 3: Optimization and Analysis of RTP Decisions

In Chapter 7, we aim to develop a model which optimizes the timing of RTP while accounting for symptom-reporting behavior. Specifically, we formulate the Behavior-Learning Multi-agent POMDP (BLM-POMDP) — a multi-agent, multi-period, stochastic dynamic programming model which incorporates uncertainty in the patient's health and symptom-reporting behavior. We apply the BLM-POMDP to optimize the timing of RTP from sports-related concussion by incorporating CARE Consortium data and published values from the literature to parameterize and validate our model. Compared to benchmark RTP policies which fail to account for symptom-reporting behavior, the BLM-POMDP increase athlete's health utilities, reduces the likelihood of premature RTP, and increases their post-RTP

athletic exposures — especially in the presence of symptom under-reporting. The main contributions of this chapter are as follows:

- We formulate a novel BLM-POMDP framework, a multi-agent stochastic dynamic programming model which naturally fits patient-doctor interactions in many healthcare settings.
- We characterize the structure of the BLM-POMDP's optimal value function and optimal policy. Analyzing and solving the BLM-POMDP is a formidable task due to the curse of dimensionality. Furthermore, by explicitly modeling patient-doctor interactions, the BLM-POMDP negates classic results for the POMDP such as convexity of the value function. Nevertheless, we leverage decomposition and Blackwell dominance to derive structural properties of the BLM-POMDP which give way to an approximation solution method and reveal insights on optimizing treatment decisions with potentially strategic patients.
- We are one of the first groups to approach RTP from concussion with a decision-theoretic and data-driven approach by applying and evaluating the BLM-POMDP to RTP from concussion among collegiate athletes.
- We provide a data-driven framework to tailor athlete-specific RTP criteria and demonstrate its improvement over current practice.
- This research will be submitted to an engineering journal.

Chapter 2

Quantifying the Value of Multidimensional Assessment Models for Acute Concussion

2.1 Introduction

In sporting environments, concussion diagnosis is critical for proper injury management and must be both timely and accurate. To this end, many challenges remain for physicians and athletic trainers. One major challenge lies in the simultaneous use and interpretation of various concussion assessment tools. Multiple domestic and international organizations recommend clinicians implement multiple tests to evaluate several domains to support the clinical examination of the injured athlete (Broglio et al., 2014; Giza et al., 2013; Harmon et al., 2013; McCrory et al., 2017). Commonly used assessments include the Standard Assessment of Concussion (SAC), athlete-reported symptoms, and the Balance Error Scoring System (BESS). Unfortunately, few studies have quantified the performance of these methods to determine which are most accurate in identifying concussions alone or in combination. Among the studies which have analyzed these assessments for the evaluation of concussion, many of them consider only one assessment in isolation (Barr and McCrea, 2001; Broglio et al., 2008; Broglio, Macciocchi, and Ferrara, 2007; Lovell et al., 2006; McCrea et al., 2005). Moreover, most of these studies analyzed a small sample of male football athletes who have experienced concussion, limiting the generalizability of the results across sexes and sports.

Additionally, while it is widely accepted that combining multiple assessments (i.e., a testing battery such as the Sport Concussion Assessment Tool 5th Edition (SCAT5) (Echemendia et al., 2017)) improves concussion assessment capability (Broglio et al., 2014; Harmon et al., 2013; McCrory et al., 2017), no method combines the results of multiple tests into a single measure for guiding injury assessment – making the simultaneous interpretation of test results difficult. Another challenge in acute concussion assessment lies in the importance of change scores, i.e., the differences between an athlete's performance prior to and following injury. While many assessment methods require the use of change scores, these data may not always be available to clinicians and when they are, their utility has been questioned (Echemendia et al., 2012; Randolph, 2011; Schmidt et al., 2012). Similarly, another challenge lies in determining which clinical measures can be used for concussion assessment when data is unavailable or when self-reported symptoms are unreliable. Finally, while a number of modifying factors (e.g., age (Covassin et al., 2012; Putukian et al., 2015; Valovich McLeod et al., 2012), sex (Covassin, Buz Swanik, and Sachs, 2003; Covassin et al., 2012; Covassin, Schatz, and Swanik, 2007; Shehata et al., 2009; Valovich McLeod et al., 2012), and concussion history (Bruce and Echemendia, 2004; Covassin, Stearne, and Elbin, 2008; Guskiewicz et al., 2003; Shehata et al., 2009)) have been suggested to influence concussion risk and affect injury presentation, how to incorporate these risk modifiers in acute concussion assessment remains unclear.

The goal of this study is to address these aforementioned challenges in acute concussion assessment through a statistical modeling approach. Specifically, this study aims to: (1) evaluate selected standard assessments for the evaluation of acute concussion; (2) determine the assessments for which change score has more clinical utility than raw score for evaluating acute concussion; (3) quantitatively evaluate concussion assessment under limited clinical data or objective clinical measures. These aims are achieved by building, analyzing, and validating logistic regression models using data from a nationwide and multi-site study on sports-related concussions. This approach not only provides insight into which combinations of standard assessments are most important in acute concussion assessment, but it also combines these measures into a single risk estimate which can be used to guide clinical decision-making in the evaluation of acute concussion.

2.2 Methodology

This analysis used data from the CARE Consortium (see Section 1.1.1)

2.2.1 Study Measures

The timepoints considered in this study included baseline assessments and post-injury assessments from <6 hours (<6h), 24-48 hours (24-48h), and at the time of unrestricted RTP. Concussions without a baseline assessment prior to the injury were excluded from the analysis (n=62). Variables in this study include concussion risk modifiers (e.g., age and sex), the SAC, the Standard Concussion Assessment Tool (SCAT) symptom evaluations, and the BESS. The risk modifiers selected for this study have been previously identified as potential indicators for increased risk of concussion. Similarly, the selected assessments have been identified as practical and effective methods for evaluating concussion (Broglio et al., 2014; McCrory et al., 2017; McCrory et al., 2013). Furthermore, the individual assessments considered in this study are also part of existing, widely-used, and easily available testing batteries such as the SCAT5. Finally, as a clinical benefit, these study variables can also be obtained within the time constraints of athletics – making them useful for sideline assessment. For the SAC, SCAT symptom evaluations, and BESS, both raw score and change score were considered. Change score for a timepoint is defined as the difference between the raw score at that timepoint and the raw score at baseline. That is, a positive change score implies that an athlete scored higher post-injury compared to baseline and a negative change score implies that an athlete scored lower post-injury compared to baseline. The variables considered in this analysis are described in more detail below.

Select variables thought to influence concussion risk were selected for modeling. Younger athletes, females, and those with previous concussions may be at higher risk for concussion (Broglio et al., 2014; Guskiewicz et al., 2005; Harmon et al., 2013; McCrory et al., 2017). Additionally, injury characteristics such as whether the athlete experienced loss of consciousness (LOC), post-traumatic amnesia (PTA), and retrograde amnesia (RGA) are associated with concussion (Broglio et al., 2014; Harmon et al., 2013; McCrory et al., 2017). Whether the athlete was removed from play immediately and reported the injury immediately can modify the symptom presentation at the time of assessment (Asken et al., 2016; Broglio et al., 2014; Elbin et al., 2016; Harmon et al., 2013; McCrory et al., 2017).

The SAC is an instrument which includes measures of orientation, immediate memory, concentration, and delayed recall (McCrea et al., 1998). The SAC total score is considered to provide a holistic measure of neurological status, so the SAC total score and change score were considered in this study.

The SCAT symptom list is a standardized tool for evaluating symptom presentation among injured athletes (Concussion in Sport Group, 2013). Hence, the total number of symptoms reported and the total symptom severity score were included as variables, along with their change scores.

The BESS is a physical examination consisting of a double-leg stance, single-leg stance, and a tandem stance (Bell et al., 2011). Since impaired balance is believed to indicate concussion, the BESS total score and change score were considered.

2.2.2 Data Analysis

Most study variables were missing at <6% across all timepoints. The BESS assessments were missing at <10% for all timepoints except <6h, which was missing at 26.61%. Multiple imputation by chained equations was used to fill missing data (Van Buuren, Boshuizen, and Knook, 1999). This method is common in medical literature, including previous concussion studies (McCrea et al., 2013; McCrea et al., 2003; McCrea et al., 2005). Imputation was performed using the software R version 3.2.2.

2.2.3 Quantifying Differences Between Acute Concussions and Normal Performances

The probability of belonging to the "acute concussion" group was estimated using logistic regression. The "acute concussion" group consisted of post-injury assessments from <6h or 24-48h and the "normal performance" group consisted of assessments from baseline and unrestricted RTP. Separate models were created for the <6h and 24-48h timepoints. For the baseline data, it was assumed that all change scores were 0, the injury was reported immediately, the athlete was removed from play immediately, and the athlete did not experience LOC, PTA, or RGA. These assumptions were made since these variables were not measured during baseline evaluations and values were needed for logistic regression. To perform the analysis, all data, excluding baseline data, were randomly split into training

(75%) and testing (25%) sets. Baseline data were excluded from the training set to prevent model fitting from being influenced by baseline data assumptions. Five-fold cross-validation was performed on the training set to select and parameterize the models. Models were first validated using the testing set, which consisted of data from the unrestricted RTP group and the appropriate acute concussion group. Then, baseline data were used to validate the models separately.

Univariate logistic regression analysis was performed to understand how each study variable, individually, is associated with acute concussion. Then, multivariate logistic regression was performed using backward variable selection with Akaike Information Criteria (Bozdogan, 1987). The goal of the multivariate logistic regression was to understand how the study variables, in combination, are associated with acute concussion (full models). If the resulting variables contained both change score and raw score for the SAC, SCAT symptom assessments, or BESS, then two models were created for each instance. One model contained all variables resulting from backward variable selection except change score for that assessment and the other model contained all variables except raw score. The model with better performance measures on the training set was kept.

To assess multivariate model performance under limited data, full models were modified to estimate the impact of deleting one variable (limited models). These limited models highlight which variables most drastically affect the full model's performance when certain measures are unavailable. To create limited models, the full model was recreated on the training set with all variables except one. This procedure was repeated until a limited model was created for every variable in the full model. The impact on concussion assessment when baseline data is unavailable was evaluated by creating models which replaced all change score assessments in the full model with raw scores (raw score models).

Finally, since symptom under-reporting is a major concern within concussion management (Williamson and Goodman, 2006), multivariate models were created to estimate model performance without self-reported symptoms (objective models). Objective models were created using the same procedure as the full model except SCAT total symptoms and SCAT symptom severity were excluded from the initial set of variables. Models were created and analyzed using Python 3.5.2.

2.2.4 Performance Measures

Models were evaluated on sensitivity, specificity, and area under the curve (AUC). Sensitivity is the percentage of acute concussions at <6h or 24-48h correctly classified as acute concussions and specificity is the percentage of normal performance assessments from the baseline or unrestricted RTP timepoints correctly classified as normal performance. AUC is the likelihood that a model will estimate the probability of acute concussion to be higher for a randomly chosen acute concussion than a randomly chosen normal performance. Additionally, the limited models, raw score models, and objective models were compared to the full models using a bootstrap test for AUC on the testing set (bootstrap test) (Pepe, Longton, and Janes, 2009). A significant p-value suggests that a given model's AUC is less than the full model's.

2.3 Results

The study data with respect to selected concussion risk modifiers and standard assessments are summarized in Table 2.1. The groups do not differ significantly in terms of height, weight, age, sex, and previous number of concussions. However, the raw scores for the groups at <6h, 24-48h, and unrestricted RTP differ significantly from the baseline group ($P<0.01$).

Table 2.1: Characteristics of the study data with reference to selected variables by timepoint

Variable	Baseline	<6h	24-48h	Unrestricted RTP
n*	842	560	733	707
Days since injury (SD)*	NA	0.39 (7.66)	0.92 (1.80)	14.47 (12.66)
Height in meters (SD)*	1.80 (0.12)	1.80 (0.10)	1.79 (0.12)	1.79 (0.12)
Weight in kg (SD)*	84.04 (21.77)	86.24 (22.80)	83.68 (21.93)	83.22 (21.57)
Age in years (SD)	19.40 (1.30)	19.37 (1.31)	19.33 (1.27)	19.37 (1.30)
Male Sex (% yes)	61.52%	64.46%	60.03%	59.83%
Number of previous concussions (SD)	0.75 (1.02)	0.79 (1.04)	0.75 (1.05)	0.71 (0.95)
Report injury immediately? (% yes)	NA	53.57%	39.29%	39.75%
Removed from play immediately? (% yes)	NA	56.07%	45.98%	47.52%
LOC? (% yes)	NA	5.71%	4.50%	5.23%
PTA? (% yes)	NA	11.79%	11.19%	11.32%
RGA? (% yes)	NA	5.89%	6.14%	5.94%
SAC change score (SD)	NA	-0.83 (3.19)	-0.42 (2.60)	0.96 (2.13)
SAC raw score (SD)	27.05 (2.01)	26.18 (2.92)**	26.61 (2.42)**	27.93 (1.75)**
SCAT symptom severity change score (SD)	NA	23.47 (20.90)	19.53 (21.87)	-4.92 (8.81)
SCAT symptom severity raw score (SD)	5.08 (8.44)	28.79 (20.87)**	25.16 (21.64)**	0.63 (1.99)**
SCAT total symptoms change score (SD)	NA	8.07 (5.96)	7.48 (6.59)	-2.51 (4.02)
SCAT total symptoms raw score (SD)	2.77 (3.82)	10.89 (5.42)**	10.49 (6.02)**	0.47 (1.40)**
BESS change score (SD)	NA	3.68 (8.66)	1.43 (7.42)	-2.31 (6.27)
BESS raw score (SD)	12.62 (6.29)	16.41 (8.74)**	14.32 (7.87)**	10.40 (5.76)**

n, number of data points; SD, standard deviation; NA implies that the measure was not taken at baseline; LOC, loss of consciousness; PTA, post-traumatic amnesia; RGA, retrograde amnesia; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; Change score at a time point is computed as: raw score at time point - raw score at baseline; *variable was not considered in the models; **P<0.01 when compared to mean value at baseline using t-test

2.3.1 Univariate Logistic Regression Analysis

The relationships between individual variables and acute concussion are summarized in Table 2.2. These results address aim (1). For both <6h and 24-48h models, the SAC, SCAT symptom assessment, and BESS had significant p-values (<0.05). In contrast, age and sex both had p-values which were not significant (>0.05). Within the 24-48h univariate analysis, the variables PTA, RGA, and LOC had negative coefficients suggesting that these factors reduce one's likelihood of having acute concussion when present. This contradictory finding may be due to variance in the estimates; their p-values suggest that their coefficients are not significantly different from 0. To avoid implications that PTA, RGA, and LOC are "protective", these variables were removed from multivariate analysis. The performance measures for all univariate models are shown in Table 2.3. These results address aim (2). For both <6h and 24-48h models, raw scores for SCAT symptom severity and SCAT total number of symptoms resulted in the greatest combination of AUC and sensitivity. For the SAC and BESS, the raw score univariate models performed similarly to their change score counterparts.

Table 2.2: Results of the univariate logistic regression for association between risk factors and concussion for <6h and 24-48h hours

	<6 hours			**24-48h**		
Variable	*Intercept (SE)*	*Coefficient (SE)*	*p-value*	*Intercept (SE)*	*Coefficient (SE)*	*p-value*
Age in years	-1.46 (0.98)	0.07 (0.05)	0.19	0.11 (0.91)	0.00 (0.05)	0.96
Male Sex	-0.31 (0.11)	0.22 (0.14)	0.11	0.11 (0.10)	-0.06 (0.13)	0.61
Number of previous concussions	-0.24 (0.08)	0.10 (0.07)	0.14	0.04 (0.08)	0.04 (0.06)	0.56
Report injury immediately?	-0.36 (0.09)	0.41 (0.13)	0.00	0.07 (0.08)	0.00 (0.13)	0.97
Removed from play immediately?	-0.36 (0.10)	0.36 (0.13)	0.01	0.09 (0.08)	-0.03 (0.12)	0.78
LOC?	-0.18 (0.07)	0.14 (0.28)	0.62	0.08 (0.06)	-0.12 (0.29)	0.69
PTA?	-0.20 (0.07)	0.25 (0.20)	0.21	0.08 (0.06)	-0.06 (0.19)	0.76
RGA?	-0.17 (0.07)	-0.04 (0.27)	0.88	0.08 (0.06)	-0.17 (0.26)	0.50
Number of hours of sleep last night?	-0.37 (0.33)	0.03 (0.04)	0.54	-1.16 (0.28)	0.16 (0.04)	0.00
SAC change score	-0.12 (0.07)	-0.25 (0.03)	0.00	0.13 (0.06)	-0.25 (0.03)	0.00
SAC raw score	8.90 (0.93)	-0.33 (0.03)	0.00	8.76 (0.92)	-0.32 (0.03)	0.00
SCAT symptom severity change score	-1.17 (0.11)	0.27 (0.02)	0.00	-0.59 (0.09)	0.20 (0.01)	0.00
SCAT symptom severity raw score	-3.19 (0.21)	0.58 (0.04)	0.00	-2.18 (0.14)	0.50 (0.04)	0.00
SCAT total symptoms change score	-1.16 (0.12)	0.60 (0.04)	0.00	-0.67 (0.09)	0.46 (0.03)	0.00
SCAT total symptoms raw score	-3.30 (0.22)	0.94 (0.07)	0.00	-2.32 (0.15)	0.76 (0.05)	0.00
BESS change score	-0.23 (0.07)	0.12 (0.01)	0.00	0.10 (0.06)	0.09 (0.01)	0.00
BESS raw score	-1.70 (0.16)	0.12 (0.01)	0.00	-0.98 (0.14)	0.09 (0.01)	0.00

SE, standard error; LOC, loss of consciousness; PTA, post-traumatic amnesia; RGA, retrograde amnesia; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; Change score at a timepoint is computed as: raw score at timepoint - raw score at baseline

Table 2.3: Testing and training set estimates for performance measures of univariate models at <6h and 24-48h hours

	<6h			24-48h		
Variable	*Sensitivity*	*Specificity**	*AUC*	*Sensitivity*	*Specificity**	*AUC*
Age in years	0.05 (0.04)	0.92 (0.95)	0.41 (0.52)	1.00 (0.95)	0.00 (0.03)	0.54 (0.49)
Male Sex	0.00 (0.00)	1.00 (1.00)	0.51 (0.53)	1.00 (0.87)	0.00 (0.12)	0.48 (0.49)
Number of previous concussions	0.07 (0.06)	0.93 (0.94)	0.50 (0.52)	1.00 (0.75)	0.00 (0.21)	0.48 (0.48)
Report injury immediately?	0.56 (0.42)	0.68 (0.66)	0.62 (0.55)	1.00 (1.00)	0.00 (0.00)	0.51 (0.47)
Removed from play immediately?	0.51 (0.21)	1.00 (0.80)	0.53 (0.54)	1.00 (0.90)	0.00 (0.08)	0.52 (0.46)
LOC?	0.00 (0.02)	1.00 (0.99)	0.50 (0.50)	0.95 (0.97)	0.06 (0.03)	0.51 (0.50)
PTA?	0.07 (0.13)	0.86 (0.93)	0.47 (0.51)	1.00 (0.93)	0.00 (0.05)	0.49 (0.48)
RGA?	0.00 (0.00)	1.00 (1.00)	0.50 (0.49)	0.92 (0.94)	0.04 (0.07)	0.48 (0.51)
Number of hours of sleep last night?	0.00 (0.01)	1.00 (1.00)	0.45 (0.48)	0.60 (0.62)	0.53 (0.45)	0.59 (0.58)
SAC change score	0.54 (0.47)	0.79 (0.76)	0.67 (0.66)	0.66 (0.65)	0.57 (0.55)	0.65 (0.66)
SAC raw score	0.48 (0.48)	0.78 (0.82)	0.69 (0.69)	0.55 (0.62)	0.70 (0.63)	0.66 (0.67)
SCAT symptom severity change score	0.83 (0.84)	0.97 (0.98)	0.95 (0.95)	0.75 (0.80)	0.98 (0.97)	0.92 (0.91)
SCAT symptom severity raw score	0.92 (0.91)	0.97 (0.97)	0.97 (0.98)	0.82 (0.87)	0.96 (0.95)	0.96 (0.96)
SCAT total symptoms change score	0.84 (0.87)	0.94 (0.97)	0.95 (0.95)	0.79 (0.82)	0.97 (0.96)	0.93 (0.92)
SCAT total symptoms raw score	0.93 (0.92)	0.95 (0.96)	0.97 (0.98)	0.83 (0.86)	0.96 (0.95)	0.96 (0.96)
BESS change score	0.57 (0.56)	0.74 (0.80)	0.71 (0.72)	0.63 (0.66)	0.51 (0.59)	0.62 (0.67)
BESS raw score	0.57 (0.53)	0.73 (0.80)	0.67 (0.72)	0.55 (0.57)	0.66 (0.65)	0.66 (0.65)

AUC, area under the receiver operating characteristic curve; LOC, loss of consciousness; PTA, post-traumatic amnesia; RGA, retrograde amnesia; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; All values are listed as: testing set estimate (training set estimate); Change score at a timepoint is computed as: raw score at timepoint - raw score at baseline; *Specificity estimates do not include the baseline data

2.3.2 Multivariate Logistic Regression Analysis

Final variables in each full model are summarized in Table 2.4. These results address aims (1). Both full models contained sex, whether the injury was reported immediately, SAC change score, SCAT symptom severity change score, and SCAT total symptoms raw score. The <6h model contained BESS change score while the 24-48h model contained BESS raw score. Every variable in the full model was significant except for whether the athlete was removed from play immediately in the <6h model and the BESS raw score in the 24-48h model.

Table 2.4: Factors in multivariate logistic regression (full model) associated with acute concussion at <6h and 24-48h

Variable	<6 hours		24-48h	
	Coefficient (SE)	*p-value*	*Coefficient (SE)*	*p-value*
Intercept	-4.51 (0.53)	0.000	-2.67 (0.35)	0.00
Male Sex	1.02 (0.42)	0.01	0.56 (0.26)	0.03
Report injury immediately?	1.85 (0.44)	0.00	0.74 (0.24)	0.00
Removed from play immediately?	-0.64 (0.41)	0.12	NA	NA
SAC change score	-0.16 (0.08)	0.04	-0.13 (0.05)	0.01
SCAT symptom severity change score	0.13 (0.03)	0.00	0.07 (0.02)	0.00
SCAT total symptoms raw score	1.01 (0.09)	0.00	0.73 (0.06)	0.00
BESS change score	0.09 (0.03)	0.00	NA	NA
BESS raw score	NA	NA	-0.01 (0.02)	0.73

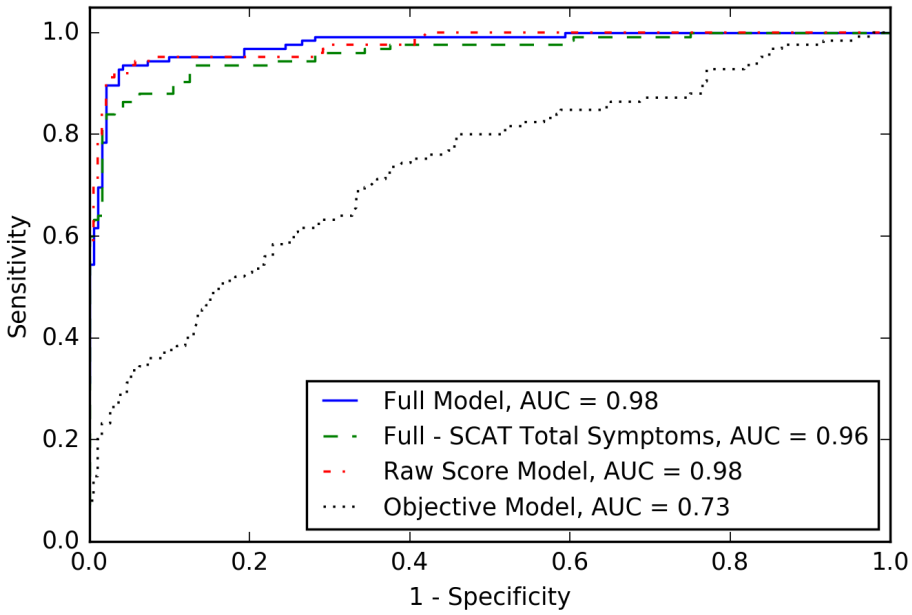
SE, standard error; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; Change score at a time point is computed as: raw score at time point - raw score at baseline; NA implies that the variable was not included in the model

The sensitivity, specificity, and AUC for all multivariate models are summarized in Table 2.5 and Receiver Operating Characteristic (ROC) curves for selected models are shown in Figure 2.1. These results address aims (2) and (3). Full models outperformed all univariate models. The coefficients for the limited models, raw score models, and objective models are presented in Table 2.6, Table 2.7, and Table 2.8, respectively. At <6h and 24-48h, removing SCAT total symptoms raw score from the full model resulted in the greatest decrease in performance. At <6h, removing whether the injury was reported immediately, whether the athlete was removed from play immediately, or the BESS change score from the full model does not reduce AUC significantly ($P>0.10$). Similarly, removing BESS raw score from the full model at 24-48h does not reduce AUC significantly. Full models achieved greater AUC than raw score models ($P<0.001$).

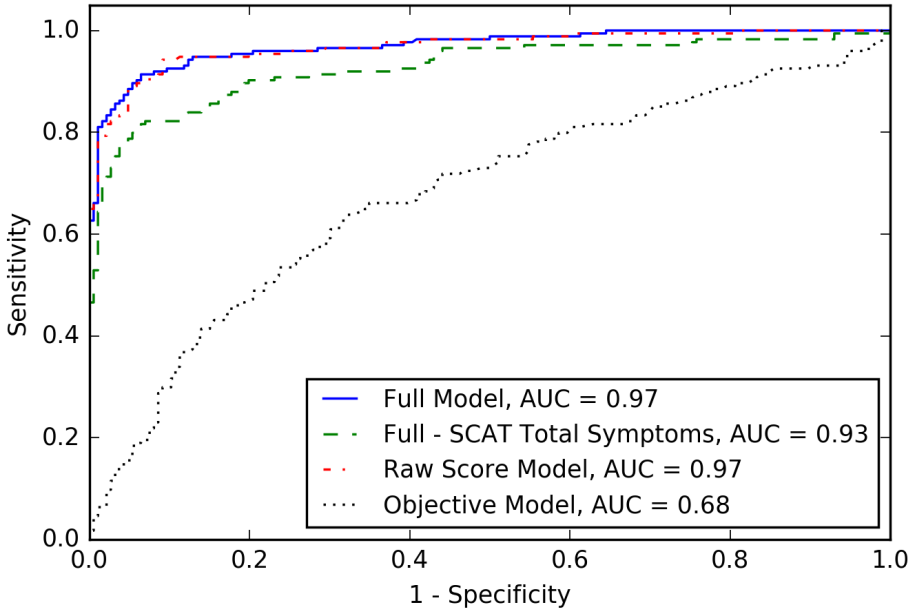
Table 2.5: Testing set and training set estimates for performance measures of multivariate models at <6h and 24-48h

		<6h	
Model	*Sensitivity*	*Specificity**	*AUC*
Full model	0.93 (0.94)	0.96 (0.97)	0.98 (0.99)
Limited models: Full (-removed variable)			
(-Male Sex)[1]	0.92 (0.94)	0.95 (0.97)	0.98 (0.99)
(-Report injury immediately?)	0.92 (0.93)	0.96 (0.97)	0.98 (0.99)
(-Removed from play immediately?)	0.92 (0.93)	0.96 (0.97)	0.98 (0.99)
(-SAC change score)1	0.93 (0.93)	0.96 (0.97)	0.98 (0.99)
(-SCAT symptom severity change score)[1]	0.91 (0.92)	0.96 (0.97)	0.97 (0.99)
(-SCAT total symptoms raw score)[1]	0.86 (0.88)	0.96 (0.97)	0.96 (0.95)
(-BESS change score)	0.93 (0.93)	0.95 (0.98)	0.98 (0.99)
Raw score model[1]	0.92 (0.93)	0.96 (0.97)	0.98 (0.98)
Objective model[1]	0.61 (0.60)	0.74 (0.81)	0.73 (0.76)
		24-48h	
Model	*Sensitivity*	*Specificity**	*AUC*
Full model	0.85 (0.88)	0.97 (0.96)	0.97 (0.97)
Limited models: Full (-removed variable)			
(-Male Sex)[2]	0.86 (0.88)	0.96 (0.95)	0.97 (0.96)
(-Report injury immediately?)[2]	0.86 (0.88)	0.97 (0.95)	0.97 (0.96)
(-SAC change score)	0.84 (0.89)	0.96 (0.95)	0.97 (0.96)
(-SCAT symptom severity change score)[2]	0.85 (0.88)	0.95 (0.95)	0.96 (0.96)
(-SCAT total symptoms raw score)[2]	0.75 (0.81)	0.96 (0.94)	0.93 (0.91)
(-BESS raw score)	0.84 (0.88)	0.97 (0.96)	0.97 (0.97)
Raw score model[2]	0.84 (0.87)	0.95 (0.95)	0.97 (0.96)
Objective model[2]	0.60 (0.66)	0.70 (0.68)	0.68 (0.72)

All values are reported as: testing set estimate without baselines (training set estimate); AUC, area under the receiver operating characteristic curve; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; Change score at a time point is computed as: raw score at time point - raw score at baseline; NA implies that the variable was not included in the model; [1]P<0.001 in bootstrap test for AUC, <6h; [2]P<0.001 in bootstrap test for AUC, 24-48h; *Specificity estimates do not include the baseline data.



(a) <6h



(b) 24-48h

Figure 2.1: Receiver operating characteristic curves for selected multivariate models based on validation against testing sets at <6h and 24–48 hours. AUC, area under the receiver operating characteristic curve; SCAT, Sport Concussion Assessment Tool

Table 2.6: Association in limited models between risk factors and concussion for <6 hours and 24-48 hours

(a) <6 hours

Full model - :	**Male Sex**	**Report injury immediately?**	**Removed from play immediately?**	**SAC change score**	**SCAT symptom severity change score**	**SCAT total symptoms raw score**	**BESS change score**
Variable	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*
Intercept	-3.76 (0.40)*	-3.84 (0.47)*	-4.68 (0.53)*	-4.55 (0.53)*	-5.15 (0.51)*	-1.45 (0.25)*	-4.52 (0.52)*
Male Sex	NA	1.08 (0.42)*	1.01 (0.42)*	0.99 (0.41)*	1.33 (0.39)*	0.01 (0.24)	1.11 (0.41)*
Report injury immediately?	1.86 (0.43)*	NA	1.57 (0.40)*	1.82 (0.44)*	1.90 (0.41)*	0.76 (0.25)*	1.75 (0.42)*
Removed from play immediately?	-0.60 (0.40)	0.14 (0.35)	NA	-0.54 (0.40)	-0.42 (0.38)	-0.04 (0.26)	-0.69 (0.39)
SAC change score	-0.15 (0.08)*	-0.15 (0.08)	-0.15 (0.08)	NA	-0.19 (0.07)*	-0.09 (0.05)	-0.18 (0.08)*
SCAT symptom severity change score	0.14 (0.03)*	0.13 (0.02)*	0.13 (0.03)*	0.14 (0.03)*	NA	0.26 (0.02)*	0.13 (0.03)*
SCAT total symptoms raw score	0.98 (0.09)*	0.94 (0.08)*	1.00 (0.09)*	1.00 (0.09)*	1.04 (0.08)*	NA	1.02 (0.09)*
BESS change score	0.10 (0.03)*	0.08 (0.03)*	0.09 (0.03)*	0.10 (0.03)*	0.07 (0.03)*	0.10 (0.02)*	NA

(b) 24-48 hours

Full model -	**Male Sex**	**Report injury immediately?**	**SAC change score**	**SCAT symptom severity change score**	**SCAT total symptoms raw score**	**BESS raw score**
Variable	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*	*Coefficient (SE)*
Intercept	-2.30 (0.30)*	-2.28 (0.32)*	-2.79 (0.35)*	-2.96 (0.34)*	-0.97 (0.23)*	-2.74 (0.29)*
Male Sex	NA	0.55 (0.26)*	0.59 (0.26)*	0.66 (0.26)*	-0.12 (0.18)	0.55 (0.26)*
Report injury immediately?	0.73 (0.24)*	NA	0.76 (0.24)*	0.80 (0.24)*	0.36 (0.17)*	0.75 (0.24)*
SAC change score	-0.14 (0.05)*	-0.14 (0.05)*	NA	-0.14 (0.05)*	-0.12 (0.04)*	-0.13 (0.05)*
SCAT symptom severity change score	0.08 (0.02)*	0.07 (0.02)*	0.07 (0.02)*	NA	0.20 (0.01)*	0.07 (0.02)*
SCAT total symptoms raw score	0.71 (0.05)*	0.71 (0.06)*	0.72 (0.06)*	0.80 (0.06)*	NA	0.72 (0.06)*
BESS raw score	0.00 (0.02)	-0.01 (0.02)	-0.01 (0.02)	-0.01 (0.02)	0.04 (0.01)*	NA

SE, standard error; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; Change score at a timepoint is computed as: raw score at timepoint - raw score at baseline; NA implies that the variable was not included in the model; *This variable had a p-value <0.05

Table 2.7: Raw score analog of factors in multivariate logistic regression (raw score model) associated with concussion diagnosis at <6 hours and 24-48 hours

	<6 hours		**24-48 hours**	
Variable	*Coefficient (SE)*	*p-value*	*Coefficient (SE)*	*p-value*
Intercept	-2.01 (2.44)	0.41	-1.49 (1.76)	0.40
Male Sex	1.45 (0.42)	0.00	0.75 (0.26)	0.00
Report injury immediately?	1.67 (0.41)	0.00	0.82 (0.24)	0.00
Removed from play immediately?	-0.38 (0.38)	0.31	NA	NA
SAC raw score	-0.13 (0.09)	0.14	-0.06 (0.06)	0.34
SCAT symptom severity raw score	0.33 (0.10)	0.00	0.16 (0.08)	0.05
SCAT total symptoms raw score	0.49 (0.17)	0.00	0.55 (0.13)	0.00
BESS raw score	0.03 (0.03)	0.30	-0.01 (0.02)	0.57

SE, standard error; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; NA implies that the variable was not included in the model

Table 2.8: Objective factors in multivariate logistic regression (objective model) associated with concussion diagnosis at <6 hours and 24-48 hours

	<6 hours		**24-48 hours**	
Variable	*Coefficient (SE)*	*p-value*	*Coefficient (SE)*	*p-value*
Intercept	-3.59 (1.12)	0.00	8.35 (1.36)	0.00
Age in years	0.09 (0.06)	0.10	-0.03 (0.05)	0.59
Report injury immediately?	0.49 (0.15)	0.00	NA	NA
SAC change score	-0.22 (0.03)	0.00	NA	NA
SAC raw score	NA	NA	-0.28 (0.03)	0.00
BESS change score	NA	NA	0.08 (0.01)	0.00
BESS raw score	0.12 (0.01)	0.00	NA	NA

SE, standard error; SAC, Standard Assessment Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; Change score at a timepoint is computed as: raw score at timepoint - raw score at baseline; NA implies that the variable was not included in the model

The objective models for <6h and 24-48h contained the SAC and BESS assessments along with age. Using objective models instead of full models can result in losses up to 0.34 in sensitivity, 0.28 in specificity, and 0.29 in AUC. This loss in AUC is significant (P<0.001).

Finally, all models were validated against baseline data. The best univariate models achieved specificities ranging from 0.67-0.71 across <6h and 24-48h timepoints. Multivariate models did not demonstrate improved performance.

2.4 Discussion

While logistic regression has been applied to other aspects of concussion management (Meehan et al., 2013; Sufrinko et al., 2017a; Sufrinko et al., 2017b), this analysis is among the first applications to acute concussion assessment. This methodology combines individual risk modifiers and standard assessments to detect the acute effects of concussion – providing a single measure to guide injury assessment. The variables in both full models can be obtained within the time constraints of athletics, suggesting their potential application in sideline concussion management. These models were trained and validated on a larger sample of concussed athletes (n=560 for <6h and n=733 for 24-48h) compared to similar studies (n=40 to 166) (Broglio, Macciocchi, and Ferrara, 2007; Chin et al., 2016; McCrea et al., 2005; Putukian et al., 2015; Register-Mihalik et al., 2013b; Resch et al., 2016).

The full models identify the effects of concussion more accurately than univariate models. This result supports previous studies, demonstrating that testing batteries provide more utility in acute concussion evaluation than any single assessment (Broglio, Macciocchi, and Ferrara, 2007; McCrea et al., 2005; Putukian et al., 2015; Register-Mihalik et al., 2013b; Resch et al., 2016). Both full models contained SAC, SCAT symptom assessments, and BESS. However, removing BESS from these models does not reduce AUC ($P<0.001$), suggesting that it provides little additional value beyond the SAC and SCAT symptom assessments. Conversely, removing SCAT total symptoms raw scores from each full model results in the greatest reduction in model performance, suggesting that symptoms better indicate acute concussion than neurological status and balance assessments. These results support previous studies which found that symptom assessments have higher sensitivity and specificity compared to neurocognitive and postural stability assessments (Chin et al., 2016; McCrea et al., 2005; Putukian et al., 2015; Register-Mihalik et al., 2013b; Resch et al., 2016).

These findings differ from Broglio, Macciocchi, and Ferrara, 2007, who found that neurocognitive assessments had higher sensitivity than symptom assessments. These differences may be attributed to methodology and sample size. Broglio, Macciocchi, and Ferrara, 2007

used significant change from baseline (1 SD) to indicate concussion whereas this study used the logistic regression's estimates. The choice of 1 SD may create classification thresholds having higher sensitivity but also higher false-positive rates. Conversely, logistic regression models are optimized to minimize prediction error, leading to more balanced classification thresholds. Furthermore, a neurological status examination (i.e., SAC) was used in this study whereas Broglio, Macciocchi, and Ferrara, 2007 used computer-based neurocognitive examinations (i.e., Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT), HeadMinder Concussion Resolution Index), and a pencil-and-paper neurocognitive battery. This study also used a 22-item SCAT symptom assessment compared to the 9-item symptom assessment used by Broglio, Macciocchi, and Ferrara, 2007, potentially explaining differences in sensitivity for symptom assessments. Finally, sensitivity estimates in this study are based on a much larger sample compared to Broglio, Macciocchi, and Ferrara, 2007 (n=75).

Unfortunately, relying on self-reported symptoms raises concern for symptom under-reporting, which may occur at rates up to 50% (Williamson and Goodman, 2006). To account for this possibility in clinical settings, objective models were created by removing self-reported symptoms. Ultimately, these models were outperformed by all other multivariate models and the univariate models for SCAT symptom assessments, emphasizing the need for objective clinical measures which can better assess acute concussion. However, these findings may apply only to acute concussions; other studies found that symptom presentation and severity decline days after the injury while impairments in cognitive function and postural stability remain (McCrea et al., 2013; McCrea et al., 2005). Differences in concussion presentation by timepoint motivated the separate analyses at <6h and 24-48h. Nonetheless, both full models were nearly identical so the time difference between <6h and 24-48h may be insufficient for detecting changes in concussion presentation.

Change scores for SAC and SCAT symptom severity appeared in both full models, suggesting their importance in a battery. Yet, BESS may be removed from full models with no significant loss in AUC. Full models have significantly greater AUC than raw score models, though this difference may not be practically significant. Univariate analysis showed similar performance between raw score and change score for the SAC and BESS. However, SCAT symptom assessment raw scores outperformed their change score counterparts. These results provide some quantitative support for the notion that neuropsychological assessments can be performed without baseline information (Randolph, 2011), as expressed in the most recent

consensus statement on concussion in sport (McCrory et al., 2017). Overall, these results mirror findings in similar studies. For instance, previous research has shown that incorporating baseline information leads to improved diagnostic accuracy on the SCAT battery, but these batteries can still perform at clinically acceptable levels when baseline information is unavailable (Chin et al., 2016; Putukian et al., 2015). Similar results were also found for other concussion assessment batteries, where raw scores and change scores typically agreed for acute concussion assessment (Echemendia et al., 2012; Schmidt et al., 2012). Clinically speaking, these results imply that acute concussion assessment can still be performed accurately if baseline information is unavailable. This point is echoed in the most recent consensus statement (McCrory et al., 2017), which states that "baseline testing may be useful, but is not necessary for interpreting post-injury scores." While our results provide evidence to support this statement at <6h and 24-48h, future studies should analyze the utility of baseline information in concussion assessment at timepoints beyond the acute stage, e.g., 1-2 weeks post-injury.

The analysis on age and number of previous concussions was largely inconclusive, as neither variable was significant in univariate analysis nor included in the full model. However, male sex was found to be significant in the full model ($P<0.05$) and removing this variable from the full model results in reduced AUC ($P<0.001$), suggesting the importance of considering sex differences in acute concussion assessment. The positive coefficient value associated with male sex suggests that, all else held equal, males have increased risk of acute concussion. Initially, this finding seems to contradict previous research which found female athletes to be at higher risk for concussion (Covassin, Buz Swanik, and Sachs, 2003; Dick, 2009; Gessel et al., 2007; Lincoln et al., 2011). However, a post-hoc analysis of sex differences in the SAC, SCAT symptom assessments, and BESS at <6h and 24-48h (see Table 2.9) shows that this result may instead suggest that males and females could achieve the same risk level even if a male athlete reports "better" performance on the assessments considered in the full model. For instance, the full model may consider a male athlete and female athlete to have the same acute concussion risk level even if the male reported fewer symptoms with less severity compared to the female. This interpretation supports previous findings which found female athletes to exhibit a greater symptom onset and cognitive decline compared to male athletes (Broshek et al., 2005; Chin et al., 2016; Covassin et al., 2012; Covassin, Schatz, and Swanik, 2007; Dick, 2009). Clinically, these results suggest that males may still be concussed

despite reporting lower symptom presentation and closer-to-normal neurocognitive deficits compared to females. This result also provides support for concussion assessment guidelines which are tailored by sex, as has been suggested by previous studies (Broshek et al., 2005; Covassin et al., 2012).

Table 2.9: Comparison of standard assessments for males and females at <6 hours and 24-48 hours

	<6 hours		**24-48 hours**	
Variable	*Male*	*Female*	*Male*	*Female*
n	361	199	440	293
SAC change score (SD)[1]	-1.01 (3.42)	-0.50 (2.67)	-0.49 (2.78)	-0.32 (2.31)
SAC raw score (SD)[1]	25.89 (3.13)	26.72 (2.40)	26.52 (2.54)	26.75 (2.21)
SCAT symptom severity change score (SD)[2]	22.45 (21.22)	25.28 (20.19)	18.25 (21.55)	21.43 (22.19)
SCAT symptom severity raw score (SD)[1,2]	27.37 (21.20)	31.35 (19.99)	23.37 (21.03)	27.84 (22.25)
SCAT total symptoms change score (SD)	8.01 (6.17)	8.18 (5.55)	7.42 (6.76)	7.57 (6.33)
SCAT total symptoms raw score (SD)[1,2]	10.46 (5.57)	11.68 (5.05)	10.01 (6.04)	11.20 (5.92)
BESS change score (SD)	3.95 (9.06)	3.17 (7.84)	1.73 (7.83)	0.97 (6.72)
BESS raw score (SD)[1,2]	17.01 (9.02)	15.31 (8.09)	14.83 (8.21)	13.52 (7.25)

SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; Change score at a timepoint is computed as: raw score at timepoint - raw score at baseline; [1]p-value <0.05 when comparing males and females at <6 hours; [2]p-value <0.05 when comparing males and females at 24-48 hours

Multivariate models classified normal performances from baseline less accurately than those from unrestricted RTP. Symptom under-reporting during the RTP protocol may explain these results. Table 2.1 shows that mean SCAT symptom scores are higher at baseline than at unrestricted RTP (P<0.01). Since the models were trained on unrestricted RTP data, concussion probability was inflated for athletes reporting symptoms at baseline. Additionally, Table 2.1 shows that mean SAC and BESS scores were "worse" at baseline than at unrestricted RTP (P<0.01). Learning effects, i.e., improved performance from repetition, might explain these findings (Barr and McCrea, 2001; Moreau, Langdon, and Buckley, 2014; Valovich McLeod et al., 2004).

2.5 Conclusion

This study demonstrates the value of multidimensional assessment models such as the SCAT5, which incorporates the standard assessments considered in this study (Echemendia et al., 2017). Logistic regression models were analyzed to determine which combinations of standard assessments best assess acute concussion. These models also provide a means to combine information from multiple concussion risk modifiers and standard assessments into a single measure which can be used to supplement clinical assessment decisions. Results suggest the importance of SCAT symptom assessments in acute concussion evaluation and support the use of assessment batteries over isolated assessments, though the BESS' value may be limited when SAC and SCAT symptom assessments are available. Additionally, change scores provide some clinical utility over raw scores, but acute concussion assessment can still be performed with sufficient accuracy when baseline information is unavailable. While further studies should generalize these models beyond NCAA athlete populations, this analysis can aid the future design of data-driven concussion assessment.

Chapter 3

Optimizing Components of the Sport Concussion Assessment Tool for Acute Concussion Assessment

3.1 Introduction

In Chapter 2, we developed models which quantified the value of multi-dimensional testing batteries for acute concussion assessment. While such multi-faceted approaches are supported by international guidelines (McCrory et al., 2017), the resulting batteries may inadvertently incorporate low information elements, increasing total administration time. Therefore, these batteries could be optimized by identifying critical subsets which maximizes diagnostic accuracy while eliminating low information elements, ultimately reducing the time required to administer that battery (Patricios et al., 2017).

Among the most widely used assessment batteries is the Sport Concussion Assessment Tool (SCAT) (Echemendia et al., 2017), which has been used to evaluate concussion in sideline, clinic, and hospital settings (Lebrun et al., 2013; Yengo-Kahn et al., 2016). The SCAT combines several assessments, including a 22-item graded symptom checklist, the Standard Assessment for Concussion (SAC), and a modified Balance Error Scoring System (mBESS). These components, initially chosen through consensus based on clinical experience and existing evidence, represent a robust combination of tests which are sensitive to concussion (Guskiewicz et al., 2013). Since then, several studies confirmed that the SCAT can identify

acute concussion with reasonable accuracy (Chin et al., 2016; Garcia et al., 2018; Resch et al., 2016), but cannot be completed appropriately in under ten minutes (Echemendia et al., 2017). Given the time-sensitive nature of athletics (Putukian et al., 2013) (e.g., time limits within the rules of competition), the SCAT could be improved by identifying a subset of components which could be administered more quickly than the current testing battery without sacrificing its ability to detect acute concussion.

Therefore, our study aims to identify subsets of the SCAT symptom checklist, SAC, and Balance Error Scoring System (BESS) which can accurately identify acute concussion when baseline data are and are not available.

3.2 Methodology

3.2.1 Study Design

We analyzed data from the NCAA-DoD CARE Consortium (see Section 1.1.1).

3.2.2 Test Methods

To develop models which can generalize well, we only excluded participants who did not complete a pre-season baseline assessment. Then, we analyzed the subset who experienced concussion, focusing on their assessments (where available) at baseline, <6h, 24-48h, and unrestricted RTP. We refer to assessments from the <6h and 24-48h timepoints as acute concussions and those assessments taken at baseline as normal performances.

We included demographic features and standard assessment scores from the SCAT symptom checklist, SAC, and BESS. While the SCAT contains additional assessments (e.g., Glasgow Coma Scale), the symptoms, SAC, and BESS are among the most useful for acute concussion assessment (Echemendia et al., 2017). For these assessments, we included scores from the time of assessment (i.e., raw score) and the difference between the scores obtained at baseline and at the time of assessment (i.e., change score). For <6h and 24-48h, we computed the change score as the post-injury raw score minus the score obtained at baseline. For baselines, we computed the change score as the raw score at baseline minus the score at unrestricted RTP. This approach mimics what we did in Chapter 2. Most variables were missing at less than 5% except for the BESS Firm Score (15.2% at <6h; 7.1% at 24-48h),

BESS Foam Score (24.5% at <6h; 10.2% at 24-48h), BESS Total Score (24.5% at <6h; 10.3% at 24-48h), and SCAT Trouble Falling Asleep (8.3% at <6h). Missing values for study variables were filled using multiple imputation with chained equations (Royston et al., 2009), which has been utilized in previous concussion research (Garcia et al., 2018; Garcia et al., 2019; McCrea et al., 2013; McCrea et al., 2003; McCrea et al., 2005).Imputation was performed using the statistical software R, Version 3.5.0 (R Foundation for Statistical Computing, Vienna, Austria).

We included the age, sex, and the number of previous concussions in our analyses. In previous studies, younger athletes, females, and those with greater concussion history were suggested to be at increased risk for concussion (Broshek et al., 2005; Covassin, Buz Swanik, and Sachs, 2003; Covassin et al., 2012; Covassin, Schatz, and Swanik, 2007; Gessel et al., 2007; Kutcher and Eckner, 2010).

The SCAT symptom checklist is a 22-symptom graded checklist for evaluating symptom presentation (Echemendia et al., 2017). Each symptom is rated on a scale of 0-6 based on severity, where greater numbers indicate greater severity. We included raw and change scores for each of the 22 symptoms, along with the total symptom severity score and the total number of symptoms reported. We also included whether physical activity worsens symptoms and whether mental activity worsens symptoms – both are yes/no questions.

The SAC is a brief neurocognitive assessment measuring impairments in orientation, immediate memory, concentration, and delayed recall (McCrea et al., 1998). Each domain except for immediate memory is scored from 0 to 5, with 5 indicating a perfect score. Immediate memory is scored from 0 to 15, with 15 indicating a perfect score. We included raw scores and change scores for orientation, immediate memory, concentration, and delayed recall components, along with their total score.

The BESS is a measure of postural control that counts the number of "errors" committed by an athlete across three stances. Although SCAT only includes an evaluation on a firm surface (i.e., the mBESS), the CARE protocol includes both firm and foam surfaces (Riemann, Guskiewicz, and Shields, 1999), allowing us to evaluate raw scores and change scores for both surfaces, along with the sum of these scores.

3.2.3 Statistical Analysis

We randomly divided our data into training (60%) and testing (40%) sets. The training set was used to develop our models while the testing set was used to validate their performance. Modeling variables were compared across training and testing sets, by timepoint, using two-sample non-parametric bootstrap tests (Efron and Tibshirani, 1993). Effect sizes were computed for significant differences using Cohen's d. The following analysis was replicated using 10-fold cross-validation to check the robustness of findings.

We used multivariate logistic regression to quantify the differences between acute concussions and normal performances. Our analysis yielded 65 total modeling variables. To determine which variables to include in our models, we applied Mixed Integer Programming (MIP) to optimize variable selection (Sato et al., 2016). Let our training data be represented by $\{(y_i, \mathbf{x}_i)\}_{i=1}^N$, where $y_i = 1$ if data point i is a concussion and $y_i = -1$ otherwise, $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of values for all $p = 65$ modeling variables associated with data point i, and N is the total number of data points in our training set. Our optimal variable selection MIP (VS-MIP) is given by

$$\begin{aligned}
\text{(VS-MIP)} \quad \min_{b,\mathbf{w},\mathbf{z}} \quad & \sum_{i=1}^{N} \mathcal{L}(y_i(\mathbf{w}^\top \mathbf{x}_i + b)) && (3.1)\\
\text{s.t.} \quad & z_j \leq M w_j && (3.2)\\
& z_j \geq -M w_j \text{ for all } j = 1, ..., p && (3.3)\\
& \sum_{j=1}^{p} z_j \leq k \text{ for all } j = 1, ..., p && (3.4)\\
& z_j \in \{0, 1\}, \text{ for all } j = 1, ..., p && (3.5)
\end{aligned}$$

where

- $\mathcal{L}(v) := 1 + e^{-v}$ is the log-loss function,
- the decision variable b is the logistic regression model's intercept term,
- the decision variables $\mathbf{w} \in \mathbb{R}^p$ represent logistic regression coefficients for all $p = 65$ modeling variables,

- the binary decision variable $z_j = 1$ if variable j is included in the model and $z_j = 0$ otherwise,
- the parameter k is the maximum size for the logistic regression model, and
- the parameter M is a large positive number.

In VS-MIP, the objective function (3.1) minimizes the log-loss of our logistic regression model based on the training data. Constraints (3.2)-(3.3) ensure that $w_j = 0$ if variable j is not included in the model (i.e., $z_j = 0$) while constraint (3.4) ensures that there are no more than k variables in the model. Finally, constraint (3.5) ensures that the variables z_j are binary. We remark that the objective function (3.1) is nonlinear, although convex. To facilitate the numerical solution of this MIP, we replace (3.1) with its piecewise linear approximation (see Section 3.2 in Sato et al., 2016) during our implementation.

Optimization theory has mathematically proven that traditional variable selection methods based on information criteria (e.g., forward-backward selection) or regularization (e.g., Lasso) are suboptimal compared to MIP (Bertsimas and King, 2016; Bertsimas, King, and Mazumder, 2016). MIP creates the logistic regression model which best fits the training data using at most k modeling variables. We set k=4, 8, 12, and 16 to determine which combination of 4, 8, 12, and 16 distinct variables best identify acute concussion. Variable Inflation Factors (VIFs) were computed for all models to assess multicollinearity.

To assess which variables best identify acute concussion when baseline information is and is not available, we performed two sets of analyses. First, we assumed that raw scores and change scores could both be included in the models (i.e., *Opt-k* models). Second, we assumed that only raw scores could be included (i.e., *Opt-RS-k* models). For example, Opt-4 refers to the optimized model with at most 4 variables and Opt-RS-8 refers to the optimized model using only raw scores with at most 8 variables. For comparison, we created a model composed of only the SCAT total symptom severity score, SCAT total number of symptoms, SAC total score, and BESS total score, i.e., the *Summary Scores* model. We also assessed a modified Summary Scores model which replaces the BESS total score with the BESS firm surface score.

We evaluated our models against the testing set. First, we computed the Brier score (BS), which ranges from 0 to 0.25 and measures a model's classification error. A BS of 0 indicates a perfect model while 0.25 indicates that a model is uninformative. We also evaluated

models using the receiver operating characteristic (ROC) curve by reporting sensitivity and specificity (at the threshold which maximizes their sum) and area under the ROC curve (AUC). Statistical analyses were performed using Python 3.6.5 (Python Software Foundation, Beaverton, Orgeon, USA).

3.3 Results

Study data are described in Table 3.1 and Table 3.2. Training (resp., testing) data consisted of 1337, 876, and 1473 (resp., 841, 580, and 921) assessments at baseline, <6h, and 24-48h, respectively. There were fewer assessments at baseline than 24-48h due to participants with same-season concussions. Between training and testing sets, participants did not differ significantly in age, sex, and previous concussions across timepoints. The only significant differences between training and testing sets were pressure in head ($P<0.05$, $d=0.11$) at <6h and BESS foam surface score ($P<0.05$, $d=0.09$) at 24-48h. Compared to baseline, the <6h and 24-48h assessments were significantly different for each component of the SCAT symptom checklist, SAC, and BESS except for SAC Concentration Score at <6h ($P=0.17$), trouble falling asleep at <6h ($P=0.33$), and BESS foam surface score at 24-48h ($P=0.05$) in the testing set.

3.3.1 Model Descriptions

The Opt-k models at <6h are described in Table 3.3. The Opt-4 model contained change scores for SAC Concentration and headache, along with raw scores for dizziness and don't feel right ($P<0.001$ for all). The Opt-8 model added change scores for SAC Delayed Recall and sensitivity to noise along with whether symptoms get worse with mental activity and physical activity to the Opt-4 model. Change scores for SAC Delayed Recall ($P=0.78$) and sensitivity to noise ($P=0.11$) were not significant. For Opt-12, change scores for pressure in head and sensitivity to light, along with the raw score for blurred vision were added to the Opt-8 model. The change score for sensitivity to light was not significant ($P=0.68$). Opt-16 was identical to Opt-12.

Table 3.1: Characteristics of study data by timepoint

	Training						Testing					
	Baseline		<6h		24-48h		Baseline		<6h		24-48h	
	mean	*SD*	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*
n	1337		876		1473		841		580		921	
Height, m	1.79	0.11	1.79	0.11	1.79	0.12	1.78	0.12	1.79	0.11	1.78	0.11
Weight, kg	81.96	21.08	83.96**	22.08	82.9	21.88	82.33	22.07	84.84**	22.09	82.25	21.93
Age, years	19.1	1.3	19.2	1.3	19.1	1.3	19.1	1.3	19.2**	1.3	19.1	1.2
Male Sex, % yes	58.0%		63.5%*		58.7%		57.3%		62.2%		59.0%	
Previous concussions	0.6	0.9	0.6	0.9	0.6	0.9	0.6	0.8	0.7	0.9	0.6	0.9
SCAT Symptom Raw Scores												
Balance Problems	0.1	0.5	0.9*	1.3	0.6*	1.1	0.1	0.4	0.8*	1.3	0.6*	1.1
Blurred Vision	0.1	0.3	0.9*	1.4	0.6*	1.1	0.1	0.4	0.8*	1.3	0.7*	1.1
Difficulty Concentrating	0.3	0.8	1.7*	1.7	1.6*	1.6	0.4	0.9	1.6*	1.7	1.6*	1.6
Difficulty Remembering	0.2	0.7	1.0*	1.5	0.9*	1.3	0.2	0.8	0.9*	1.4	0.8*	1.3
Dizziness	0.1	0.5	1.6*	1.6	1.1*	1.4	0.1	0.5	1.5*	1.6	1.1*	1.4
Don't Feel Right	0.2	0.7	2.4*	1.8	2.0*	1.7	0.2	0.6	2.3*	1.7	2.0*	1.7
Trouble Falling Asleep	0.3	0.9	0.5*	1.2	1.0*	1.6	0.3	0.9	0.4	1.0	1.0*	1.6
Drowsiness	0.5	1.2	1.5*	1.7	1.4*	1.6	0.5	1.1	1.4*	1.6	1.4*	1.5
More Emotional	0.2	0.7	0.7*	1.3	0.7*	1.3	0.2	0.8	0.7*	1.4	0.7*	1.2
Fatigue or Low Energy	0.9	1.3	1.7*	1.8	1.7*	1.6	0.9	1.3	1.6*	1.7	1.8*	1.7
Feeling Slowed Down	0.3	0.8	2.0*	1.8	1.8*	1.6	0.3	0.8	1.9*	1.7	1.8*	1.6
Feeling Like in a Fog	0.2	0.7	1.9*	1.8	1.6*	1.6	0.2	0.6	1.8*	1.7	1.6*	1.6
Headache	0.3	0.8	2.8*	1.6	2.4*	1.5	0.3	0.8	2.7*	1.5	2.4*	1.5
Irritability	0.2	0.8	0.7*	1.3	0.8*	1.3	0.2	0.8	0.6*	1.2	0.7*	1.3
Nausea	0.1	0.4	0.8*	1.3	0.7*	1.2	0.1	0.4	0.9*	1.4	0.7*	1.2
Neck Pain	0.3	0.7	1.1*	1.5	1.2*	1.5	0.3	0.8	1.0*	1.5	1.2*	1.6
Nervous or Anxious	0.3	0.9	0.6*	1.2	0.5*	1.1	0.4	1.0	0.6*	1.3	0.5**	1.1
Pressure in Head	0.3	0.7	2.5*	1.6	2.1*	1.6	0.3	0.7	2.3*‡	1.6	2.1*	1.6
Sadness	0.2	0.7	0.6*	1.2	0.5*	1.1	0.2	0.8	0.5*	1.2	0.5*	1.1
Sensitivity to Light	0.1	0.5	1.3*	1.6	1.4*	1.5	0.1	0.5	1.3*	1.6	1.4*	1.4
Sensitivity to Noise	0.1	0.3	1.0*	1.4	1.2*	1.4	0.1	0.3	1.0*	1.4	1.1*	1.3
Total Severity Score	5.4	9.5	28.9*	21.5	26.3*	21.7	5.4	9.3	27.4*	20.5	26.2*	20.7
Total Number of Symptoms	2.9	4.0	10.9*	5.6	10.9*	5.9	2.9	3.9	10.6*	5.4	10.9*	6.0
Symptoms Worsen with Physical Activity, % yes	2.8%		59.4%*		59.4%*		1.9%		60.0%*		60.4%*	
Symptoms Worsen with Mental Activity, % yes	2.6%		43.8%*		53.3%*		2.1%		42.2%*		53.7%*	
SAC Raw Scores												
Concentration Score	3.8	1.2	3.6*	1.2	3.9**	1.2	3.7	1.2	3.6	1.2	3.8**	1.2
Delayed Score	3.9	1.2	3.4*	1.4	3.4*	1.3	3.9	1.2	3.4*	1.4	3.5*	1.3
Immediate Memory Score	14.7	0.7	14.3*	1.4	14.4*	1.1	14.7	0.7	14.3*	1.6	14.4*	1.3
Orientation Score	4.9	0.3	4.7*	0.6	4.8*	0.5	4.9	0.4	4.7*	0.7	4.8*	0.5
Total Score	27.3	2.0	26*	3.0	26.5*	2.6	27.2	2.0	26*	3.1	26.5*	2.8
BESS Raw Scores												
Firm Surface Score	3.4	3.2	5.3*	4.4	4.7*	4.2	3.4	3.2	5.5*	4.6	4.7*	4.0
Foam Surface Score	10.2	4.4	11.4*	5.5	10.7*	5.2	9.8	4.4	11.7*	5.8	10.3‡	4.8
Total Score	13.6	6.4	16.7*	8.2	15.4*	8.2	13.2	6.3	17.2*	8.6	14.9*	7.5

n, number of data points; SD, standard deviation; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; *P<0.01 **P<0.05 compared to baseline; †p-value<0.01 ‡p-value<0.05 compared to training data at the same timepoint; Effect sizes (magnitude) for significant differences between training and testing sets: Pressure in Head (d=0.11 at <6h); BESS Foam Surface Score (d=0.09 at 24-48h)

Table 3.2: Change scores for the SCAT symptom checklist, SAC, and BESS by time-point

	Training						Testing					
	Baseline		<6h		24-48h		Baseline		<6h		24-48h	
	mean	*SD*	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*
n	1337		876		1473		841		580		921	
SCAT Symptom Checklist												
Balance Problems	0.1	0.5	0.8*	1.4	0.5*	1.2	0.1	0.4	0.7*	1.4	0.5*	1.1
Blurred Vision	0.1	0.3	0.9*	1.4	0.6*	1.2	0.1	0.4	0.7*	1.3	0.6*	1.1
Difficulty Concentrating	0.3	0.9	1.4*	1.9	1.2*	1.9	0.3	0.9	1.3*	1.9	1.3*	1.8
Difficulty Remembering	0.2	0.7	0.8*	1.7	0.6*	1.5	0.2	0.8	0.7*	1.6	0.5*	1.5
Dizziness	0.1	0.4	1.5*	1.6	1*	1.4	0.1	0.5	1.5*	1.6	1.0*	1.4
Don't Feel Right	0.2	0.7	2.2*	1.9	1.8*	1.8	0.2	0.6	2.2*	1.9	1.8*	1.8
Trouble Falling Asleep	0.3	0.9	0.1**	1.5	0.7*	1.8	0.3	1.0	0.0*	1.5	0.7*	1.8
Drowsiness	0.5	1.2	1.0*	2.0	0.9*	1.9	0.5	1.1	0.9*	2.0	0.9*	1.9
More Emotional	0.2	0.7	0.5*	1.5	0.5*	1.5	0.2	0.8	0.5*	1.5	0.5*	1.3
Fatigue or Low Energy	0.8	1.4	0.9	2.3	0.9	2.1	0.8	1.4	0.7	2.2	0.9	2.0
Feeling Slowed Down	0.3	0.8	1.8*	1.9	1.5*	1.8	0.2	0.8	1.7*	1.9	1.5*	1.7
Feeling Like in a Fog	0.2	0.7	1.7*	1.8	1.4*	1.7	0.2	0.6	1.6*	1.7	1.4*	1.7
Headache	0.3	0.9	2.5*	1.6	2.1*	1.7	0.3	0.8	2.5*	1.7	2.1*	1.6
Irritability	0.2	0.8	0.5*	1.4	0.5*	1.5	0.2	0.8	0.4**	1.5	0.5*	1.4
Nausea	0.1	0.4	0.7*	1.4	0.6*	1.3	0.1	0.4	0.8*	1.4	0.6*	1.2
Neck Pain	0.2	0.8	0.8*	1.6	0.9*	1.6	0.2	0.9	0.7*	1.6	0.9*	1.7
Nervous or Anxious	0.3	0.9	0.3	1.3	0.2*	1.4	0.4	0.9	0.3	1.5	0.1*	1.3
Pressure in Head	0.2	0.7	2.2*	1.7	1.8*	1.7	0.2	0.7	2.1*	1.7	1.9*	1.6
Sadness	0.2	0.7	0.4*	1.4	0.3**	1.3	0.2	0.8	0.3**	1.4	0.3**	1.1
Sensitivity to Light	0.1	0.5	1.2*	1.6	1.3*	1.6	0.1	0.5	1.2*	1.6	1.3*	1.5
Sensitivity to Noise	0.1	0.4	0.9*	1.4	1.1*	1.5	0.1	0.3	0.9*	1.4	1*	1.4
Total Severity Score	4.8	9.4	23.9*	22.2	20.8*	22.9	4.8	9.3	22.3*	22.1	20.9*	21.2
Total Number of Symptoms	2.5	4.0	8.2*	6.6	8.0*	6.8	2.5	3.8	7.9*	6.5	8.1*	6.5
SAC												
Concentration Score	-0.6	1.2	-0.2*	1.4	0.1*	1.2	-0.6	1.1	-0.1*	1.3	0.1*	1.3
Delayed Score	-0.3	1.3	-0.5*	1.6	-0.5*	1.6	-0.3	1.3	-0.5*	1.7	-0.4**	1.5
Immediate Memory Score	0.0	0.9	-0.4*	1.5	-0.2*	1.2	0.0	0.9	-0.4*	1.7	-0.3*	1.5
Orientation Score	0.0	0.5	-0.2*	0.7	-0.1*	0.6	0.0	0.5	-0.2*	0.7	-0.1**	0.6
Total Score	-0.8	2.0	-1.2*	3.1	-0.7	2.6	-0.9	2.0	-1.3**	3.4	-0.7	2.8
BESS												
Firm Surface Score	0.8	3.7	2.1*	4.4	1.3*	4.3	0.8	3.7	2.2*	4.5	1.4*	4.0
Foam Surface Score	1.9	4.8	1.3**	6.1	0.5*	5.4	1.9	4.8	1.3	6.1	0.0*‡	5.3
Total Score	2.7	6.6	3.4**	8.5	1.8*	8.2	2.7	6.6	3.5	8.5	1.4*	7.6

n, number of data points; Change score <6h and 24-48h computed as: raw score at time point - raw score at baseline; Change score at baseline computed as: raw score at unrestricted RTP – raw score at baseline (unrestricted RTP data not shown); SD, standard deviation; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; *P<0.01 **P<0.05 compared to baseline; †P<0.01 ‡P<0.05 compared to training data at the same timepoint; Effect sizes (magnitude) for significant differences between training and testing sets: BESS Foam Surface Score (d=0.08 at 24-48h)

Table 3.3: Model variables, coefficient values, and performance measures for Opt-k models at <6h

Model	*k=4*		*k=8*		*k=12,16*	
Model Variables and Coefficients						
Variable	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*
Intercept	-2.28 (0.10)*	(-2.47, -2.09)	-2.68 (0.11)*	(-2.91, -2.46)	-2.71 (0.12)*	(-2.96, -2.51)
SAC Concentration CS	0.38 (0.06)*	(0.26, 0.50)	0.40 (0.07)*	(0.27, 0.53)	0.40 (0.07)*	(0.26, 0.52)
SCAT Headache CS	0.85 (0.06)*	(0.74, 0.96)	0.77 (0.06)*	(0.65, 0.90)	0.67 (0.08)*	(0.53, 0.83)
SCAT Dizziness RS	0.72 (0.10)*	(0.52, 0.93)	0.53 (0.12)*	(0.31, 0.76)	0.44 (0.12)*	(0.15, 0.62)
SCAT Don't Feel Right RS	0.77 (0.07)*	(0.63, 0.91)	0.58 (0.08)*	(0.43, 0.74)	0.54 (0.08)*	(0.38, 0.69)
SAC Delayed Recall CS			-0.02 (0.06)	(-0.13, 0.09)	-0.02 (0.06)	(-0.13, 0.09)
SCAT Sensitivity to Noise CS			0.22 (0.14)	(-0.05, 0.49)	0.21 (0.16)	(-0.08, 0.53)
SCAT Symptoms Get Worse with Mental Activity			1.75 (0.26)*	(1.23, 2.27)	1.69 (0.27)*	(1.15, 2.20)
SCAT Symptoms Get Worse with Physical Activity			2.36 (0.23)*	(1.90, 2.82)	2.35 (0.24)*	(1.89, 2.82)
SCAT Pressure in Head CS					0.20 (0.09)***	(0.01, 0.35)
SCAT Blurred Vision RS					0.29 (0.13)***	(0.18, 0.76)
SCAT Sensitivity to Light CS					-0.05 (0.12)	(-0.28, 0.18)
Performance Measures						
Brier Score	0.089		0.072		0.072	
AUC	0.95		0.96		0.96	
Sensitivity[1]	0.87		0.87		0.87	
Specificity[1]	0.90		0.94		0.93	

CS, change score; RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; SE, Standard error; CI, Confidence Interval; AUC, Area under the receiver operating characteristic curve; [1]Sensitivity and specificity values reported in table correspond to the threshold which maximizes the sum of sensitivity and specificity; *P<0.001 **P<0.01 ***P<0.05 coefficient is significantly different from 0

The Opt-k models at 24-48h are described in Table 3.4. The Opt-4 model included the change score for SAC Concentration, the raw score for headache, and whether symptoms get worse with mental and physical activity (P<0.001 for all). For Opt-8, the change score for SAC Delayed Recall and the raw scores for don't feel right, pressure in head, and sensitivity to noise were added to all the variables in Opt-4. Only the change score for SAC Delayed Recall (P=0.76) was not significant. The Opt-12 model builds on Opt-8 by adding raw scores for nausea, dizziness, and sensitivity to light. Raw scores for dizziness (P=0.36) and sensitivity to light (P=0.08) were not significant. The Opt-16 and Opt-12 models were identical. For the Summary Score model, all variables were significant (P<0.05 for all).

Table 3.4: Model variables, coefficient values, and performance measures for Opt-k models at 24-48h

Model	*k=4*		*k=8*		*k=12,16*	
Model Variables and Coefficients						
Variable	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*
Intercept	-1.84 (0.08)*	(-2.00, -1.68)	-1.98 (0.09)*	(-2.15, -1.81)	-2.01 (0.09)*	(-2.19, -1.84)
SAC Concentration CS	0.64 (0.06)*	(0.53, 0.75)	0.65 (0.06)*	(0.54, 0.77)	0.65 (0.06)*	(0.54, 0.77)
SCAT Headache RS	1.05 (0.06)*	(0.94, 1.17)	0.68 (0.08)*	(0.53, 0.83)	0.70 (0.08)*	(0.54, 0.85)
SCAT Symptoms Get Worse with Mental Activity	2.13 (0.22)*	(1.70, 2.56)	1.85 (0.23)*	(1.40, 2.29)	1.82 (0.23)*	(1.38, 2.27)
SCAT Symptoms Get Worse with Physical Activity	2.41 (0.20)*	(2.01, 2.80)	2.10 (0.21)*	(1.68, 2.51)	2.17 (0.22)*	(1.74, 2.59)
SAC Delayed Recall CS			0.01 (0.05)	(-0.08, 0.10)	0.01 (0.05)	(-0.08, 0.10)
SCAT Don't Feel Right RS			0.46 (0.08)*	(0.31, 0.60)	0.49 (0.08)*	(0.34, 0.65)
SCAT Pressure in Head RS			0.23 (0.09)**	(0.06, 0.40)	0.21 (0.09)**	(0.04, 0.39)
SCAT Sensitivity to Noise RS			0.55 (0.13)*	(0.29, 0.81)	0.54 (0.14)*	(0.27, 0.82)
SCAT Nausea RS					-0.56 (0.13)*	(-0.81, -0.31)
SCAT Dizziness RS					0.11 (0.12)	(-0.13, 0.35)
SCAT Sensitivity to Light RS					0.20 (0.11)	(-0.02, 0.41)
Performance Measures						
Brier Score	0.093		0.086		0.085	
AUC	0.94		0.95		0.95	
Sensitivity[1]	0.85		0.83		0.83	
Specificity[1]	0.90		0.95		0.95	

CS, change score; RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; SE, Standard error; CI, Confidence Interval; AUC, Area under the receiver operating characteristic curve; [1]Sensitivity and specificity values reported in table correspond to the threshold which maximizes the sum of sensitivity and specificity; *P<0.001 **P<0.01 ***P<0.05 coefficient is significantly different from 0

The Summary Scores models at <6h and 24-48h are described in Table 3.5. All variables were significant (P<0.05 for all) for both timepoints. The Opt-RS-k models at <6h and 24-48h are described in Table 3.6 and Table 3.7, respectively. All variables within Opt-k and Opt-RS-k models at <6h and 24-48h had low-moderate multicollinearity (see Table 3.8).

3.3.2 Model Performance

Performance measures for the Opt-k models and the Summary Score model at <6h are summarized in Table 3.3 and Table 3.5, respectively. Their ROC curves are plotted in Figure 3.1a. The Opt-k models (BS=0.072-0.089) better identified concussion in the testing

Table 3.5: Model variables, coefficient values, and performance measures for Summary Scores models at <6h and 24-48h

Timepoint	*<6h*		*24-48h*	
Model variables and coefficients				
Variables	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*
Intercept	1.25 (0.72)	(-0.16, 2.67)	-0.25 (0.66)	(-1.54, 1.04)
SAC Total Score RS	-0.14 (0.03)*	(-0.19, -0.09)	-0.05 (0.02)***	(-0.10, -0.01)
BESS Total Score RS	0.03 (0.01)*	(0.01, 0.04)	-0.01 (0.01)	(-0.02, 0.01)
SCAT Total Symptom Severity RS	0.02 (0.01)***	(0.01, 0.04)	-0.03 (0.01)*	(-0.04, -0.01)
SCAT Total Number of Symptoms RS	0.24 (0.03)*	(0.18, 0.29)	0.37 (0.02)*	(0.32, 0.41)
Performance Measures				
Brier Score	0.14		0.15	
AUC	0.89		0.87	
Sensitivity[1]	0.79		0.74	
Specificity[1]	0.85		0.86	

RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; SE, Standard error; CI, Confidence Interval; AUC, Area under the receiver operating characteristic curve; [1]Sensitivity and specificity values reported in table correspond to the threshold which maximizes the sum of sensitivity and specificity; *P<0.001 **P<0.01 ***P<0.05 coefficient is significantly different from 0

data than the Summary Scores model (BS=0.14). This result is supported by our analysis of the ROC curve. Specifically, each Opt-k model (AUC=0.95-0.96) achieved a greater AUC than the Summary Score model (AUC=0.89). Furthermore, Figure 1(a) shows that the ROC curve for the Summary Scores model is "inside" the ROC curve for all Opt-k models, indicating that the Opt-k models achieve greater sensitivity and specificity for every possible cutoff used to designate acute concussion versus normal performance. The Opt-RS-k models also achieved improved BS (BS=0.082-0.087) and AUC (AUC=0.93-0.95) over the Summary Scores model (see Figure 3.2).

We summarize performance measures for the Opt-k and Summary Scores models at 24-48h in Table 3.4 and plot their ROC curves in Figure 3.1b. The Opt-k models achieved a lower BS (BS=0.085-0.093) and greater AUC (AUC=0.94-0.95) than the Summary Scores model (BS=0.15, AUC=0.87). Figure 1(b) also shows that the ROC curve for the Summary Scores model is dominated by the ROC curve for each Opt-k model, suggesting that the Opt-k models can better identify acute concussion at 24-48h than the Summary Scores model. The Opt-RS-k models also achieved lower BS (BS=0.095-0.099) and greater AUC (AUC=0.92-0.93) than the Summary Scores model (see Figure 3.2).

A modified Summary Scores model (BS=0.14, AUC=0.89 at <6h; BS=0.15, AUC=0.87 at

Table 3.6: Model variables, coefficient values, and performance measures for Opt-RS-k models at <6h

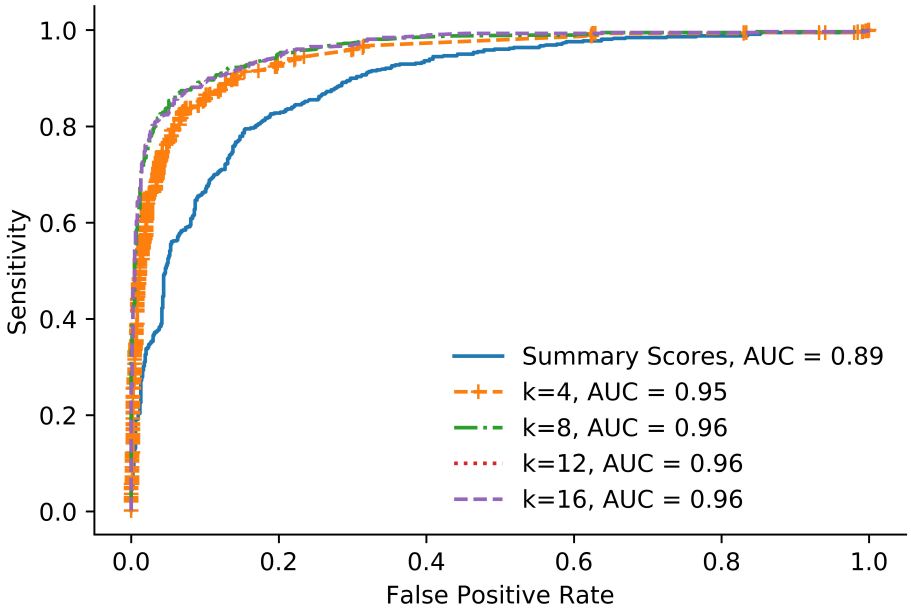
Model	*k=4*		*k=8, 12, 16*	
Model Variables and Coefficients				
Variable	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*
Intercept	-2.31 (0.09)*	(-2.49, -2.11)	-2.69 (0.11)*	(-2.90, -2.48)
SCAT Dizziness RS	0.82 (0.11)*	(0.61, 1.02)	0.44 (0.11)*	(0.21, 0.66)
SCAT Don't Feel Right RS	0.84 (0.07)*	(0.70, 0.98)	0.55 (0.08)*	(0.40, 0.70)
SCAT Symptoms Worsen with Mental Activity	1.65 (0.24)*	(1.18, 2.11)	1.52 (0.25)*	(1.02, 2.01)
SCAT Symptoms Worsen with Physical Activity	2.61 (0.21)*	(2.20, 3.01)	2.31 (0.22)*	(1.87, 2.75)
SCAT Blurred Vision RS			0.48 (0.14)*	(0.21, 0.75)
SCAT Pressure in Head RS			0.75 (0.08)*	(0.61, 0.90)
SCAT Sensitivity to Noise			0.20 (0.14)	(-0.07, 0.47)
Performance Measures				
Brier Score	0.087		0.082	
AUC	0.93		0.95	
Sensitivity[1]	0.83		0.88	
Specificity[1]	0.93		0.90	

RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; SE, Standard error; CI, Confidence Interval; AUC, Area under the receiver operating characteristic curve; [1]Sensitivity and specificity values reported in table correspond to the threshold which maximizes the sum of sensitivity and specificity; *P<0.001 **P<0.01 ***P<0.05 coefficient is significantly different from 0

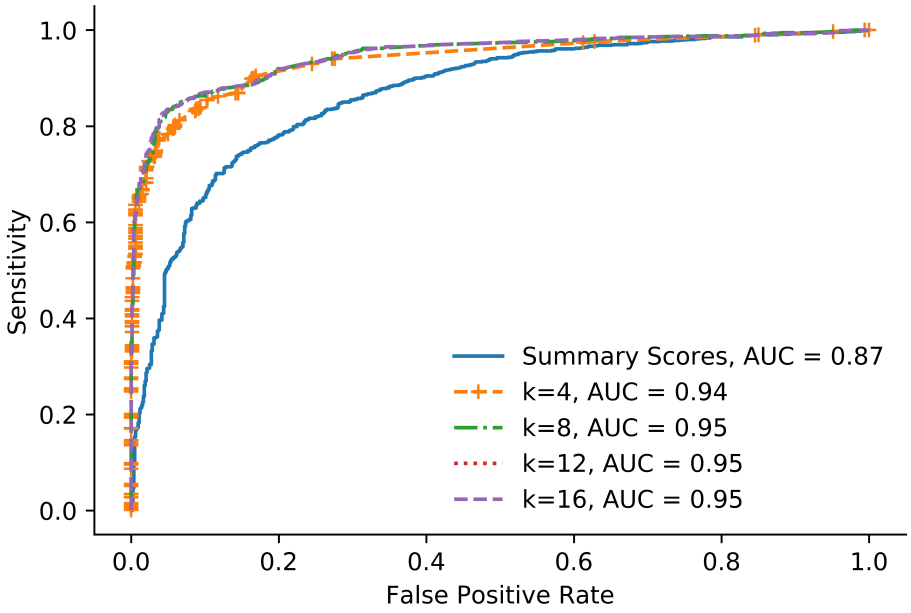
Table 3.7: Model variables, coefficient values, and performance measures for Opt-RS-k models at 24-48h

Model	*k=4*		*k=8, 12, 16*	
Model Variables and Coefficients				
Variable	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*
Intercept	-1.65 (0.09)*	(-1.79, -1.52)	-2.02 (0.08)*	(-2.18, -1.86)
SCAT Don't Feel Right RS	0.79 (0.07)*	(0.66, 0.92)	0.51 (0.08)*	(0.36, 0.65)
SCAT Symptoms Get Worse with Mental Activity	2.09 (0.21)*	(1.68, 2.49)	1.83 (0.21)*	(1.42, 2.25)
SCAT Symptoms Get Worse with Physical Activity	2.42 (0.19)*	(2.04, 2.79)	2.13 (0.20)*	(1.73, 2.53)
SCAT Sensitivity to Noise	1.04 (0.13)*	(0.78, 1.29)	0.62 (0.13)*	(0.35, 0.88)
SCAT Dizziness RS			-0.01 (0.11)	(-0.23, 0.21)
SCAT Headache RS			0.76 (0.06)*	(0.64, 0.88)
Performance Measures				
Brier Score	0.099		0.095	
AUC	0.92		0.93	
Sensitivity[1]	0.81		0.84	
Specificity[1]	0.94		0.92	

RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; SE, Standard error; CI, Confidence Interval; AUC, Area under the receiver operating characteristic curve; [1]Sensitivity and specificity values reported in table correspond to the threshold which maximizes the sum of sensitivity and specificity; *P<0.001 **P<0.01 ***P<0.05 coefficient is significantly different from 0

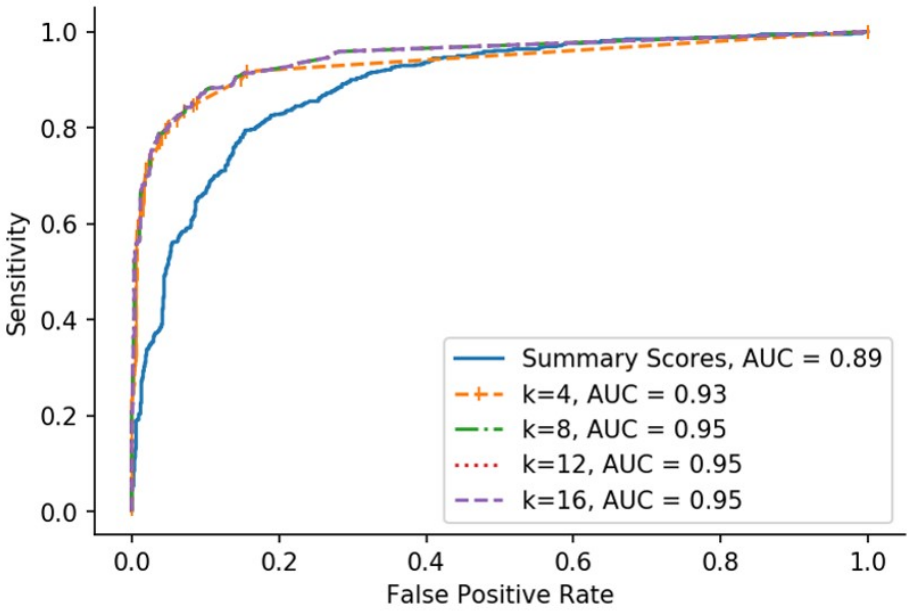


(a) <6 hours

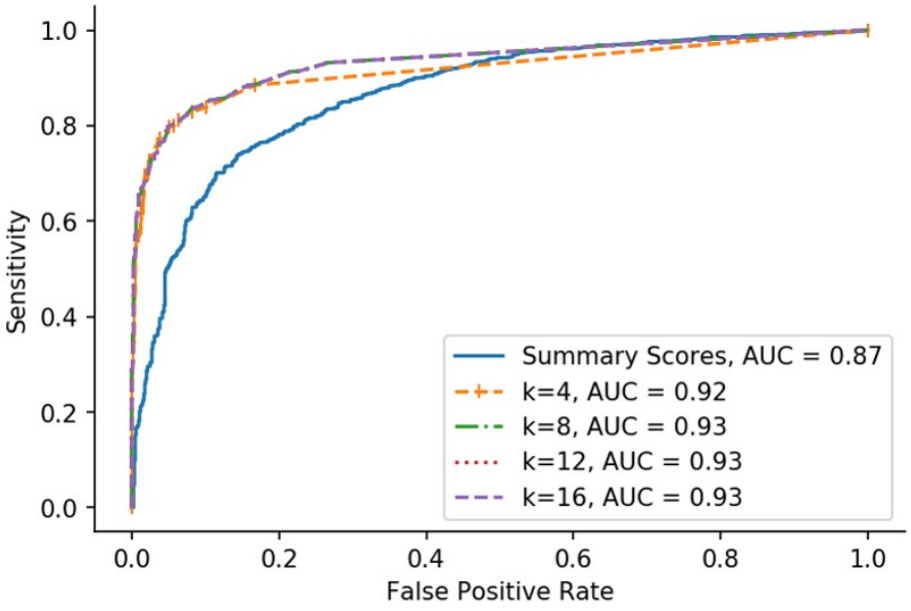


(b) 24-48 hours

Figure 3.1: Receiver operating characteristic curves for Opt-k models and Summary Scores models at <6h and 24-48h. AUC, area under the receiver operating characteristic curve



(a) <6 hours



(b) 24-48 hours

Figure 3.2: Receiver operating characteristic curves for Opt-RS-k models and Summary Scores models at <6h and 24-48h. AUC, area under the receiver operating characteristic curve

Table 3.8: Variable Inflation Factors for Opt-k Models at <6h and 24-48h

Opt-k Models at <6h	*k=4*	*k=8*	*k=12,16*	**Opt-k Models at 24-48h**	*k=4*	*k=8*	*k=12,16*
Variable	*VIF*			*Variable*	*VIF*		
Intercept	1.67	1.75	1.75	Intercept	1.92	1.96	1.97
SAC Concentration CS	1.00	1.01	1.01	SAC Concentration CS	1.01	1.02	1.02
SCAT Headache CS	1.77	1.92	2.98	SCAT Headache RS	1.63	3.57	3.69
SCAT Dizziness RS	2.15	2.27	2.60	SCAT Symptoms Worsen with Mental Activity	1.67	1.73	1.73
SCAT Don't Feel Right RS	2.45	2.71	2.86	SCAT Symptoms Worsen with Physical Activity	1.77	1.83	1.85
SAC Delayed Recall CS		1.02	1.02	SAC Delayed Recall CS		1.02	1.02
SCAT Sensitivity to Noise CS		1.60	2.03	SCAT Don't Feel Right RS		2.49	3.02
SCAT Symptoms Worsen with Mental Activity		1.56	1.59	SCAT Pressure in Head RS		3.74	3.78
SCAT Symptoms Worsen with Physical Activity		1.81	1.83	SCAT Sensitivity to Noise RS		1.94	2.55
SCAT Pressure in Head CS			3.25	SCAT Nausea RS			1.66
SCAT Blurred Vision RS			1.75	SCAT Dizziness RS			2.32
SCAT Sensitivity to Light CS			2.22	SCAT Sensitivity to Light RS			2.88

VIF, variable inflation factor; CS, change score; RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool

24-48h) which resembles the SCAT performs similarly to the Summary Scores model. Results reported in this section are consistent with results obtained using 10-fold cross-validation on the entire data sample (see Appendix 3.A).

3.4 Discussion

Previous studies have shown that the SCAT has tremendous clinical utility (Chin et al., 2016; Downey, Hutchison, and Comper, 2018; Garcia et al., 2018; Putukian et al., 2015). Our results suggest that the SCAT's symptom, SAC, and BESS composite scores may inadvertently incorporate components which add noise or unnecessary information to acute concussion assessment. Using only 4 variables and no change scores (i.e., Opt-RS-4), we identified a combination of symptoms which outperformed the full SCAT battery (i.e., Summary Scores model). Although counterintuitive, this result highlights the importance of focusing on the most critical components of the SCAT for acute concussion assessment

We identified specific symptoms and critical components of the SAC which can accurately identify acute concussion. In nearly every model, the symptom don't feel right was included,

suggesting its importance for assessing acute concussion. At <6h and 24-48h, the symptoms headache, dizziness, sensitivity to noise, physical activity makes symptoms worse, and mental activity makes symptoms worse were included in most models. These variables align with post-concussion symptoms identified in previous studies (Eisenberg, Meehan, and Mannix, 2014; Lovell et al., 2006; Randolph et al., 2009). Unfortunately, headache and dizziness are often regarded as common but not specific to concussion (Putukian, 2011). Furthermore, solely relying on symptoms can be problematic since athletes may under-report symptoms (Kerr et al., 2014b; Kroshus et al., 2015b; Williamson and Goodman, 2006). To this end, our results suggest that within the SAC, change scores for the Concentration and Delayed Recall are particularly important. Additionally, the SAC and BESS were not included in any Opt-RS-k model, suggesting that raw scores for the SAC and BESS lack clinical utility compared to symptoms. This result resembles a previous study where, after performing variable selection, the change score (instead of the raw score) for the SAC total score was included in a multidimensional acute concussion assessment model (Garcia et al., 2018). Finally, while Delayed Recall provides valuable information to acute concussion assessment, it requires Immediate Memory to be completed beforehand and therefore, may only marginally reduce the time to administer the SAC. In an auxiliary analysis using 10-fold cross-validation, the change score for SAC Delayed Recall was included in every fold of every Opt-8, Opt-12, and Opt-16 model at <6h and 24-48h (see Appendix 3.A). Yet, excluding this assessment did not appear to alter the accuracy of the Opt-k models (see Appendix 3.B). Given these mixed findings, future research should investigate the importance of Delayed Recall for acute concussion assessment.

Our analysis provides quantitative support regarding the clinical utility of change scores. The variables included in the Opt-k models suggest that baseline testing is most valuable for the symptom checklist and the SAC while the utility of baseline assessments for the BESS is unclear. Change scores for the BESS, however, are potentially important for a subset of athletes who are difficult to assess for acute concussion (e.g., Possible and Probable concussions (Garcia et al., 2019)), suggesting future studies should determine the assessment utility for different athlete subpopulations. Finally, we found that models which included change scores better identified acute concussion at both <6h and 24-48h post-injury compared to models only containing raw scores. However, these differences in performance are modest, coinciding with previous findings (Chin et al., 2016; Echemendia et al., 2012; Garcia et al.,

2018; Putukian et al., 2015; Schmidt et al., 2012).

No component of the BESS was included in any Opt-k or Opt-RS-k model at <6h and 24-48h. Previous studies have also found that among all SCAT assessments, the BESS has the lowest sensitivity and specificity (Chin et al., 2016; Downey, Hutchison, and Comper, 2018; Garcia et al., 2018; McCrea et al., 2005). Furthermore, previous research found that the BESS could be removed from the SCAT without any statistically significant change in AUC (Garcia et al., 2018). Coupled with our findings, the utility of the BESS for sideline screening is questionable. Excluding the BESS from this process could make sideline screening more efficient by reducing noisy information and overall assessment time. However, current best practices recommend the use of multidimensional testing batteries. Therefore, postural control assessments should still be considered, e.g., Tandem Gait, which has demonstrated greater sensitivity to concussion than the BESS (Oldham et al., 2018). Our findings, though, may only apply to acute concussion assessment and the BESS may be valuable at later stages of the concussion management process.

Surprisingly, demographic risk modifiers were not included in our models by the MIP algorithm, suggesting that their importance is diminished under the presence of more sensitive components of the SCAT. We also recognize that age may have been excluded from our models due to the lack of variation within the study population. Yet, risk modifiers may be of greater clinical importance beyond the acute concussion phase and should continue to be incorporated in clinical decision-making. Nevertheless, this result gives promise that an acute concussion screening tool could be developed which can be applied to all athletes.

3.4.1 Limitations

Currently, no perfect diagnostic marker for concussion exists and thus, our findings are based on the clinical diagnosis of concussion. Unfortunately, this limitation cannot be addressed until an objective method for diagnosing concussion is developed. Furthermore, our data only included the SCAT symptom checklist, SAC, and BESS so future studies should consider including other assessments (e.g. King-Devick test and Tandem Gait).

3.5 Conclusion

Sideline assessment of acute concussion must be timely and accurate. We identified optimal subsets of SCAT components which can accurately assess acute concussion. Our results highlight the importance of focusing on the most important aspects of the SCAT. The variables identified in this study build the foundation for modifying sideline screening tools. While clinical examination should continue to be the gold standard for concussion diagnosis, this research provides data-driven insights to guide the future development of concussion assessment practices.

3.A Analysis via 10-fold Cross-Validation

The results presented in Chapter 3 were based on a random split of the study data into a training and testing set. In this supplemental analysis, we recreated our analysis using 10-fold cross-validation on the combined dataset. By leveraging the entire dataset using 10-fold cross-validation, we characterize the robustness of our findings from the original analysis. Using 10-fold cross-validation to create the Opt-k models at <6h and 24-48h, we recorded the frequency with which a modeling variable was included in an Opt-k model. These results are presented in Table 3.A.1 and Table 3.A.2 for the Opt-k models at <6h and 24-48h. The frequencies at which modeling variables were included in Opt-RS-k models at <6h and 24-48h are presented in Table 3.A.3 and Table 3.A.4, respectively. Finally, estimated performance measures for Opt-k and Opt-RS-k models at both <6h and 24-48h are presented in 3.A.5. From Tables 3.A.1-3.A.4, we find that the modeling variables identified in our original analysis are consistent with the modeling variables identified in this supplemental analysis. Furthermore, from Table 3.A.5, we find that our original estimates of model performance are consistent with the estimates obtained using 10-fold cross-validation. Overall, this supplemental analysis illustrates that the results obtained in our original analysis are robust to changes in training and testing data.

Table 3.A.1: Frequency of variables in Opt-k models at <6h using 10-fold cross-validation

Variable Name	Model			
	Opt-4	*Opt-8*	*Opt-12*	*Opt-16*
SAC Concentration CS	10	10	10	10
SAC Delayed Recall CS		10	10	10
SCAT Balance Problems CS			2	2
SCAT Blurred Vision CS			1	1
SCAT Feeling Slowed Down CS			1	
SCAT Headache CS	10	10	10	10
SCAT Nausea CS			1	1
SCAT Pressure in Head CS			5	5
SCAT Sensitivity to Light CS			3	2
SCAT Sensitivity to Noise CS		8	7	6
SCAT Balance Problems RS			2	1
SCAT Blurred Vision RS			4	4
SCAT Dizziness RS	4	10	10	10
SCAT Don't Feel Right RS	6	10	10	10
SCAT Nausea RS			1	1
SCAT Symptoms Worsen with Mental Activity		10	10	10
SCAT Symptoms Worsen with Physical Activity	10	10	10	10
SCAT Sensitivity to Light RS			2	2
SCAT Sensitivity to Noise RS		2	3	4

CS, change score; RS, raw score; BESS, Balance Error Scoring System; SAC, Standardized Assessment of Concussion; SCAT, Sport Concussion Assessment Tool

Table 3.A.2: Frequency of variables in Opt-k models at 24-48h using 10-fold cross-validation

Variable Name	Model			
	Opt-4	*Opt-8*	*Opt-12*	*Opt-16*
SAC Concentration CS	10	10	10	10
SAC Delayed Recall CS		10	10	10
SCAT Dizziness CS			1	1
SCAT In a Fog CS			1	1
SCAT Headache CS		2	2	2
SCAT Nausea CS			1	1
SCAT Pressure in Head CS		1	3	3
SCAT Dizziness RS			2	2
SCAT Don't Feel Right RS	2	9	10	10
SCAT Feeling Slowed Down RS			4	4
SCAT In a Fog RS			3	3
SCAT Headache RS	3	8	8	8
SCAT Symptoms Worsen with Mental Activity	10	9	10	10
SCAT Symptoms Worsen with Physical Activity	10	10	10	10
SCAT Pressure in Head RS		2	6	6
SCAT Sensitivity to Light RS		6	10	10
SCAT Sensitivity to Noise RS	5	10	10	10

CS, change score; RS, raw score; BESS, Balance Error Scoring System; SAC, Standardized Assessment of Concussion; SCAT, Sport Concussion Assessment Tool

Table 3.A.3: Frequency of variables in Opt-RS-k models at <6h using 10-fold cross-validation

Variable Name	Model			
	Opt-RS-4	*Opt-RS-8*	*Opt-RS-12*	*Opt-RS-16*
BESS Firm Surface Score RS		1	1	1
BESS Foam Surface Score RS		8	9	9
SAC Concentration RS			1	1
SAC Delayed Score		1	1	1
SAC Immediate Memory RS		9	9	9
SAC Orientation RS		1	1	1
SCAT Dizziness RS	10	10	10	10
SCAT Don't Feel Right RS	10	10	10	10
SCAT Headache RS		2	2	2
SCAT Symptoms Worsen with Mental Activity	10	10	10	10
SCAT Symptoms Worsen with Physical Activity	10	10	10	10
SCAT Sensitivity to Noise RS		10	10	10

CS, change score; RS, raw score; BESS, Balance Error Scoring System; SAC, Standardized Assessment of Concussion; SCAT, Sport Concussion Assessment Tool

Table 3.A.4: Frequency of variables in Opt-RS-k models at 24-48h using 10-fold cross-validation

Variable Name	Model			
	Opt-RS-4	*Opt-RS-8*	*Opt-RS-12*	*Opt-RS-16*
BESS Firm Surface Score RS		2	2	2
BESS Foam Surface Score RS		7	9	9
SAC Concentration RS		1	1	1
SAC Delayed Score		1	1	1
SAC Immediate Memory RS		9	9	9
SAC Orientation RS			1	1
SCAT Dizziness RS		1	1	1
SCAT Don't Feel Right RS	10	10	10	10
SCAT Headache RS		7	7	7
SCAT Symptoms Worsen with Mental Activity	10	10	10	10
SCAT Symptoms Worsen with Physical Activity	10	10	10	10
SCAT Sensitivity to Light RS		6	6	6
SCAT Sensitivity to Noise RS	10	10	10	10

CS, change score; RS, raw score; BESS, Balance Error Scoring System; SAC, Standardized Assessment of Concussion; SCAT, Sport Concussion Assessment Tool

Table 3.A.5: Performance of Opt-k and Opt-RS-k models based on 10-fold cross-validation

Timepoint	Model	Brier Score		AUC		Sensitivity		Specificity	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
<6h	Opt-4	0.081	0.012	0.95	0.01	0.88	0.03	0.91	0.04
	Opt-8	0.073	0.011	0.96	0.01	0.89	0.03	0.93	0.02
	Opt-12	0.073	0.010	0.96	0.01	0.89	0.03	0.92	0.03
	Opt-16	0.073	0.010	0.96	0.01	0.90	0.03	0.92	0.03
24-48h	Opt-4	0.105	0.007	0.92	0.01	0.81	0.03	0.92	0.03
	Opt-8	0.090	0.006	0.94	0.01	0.86	0.03	0.92	0.04
	Opt-12, Opt-16	0.088	0.006	0.94	0.01	0.86	0.02	0.92	0.04
<6h	Opt-RS-4	0.088	0.012	0.93	0.01	0.88	0.04	0.90	0.03
	Opt-RS-8	0.085	0.014	0.94	0.02	0.89	0.03	0.90	0.03
	Opt-RS-12, Opt-RS-16	0.085	0.014	0.94	0.02	0.89	0.03	0.90	0.03
24-48h	Opt-RS-4	0.104	0.007	0.92	0.01	0.85	0.04	0.90	0.04
	Opt-RS-8	0.099	0.010	0.93	0.01	0.84	0.02	0.92	0.04
	Opt-RS-12, Opt-RS-16	0.100	0.010	0.93	0.01	0.84	0.01	0.92	0.04

SD, standard deviation; AUC, area under the curve

3.B Removal of SAC Delayed Recall

Based on our previous analysis, SAC Delayed Recall provides valuable information to acute concussion assessment in the Opt-8, Opt-12, and Opt-16 models at both <6h and 24-48h. However, Delayed Recall requires Immediate Memory to be completed beforehand and therefore, may only marginally reduce the time to administer the battery. In this supplemental analysis, we aimed to determine whether removing or replacing SAC Delayed Recall from the aforementioned models would change their accuracy in assessing acute concussion. To perform this analysis, we recreated the Opt-8, Opt-12, and Opt-16 models without including SAC Delayed Recall raw scores or change scores as a modeling variable. The results at <6h are shown below in Table 3.B.1 and the results at 24-48h are shown below in Table 3.B.2. These new models demonstrate similar levels of accuracy as those identified in our original analysis. Altogether, these results indicate that SAC Delayed Recall can be replaced by other assessments to reduce the time needed to administer the concussion assessment battery.

Table 3.B.1: Modified Opt-8, Opt-12, and Opt-16 Models without SAC Delayed Recall at <6h post-injury

Model	k=8		k=12, 16	
Model Variables and Coefficients				
Variable	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*
Intercept	-2.69 (0.11)*	(-2.91, -2.47)	-2.78 (0.12)*	(-3.01, -2.55)
SAC Concentration Score CS	0.40 (0.07)*	(0.27, 0.54)	0.38 (0.07)*	(0.25, 0.51)
SCAT Headache CS	0.66 (0.08)*	(0.52, 0.81)	0.63 (0.07)*	(0.49, 0.77)
SCAT Pressure in Head CS	0.21 (0.09)***	(0.04, 0.38)		
SCAT Sensitivity to Noise CS	0.19 (0.14)	(-0.09, 0.46)	0.19 (0.15)	(-0.11, 0.48)
SCAT Dizziness RS	0.50 (0.12)*	(0.28, 0.73)	0.32 (0.12)**	(0.08, 0.56)
SCAT Don't Feel Right RS	0.56 (0.08)*	(0.40, 0.71)	0.57 (0.10)*	(0.37, 0.76)
SCAT Symptoms Worsen with Mental Activity	1.70 (0.27)*	(1.18, 2.22)	1.67 (0.27)*	(1.14, 2.21)
SCAT Symptoms Worsen with Physical Activity	2.33 (0.24)*	(1.87, 2.79)	2.22 (0.24)*	(1.75, 2.69)
SCAT Difficulty Remembering CS			-0.28 (0.08)*	(-0.44, -0.12)
SCAT Balance Problems RS			0.05 (0.13)	(-0.19, 0.30)
SCAT Blurred Vision RS			0.52 (0.15)*	(0.23, 0.82)
SCAT Feeling Like in a Fog RS			0.04 (0.11)	(-0.18, 0.25)
SCAT Pressure in Head RS			0.37 (0.09)*	(0.19, 0.54)
Performance Measures				
Brier Score	0.072		0.072	
AUC	0.96		0.96	
Sensitivity[1]	0.86		0.90	
Specificity[1]	0.94		0.91	

CS, change score; RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; SE, Standard error; CI, Confidence Interval; AUC, Area under the receiver operating characteristic curve; [1]Sensitivity and specificity values reported in table correspond to the threshold which maximizes the sum of sensitivity and specificity; *P<0.001 **P<0.01 ***P<0.05 coefficient is significantly different from 0

Table 3.B.2: Modified Opt-8, Opt-12, and Opt-16 Models without SAC Delayed Recall at 24-48h post-injury

Model	*k=8*		*k=12, 16*	
Model Variables and Coefficients				
Variable	*Coefficient (SE)*	*95% CI*	*Coefficient (SE)*	*95% CI*
Intercept	-1.99 (0.09)*	(-2.16, -1.82)	-1.99 (0.09)*	(-2.16, -1.82)
SAC Concentration Score CS	0.65 (0.06)*	(0.54, 0.77)	0.66 (0.06)*	(0.54, 0.77)
SCAT Dizziness RS	-0.02 (0.12)	(-0.25, 0.20)	-0.03 (0.12)	(-0.26, 0.20)
SCAT Don't Feel Right RS	0.46 (0.08)*	(0.31, 0.62)	0.45 (0.08)*	(0.29, 0.61)
SCAT Headache RS	0.68 (0.08)*	(0.53, 0.83)	0.67 (0.08)*	(0.52, 0.82)
SCAT Symptoms Worsen with Mental Activity	1.85 (0.23)*	(1.40, 2.29)	1.83 (0.23)*	(1.39, 2.28)
SCAT Symptoms Worsen with Physical Activity	2.10 (0.21)*	(1.68, 2.51)	2.07 (0.21)*	(1.66, 2.49)
SCAT Pressure in Head RS	0.23 (0.09)**	(0.06, 0.40)	0.22 (0.09)***	(0.05, 0.39)
SCAT Sensitivity to Noise RS	0.55 (0.13)*	(0.29, 0.81)	0.48 (0.14)*	(0.21, 0.76)
SCAT Sensitivity to Light RS			0.17 (0.11)	(-0.05, 0.39)
Performance Measures				
Brier Score	0.086		0.086	
AUC	0.95		0.95	
Sensitivity[1]	0.85		0.85	
Specificity[1]	0.94		0.94	

CS, change score; RS, raw score; SAC, Standard Assessment of Concussion; BESS, Balance Error Scoring System; SCAT, Sport Concussion Assessment Tool; SE, Standard error; CI, Confidence Interval; AUC, Area under the receiver operating characteristic curve; [1]Sensitivity and specificity values reported in table correspond to the threshold which maximizes the sum of sensitivity and specificity; *P<0.001 **P<0.01 ***P<0.05 coefficient is significantly different from 0

Chapter 4

Data-driven Diagnosis Decision Thresholds for Risk Estimation Models

4.1 Introduction

In Chapter 2 and Chapter 3, we developed risk estimation models for acute concussion. In this chapter, we show how these risk estimation models can be used to optimize diagnosis decisions. More broadly, the growing availability of data has led to the rapid development of risk estimation models in several industries including healthcare, finance, and manufacturing. These models can be especially impactful in healthcare, where risk estimation models can be used as decision aids to supplement medical diagnosis and treatment decisions. For example, methods such as logistic regression, survival analysis, neural networks, and other machine learning models have been used to assess emergency department admission risk (Peck et al., 2012), estimate hospital readmission risk (Xue, Klabjan, and Luo, 2019), detect glaucoma progression (Schell et al., 2014), predict adverse coronary heart disease events (Anderson, 1991), and aid diagnosis of cancer (Kourou et al., 2015). Yet, applying these models in clinical practice is challenging (Degeling, Koffijberg, and IJzerman, 2017; Moons et al., 2009). One way to bridge this gap between research and practice is by determining decision thresholds to help clinicians interpret these models (Ebell, 2010).

In this research, we aim to determine suitable decision thresholds for risk estimation models. For instance, if a model estimates that a patient has a 55% chance of having a cancerous tumor, does that estimate provide sufficient evidence to diagnose the patient with

cancer? What if the model estimates a 45% chance? When the consequences associated with misdiagnoses are great, how one determines these decision boundaries is critical.

Determining appropriate decision thresholds is not straightforward. First, these decision boundaries should reflect the decision-maker's risk attitude, i.e., willingness to take on consequences associated with misdiagnoses (Felder and Mayrhofer, 2014). While some patients and physicians may be more willing to accept highly consequential outcomes, others may be far more conservative in their decision-making (based on the perceived consequences associated with each decision). Yet, traditional methods for interpreting risk estimates do not reflect such risk attitudes (Degeling, Koffijberg, and IJzerman, 2017). Even worse, some may provide arbitrary risk stratifications.

Another key issue is the application of these risk estimation models to populations for which the underlying population does not match the population used to parameterize the model (Royston et al., 2009). For example, consider risk scores from the Framingham Heart Study (Wang et al., 2003), which was parameterized on a cohort of roughly 800 participants from Framingham, Massachussettes between 1948 and 2000. This model has been used to inform clinical practice for cardiovascular disease management, which is applied to populations which are far more diverse (Perk et al., 2012). Therefore, decision boundaries should not only depend on the risk estimation model at hand but also on the population to which it will be applied.

Furthermore, traditional binary classification models may not sufficiently address uncertainty in diagnosis decisions. This sentiment is reflected in diagnosis guidelines for conditions such as multiple sclerosis (McDonald et al., 2001), Alzheimer's disease (AD) (McKhann et al., 2011), diabetes (American Diabetes Society, 2016), and concussion (Kutcher and Giza, 2014), where diagnosis is divided into risk classifications rather than a dichotomous outcome. For instance, for sports-related concussion, the diagnosis may be broken up into (1) Possible, (2) Probable, and (3) Definite concussion depending on the clinician's diagnostic certainty (Kutcher and Giza, 2014). In guidelines which use such risk classifications, intermediate risk classifications arise from conflicting diagnostic assessment results or a lack of definitive evidence to make strong diagnostic conclusions. Analogously, there may be ranges along the risk spectrum where a risk estimation model's estimates are not "certain enough" and call for more information before a diagnosis decision can be made. In particular, these intermediate ranges reflect cases in which qualitative information, which may not be easily implemented

in risk estimation models, should be used to guide diagnosis decisions. Since risk estimation models are typically designed to supplement clinical diagnosis, identifying these ranges is critical. Yet, few methods create decision boundaries which account for this risk estimation uncertainty (Degeling, Koffijberg, and IJzerman, 2017).

This research aims to bridge the gap between risk estimation models and clinical application by presenting a rigorous approach to determine diagnosis decision thresholds. These thresholds (1) reflect the decision-maker's risk attitude, (2) jointly depend on the risk estimation model and patient population to which it is applied, and (3) identify ranges in the risk continuum in which the risk estimation model is most and least accurate. We apply our method to acute concussion assessment, a field where diagnostic decisions must be made accurately and quickly to mitigate prolonged injury recovery and post-concussion symptom severity (Asken et al., 2018).

The key contributions of this work are as follows:

1. We introduce a data-driven stochastic optimization framework to determine diagnostic decision thresholds based on the application of a risk estimation model to patients from a fixed population. Compared to previous methods (see Section 4.2), we avoid the need to estimate outcome-based utilities or make distributional assumptions to account for uncertainty in risk estimates.

2. In our analytical study, we show that the optimal solution to our proposed model can be characterized by extreme-point solutions of a related linear program. Thus, our model can be solved using quantile estimation — bypassing the need for advanced optimization software. We also identify additional modeling frameworks, including utility-based and multi-class classification frameworks, for which our analytical results can be applied. Specifically, for our utility-based extensions, we formulate a model for which utilities such as quality-adjusted life-years may be used. In our extensions to multi-class classification, we develop frameworks for both multi-label and ordinal classification.

3. Through an analytical study and numerical analysis using both real and simulated data, we determine when two decision thresholds, which allow for a deferred diagnosis decision, will outperform a single decision threshold, which only allows for binary classification.

4. We perform extensive numerical analysis to determine how the modeling parameters should be chosen based on the general characteristics of the population that undergoes diagnostic testing. Our analysis also gives insight to guide the choice of data-driven solution methodology based on sample size and the quality of the underlying risk estimation model.

5. We are one of the first groups to apply an optimization framework to develop data-driven diagnostic thresholds for acute concussion based on data from the CARE Consortium — a nationwide collaboration comprising 29 National Collegiate Athletic Association (NCAA) universities and military service academies. By incorporating feedback from concussion experts across the CARE Consortium, we ensure that, in the case of acute concussion assessment, our modeling framework outperforms methods which are commonly used in practice. Furthermore, we provide a valuable framework which quantifies the uncertainty in diagnosis decisions using real data rather than subjective clinical experience. The models developed in this research have the potential to be developed into tools which can supplement clinical decision-making.

The remainder of this chapter is organized as follows. In Section 4.2, we present a review of related research literature. In Section 4.3, we present our modeling approach and the analytical properties of this model, along with model extensions to utility-based and multi-class classification frameworks. In Section 4.4, we present our data-driven solution methods and in Section 4.5, we evaluate each of these solution methods using simulation. In Section 4.6, we apply our models to concussion assessment data. We perform sensitivity analysis on the model parameters and analyze the performance of this model compared to existing methods. Finally, we present managerial insights and other concluding remarks in Section 4.7.

4.2 Relevant Literature

This research falls within the domains of (1) operations research in disease screening and diagnosis decisions and (2) determining diagnosis decision boundaries.

4.2.1 Operations Research in Disease Screening and Diagnosis Decisions

Operations Research has been applied to many areas of disease screening. Specifically, applications to cancer screening are reviewed extensively by Pierskalla and Brailer, 1994 and Alagoz, Ayer, and Erenay, 2011 while more recent works include Ayer, Alagoz, and Stout, 2012; Ayer et al., 2016; Bertsimas, Silberholz, and Trikalinos, 2018; Erenay, Alagoz, and Said, 2014; Güneş, Örmeci, and Kunduzcu, 2015; Lee et al., 2015; Li et al., 2014; Maillart et al., 2008; McLay, Foufoulides, and Merrick, 2010; Tejada et al., 2015 and Barnett et al., 2017. Other applications include obesity (Yang, Goldhaber-Fiebert, and Wein, 2013), Glaucoma (Helm et al., 2015), Chlamydia (Teng, Kong, and Tu, 2015), Ebola (Jacobson, Yu, and Jokela, 2016), blood screening (El-Amine, Bish, and Bish, 2018), and HIV (Deo et al., 2015; Deo and Sohoni, 2015; Jónasson, Deo, and Gallien, 2017). Like our study, these works consider the imperfect nature of diagnostic tests in their models. However, they determine optimal strategies based on estimated utilities while we focus on diagnostic accuracy. Further, they focus on sequential decisions while we consider the immediate diagnosis decision.

Operations Research has also been applied to medical diagnosis decisions, where such problems typically optimize pre-diagnosis decisions or follow-up decisions after an initial diagnostic test. For example, Bayati, Bhaskar, and Montanari, 2018 determine the least-cost set of biomarker tests which allow for sufficient diagnostic power while Ayvaci, Alagoz, and Burnside, 2012 and Zhang et al., 2012 optimize biopsy follow-up decisions for cancer. However, our work focuses on the actual diagnosis decision at hand instead of pre- or post-diagnosis decisions. To this end, Ayvaci et al., 2017 and Ahsen, Ayvaci, and Raghunathan, 2019 study when and how bias-inducing information should be incorporated in breast cancer diagnostic decisions. Similarly to Ahsen, Ayvaci, and Raghunathan, 2019, we also study the incorporation of clinical decision support systems in diagnostic decisions. However, they focus on the design of such systems while we focus on the interpretation of these decision support systems, i.e., a risk estimation model. Furthermore, their work focuses on balancing two sources of information (i.e., mammogram risk and clinical-risk information) and deriving one diagnostic threshold whereas our work focuses on a single source of information (i.e., risk estimates) and deriving two diagnostic decision thresholds.

4.2.2 Determining Diagnosis Decision Boundaries

A number of methods have been developed to optimize a single decision threshold. In this literature, it is typical to assign a utility to each possible diagnostic outcome and determine an optimal threshold which maximizes utilities (Deneef and Kent, 1993; Felder and Mayrhofer, 2014; Giessen et al., 2018; Jund et al., 2005; Moons et al., 1997; Pauker and Kassirer, 1975). For instance, Giessen et al., 2018 develop a stepwise method to optimize a risk threshold based on multiple criteria, including quality-adjusted life-years, cost of treatment, and net health benefit. However, the use of such utilities has been questioned (McGregor and Caro, 2006; Nord, Daniels, and Kamlet, 2009). To circumvent this difficulty, methods based on the receiver operating characteristic (ROC) have been developed (Greiner, Pfeiffer, and Smith, 2000; Greiner, Sohr, and Göbel, 1995; Odetola et al., 2016; Somoza and Mossman, 1992; Vermont et al., 1991). These methods typically optimize a single threshold based on measures of diagnostic accuracy. However, single thresholds do not specify regions for which the risk estimation models perform poorly, i.e., risk estimate ranges over which false-positive and false-negative rates are especially high and diagnostic decisions should be avoided. Therefore, we also review methods to determine multiple decision thresholds.

Utility-based methods for deriving multiple thresholds include Glasziou and Hilden, 1986; Pauker and Kassirer, 1980 and Nease, Owens, and Sox, 1989. In contrast, Hartz et al., 1986 determine thresholds based on uncertainty in different physicians' decision thresholds and Mangasarian, Street, and Wolberg, 1995 apply linear programming to determine diagnostic decision thresholds based on tumor characteristics. Zhu and Fang, 2016 and Si, Yakushev, and Li, 2017 develop tree-based approaches which categorize diagnoses as positive, negative, or uncertain, where those who are uncertain are not predicted well by their classification tree and hence should be further evaluated. We apply this same classification scheme in our work but we develop thresholds along the probability spectrum. Our work most closely relates to Si, Yakushev, and Li, 2017 since we both employ optimization frameworks to determine diagnosis thresholds based on ROC statistics. However, they determine decision boundaries for each biomarker in a sequence of biomarkers, rather than a one-time threshold in probability space which can potentially incorporate multiple biomarkers at once. Furthermore, they assume that their biomarker readings follow a Gaussian distribution while we do not make any assumptions on the distribution of risk scores. Additionally, in their treatment of non-Gaussian biomarkers, they solve the optimization problem using an iterative approximation.

In contrast, we solve tractable data-driven optimization problems to global optimality.

Several methods based on Bayesian decision theory have also been used to derive decision thresholds (Sheppard and Kaufman, 2005; Weise et al., 2006; Yao, 2010; Yao and Zhou, 2016). Among these approaches, both Sheppard and Kaufman, 2005 and Weise et al., 2006 derive thresholds based on distributional information. Specifically, Sheppard and Kaufman, 2005 compare posterior likelihood ratios to determine whether a risk estimate is more likely to come from a true-positive or a true-negative patient. However, this approach requires estimation of the full distribution of risk estimates to compute, whereas we find that our thresholds only depend on quantiles of these distributions. To this end, Weise et al., 2006 also derive thresholds based on quantiles. However, they assume that these quantiles are derived from a standard normal distribution whereas we do not make distributional assumptions. Finally, the decision thresholds derived by Yao, 2010 and Yao and Zhou, 2016 are most similar to our research since they explicitly aim to determine whether an object should be classified as either positive, negative, or on the boundary (i.e., too uncertain to make a decision). To this end, the thresholds derived in both Yao, 2010 and Yao and Zhou, 2016 are determined entirely by "risk" associated with each decision and do not rely on any distributional information at all. Moreover, all of the aforementioned approaches determine decision thresholds based on unconstrained optimization problems, whereas our approach requires solving a constrained stochastic program.

Overall, our work differs from previous research in three ways: (1) we take a constrained optimization approach, allowing us to simultaneously maximize sensitivity and specificity while limiting false-positive and false-negative rates, (2) we limit false-positive and false-negative rates based on the decision-maker's risk attitude, creating personalized decision thresholds, and (3) we consider uncertainty without making distributional assumptions. In Section 4.3, we detail our stochastic programming approach to determining these decision thresholds.

4.3 Modeling Approach

In this section, we describe our modeling approach, its related analytical properties, and several modeling extensions. Specifically, we provide our general problem setting and notation in Section 4.3.1 and the model formulation in Section 4.3.2. In Section 4.3.3, we develop an

Table 4.1: Model notation

Notation	Description
$\xi^+ \sim \mathcal{P}^+$, $\xi^- \sim \mathcal{P}^-$	Random risk estimate belonging to true-positives and true-negatives, respectively, along with their corresponding distributions
t	General diagnostic threshold
u, l	Upper and lower diagnostic decision thresholds, respectively
$se(u, \xi^+)$	Event that a true-positive patient is correctly classified as positive given upper threshold u and risk estimate ξ^+ (sensitivity)
$sp(l, \xi^-)$	Event that a true-negative patient is correctly classified as negative given lower threshold l and risk estimate ξ^- (specificity)
$fp(u, \xi^-)$	Event that a true-negative patient is incorrectly classified as positive given upper threshold u and risk estimate ξ^- (false-positive)
$fn(l, \xi^+)$	Event that a true-positive patient is incorrectly classified as negative given lower threshold l and risk estimate ξ^+ (false-negative)
λ, ϕ	Weighting parameters to balance sensitivity and specificity in TTP and TTP*, respectively
γ^{fp}, γ^{fn}	Maximum levels of false-positive and false-negative rates, respectively

approximation to our stochastic programming model and in Section 4.3.4, we characterize its optimal solution based on the extreme-point solutions of a related linear program. We then identify conditions under which the optimal solution for the approximation is also optimal for the original stochastic programming model. In Section 4.3.5 and Section 4.3.6, we formulate and analyze utility-based and multi-class diagnosis extensions, respectively, to our modeling framework.

4.3.1 Problem Setting and Notation

A summary of our notation is provided in Table 4.1. We consider a patient population for a chosen disease, which can be divided into two mutually exclusive populations of true-positives (e.g., has a concussion) and true-negatives (e.g., does not have a concussion). A randomly chosen patient is associated with the random variables (X, Y). The random vector $X \in \mathcal{X}$ represents a vector of patient characteristics (e.g., age, sex, concussion assessment results) which have been transformed into a numerical representation (e.g., using one-hot encoding for categorical variables). We let $\mathcal{X} \subseteq \mathbb{R}^p$ denote the set of p-length numerical representations of patient characteristics. The random variable $Y \in \{0, 1\}$ represents the patient's label, which

indicates whether he or she is from the true-positive (i.e., $Y = 1$) or true-negative (i.e., $Y = 0$) population. A risk estimation model $f : \mathcal{X} \to (0, 1)$ approximates the conditional probability $\mathbb{P}(Y = 1|X)$. Such models include logistic regression, classification and regression trees, and artificial neural networks.

Throughout the remainder of the chapter, we focus on the risk estimates $\xi^+ := f(X|Y = 1)$ and $\xi^- := f(X|Y = 0)$, which denote the risk estimates belonging to a patient from the population of true-positives and true-negatives, respectively. Note that ξ^+ and ξ^- are expressed as random variables since they are functions of X, a random vector. Therefore, we also write that $\xi^+ \sim \mathcal{P}^+$ (i.e., ξ^+ has distribution $\mathcal{P}^+$) and $\xi^- \sim \mathcal{P}^-$.

Given a diagnostic threshold t, a patient is classified as positive if his or her risk estimate exceeds t. Otherwise, the patient is classified as negative. Let $\mathbb{1}(\cdot)$ denote the indicator function. Then, given a threshold t and a true-positive patient with risk estimate ξ^+, we define sensitivity $se(t, \xi^+) := \mathbb{1}(\xi^+ \geq t)$ and false-negative $fn(t, \xi^+) := \mathbb{1}(\xi^+ < t)$. Similarly, given a threshold t and a true-negative patient with risk estimate ξ^-, we define specificity $sp(t, \xi^-) := \mathbb{1}(\xi^- \leq t)$ and false-positive $fp(t, \xi^-) := \mathbb{1}(\xi^- > t)$. The conditional expectations $\mathbb{E}[se(t, \xi^+)]$ and $\mathbb{E}[sp(t, \xi^-)]$ represent the expected sensitivity and specificity, respectively, under threshold t based on each population's risk estimates. The expected false-positive rate, $\mathbb{E}[fp(t, \xi^-)]$, and expected false-negative rate, $\mathbb{E}[fn(t, \xi^+)]$, are defined similarly.

4.3.2 Stochastic Programming Formulation

We consider an upper threshold u and lower threshold l such that any risk estimate above u is classified as positive and any risk estimate below l is classified as negative. We aim to determine the values of u and l which balance sensitivity and specificity while also limiting the rate of false-positive and false-negative classifications. The region between u and l defines a range of risk estimates for which diagnostic decisions may be deferred due to elevated risk of false-positive or false-negative diagnoses. This region also reflects a range over which risk scores were not estimated well and clinical judgment should be favored. Since we model patient risk estimates as random variables, we formulate the Two-Threshold Problem (TTP)

as the following stochastic program.

$$\begin{aligned}
(\text{TTP}) \quad \max_{u,l} \quad & \lambda \mathbb{E}[se(u, \xi^+)] + (1-\lambda)\mathbb{E}[sp(l, \xi^-)] && \text{(4.1a)} \\
\text{s.t.} \quad & \mathbb{E}[fp(u, \xi^-)] \leq \gamma^{fp} && \text{(4.1b)} \\
& \mathbb{E}[fn(l, \xi^+)] \leq \gamma^{fn} && \text{(4.1c)} \\
& 0 \leq l \leq u \leq 1. && \text{(4.1d)}
\end{aligned}$$

The objective function (4.1a) uses the parameter $\lambda \in (0, 1)$ to specify the relative importance of sensitivity to specificity. Higher values of λ imply that greater importance is placed on correctly classifying true-positives. Alternatively, setting λ equal to the proportion of true-positives in the overall patient population equates (4.1a) to maximizing the probability of making a correct diagnosis decision. The value of λ is chosen by the decision-maker and making an appropriate choice can be difficult. Fortunately, we show, in Section 4.3.4, that when it is optimal to use two distinct decision thresholds, the choice of λ does not affect the optimal solution. The constraints (4.1b) and (4.1c) imply that, based on the thresholds u and l, the false-positive and false-negative rates should not exceed the bounds γ^{fp} and γ^{fn}, respectively, where $\gamma^{fp}, \gamma^{fn} \in (0, 1)$. These parameters can reflect clinically acceptable levels of diagnostic accuracy. We provide guidelines for choosing these parameters in Section 4.7.1. Finally, constraint (4.1d) ensures that the upper threshold u remains above the lower threshold l, and that both are between 0 and 1.

4.3.3 Approximating the Two-threshold Problem

Very little can be said about the form of TTP's objective function (4.1a), even if the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ are known exactly. For example, in general, (4.1a) may not be concave, continuous, or differentiable everywhere. However, the functions $se(u, \xi^+)$ and $sp(l, \xi^-)$ are monotone in u and l for every $\xi^+, \xi^- \in (0, 1)$, respectively. Using these properties, we approximate TTP with TTP*:

$$\begin{aligned}
(\text{TTP}^*) \quad \min_{u,l} \quad & \phi u - (1-\phi) l && \text{(4.2)} \\
\text{s.t.} \quad & \text{(4.1b)-(4.1d)},
\end{aligned}$$

where $\phi \in (0,1)$ is also a weighting parameter which may not necessarily equal λ but also serves the purpose of defining the relative importance of one threshold to the other. In Section 4.3.4, we analyze TTP* and identify cases in which the optimal solutions to TTP and TTP* coincide.

4.3.4 Structural Properties

In this section, we highlight structural properties which are useful in understanding TTP through TTP*. First, we show that by expressing the constraints (4.1b) and (4.1c) in terms of quantiles, TTP* is equivalent to a linear program. Based on this linear program, we show that the optimal solution to TTP* is either a two-threshold solution or a one-threshold solution, and that the optimal solution can be characterized based on the parameters of TTP*. We then relate the optimal solution of TTP* with that of TTP, showing that a two-threshold solution of TTP* is optimal in TTP, but not necessarily for one-threshold solutions. These results indicate that TTP can be solved, in two-threshold cases, using quantile estimation.

Throughout this section, we assume that probabilities of any true-positive or true-negative risk estimate falling on a fixed threshold t is zero. That is,

A1 For any fixed t, $\mathbb{P}(t = \xi^+) = 0$ and $\mathbb{P}(t = \xi^-) = 0$.

Assumption A1 implies that the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ are continuous which is not too restrictive since most risk estimation models output risk scores along a continuum.

Now, we relate the optimal solution to TTP* with the parameters ϕ, γ^{fp} and γ^{fn}. Define

$$u(\gamma^{fp}) := \inf\{u \in [0,1] : g_1(u) \leq \gamma^{fp}\} \text{ and} \tag{4.3}$$

$$l(\gamma^{fn}) := \sup\{l \in [0,1] : g_2(l) \leq \gamma^{fn}\}, \tag{4.4}$$

where we define $g_1(u) := \mathbb{E}[fp(u, \xi^-)]$ and $g_2(l) := \mathbb{E}[fn(l, \xi^+)]$ for notational convenience. The quantity $u(\gamma^{fp})$ specifies the smallest value of u which ensures that the false-positive rate is no more than γ^{fp} and similarly, $l(\gamma^{fn})$ denotes the greatest value of l such that the false-negative rate is no more than γ^{fn}. Note that $u(\gamma^{fp})$ and $l(\gamma^{fn})$ are quantiles. Specifically, $u(\gamma^{fp})$ is the $(1-\gamma^{fp})$-quantile of $\mathcal{P}^-$ and $l(\gamma^{fn})$ is the negative value of the $(1-\gamma^{fn})$-quantile

of $\mathcal{P}^+$ (see Appendix 4.A). This fact forms the basis for the solution methodology described in Section 4.4.1.

Now, by Theorem 7.48 in Shapiro, Dentcheva, and Ruszczyński, 2009, the functions $g_1(u)$ and $g_2(l)$ are finite-valued and continuous over $[0,1]$, so the function $g(u,l) := \max\{g_1(u), g_2(l)\}$ is finite-valued and continuous on $[0,1]^2$. The continuity of these functions along with the fact that the interval $[0,1]$ is compact and convex implies that $u(\gamma^{fp})$ and $l(\gamma^{fn})$ are attained on $[0,1]$. Now, in Proposition 4.1, we show that TTP* is equivalent to a linear program with constraints written in terms of $u(\gamma^{fp})$ and $l(\gamma^{fn})$.

Proposition 4.1. *TTP* is equivalent to the following linear program.*

$$\begin{aligned} \min_{u,l} \quad & \phi u - (1-\phi) l \\ s.t. \quad & u \geq u(\gamma^{fp}), \quad l \leq l(\gamma^{fn}), \quad 0 \leq l \leq u \leq 1. \end{aligned}$$

Proof. It suffices to show the following equivalence:

$$u \geq u(\gamma^{fp}) \iff g_1(u) \leq \gamma^{fp} \text{ and } l \leq l(\gamma^{fn}) \Leftrightarrow g_2(l) \leq \gamma^{fn}.$$

To see the first equivalence, we note that $g_1(u)$ is continuous and (weakly) decreasing in u. It follows that $u \geq u(\gamma^{fp})$ if and only if $g_1(u) \leq \gamma^{fp}$. The second equivalence can be similarly shown, which completes the proof. □

From Proposition 4.1, we can define the following polyhedron as the feasible region of TTP*.

$$Y = \left\{(u,l) : u \geq u(\gamma^{fp}), l \leq l(\gamma^{fn}), 0 \leq l \leq u \leq 1\right\} \tag{4.5}$$

Additionally, by recasting TTP* as a linear program, we can focus our analysis on the extreme-point solutions of TTP*. The following results characterize the types of optimal solution to TTP*.

Proposition 4.2 (Two-threshold Solutions)**.** *If $\phi \in (0,1)$ and both γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) > l(\gamma^{fn})$, then $\left(u(\gamma^{fp}), l(\gamma^{fn})\right)$ is the unique optimal solution to TTP*.*

Proof. For notational convenience, we write $x = (u, l)$ and $\psi(x) = \phi u - (1 - \phi)l$. There exists an optimal solution to TTP* which is also an extreme point of its polyhedral feasible region, denoted by Y, as described in (4.5). As $u(\gamma^{fp}) > l(\gamma^{fn})$, the extreme points of Y are $(u(\gamma^{fp}), l(\gamma^{fn}))$, $(u(\gamma^{fp}), 0)$, $(1, l(\gamma^{fn}))$, and $(1, 0)$. Of all extreme points, $(u(\gamma^{fp}), l(\gamma^{fn}))$ results in the best objective function value and hence is optimal to TTP* because $\psi(x)$ is decreasing in u and increasing in l.

To show that this optimal solution is unique, suppose that there exists another optimal solution $\bar{x} = (\bar{u}, \bar{l}) \neq (u(\gamma^{fp}), l(\gamma^{fn}))$. Then, it must be the case that $\bar{u} > u(\gamma^{fp})$ or $\bar{l} < l(\gamma^{fn})$ since otherwise, $\bar{x}$ would be infeasible. However, if either case were true, we would have $\psi(\bar{x}) > \psi((u(\gamma^{fp}), l(\gamma^{fn})))$ which is a contradiction on the optimality of $\bar{x}$. Thus, $(u(\gamma^{fp}), l(\gamma^{fn}))$ is the unique optimal solution. □

Proposition 4.2 implies that the parameter $\phi \in (0, 1)$ has no effect if the optimal solution consists of two thresholds. Additionally, the conditions specified in Proposition 4.2 are satisfied if both γ^{fp} and γ^{fn} are "low enough". Finally, we have shown that the optimal solution may be obtained without solving an optimization problem if the values $u(\gamma^{fp})$ and $l(\gamma^{fn})$ can be computed exactly and the parameters γ^{fp} and γ^{fn} are chosen to satisfy the sufficient conditions. However, the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ are typically unknown, making $u(\gamma^{fp})$ and $l(\gamma^{fn})$ difficult to ascertain.

While we have identified conditions on γ^{fp} and γ^{fn} for which the optimal solution will consist of two separate thresholds, we also seek to determine conditions under which it is optimal to have a single threshold, i.e., $u^* = l^*$. We specify these conditions in Proposition 4.3.

Proposition 4.3 (One-threshold Solutions)**.** *If γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) \leq l(\gamma^{fn})$, then the optimal solution to TTP* consists of a single threshold, i.e., $u^* = l^*$. Furthermore,*

(a) if $\phi > 0.5$, the optimal threshold will be given by $t^* = u^* = l^* = u(\gamma^{fp})$,

(b) if $\phi < 0.5$, the optimal threshold will be given by $t^* = u^* = l^* = l(\gamma^{fn})$, and

(c) if $\phi = 0.5$, every $t \in [u(\gamma^{fp}), l(\gamma^{fn})]$ is optimal.

Proof of Proposition 4.3. There exists an optimal solution to TTP* which is also an extreme point of its polyhedral feasible region, denoted by Y, as described in (4.5). As $u(\gamma^{fp}) \leq l(\gamma^{fn})$, the extreme points of Y are $(u(\gamma^{fp}), u(\gamma^{fp}))$, $(l(\gamma^{fn}), l(\gamma^{fn}))$, $(u(\gamma^{fp}), 0)$, $(1, l(\gamma^{fn}))$, and $(1, 0)$. If $\phi \in (0.5, 1)$, then $(u(\gamma^{fp}), u(\gamma^{fp}))$ results in the best objective function value of all these extreme points and hence is optimal to TTP*, because $\psi(x)$ is decreasing in u and increasing in l. Similarly, if $\phi \in (0, 0.5)$, then extreme point $(l(\gamma^{fn}), l(\gamma^{fn}))$ is the optimal solution. If $\phi = 0.5$, then extreme points $(u(\gamma^{fp}), u(\gamma^{fp}))$ and $(l(\gamma^{fn}), l(\gamma^{fn}))$ are equally optimal. It follows that (t, t) is optimal to TTP* if and only if $t \in [u(\gamma^{fp}), l(\gamma^{fn})]$. □

The implication of Proposition 4.3 is that if at least one of the bounds γ^{fp} or γ^{fn} is "loose" enough, a one-threshold solution is better than a two-threshold solution. Therefore, if it is desired to constrain just one of false-positive or false-negative rates, then a one-threshold solution will be at least as good as two-threshold solutions. Alternatively, this result can be interpreted to mean that using one decision threshold will not be optimal if aiming to keep both false-positive and false-negative rates low while maximizing sensitivity and specificity. Furthermore, Proposition 4.3 implies that in parameterizing TTP*, one only needs to consider the three choices of ϕ described. This characterization is extremely useful in a practical sense, since choosing appropriate values for parameters is often challenging when implementing such models in practice.

Based on the results established in Propositions 4.2 and 4.3, we can relate the optimal solutions of TTP* and TTP based on the parameter choices.

Theorem 4.1 (Relation Between TTP* and TTP Optimal Solutions)**.** *Suppose that γ^{fp} and γ^{fn} are chosen identically for TTP* and TTP.*

1. *If γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) > l(\gamma^{fn})$, then, for any $\phi \in (0, 1)$, the optimal solution to TTP* is also optimal in TTP for any $\lambda \in (0, 1)$.*

2. *If γ^{fp} and γ^{fn} are chosen such that $u(\gamma^{fp}) \leq l(\gamma^{fn})$, then the following statements hold.*

   a) If $\phi > 0.5$, then the optimal solution to TTP is optimal in TTP with $\lambda = 1$,*

   b) If $\phi < 0.5$, then the optimal solution to TTP is optimal in TTP with $\lambda = 0$, and*

c) For any $\lambda \in (0,1)$, there exists a one-threshold optimal solution to TTP. Furthermore any one-threshold optimal solution to TTP is optimal in TTP* with $\phi = 0.5$.

Proof. **Statement 1:** From Proposition 4.2, the optimal solution to TTP* is given by $(u(\gamma^{fp}), l(\gamma^{fn}))$ for any $\phi \in (0,1)$. Since the feasible region of TTP is identical to that of TTP*, $(u(\gamma^{fp}), l(\gamma^{fn}))$ is feasible in TTP. By the monotonicity of $\mathbb{E}[se(u,\xi^+)]$ in u, $\mathbb{E}[se(u(\gamma^{fp}),\xi^+)] \geq \mathbb{E}[se(u,\xi^+)]$ for any $u \geq u(\gamma^{fp})$. Similarly, $\mathbb{E}[sp(l(\gamma^{fn}),\xi^-)] \geq \mathbb{E}[sp(l,\xi^-)]$ for any $l \leq l(\gamma^{fn})$. Thus,

$$\lambda\mathbb{E}[se(u(\gamma^{fp}),\xi^+)] + (1-\lambda)\mathbb{E}[sp(l(\gamma^{fn}),\xi^-)] \geq \lambda\mathbb{E}[se(u,\xi^+)] + (1-\lambda)\mathbb{E}[sp(l,\xi^-)],$$

for any $\lambda \in (0,1)$ and $(u,l) \in Y$, where Y is the polyhedral feasible region defined in (4.5). Therefore, the optimal solution to TTP* is also optimal in TTP for any $\lambda \in (0,1)$.

Statement 2: We begin by showing 2a. From Proposition 4.3, the optimal solution to TTP* is given by $(u(\gamma^{fp}), u(\gamma^{fp}))$. Suppose that there exists another feasible solution (u', l'), to TTP which achieves greater objective function value. Since $\lambda = 1$, it must be the case that $\mathbb{E}[se(u',\xi^+)] > \mathbb{E}[se(u(\gamma^{fp}),\xi^+)]$. Yet, by the feasibility of (u', l'), it must also be true that $u' \geq u(\gamma^{fp})$, which implies that $\mathbb{E}[se(u(\gamma^{fp}),\xi^+)] \geq \mathbb{E}[se(u',\xi^+)]$ by the monotonicity of $\mathbb{E}[se(u,\xi^+)]$ in u. We have reached a contradiction, so the optimal solution to TTP* is also optimal in TTP. The proof for 2b is similar to that of 2a and has been omitted.

To show 2c, we first establish the existence of a one-threshold optimal solution. Let (u', l') be any two-threshold solution to TTP. Then, since $u(\gamma^{fp}) \leq l(\gamma^{fn})$, the one-threshold solution $(\bar{l}, \bar{l})$ is also feasible for TTP, where $\bar{l} := \max\{l', u(\gamma^{fp})\}$. Notice that by definition, either $\bar{l} > l'$ or $\bar{l} < u'$, or both are true. By the monotonicity of $\mathbb{E}[sp(l,\xi^-)]$ in l and $\mathbb{E}[se(u,\xi^+)]$ in u, $(\bar{l}, \bar{l})$ achieves at least as high an objective function value as (u', l'). Hence, there exists an optimal one-threshold solution to TTP. Now, let (t,t) be any one-threshold optimal solution to TTP. Since (t,t) is feasible, it must be that $u(\gamma^{fp}) \leq t \leq l(\gamma^{fn})$. By Proposition 4.3, this solution is optimal to TTP* for $\phi = 0.5$, completing the proof. □

From Theorem 4.1, we find that the optimal solution to TTP* is optimal for TTP in the two-threshold case, regardless of how the parameters ϕ and λ are chosen. Specifically, for two-threshold solutions, the optimal thresholds to TTP* at some initial value of ϕ are already

tight on their constraints (4.1b)-(4.1c). Since changing the value of ϕ will not change the direction of improvement for the objective function (4.2), the optimal thresholds will also remain unchanged. Therefore, the optimal solution to TTP* does not depend on the value of ϕ. Since the optimal solution to TTP* is optimal in TTP (regardless of λ) for the two-threshold solution case, the optimal solution also does not depend on λ.

However, the one-threshold case for TTP* does not translate well to TTP. To derive stronger results for the one-threshold case, we need additional assumptions regarding the form of the objective function (4.1a). For instance, if (4.1a) is monotone over the interval $[u(\gamma^{fp}), l(\gamma^{fn})]$, then the optimal solution will coincide with one of the endpoints. Unfortunately, the form of (4.1a) is difficult to know *a priori* and it may be the case that (4.1a) has many maxima over $[u(\gamma^{fp}), l(\gamma^{fn})]$. However, if it is known that TTP has a one-threshold solution, then it suffices to use existing one-threshold methods (see Section 4.2). Since such instances are well-studied, we emphasize our analysis of TTP over two-threshold solutions.

Propositions 4.2 and 4.3 characterize the optimal solution to TTP* based on ϕ, γ^{fp}, and γ^{fn}. Specifically, the optimal thresholds can be constructed if $u(\gamma^{fp})$ and $l(\gamma^{fn})$ are known or estimated. Practically speaking, this result implies that one can solve TTP without sophisticated optimization software by using quantile estimation methods instead (see Section 4.4.1).

4.3.5 Utility-based Frameworks

Our formulation of TTP does not take into account any utilities because we aimed to determine a method which generalizes well across many applications, including those for which utilities are poorly estimated. However, utilities are an important component of decision-theoretic frameworks. In this section, we extend our analysis and analyze two utility-based frameworks for identifying diagnosis decision thresholds. Specifically, in Section 4.3.5, we provide a general set of conditions which ensure that the analytical results for TTP (as derived in Section 4.3.4) still hold. Based on these conditions, we provide a specific utility-based framework for TTP. Then, in Section 4.3.5, we study a utility-based objective function which does not fit the conditions provided in Section 4.3.5. Nevertheless, we derive conditions which must hold such that a two-threshold solution (e.g., one obtained by solving TTP*) is better than a fixed one-threshold solution.

Utility-based Two-Threshold Problem

In this section, we identify a general class of objective functions and constraints which, when used in the TTP framework, ensure that the structural results from Section 4.3.4 still hold.

Remark 4.1. *The results shown in Section 4.3.4 hold for general optimization problems of the form*

$$\begin{aligned} \max_{u,l} \quad & g(\mathbb{E}[se(u,\xi^+)], \mathbb{E}[sp(l,\xi^-)]) \\ s.t. \quad & h_1(\mathbb{E}[fp(u,\xi^-)]) \leq b_1 \\ & h_2(\mathbb{E}[fn(l,\xi^+)]) \leq b_2 \\ & 0 \leq l \leq u \leq 1, \end{aligned}$$

where $g(\cdot)$ is non-increasing in u and non-decreasing in l over the feasible region, $h_1(\cdot)$ is non-increasing in u, $h_2(\cdot)$ is non-decreasing in l, and $b_1, b_2 > 0$.

Through Remark 4.1, we now show how to incorporate utilities into the TTP framework. Let

- q be the proportion of true-positives in the population,
- $r^+ > 0$ and $r^- > 0$ be the utilities associated with correctly classifying true-positives and true-negatives, respectively,
- $c^+ > 0$ and $c^- > 0$ be the costs associated with false-negatives and false-positives, respectively, and
- $b^+ > 0$ and $b^- > 0$ be the maximum allowable costs associated with false-negatives and false-positives, respectively.

Examples of utilities for r^+, r^-, c^+ and c^- include quality-adjusted life-years — a common utility measure used in medical decision-making and public health. Now, we can formulate

the Utility-based Two-Threshold Problem (UTTP) as:

$$\begin{aligned}(\text{UTTP}) \quad \max_{u,l} \quad & r^+ q\mathbb{E}[se(u,\xi^+)] + r^-(1-q)\mathbb{E}[sp(l,\xi^-)] \\ \text{s.t.} \quad & c^-(1-q)\mathbb{E}[fp(u,\xi^-)] \leq b^- \\ & c^+ q\mathbb{E}[fn(l,\xi^+)] \leq b^+ \\ & 0 \leq l \leq u \leq 1.\end{aligned}$$

UTTP aims to determine thresholds u and l which maximize the expected utility associated with correct classification without exceeding expected misclassification costs. Since the monotonicity of the objective function and constraints for UTTP match the monotonicity for TTP, Theorem 4.1 applies to UTTP.

Utility-based Cost Function

In this section, we consider a more general objective function than the one considered in Section 4.3.5. For a fixed set of thresholds (u, l), define

$$\begin{aligned}p^+(u,l) &:= \mathbb{P}(\text{Correct Classification}|u,l) = \mathbb{E}[se(u,\xi^+)]q + \mathbb{E}[sp(l,\xi^-)](1-q) \\ p^-(u,l) &:= \mathbb{P}(\text{Misclassification}|u,l) = \mathbb{E}[fn(l,\xi^+)]q + \mathbb{E}[fp(u,\xi^-)](1-q) \\ p^D(u,l) &:= \mathbb{P}(\text{Defer}|u,l) = 1 - p^+(u,l) - p^-(u,l),\end{aligned}$$

where q specifies the proportion of people who are true-positives in the population subject to diagnosis. Furthermore, consider three utilities c^+, c^-, and c^D associated with a correct classification, misclassification, and deferred diagnosis decision, respectively. We assume that $c^+ \geq c^D \geq c^-$, i.e., it is better to correctly classify a patient than to defer a patient and it is better to defer a patient than to misclassify a patient. Given this ordering of utilities, we can express c^D as $c^D = \delta c^- + (1-\delta)c^+$ for some $\delta \in [0,1]$. Then, we can define an expected utility function of the form

$$J(u,l) = p^+(u,l)c^+ + p^-(u,l)c^- + p^D(u,l)c^D. \tag{4.6}$$

In general, $J(u,l)$ does not satisfy the conditions in Remark 4.1. More specifically, $J(u,l)$ is not generally non-increasing in u and non-decreasing in l. We note that (4.6) is nearly

identical to the regret-based framework in Section 5.1 of Shapiro, 1999, except we include a utility associated with deferred diagnosis decisions. With the function (4.6), one may be interested in comparing a fixed two-threshold solution (u, l) with a one-threshold solution (t, t). In particular, how high must the utility c^D be for (u, l) to be a better solution than (t, t)? We provide intuitive conditions to answer this very question in Proposition 4.4.

Proposition 4.4. *Consider a one-threshold solution (t, t) and a two-threshold solution (u, l) such that $p^D(u, l) > 0$. Then, $J(u, l) \geq J(t, t)$ if and only if*

$$\delta \leq \frac{p^-(t,t) - p^-(u,l)}{p^D(u,l)}. \tag{4.7}$$

Proof. We will first show the " $\Longrightarrow$ " direction. Suppose that δ satisfies (4.7). Then, we have

$$\begin{aligned}
J(u,l) &= p^+(u,l)c^+ + p^-(u,l)c^- + p^D(u,l)c^D \\
&= p^+(u,l)c^+ + p^-(u,l)c^- + p^D(u,l)(\delta c^- + (1-\delta)c^+) \\
&= \left(p^+(u,l) + p^D(u,l)\right)c^+ + p^-(u,l)c^- + \delta p^D(u,l)(c^- - c^+) \\
&\geq \left(p^+(u,l) + p^D(u,l)\right)c^+ + p^-(u,l)c^- + (p^-(t,t) - p^-(u,l))(c^- - c^+) \qquad (4.8) \\
&= \left(p^+(u,l) + p^D(u,l) + p^-(u,l) - p^-(t,t)\right)c^+ + p^-(t,t)c^- \\
&= (1 - p^-(t,t))c^+ + p^-(t,t)c^- \qquad (4.9) \\
&= p^+(t,t)c^+ + p^-(t,t)c^- \qquad (4.10) \\
&= J(t,t),
\end{aligned}$$

where (4.8) comes from the fact that $c^+ \geq c^-$ and δ satisfies (4.7), (4.9) comes from the fact that $p^+(u, l)+p^D(u, l)+p^-(u, l) = 1$, and (4.10) uses the fact that $p^+(t, t)+p^-(t, t) = 1$ since a one-threshold solution has no deferred decisions. Thus, we have shown that $J(u, l) \geq J(t, t)$.

We will now show that " $\Longleftarrow$ " direction. Suppose that $J(u, l) \geq J(t, t)$. Then,

$$p^+(u,l)c^+ + p^-(u,l)c^- + p^D(u,l)c^D \geq p^+(t,t)c^+ + p^-(t,t)c^-.$$

By expanding $p^D(u, l)$ and c^D, rearranging terms, using the definition $p^D(u, l)$, and using the

fact that $p^+(t,t) + p^-(t,t) = 1$, we have

$$\begin{aligned}
\delta p^D(u,l)(c^- - c^+) &\geq [p^+(t,t) - p^+(u,l) - p^D(u,l)]c^+ + [p^-(t,t) - p^-(u,l)]c^- \qquad (4.11)\\
&= [p^+(t,t) - (1 - p^-(u,l))]c^+ + [p^-(t,t) - p^-(u,l)]c^-\\
&= [p^-(t,t) - p^-(u,l)](-c^+) + [p^-(t,t) - p^-(u,l)]c^-\\
&= [p^-(t,t) - p^-(u,l)](c^- - c^+). \qquad (4.12)
\end{aligned}$$

If $c^+ = c^-$, the inequality holds trivially. If $c^+ > c^-$, the desired inequality is shown by dividing the LHS of (4.11) and (4.12) by $p^D(u,l)(c^- - c^+)$. □

Proposition 4.4 provides insights in the trade-off between two-threshold and one-threshold solutions. Let us first examine the expression on the right-hand side of (4.7). Note that the numerator of (4.7) specifies the misclassifications which are saved by switching from (t,t) to (u,l) and the denominator specifies the overall probability of being deferred under (u,l). Thus, we can interpret the right-hand side of (4.7) as the proportion of deferred decisions under (u,l) which consist of would-be misclassifications under (t,t). Therefore, Proposition 4.4 tells us that the utility associated with deferred decisions c^D can only be as low as the proportion of saved misclassifications by switching from (t,t) to (u,l). Furthermore, if we interpret δ to be the expected proportion of deferred decisions which end up being misclassified, Proposition 4.4 tells us that it is better to switch from (t,t) to (u,l) if the expected proportion of misclassifications after a deferred decision is no more than the misclassifications saved by switching from (t,t) to (u,l). To summarize, this analysis shows that two-threshold solutions are particularly useful when either: (i) the utility c^D is relatively high compared to c^+ and c^-, (ii) if most of deferred decisions consist of would-be misclassifications, or (iii) if ultimately, few deferred decisions end up being misclassified.

4.3.6 Extensions to Multi-class Diagnosis

In our formulation of TTP, we focused on the case in which the population consists of two subpopulations. In this section, we extend our modeling framework to handle the case when there are three or more subpopulations, i.e., multi-class diagnosis. We specifically focus on the cases of (1) multi-label classification and (2) ordinal classification.

Multi-label Classification via Binary Relevance

In this section, we extend TTP to address multi-label classification problems through the Binary Relevance (BR) method. In multi-label classification through BR, one aims to determine which labels should be assigned to a patient given an estimate for the likelihood that the patient should be assigned each label. Examples of this problem include determining which concussion subtypes a patient may have based on a multi-dimensional clinical assessment (Collins et al., 2014; Maruta, Lumba-Brown, and Ghajar, 2018) or determining which chronic diseases a person may have based on electronic health record information (Zufferey et al., 2015).

The specific problem setting is as follows. The patient population consists of patients who may belong to any subset of $K \geq 2$ classes. Hence, in this patient population, there are 2^K mutually exclusive subpopulations. A randomly drawn patient from this population is associated with random vectors (X, Y), where $X \in \mathcal{X}$ consists of patient characteristics (and $\mathcal{X}$ is the set of patient characteristics) and $Y \in \{0,1\}^K$ is a vector where the k^{th} entry y_k is equal to 1 if the patient is associated with class k and 0 otherwise. Furthermore, there are K risk estimation models $f_1, ..., f_K$ where each risk estimation model $f_k(X) : \mathcal{X} \to (0,1)$ approximates $\mathbb{P}(y_k = 1|X)$. Examples of such models include multinomial regression, multi-class support vector machines, and multi-class perceptrons. The BR method associates a patient with class label k if $f_k(X) \geq t_k$, where $t_k \in [0,1]$ is a class k decision threshold.

Note that BR treats risk estimates for each class label as if they were independent of risk estimates from all other class labels. This simplifying assumption is a known limitation to the BR framework (Zhang et al., 2018), though it has been shown to outperform more intricate modeling approaches for some applications (Zufferey et al., 2015). To this end, ensemble learning methods have been developed which account for dependence between labels in the creation of risk estimation models $f_1, ..., f_k$ (Godbole and Sarawagi, 2004; Read et al., 2011; Zhang and Zhang, 2010). These existing methods may be applied prior to determining each of the k decision thresholds.

The performance of a set of thresholds $t_1, ..., t_K$ can be evaluated similarly to the binary classification problem which is the focus of this manuscript. For brevity, we denote class k

risk estimates as

$$\xi_k^+ := f_k(X|y_k = 1) \text{ and}$$
$$\xi_k^- := f_k(X|y_k = 0),$$

for each $k = 1, ..., K$. Given a patient with label $y_k = 1$ and some threshold t_k, we define the functions

$$se_k(t_k, \xi_k^+) := \mathbb{1}\{\xi_k^+ \geq t_k\}$$
$$fn_k(t_k, \xi_k^+) := \mathbb{1}\{\xi_k^+ < t_k\}$$

as class k sensitivity and false-negative, respectively. Similarly, for patients with $y_k = 0$ and a given threshold t_k, we define

$$sp_k(t_k, \xi_k^-) := \mathbb{1}\{\xi_k^- \leq t_k\}$$
$$fp_k(t_k, \xi_k^-) := \mathbb{1}\{\xi_k^- > t_k\}$$

as class k sensitivity and false-positive, respectively.

In our extension to BR, we consider an upper threshold u_k and lower threshold l_k such that a risk estimate above u_k labels a patient as class k and below l_k does not label a patient as class k. For patients with risk estimates between u_k and l_k, their association with label k is inconclusive. For each class k, we aim to determine the thresholds u_k and l_k by maximizing the expected class k sensitivity and specificity, respectively, while restricting the expected false-positive and false-negative rates. This problem can be represented with the Multi-label Two Threshold Problem (MTTP), which we define as

$$\text{(MTTP)} \quad \max_{u_1,...,u_K,l_1,...,l_K} \quad \sum_{k=1}^{K} \left(\lambda_k \mathbb{E}[se_k(u_k, \xi_k^+)] + (1 - \lambda_k)\mathbb{E}[sp_k(l_k, \xi_k^-)] \right) \tag{4.13a}$$

$$\text{s.t.} \quad \mathbb{E}[fp_k(u_k, \xi_k^-)] \leq \gamma_k^{fp} \quad \text{for all } k = 1, ..., K \tag{4.13b}$$

$$\mathbb{E}[fn_k(l_k, \xi_k^+)] \leq \gamma_k^{fn} \quad \text{for all } k = 1, ..., K \tag{4.13c}$$

$$0 \leq l_k \leq u_k \leq 1, \text{ for } k = 1, ..., K, \tag{4.13d}$$

where the parameters $\lambda_k \in (0,1)$ indicate the preference for sensitivity over specificity in class k for all $k = 1, ..., K$. If we take λ_k to be the proportion of patients with $y_k = 1$ in the overall population, then the objective function (4.13a) can be interpreted as maximizing the labeling accuracy.

Following the BR framework, the thresholds in each class k are treated independently of thresholds in class $j \neq k$. Through this simplification, one could decompose MTTP into solving TTP K times. Therefore, the results derived for TTP can be applied directly to MTTP. Specifically, we show in Corollary 4.1 that the optimal solution can be written in terms of the following quantiles:

$$u_k(\gamma_k^{fp}) := \min\{u : \mathbb{E}[fp(u, \xi_k^-)] \leq \gamma_k^{fp}\} \text{ for all } k = 1, ..., K$$
$$l_k(\gamma_k^{fn}) := \max\{l : \mathbb{E}[fn(l, \xi_k^k)] \leq \gamma_k^{fn}\} \text{ for all } k = 1, ..., K.$$

Corollary 4.1. *If $u_k(\gamma_k^{fp}) > l_k(\gamma_k^{fn})$ for all $k = 1, ..., K$, the optimal solution to MTTP is given by $u_k^* = u_k(\gamma_k^{fp})$ and $l_k^* = l_k(\gamma_k^{fn})$ for all $k = 1, ..., K$.*

Proof. Notice that since the thresholds (u_k, l_k) are not coupled with any other set of thresholds (u_j, l_j), $j \neq k$, then MTTP can be decomposed into K TTP problems, where TTP_k denotes TTP for the k^{th} class and is defined as follows:

$$\begin{aligned}
(\text{TTP}_k) \quad \max_{u_k, l_k} \quad & \lambda_k \mathbb{E}[se(u_k, \xi_k^+)] + (1 - \lambda_k)\mathbb{E}[sp(l_k, \xi^-)] \\
\text{s.t.} \quad & \mathbb{E}[fp_k(u_k, \xi_k^-)] \leq \gamma_k^{fp} \\
& \mathbb{E}[fn_k(l_k, \xi_k^+)] \leq \gamma_k^{fn} \\
& 0 \leq l_k \leq u_k \leq 1.
\end{aligned}$$

Now, since TTP_k follows the form of TTP exactly, it follows from Theorem 4.1 that if $u_k(\gamma_k^{fp}) > l_k(\gamma_k^{fn})$, then the optimal solution is given by $\left(u_k(\gamma_k^{fp}), l_k(\gamma_k^{fn})\right)$. Hence, the proof is complete. □

Remark 4.2. *If it were the case that each patient belongs to only 1 of the K classes, then this classification problem is no longer a multi-label classification problem. Instead, it is a multi-class classification problem. To this end, the One-versus-All approach to multi-class classification is similar to the BR approach studied in this section. Specifically, both methods*

create K risk estimation models $f_1, ..., f_K$ with one model for each class. The key distinction between these two methods is that in the One-versus-All framework, each patient may only be assigned a single label. For example, one may wish to classify which type of skin lesion a patient has among 7 different possibilities (Tschandl et al., 2019). While MTTP could certainly be applied directly to this related problem, it may not be well-suited for this scenario since there exists the possibility that MTTP recommends a patient for more than one class. This result would be difficult to interpret and could limit its application in practice.

Ordinal Classification

In this section, we present a model similar to TTP which can be used to identify thresholds for ordinal classification problems. Ordinal classification aims to determine which severity level a patient belongs to for a certain disease. Currently, there is not consensus on severity ratings for concussion, though one may try to predict whether a patient has no concussion, a concussion with normal recovery, or a concussion with pro-longed recovery (Lau et al., 2011). Another example of ordinal classification is in predicting a patient's degree of recovery six months after traumatic brain injury (Roozenbeek et al., 2011).

In this problem setting, there are $K \geq 2$ mutually exclusive patient classes ordered by severity. That is, class 1 is the least severe class and class K is the most severe. A randomly chosen patient is associated with the tuple (X, Y), where $X \in \mathcal{X}$ is a random vector of patient characteristics from the set of patient characteristics $\mathcal{X}$ and the random variable $Y \in \{1, ..., K\}$ describes his or her class. Rather than directly estimating Y based on patient characteristics X, ordinal classification methods estimate a continuous latent variable $\xi \in [0, 1]$. Then, given $K - 1$ ordered thresholds $t_1, ..., t_{K-1}$ such that $t_0 := 0 \leq t_1 \leq ... \leq t_{K-1} \leq t_K := 1$, a patient is classified as being in class k if $t_{k-1} \leq \xi \leq t_k$. Specifically, we assume that there is a severity score model $f : \mathcal{X} \to [0, 1]$ which estimates the latent variable ξ based on patient characteristics X. Such models include ordinal logistic regression and ordinal support vector machines. Throughout the remainder of this section, we define $\xi^k := f(X|Y = k)$ to be the severity score for a random patient from class $Y = k$ with characteristics X.

We aim to identify the thresholds $t_1, ..., t_{K-1}$ which correctly classify as many patients as possible into the correct severity class while limiting the number of patients misclassified into lower and higher severity classes. Using the same functions defined in Section 4.3.1, we

define the Ordinal Threshold Problem (OTP) as follows:

$$\text{(OTP)} \quad \max_{t_1,...,t_{K-1}} \quad \sum_{k=0}^{K-1} \lambda_{k+1} \mathbb{E}[se(t_k, \xi^{k+1})] \tag{4.14a}$$

$$\text{s.t.} \quad \mathbb{E}[fp(t_k, \xi^j)] \leq \gamma_{k,j}^{fp} \text{ for all } j \leq k, k = 1, ..., K-1 \tag{4.14b}$$

$$\mathbb{E}[fn(t_k, \xi^j)] \leq \gamma_{k,j}^{fn} \text{ for all } j > k, k = 1, ..., K-1 \tag{4.14c}$$

$$t_k \leq t_{k+1} \text{ for all } k = 1, ..., K-2, \tag{4.14d}$$

where $t_0 = 0$ and the weight parameters $\lambda_k \in (0, 1)$ satisfy $\sum_{k=1}^{K} \lambda_k = 1$. These weight parameters can be interpreted as the relative importance of correctly classifying one class to the others. In the objective function, we focus on modeling sensitivity — rather than combining sensitivity and specificity — since there is generally more urgency in correctly identifying conditions which are more severe compared to those which are less severe. For example, this urgency may arise in classification of traumatic brain injury since the health-related costs (resp., quality of life) for patients with mild or moderate traumatic brain injury are estimated to be less (resp., greater) compared to patients with severe traumatic brain injury (Andelic et al., 2009; Humphreys et al., 2013). The constraint parameters $\gamma_{k,j}^{fp}$ (resp., $\gamma_{k,j}^{fn}$) specify the maximum expected false-positive rate (resp., false-negative rate) for class j against threshold t_k. We assume that for each $k = 1, ..., K$, the constraint parameters are chosen such that $\gamma_{k+1,j}^{fp} \leq \gamma_{k,j}^{fp}$ for all $j \leq k$ and $\gamma_{k+1,j}^{fn} \geq \gamma_{k,j}^{fn}$ for all $j > k$. If these parameters are not chosen in this way, then OTP is guaranteed to be infeasible.

We now show that, much like TTP, the optimal solution to OTP can be defined in terms of quantiles. Specifically, for all $k = 1, ..., K$, we define the following quantiles:

$$\underline{t}^k(\gamma^{fp}) := \min\{t : \mathbb{E}[fp(t, \xi^k)] \leq \gamma^{fp}\}$$
$$\bar{t}^k(\gamma^{fn}) := \max\{t : \mathbb{E}[fn(t, \xi^k)] \leq \gamma^{fn}\}.$$

For ease of notation, we also define

$$\underline{t}_k := \max_{j \leq k} \underline{t}^j(\gamma_{k,j}^{fp}) \tag{4.15}$$

$$\bar{t}_k := \min_{j > k} \bar{t}^j(\gamma_{k,j}^{fn}), \tag{4.16}$$

for all $k = 1, ..., K-1$. Notice that $\bar{t}_k \leq \bar{t}_{k+1}$ and $\underline{t}_k \geq \underline{t}_{k+1}$ based on how the parameters $\gamma^{fp}_{k,j}$ and $\gamma^{fn}_{k,j}$ need to be chosen.

Now, we show that if OTP is feasible, the optimal solution to OTP can be stated in terms of the quantiles defined in (4.15)-(4.16).

Theorem 4.2. *OTP is feasible if and only if $\underline{t}_k \leq \bar{t}_k$ for all $k = 1, ..., K$. Furthermore, if OTP is feasible, then the optimal solution is given by $t^*_k = \underline{t}_k$ for all $k = 1, ..., K-1$.*

Proof. We break this proof into two parts. In the first part, we show that OTP is feasible if and only if $\underline{t}_k \leq \bar{t}_k$ for all $k = 1, ..., K$. Then, in the second part, we show that if OTP is feasible, then the optimal solution is given by $t^*_k = \underline{t}_k$ for all $k = 1, ..., K$.

Part 1: " $\Longrightarrow$ ": Consider any $t_k \geq \underline{t}_k$. It follows that

$$\begin{aligned} t_k \geq \underline{t}_k &= \max_{j \leq k} \underline{t}^j(\gamma^{fp}_{k,j}) \\ &\geq \underline{t}^j(\gamma^{fp}_{k,j}) \text{ for all } j \leq k \\ \iff \mathbb{E}[fp(t_k, \xi^j)] &\leq \gamma^{fp}_{k,j} \text{ for all } j \leq k. \end{aligned}$$

Similarly, any $t_k \leq \bar{t}_k$ implies that $\mathbb{E}[fn(t_k, \xi^j)] \leq \gamma^{fn}_{k,j}$ for all $j > k$. Therefore, if $\underline{t}_k \leq \bar{t}_k$, we can pick any $t_k \in [\max_{j \leq k} \underline{t}_j, \bar{t}_k]$ and it will satisfy constraints (4.14b)-(4.14c). Furthermore, since we assume that $\gamma^{fp}_{k,j} \geq \gamma^{fp}_{k+1,j}$ for all $j \leq k$, it follows that $\underline{t}_{k+1} \geq \underline{t}_k$ for all $k = 1, ..., K-1$ and therefore, $t_{k+1} \geq t_k$ for all $k = 1, ..., K-1$ implying that constraint (4.14d) is satisfied. Thus, OTP is feasible.

" $\Longleftarrow$ ": For any feasible solution to OTP, consider any threshold t_k. It follows that $\mathbb{E}[fp(t_k, \xi^j)] \leq \gamma^{fp}_{k,j}$ for all $j \leq k$, implying that $t_k \geq \underline{t}_k$. Similarly, $\mathbb{E}[fn(t_k, \xi^j)] \leq \gamma^{fn}_{k,j}$ for all $j > k$. Therefore, $\underline{t}_k \leq t_k \leq \bar{t}_k$ implying that $\underline{t}_k \leq \bar{t}_k$ for all $k = 1, ..., K-1$.

Part 2:

Consider any feasible solution $(t_1, ..., t_K)$. From Part 1, it follows that $\underline{t}_k \leq t_k \leq \bar{t}_k$ for all $k = 1, ..., K$. Since the function $\mathbb{E}[se(t, \xi^k)]$ is decreasing in t for all $k = 1, ..., K$, it follows that

$$\sum_{k=0}^{K-1} \lambda_{k+1} \mathbb{E}[se(\underline{t}_k, \xi^{k+1})] \geq \sum_{k=0}^{K-1} \lambda_{k+1} \mathbb{E}[se(t_k, \xi^{k+1})].$$

Since the set of thresholds $(\underline{t}_1, ..., \underline{t}_{K-1})$ is feasible for OTP, it must also be optimal. □

Theorem 4.2 tells us that the optimal solution to OTP can be stated in terms of quantiles

and therefore, computed via quantile estimation. Furthermore, OTP is guaranteed to be feasible if we set $\gamma_{k,j}^{fn} = 1$ for all $j > k$. Thus, this formulation of OTP actually emphasizes constraints on the false-positive rate more than it does the false-negative rate. However, one could also formulate OTP based on maximizing specificity with more emphasis on constraining false-negative rates. Specifically, if (4.14a) was replaced with

$$\max_{t_1,...,t_K} \quad \sum_{k=1}^{K} \lambda_k \mathbb{E}[sp(t_k, \xi^k)],$$

where $t_K = 1$, then it is straightforward to show that the feasibility result in Theorem 4.2 holds and the optimal solution to this new problem is given by $t_k^* = \bar{t}_k$ for $k = 1, ..., K-1$. That is, with this alternative objective function, the constraints on false-negative rates are emphasized. Using the results from Theorem 4.2 to develop solutions for OTP, we use simulation to assess the performance of OTP in Section 4.5.4.

While restricting the objective function to containing only sensitivity (or specificity) was motivated by practical reasons, it serves a technical purpose as well. Specifically, the objective function would no longer be monotone in the thresholds $t_1, ..., t_K$ if both sensitivity and specificity were included. As a result, little can be said about the form taken by the optimal solution to OTP without making assumptions about the distributions of risk estimates $\xi^1, ..., \xi^K$. To this end, finding ways to incorporate both sensitivity and specificity in OTP would be an interesting direction for future research.

4.4 Data-driven Solution Methods

In Section 4.3.4, we showed important analytical properties which relate TTP* and TTP. In particular, TTP* provides an optimal solution to TTP if the optimal solution is a two-threshold solution. However, solving TTP* is not straightforward since obtaining the functions (4.1b) and (4.1c) may not be possible, even if the distribution $\mathcal{P}$ or $\mathcal{P}^+, \mathcal{P}^-$ are known exactly. On the other hand, data may be available which can be used to estimate (4.1b) and (4.1c). In this section, we propose two data-driven methods for solving TTP* tractably: quantile estimation and distributionally robust optimization.

In the remainder of this section, we denote a data sample of size N as $\hat{P}_N = \{\hat{\xi}_1, .., \hat{\xi}_N\}$.

Let the set $\hat{P}_N^+ = \{\hat{\xi} \in \hat{P}_N : \text{true-positive}\} = \{\hat{\xi}_1^+, ..., \hat{\xi}_{N^+}^+\}$ be the part of the data sample consisting of true-positives. Similarly, let $\hat{P}_N^- = \{\xi \in \hat{P}_N : \text{true-negative}\} = \{\hat{\xi}_1^-, ..., \hat{\xi}_{N^-}^-\}$ be the true-negatives. The sets $\hat{P}_N^+$ and $\hat{P}_N^-$ are mutually exclusive and $\hat{P}_N = \hat{P}_N^+ \cup \hat{P}_N^-$, so $N^+ + N^- = N$.

4.4.1 Quantile Estimation

In Section 4.3.4, we have shown that the optimal solution to TTP* is characterized by the quantiles $u(\gamma^{fp})$ and $l(\gamma^{fn})$. Thus, we propose to solve TTP* using Harrell-Davis quantile estimation (Harrell and Davis, 1982), as it has been shown to outperform the standard quantile estimation method across various distributions. We also performed auxiliary analyses and found that the Harrell-Davis quantile estimation method outperformed standard sample average approximation techniques. Once we have estimated $u(\gamma^{fp})$ and $l(\gamma^{fn})$, we apply Propositions 4.2 and 4.3 to estimate the optimal solution to TTP*. We refer to this solution as TTP*-Q.

Remark 4.3. *Different application needs may call for different quantile estimation methods. For instance, if γ^{fp} or γ^{fn} are set very close to* 0*, it may be more appropriate to use a method for estimating extreme quantiles (Daníelsson and Vries, 1997).*

An important factor in obtaining high out-of-sample feasibility for TTP*-Q is the choice of constraint parameters γ^{fp} and γ^{fn}. Specifically, solving TTP*-Q with constraint parameters $\gamma_\tau^{fp} < \gamma^{fp}$ and $\gamma_\tau^{fn} < \gamma^{fn}$ can improve its out-of-sample feasibility. However, choosing values of γ_τ^{fp} and γ_τ^{fn} which are too small can result in overly conservative thresholds. In Section 4.4.3, we describe how the values of γ_τ^{fp} and γ_τ^{fn} can be calibrated using data.

4.4.2 Data-driven Distributionally Robust Optimization

A drawback to the quantile estimation approach in Section 4.4.1 is that it may require a very large sample size N to sufficiently approximate the distribution $\mathcal{P}$. However, obtaining a large enough sample of data may not be possible, e.g., when data collection is very costly or time-consuming. Hence, we consider the context in which the data samples $\hat{P}_N$ do not sufficiently represent the true population distribution $\mathcal{P}$ and we wish to create thresholds

which are robust to the worst-case differences between $\hat{P}_N$ and $\mathcal{P}$. We consider the following distributionally robust TTP* (TTP*-DR) model:

$$\begin{aligned}
\text{(TTP*-DR)} \quad \min_{u,l} \quad & \phi u - (1-\phi) l \\
\text{s.t.} \quad & \sup_{\mathcal{Q}^- \in D^-} \mathbb{E}_{\mathcal{Q}^-}[fp(u, \xi^-)] \leq \gamma^{fp} && \text{(4.17a)} \\
& \sup_{\mathcal{Q}^+ \in D^+} \mathbb{E}_{\mathcal{Q}^+}[fn(l, \xi^+)] \leq \gamma^{fn} && \text{(4.17b)} \\
& 0 \leq l \leq u \leq 1,
\end{aligned}$$

where D^- (respectively, D^+) represents a family of probability distributions that are plausible candidates of $\mathcal{P}^-$ (respectively, $\mathcal{P}^+$) and is termed the ambiguity set. For TTP*-DR, we propose to construct D^- and D^+ in a data-driven fashion by considering all distributions which are "close" to the data sample $\hat{P}_N$ based on the Wasserstein distance metric. We opt to construct our ambiguity set based on the Wasserstein distance metric since such ambiguity sets have been shown, under mild conditions, to satisfy strong finite sample guarantees, asymptotic consistency, and tractability (Mohajerin Esfahani and Kuhn, 2018).

To define the Wasserstein-based ambiguity set, we must first define some preliminary notation. Let $\mathcal{M}(\Xi)$ represent the set of all probability distributions over the support $\Xi = [0, 1]$, i.e., all distributions which generate risk scores. Then, the Wasserstein distance between any two distributions $\mathcal{Q}_1, \mathcal{Q}_2 \in \mathcal{M}(\Xi)$ is defined by

$$W(\mathcal{Q}_1, \mathcal{Q}_2) := \inf_{\Pi \in \Xi \times \Xi} \left\{ \int_{\Xi^2} |\xi - \xi'| \Pi(d\xi, d\xi') : \begin{array}{l} \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\ \text{with marginals } \mathcal{Q}_1 \text{ and } \mathcal{Q}_2, \text{ respectively,} \end{array} \right\} \tag{4.18}$$

where $\xi, \xi' \in \Xi$. Now, let $\hat{\mathcal{P}}_N$ be an empirical uniform distribution centered on the data sample $\hat{P}_N$. Then, for any Wasserstein radius $\epsilon > 0$, we define our ambiguity set as

$$D_\epsilon(\hat{\mathcal{P}}_N) := \{\mathcal{Q} : W(\mathcal{Q}, \hat{\mathcal{P}}_N) \leq \epsilon\}. \tag{4.19}$$

The ambiguity set (4.19) represents the set of all probability distributions which are within an ϵ distance, based on (4.18), to the data-driven empirical distribution $\hat{\mathcal{P}}_N$. Intuitively,

larger choices of ϵ lead to more robust (i.e., conservative) solutions. Now, let ϵ^+ and ϵ^- denote chosen Wasserstein radii for the data samples $\hat{P}_N^+$ and $\hat{P}_N^-$, respectively. In Section 4.4.3, we detail a cross-validation scheme to calibrate the selection of Wasserstein radii ϵ^+ and ϵ^- based on the data $\hat{P}_N^+$ and $\hat{P}_N^-$. Setting the data-driven ambiguity sets as $D^+ := D_{\epsilon^+}(\hat{\mathcal{P}}_N^+)$ and $D^- := D_{\epsilon^-}(\hat{\mathcal{P}}_N^-)$, TTP*-DR becomes

$$\begin{aligned}
\min_{u,l} \quad & \phi u - (1-\phi) l \\
\text{s.t.} \quad & \sup_{\mathcal{Q} \in D^-} \mathbb{E}_{\mathcal{Q}}[fp(u, \xi^-)] \leq \gamma^{fp} && \text{(4.20a)} \\
& \sup_{\mathcal{Q} \in D^+} \mathbb{E}_{\mathcal{Q}}[fn(l, \xi^+)] \leq \gamma^{fn} && \text{(4.20b)} \\
& 0 \leq l \leq u \leq 1.
\end{aligned}$$

Due to the optimization problems embedded in the constraints (4.20a)-(4.20b), solving TTP*-DR is not straightforward. Furthermore, this optimization occurs over the space of all probability distributions in D^+ and D^-. To this end, we show in Appendix 4.B that the constraints (4.20a) and (4.20b) can be reformulated into tractable problems. Specifically, when the thresholds u and l are fixed, these reformulations are linear programs which scale in size based on the size of the data sample P_N and can be handled efficiently by most modern commercial solvers (e.g., GUROBI, AMPL, and MOSEK). However, since u and l are not fixed in TTP*-DR, the reformulations present bilinear constraints. Nevertheless, since the left-hand sides of (4.20a) and (4.20b) are monotone in u and l, respectively, the optimal values of u and l can still be determined in polynomial time using algorithms such as bisection line search.

4.4.3 Model Calibration

In data-driven settings, it is important to calibrate the parameters of stochastic optimization models to ensure that the model performs well against unseen data. First, we describe how the parameters γ^{fp} and γ^{fn} can be calibrated for TTP*-Q to ensure that the constraints are satisfied a greater proportion of the time. Then, we describe how the Wasserstein radii ϵ^+ and ϵ^- can be calibrated based on the data samples $\hat{P}_N^+$ and $\hat{P}_N^-$ so that the solutions to TTP*-DR are not overly conservative.

Calibrating constraints for TTP*-Q: For stochastic programming models, data-driven solution methodologies (e.g., sample average approximation) have been shown to suffer in out-of-sample feasibility when the data-driven problem is solved using the same constraint parameters as the desired level of feasibility (Ahmed and Shapiro, 2008; Luedtke and Ahmed, 2008; Pagnoncelli, Ahmed, and Shapiro, 2009). To alleviate this issue for TTP*-Q, we aim to rewrite the right-hand sides of (4.1b) and (4.1d) as $\gamma_\tau^{fp} := \gamma^{fp} - \tau^{fp}$ and $\gamma_\tau^{fn} := \gamma^{fn} - \tau^{fn}$, respectively, where $\tau^{fp} \in (0, \gamma^{fp})$ and $\tau^{fn} \in (0, \gamma^{fn})$. In principle, the parameters τ^{fp} and τ^{fn} act as safety factors for the constraints. Intuitively, larger values of τ^{fp} and τ^{fn} should correspond to greater feasibility. To determine τ^{fn}, we applied the following procedure.

1. Fix γ^{fn} and created a grid $G = \{0, ..., \gamma^{fn}\}$ of $|G|$ evenly spaced potential units between 0 and γ^{fn}.

2. Following the procedure for k-fold cross-validation, divide the data $\hat{P}_N^+$ into k equally sized subsets $\{\hat{P}_{N,i}^+\}_{i=1}^k$, where the i^{th} fold involves a training subset $\hat{P}_{N,i}^{+,\text{train}} := \bigcup_{j \neq i} \hat{P}_{N,j}^+$ and a held-out testing set $\hat{P}_{N,i}^+$.

3. For each $i = 1, ..., k$, set $l_\tau^i = l(\gamma^{fn} - \tau)$ for each $\tau \in G$ where the quantile $l(\gamma^{fn} - \tau)$ is determined using the Harrell-Davis quantile estimator over the training data $\hat{P}_{N,i}^{+,\text{train}}$.

4. Evaluate the quality of l_τ^i using

$$\mathcal{L}_i^{fn}(l_\tau^i) = \mathbb{E}[sp(l_\tau^i, \xi^-)] - \rho \mathbb{E}[fn(l_\tau^i, \xi^+)],$$

   where the first expectation is approximated by the data sample $\hat{P}_N^-$, the second expectation is approximated by the held-out testing set $\hat{P}_i^+$, and the penalty term ρ is set sufficiently high (e.g., $\rho = 100$) to ensure that infeasible solutions are generally outperformed by feasible solutions.

5. For each fold $i = 1, ..., k$, set $\tau_i^* = \arg\max_\tau \mathcal{L}_i^{fn}(l_\tau^i)$.

6. Set τ^{fn} as the q^{th} quantile of the set $\{\tau_i^*\}_{i=1}^k$. Higher values of q correspond to more conservative estimates of τ^{fn}.

The procedure is similar for determining τ^{fp}. The main differences are as follows.

1. The maximum element in the grid is given by γ^{fp}.

2. Cross-validation is performed over $\hat{P}_N^-$, with each fold i having a training subset $\hat{P}_{N,i}^{-,\text{train}} := \bigcup_{j \neq i} \hat{P}_{N,j}^-$ and held-out testing set $\hat{P}_{N,i}^-$.

3. For each $\tau \in G$, we set $u_\tau^i = u(\gamma^{fp} - \tau)$, where the quantile $u(\gamma^{fp} - \tau)$ is determined by the Harell-Davis quantile estimator over the training data $\hat{P}_{N,i}^{-,\text{train}}$.

4. Each u_τ^i is evaluated under the function

$$\mathcal{L}_i^{fp}(u_\tau^i) = \mathbb{E}[se(u_\tau^i, \xi^+)] - \rho\mathbb{E}[fp(u_\tau^i, \xi^-)],$$

   where the first expectation is approximated by the data sample $\hat{P}_N^+$ and the second expectation by the held-out sample $\hat{P}_{N,i}^-$.

Calibrating Wasserstein radii for TTP*-DR: The choice of Wasserstein radii ϵ^+ and ϵ^- can greatly influence the performance of TTP*-DR. For example, setting ϵ^+ equal to

$$\epsilon_{LD}^+ = \sqrt{2/N^+ \log(1/(1-\beta))}$$

guarantees that the ambiguity set $\hat{D}_N^+(\epsilon_{LD}^+)$ contains the true data-generating distribution $\mathcal{P}^+$ with probability $1-\beta$ (Mohajerin Esfahani and Kuhn, 2018). Unfortunately, in practice, this choice of ϵ^+ results in an overly conservative optimal threshold. Fortunately, cross-validation can be used to calibrate the choice of ϵ^+, resulting less conservative optimal thresholds. To determine ϵ^+, we apply the following procedure.

1. Fix γ^{fn} and created a grid $G = \{0, ..., \epsilon_{LD}^+\}$ of $|G|$ evenly spaced potential radii between 0 and ϵ_{LD}^+.

2. Following the procedure for k-fold cross-validation, divide the data $\hat{P}_N^+$ into k equally sized subsets $\{\hat{P}_{N,i}^+\}_{i=1}^k$, where the i^{th} fold involves a training subset $\hat{P}_{N,i}^{+,\text{train}} = \bigcup_{j \neq i} \hat{P}_{N,j}^+$ and a held-out testing set $\hat{P}_{N,i}^+$.

3. For each $i = 1, ..., k$, solve

$$l_\epsilon^i = \max \left\{ l \in [0,1] : \begin{array}{c} \text{(4.20b) is satisfied with} \\ D^+ = D_\epsilon(\hat{\mathcal{P}}_{N,i}^{+,\text{train}}) \end{array} \right\},$$

where $\hat{\mathcal{P}}_{N,i}^{+,\text{train}}$ is the empirical uniform distribution centered on $\hat{P}_{N,i}^{+,\text{train}}$.

4. Evaluate the quality of l_ϵ^i using

$$\mathcal{L}_i^{fn}(l_\epsilon^i) = \mathbb{E}[sp(l_\epsilon^i, \xi^-)] - \rho\mathbb{E}[fn(l_\epsilon^i, \xi^+)],$$

where the first expectation is approximated by the data sample $\hat{P}_N^-$, the second expectation is approximated by the held-out testing set $\hat{P}_i^+$, and the penalty term ρ is set sufficiently high (e.g., $\rho = 100$) to ensure that infeasible solutions are generally outperformed by feasible solutions.

5. For each fold, we set $\epsilon_i^* = \arg\max_\epsilon \mathcal{L}_i^{fn}(l_\epsilon^i)$.

6. Set the Wasserstein radius as $\epsilon^+ = 1/k \sum_{i=1}^k \epsilon_i^*$.

The procedure is similar for determining ϵ^-. The main differences are as follows.

1. The maximum radius in the grid is given by $\epsilon_{LD}^- = \sqrt{2/N^- \log(1/(1-\beta))}$.

2. Cross-validation is performed over $\hat{P}_N^-$, with each fold i having a training subset $\hat{P}_{N,i}^{-,\text{train}} := \bigcup_{j \neq i} \hat{P}_{N,j}^-$ and held-out testing set $\hat{P}_{N,i}^-$.

3. The upper threshold is determined by solving

$$u_\epsilon^i = \min \left\{ u \in [0,1] : \begin{array}{c} \text{(4.20a) is satisfied with} \\ D^- = D_\epsilon(\hat{\mathcal{P}}_{N,i}^{-,\text{train}}) \end{array} \right\},$$

where $\hat{\mathcal{P}}_{N,i}^{-,\text{train}}$ is the empirical uniform distribution centered on $\hat{P}_{N,i}^{-,\text{train}}$.

4. Each u_ϵ^i is evaluated under the function

$$\mathcal{L}_i^{fp}(u_\epsilon^i) = \mathbb{E}[se(u_\epsilon^i, \xi^+)] - \rho\mathbb{E}[fn(u_\epsilon^i, \xi^-)],$$

where the first expectation is approximated by the data sample $\hat{P}_N^+$ and the second expectation by the held-out sample $\hat{P}_{N,i}^-$.

4.5 Simulation Analysis

In this section, we use simulation to analyze the performance of TTP and its extensions under various conditions. Specifically, we first analyze the computational time required for TTP*-Q and TTP*-DR in Section 4.5.1. Then, we analyze the feasibility and optimality of TTP*-Q and TTP*-DR with respect to known distributions in Section 4.5.2. In Section 4.5.3, we perform a numerical analysis to compare the performance of TTP*-Q with an optimized single threshold. Finally, in Section 4.5.4 we analyze the OTP model of Section 4.3.6.

4.5.1 Computational Time for TTP*-Q and TTP*-DR

In this section, we analyze the computational time required to solve TTP*-Q and TTP*-DR. We assumed that $\mathcal{P}^+ = \text{Beta}(55, 45)$ and $\mathcal{P}^- = \text{Beta}(45, 55)$. This distributional setup is equivalent to the distribution in Section 4.5.2 with AUROC = 0.922. For each sample size $N^+ = N^- \in \{100, 500, 1000, 2000, 5000\}$, we solved TTP*-Q and TTP*-DR 100 times under the constraints $\gamma^{fp} = 0.10$ and $\gamma^{fn} = 0.05$. Models were implemented using Python 3.6 and optimization models were solved using Gurobi version 8.0. Computational times are based on a computer with 16 GB of RAM and an Intel Xeon E3-1241 3.50 GHz CPU. The results of this analysis are shown in Figure 4.1.

This analysis reveals that solving TTP*-Q takes several orders of magnitude less computational time compared to TTP*-DR. Specifically, even at 5000 samples, the median solution time for TTP*-Q is 0.04 seconds. In contrast, the computational burden of TTP*-DR rises sharply as the sample size increases. With only 100 samples, the median solution time for TTP*-DR is 0.64 seconds. As the sample size rises to 1000, the median solution time increases to 9.45 seconds. At 5000 samples, the median computational time is 74.20 seconds.

4.5.2 Feasibility and Optimality of TTP*-Q and TTP*-DR

In this section, we use simulation to estimate the performance of TTP*-Q and TTP*-DR under different underlying risk estimation distributions and sample sizes. We aim to identify the situations for which each solution methodology performs well.

We performed our simulation analysis under the following settings.

Underlying risk estimation distributions: The Beta distribution is commonly used to

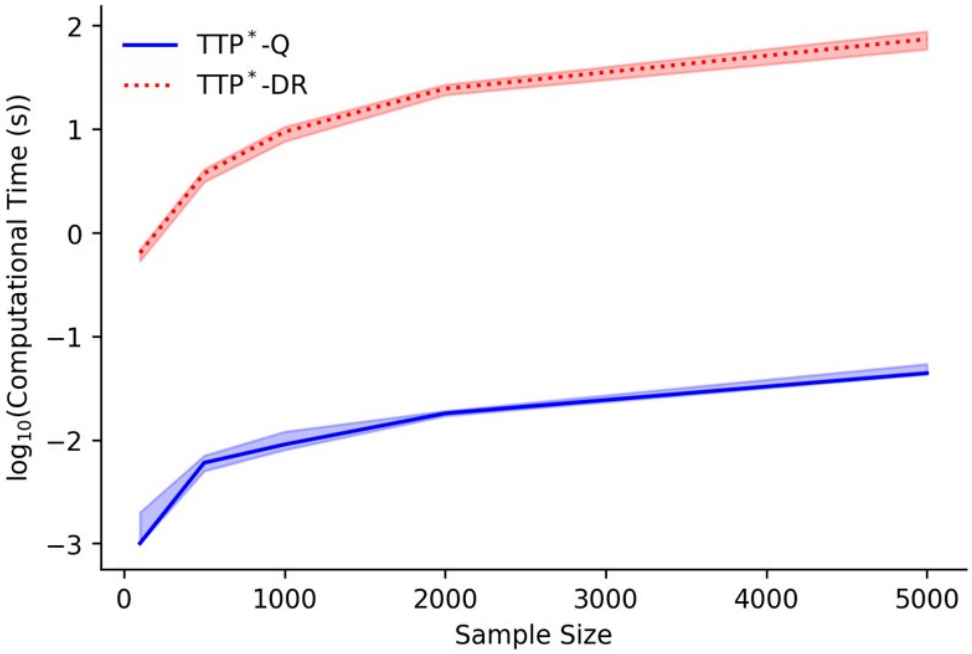


Figure 4.1: Median $\log_{10}$(computational time in seconds) for solving TTP*-Q and TTP*-DR. Shaded area represents the 5^{th} and 95^{th} percentiles.

simulate random variables with the support $[0, 1]$ and its shape parameters can be manipulated to change the distribution's mean, variance, and skewness. Therefore, we assumed that $\mathcal{P}^+ = \text{Beta}(0.55v, 0.45v)$ and $\mathcal{P}^- = \text{Beta}(0.45v, 0.55v)$ where $v \in \{1, 10, 50, 100\}$. Larger values of v reflect a higher quality of risk estimation model as measured by area under the receiver operating characteristic curve (AUROC). Specifically, v being equal to $1, 10, 50$, and 100 is equivalent to an AUROC of $0.585, 0.680, 0.844$, and 0.922, respectively.

Sample size: We assumed that N^+ and N^- are equal, with $N^+, N^- \in \{100, 500, 1000\}$.

For each scenario given by AUROC and sample size, we calibrated and solved TTP*-Q and TTP*-DR 100 times under the constraints $\gamma^{fp} = 0.10$ and $\gamma^{fn} = 0.05$.

Since this simulation analysis utilized Beta distributions with known parameters, we were able to compute the true optimal solution to TTP* for each simulation scenario. We then compared these true optimal solutions to the solutions obtained by solving TTP*-Q and TTP*-DR. Specifically, we compared these methods on the basis of optimality and feasibility.

To measure optimality, we computed the optimality gap, which we define by

$$1 - \frac{\mathbb{E}[se(u, \xi^+)] + \mathbb{E}[sp(l, \xi^-)]}{\mathbb{E}[se(u^*, \xi^+)] + \mathbb{E}[sp(l^*, \xi^-)]}, \tag{4.21}$$

where (u, l) denotes a candidate solution from TTP*-Q and (u^*, l^*) denotes the true optimal solution. From (4.21), it can be seen that a set of thresholds (u, l) achieving a positive optimality gap near 0 indicates near-optimal performance and near 1 indicates poor performance. By definition, optimality gaps below 0 imply that the thresholds are infeasible to the true underlying distribution.

To measure feasibility, we computed the maximum constraint violation, which is defined by

$$\max \left\{ \mathbb{E}[fp(u, \xi^-)] - \gamma^{fp}, \mathbb{E}[fn(l, \xi^-)] - \gamma^{fn} \right\}. \tag{4.22}$$

From (4.22), it can be seen candidate thresholds (u, l) are feasible for the true problem only if their maximum constraint violations are non-positive. For both TTP*-Q and TTP*-DR, we plot joint distribution of optimality gap and maximum constraint violation in Figure 4.2.

Several insights can be drawn from this analysis. For TTP*-Q, we find that its optimality and feasibility vary greatly based on sample size and AUROC. Specifically, TTP*-Q has low feasibility when the sample size is small (i.e., 17%-34% of solutions are feasible when $N^+, N^- = 100$). However, feasibility greatly improves when the sample size reaches 500 (i.e., 78%-87% of solutions are feasible) and 1000 (i.e., 93%-95% of solutions are feasible). For fixed sample sizes, we find that feasibility is the lowest when the AUROC is 0.585 and rises as AUROC increases to 0.680. However, additional increases in AUROC do not result in additional improvements in feasibility. Additionally, we find that the optimality gap (regardless of whether it is conditional on being feasible or not) decreases as the AUROC increases. On the other hand, increasing sample size does not seem to provide a noticeable decrease in the optimality gap.

The feasibility of TTP*-DR is largely driven by sample size. Specifically, 76%-96% of solutions are feasible when $N^+, N^- = 100$ and the feasibility rises between 96%-100% once the sample size increases to 500 and 1000. AUROC does not have a discernible effect on feasibility but much like for TTP*-Q, the optimality gap for TTP*-DR decreases as AUROC increases. When AUROC is at least 0.680, the optimality gap decreases as the sample size

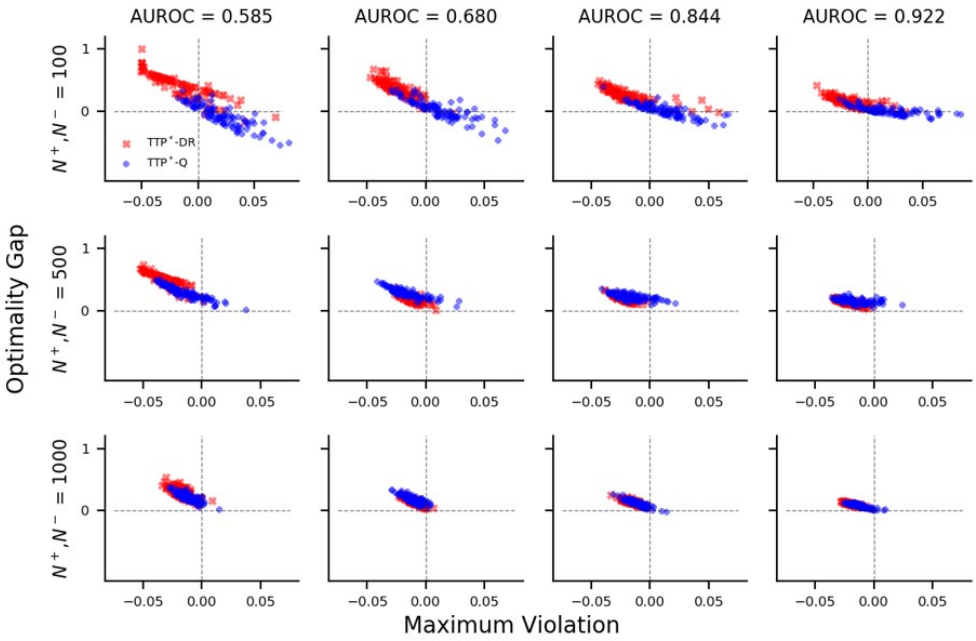


Figure 4.2: Distribution of optimality gap and maximum constraint violation under varying quality of risk estimation model (AUROC) and sample size (N^+, N^-).

increases.

Comparing the two solution methods, TTP*-DR achieves a greater level of feasibility than TTP*-Q in every combination of AUROC and sample size tested. Furthermore, the difference in feasibility between the two approaches is greatest when the sample size is small. Conditional on being feasible, TTP*-Q achieves a smaller optimality gap than TTP*-DR when the sample size is small and the AUROC is low. However, as the sample sizes increase and the AUROC increases, results are mixed in terms of optimality gap. However, the two methods appear to achieve similar optimality gaps as the AUROC and sample size increase.

Overall, sample size and AUROC play an important role in feasibility and optimality. Increasing sample size greatly improves feasibility, especially for TTP*-Q. Increasing AUROC decreases the optimality gap for both methods, and may slightly improve feasibility for TTP*-Q. TTP*-DR appears to be superior to TTP*-Q for small sample sizes due to its greater rate of feasibility. Since both methods perform similarly for large sample sizes and sufficiently high AUROC, computational time can be an important deciding factor in choosing which method to employ in practice. In Section 4.5.1, we found that TTP*-DR

is several orders of magnitude higher in computational time compared to TTP*-Q. While the overall computation time is not prohibitive for small sample sizes, TTP*-DR takes a median time of 74.20 seconds once the sample size reaches 5000 (compared to 0.04 seconds for TTP*-Q). Hence, TTP*-Q appears to be more practical for large sample sizes given the need to solve each model several times through the calibration procedure and their similarity in out-of-sample performance at large sample sizes. While our simulation study assumed that the underlying distributions for $\mathcal{P}^+$ and $\mathcal{P}^-$ were Beta distributions, we performed the same analysis under different distributional assumptions and found that similar trends follow (see Appendix 4.C). However, this auxiliary analysis does suggest that the variance of the distributions $\mathcal{P}^+$ and $\mathcal{P}^-$ play an important role in the feasibility of TTP*-Q.

4.5.3 Comparing One- and Two-threshold Classification Schemes

In this section, we use simulation to compare the accuracy of a one-threshold classification scheme and the two-threshold classification scheme determined by TTP*. Specifically, we obtain an optimal one-threshold solution (denoted 1T*) by using sample average approximations to solve

$$(1\text{T}^*) \quad \max_t \lambda \mathbb{E}[se(t, \xi^+)] + (1-\lambda)\mathbb{E}[sp(t, \xi^-)], \tag{4.23}$$

where $\lambda \in (0,1)$ plays the same role as the λ used in the objective function of TTP. Notice that (4.23) is the same as solving an unconstrained version of TTP. Furthermore, the resulting threshold from solving (4.23) is akin to one chosen from the ROC curve, as is commonly done in practice (see Section 4.2). We evaluated both of these models based on classification accuracy, which can be interpreted as the probability of correctly classifying a randomly chosen patient. Specifically, we define accuracy as

$$q\left(\mathbb{E}[se(u, \xi^+)] + \eta^+(u,l)\right) + (1-q)\left(\mathbb{E}[sp(u,l)|\xi^-] + \eta^-(u,l)\right),$$

where q represents the proportion of true-positives in the overall population and the functions $\eta^+(u,l)$ and $\eta^-(u,l)$ denote the probabilities of correctly classifying a true-positive and true-negative patient, respectively, after they are deferred based on the thresholds u and l. Without the functions $\eta^+(u,l)$ and $\eta^-(u,l)$, 1T* would always have a higher accuracy since

no proportion of the population is deferred. Note that $\eta^+(t,t) = 0$ and $\eta^-(t,t) = 0$ for 1T* since the upper and lower thresholds are equal.

Throughout this simulation analysis, we set

$$\eta^+(u,l) := \int_l^u \mathcal{L}^+_{k,\xi_0^+} f_{\mathcal{P}^+}(\xi^+)d\xi^+ \text{ and } \eta^-(u,l) := \int_l^u \mathcal{L}^-_{k,\xi_0^-} f_{\mathcal{P}^-}(\xi^-)d\xi^+,$$

where $f_{\mathcal{P}^+}$ and $f_{\mathcal{P}^-}$ are density functions based on the Beta distributions specified in Section 4.5.2 and the logistic functions

$$\mathcal{L}^+_{k,\xi_0} = \left(1 + \exp\Big(-k(\xi^+ - \xi_0)\Big)\right)^{-1} \text{ and } \mathcal{L}^-_{k,\xi_0} = \left(1 + \exp\Big(-k(\xi_0 - \xi^-)\Big)\right)^{-1}$$

are parameterized by scale $k \geq 0$ and centers $\xi_0^+, \xi_0^- \in [0,1]$. We use $\mathcal{L}^+_{k,\xi_0}$ (resp., $\mathcal{L}^-_{k,\xi_0}$) to approximate the likelihood of correctly classifying a deferred true-positive (resp., true-negative) patient given risk estimate ξ^+ (resp., ξ^-).

In our analysis, we solved TTP*-Q with $\gamma^{fp} = \gamma^{fn} = 0.10$ and 1T* with $\lambda = 0.5$ by sampling from each distribution specified in Section 4.5.2 with sample sizes $N^+ = N^- = 1000$. To compute accuracy, we set $q = 0.5$ and parameterized the functions L_{k,ξ_0^+} and L_{k,ξ_0^-} with fixed centers $\xi_0^+ = 0.45, \xi_0^- = 0.55$ and varying $k \in [0, 20]$. Higher values of k indicate greater likelihood of correctly classifying a patient after they are initially deferred.

The results of our simulation analysis are shown in Figure 4.3. In all cases, the accuracy of TTP*-Q increases quadratically in k, indicating that small improvements in the accuracy of post-defer classification will initially result in large gains of overall accuracy. However, there is an inflection point at which there are diminishing returns on the gains in overall accuracy. To this end, the inflection point appears to be increasing in the AUROC of the underlying risk estimation model. This result implies that the marginal value of increasing the accuracy of post-defer classification remains high for risk estimation models which are already accurate. Practically speaking, if the underlying risk estimation model is accurate, then increasing the accuracy of post-defer classification has high utility if the marginal cost of increasing post-defer accuracy is low (e.g., ordering low-cost lab tests).

Comparing TTP*-Q and 1T*, our results indicate that when few patients are correctly classified after they are first deferred (i.e., k is near 0), 1T* achieves a greater classification accuracy. However, as the likelihood of correctly classifying deferred patients increases,

TTP*-Q begins to achieve a greater level of accuracy than 1T*. To this end, the minimum classification accuracy for deferred patients (i.e., k) needed such that TTP*-Q outperforms 1T* also increases in the quality of the underlying risk estimation model (i.e., AUROC) improves. This result indicates that if the underlying risk estimation model can already distinguish between true-positives and true-negatives with high accuracy, more effort is required in ensuring that deferred patients can actually be correctly identified after they are initially deferred. Practically speaking, a one-threshold classification scheme may be more useful when the underlying risk estimation model is already accurate and there is a high cost associated with increasing the accuracy of classifying patients who are deferred (e.g., ordering assessments which require expensive equipment). This analysis illustrates that the utility of TTP increases as the accuracy of a risk estimation model decreases. This result stems from the fact that the underlying advantage to using TTP is its ability to identify patients who cannot be accurately diagnosed using the risk estimation model. Therefore, these patients are better served by the expert judgment of clinicians.

4.5.4 Analysis of Ordinal Threshold Problem

In this section, we use simulation to analyze the performance of the Ordinal Threshold Problem (OTP) model derived in Section 4.3.6 for an ordinal classification problem with three different populations. We performed our simulation analysis under the following settings.

Underlying severity score distributions: Let $\mathcal{P}^k$ denote the distribution of the severity scores for class k, $k = 1, 2, 3$. We assumed that $\mathcal{P}^1 = \text{Beta}(0.25v, 0.75v)$, $\mathcal{P}^2 = \text{Beta}(0.50v, 0.50v)$ and $\mathcal{P}^3 = \text{Beta}(0.75v, 0.25v)$ where $v \in \{10, 50, 100\}$. Larger values of v reflect a higher quality of severity score model. That is, there is a larger degree of "separation" between the distributions as v increases, reflecting a severity score model which can better discriminate between the three populations.

Sample size: We assumed that the sample sizes of class 1, 2, and 3 (denoted by N^1, N^2, and N^3, respectively) are equal, with $N^1, N^2, N^3 \in \{100, 500, 1000\}$.

We solved OTP using quantile estimation under each of these settings 100 times under different parameterizations. Specifically, we set $\gamma^{fn}_{k,j} = 1$ for all $k = 1, 2$ and $j = 2, 3$ to guarantee feasibility. We also tested all combinations of $\gamma^{fp}_{1,1} \in \{0.01, 0.03, ..., 0.13\}$, $\gamma^{fp}_{2,1} \in \{\gamma^{fp}_{1,1}, ..., 0.13\}$ and $\gamma^{fp}_{2,2} \in \{0.01, ..., 0.13\}$. For a given set of thresholds t_1 and t_2, we measured

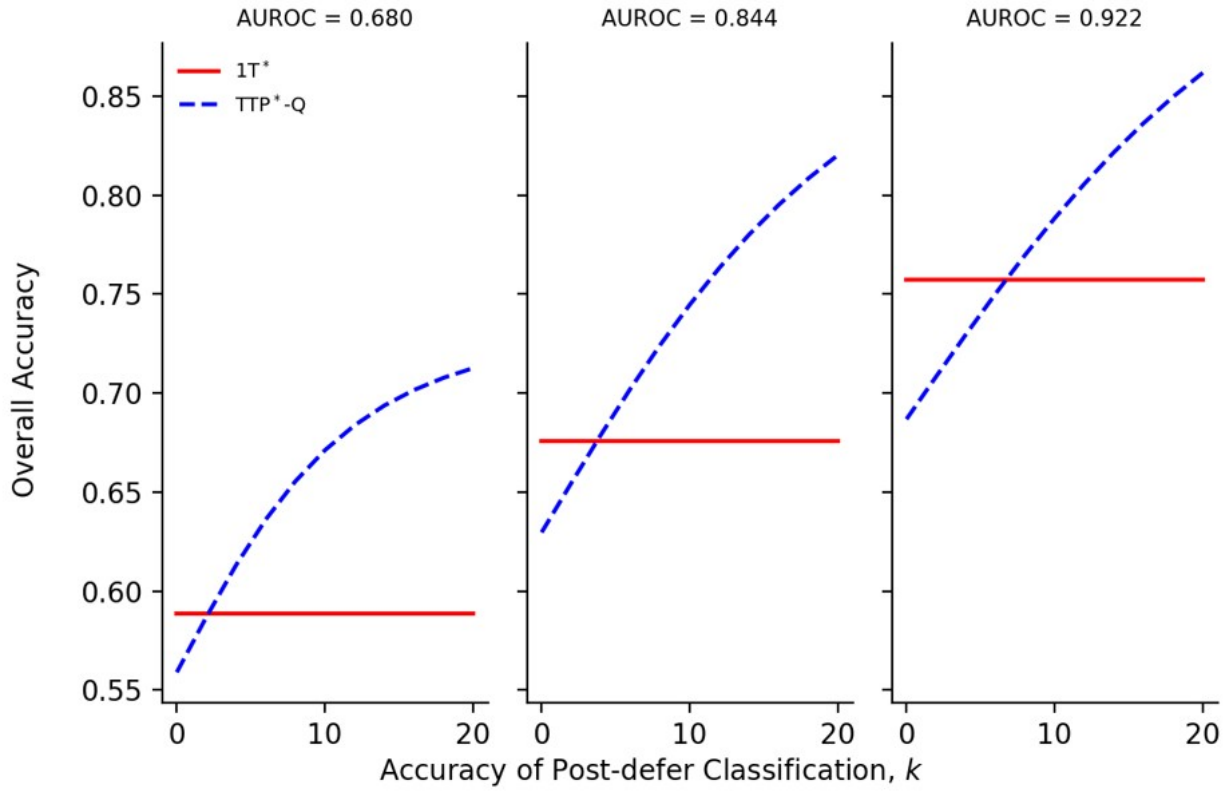


Figure 4.3: Comparison of overall accuracy between one-threshold (1T*) and two-threshold (TTP*-Q) classification schemes under varying quality of risk estimation model (AUROC) and post-defer classification accuracy (k). AUROC, area under the receiver operating characteristic curve.

the performance of OTP based on accuracy, which is defined as

$$\frac{1}{3}\Bigg(\mathbb{P}(\xi^1 < t_1) + \mathbb{P}(t_1 < \xi^2 < t_2) + \mathbb{P}(\xi^3 > t_2)\Bigg). \tag{4.24}$$

Accuracy can be interpreted as the proportion of patients who are correctly identified based on the thresholds t_1 and t_2 (e.g., $\xi^1 < t_1$). Values of accuracy close to 1 indicate near-perfect classification. We also computed mean squared error (MSE), which is defined as

$$\begin{aligned}&\Bigg(\mathbb{P}(t_1 < \xi^1 < t_2) + \mathbb{P}(\xi^2 < t_1) + \mathbb{P}(\xi^2 > t_2) + \mathbb{P}(t_1 < \xi^3 < t_2)\Bigg) \\ &+ 4\Bigg(\mathbb{P}(\xi^1 > t_2) + \mathbb{P}(\xi^3 < t_1)\Bigg).\end{aligned} \tag{4.25}$$

MSE is a measure of classification error which more harshly penalizes greater deviations from the true class. For example, patients from class 1 who are classified as class 2 (i.e., $t_1 < \xi^1 < t_2$) are penalized less harshly than those who are classified as class 3 (i.e., $\xi^1 > t_2$). Thus, higher values of MSE imply a greater degree of misclassification.

Based on (4.24) and (4.25), we compared OTP to Equidistant Thresholds (ET) which assume that all thresholds are the same distance apart. In general, ET assumes that $t_k = t_1 + (k-1)\delta$ for all k. Such models are commonly referred to as Rating Scale Models (Andersen, 1977) and are practically useful because they only require one to determine the values of t_1 and δ. To vary the parameters of ET, we tried every combination of t_1 and δ such that $t_1 = \min\{t : fp(t, \xi^1) \leq \gamma^{fp}\}$ for $\gamma^{fp} \in \{0.01, 0.03, ..., 0.13\}$ and $\delta \in \{0.10, 0.15, 0.20, 0.25\}$.

Figure 4.4 illustrates the comparison between OTP and ET, with only the pareto optimal parameter combinations shown for clarity.

In general, we find that the performance of OTP in terms of accuracy and MSE improve as both the quality of the underlying severity score model (i.e., v) and the sample size (i.e., N^1, N^2, N^3) increase. However, improving the quality of the severity score model has a much greater effect. In comparing OTP to ET, we find that several parameter settings for OTP which achieve an equal or similar accuracy level as ET at lower levels of MSE. This result seems to hold across all simulation settings.

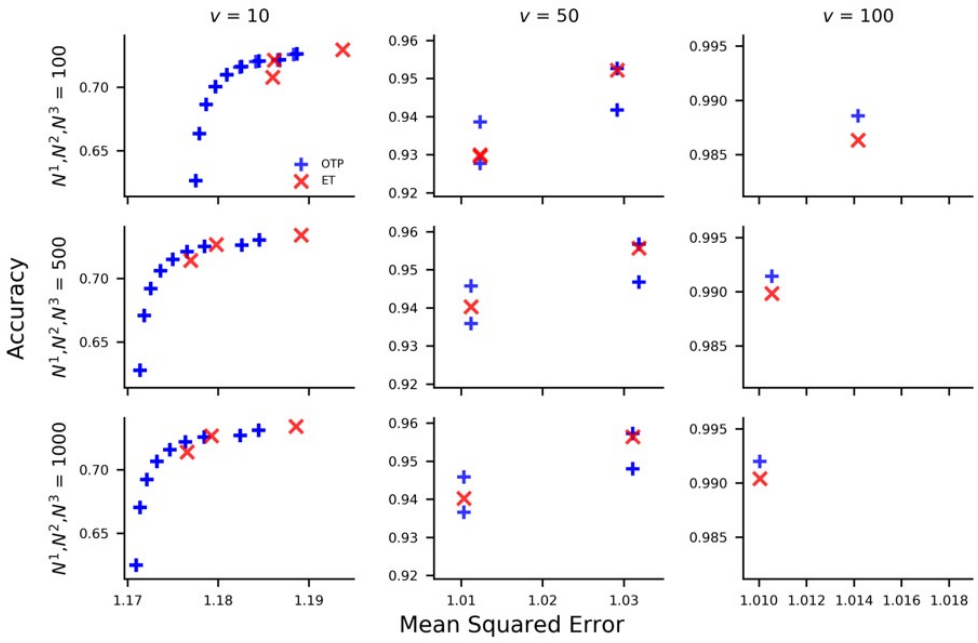


Figure 4.4: Comparison of accuracy and mean squared error in ordinal classification for the Ordinal Threshold Problem (OTP) and Equidistant Thresholds (ET)

4.6 Case Study: Acute Concussion Assessment

In this section, we present numerical results based on the application of TTP* to concussion assessment data. We first describe our data and choice of risk estimation model. Then, we define three efficiency measures and analyze the relationship between modeling parameters and these efficiency measures. Finally, we compare the performance of TTP* to (1) an optimized one-threshold solution which is commonly used to determine decision thresholds for risk estimation models and (2) a normative value comparison method which is commonly used in acute concussion assessment.

4.6.1 Concussion Assessment Data

Our numerical analysis uses multi-center longitudinal data provided by the CARE Consortium (see Section 1.1.1). We focus on evaluations from < 6 hours, which we denote as "concussion" (i.e., true-positive), and the first time at which the athlete is cleared to return to play, which we denote "non-concussion" (i.e., true-negative).

Data was received in two batches. Because some data were missing at the < 6 hours time-point, the first batch (i.e., training data) contained 560 concussions and 707 non-concussions. Similarly, the second batch contained 539 concussions and 629 non-concussions. We used a randomly chosen 40% of the training data to form a risk estimation model and the remaining 60% to solve TTP* using each method in Section 4.4. We validated our thresholds and assessed out-of-sample performance using the second batch of data (i.e., validation data). We summarize this data in Table 4.1 for males and females in the training and validation set with respect to the Standard Assessment of Concussion (SAC), the Sport Concussion Assessment Tool (SCAT) symptom assessments, and the Balance Error Scoring System (BESS).

Table 4.1: Description of training and validation data

	Training Data							
	Male				**Female**			
	Concussion (n=361)		*Non-concussion* (n=423)		*Concussion* (n=199)		*Non-concussion* (n=284)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
SAC Total Score	25.90	3.12	27.87	1.85	26.74	2.42	28.00	1.60
SCAT Symptom Severity	27.33	20.88	0.43	1.55	31.67	20.02	0.93	2.46
SCAT Total Number of Symptoms	10.53	5.50	0.33	1.17	11.81	5.02	0.67	1.65
BESS Total Score	16.58	8.88	10.92	5.93	14.86	7.90	9.86	5.78
	Validation Data							
	Male				**Female**			
	Concussion (n=332)		*Non-concussion* (n=340)		*Concussion* (n=207)		*Non-concussion* (n=289)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
SAC Total Score	25.52	3.52	28.07	1.91	26.38	2.86	28.31	1.62
SCAT Symptom Severity	27.42	21.54	0.29	0.92	30.31	21.2	0.69	2.24
SCAT Total Number of Symptoms	10.62	5.75	0.22	0.70	11.03	5.17	0.51	1.46
BESS Total Score	17.86	8.97	11.12	5.02	16.99	8.67	10.76	5.73

SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; SD = standard deviation

4.6.2 Risk Estimation Model

Multivariate logistic regression can be used to create a risk estimation model of the form

$$f(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x}^T\mathbf{b})},$$

where $f(\mathbf{x})$ is a risk estimate, $\mathbf{x}$ is a vector of patient characteristics, and $\mathbf{b}$ is a vector of corresponding coefficients. We used the multivariate logistic regression model developed in Chapter 2 for acute concussion assessment, in which the relevant patient characteristics include sex, whether the injury was reported immediately, whether the athlete was removed from play immediately, and scores from the SAC, SCAT symptom assessment, and BESS. These variables are described in Chapter 2.

4.6.3 Example Solution and Post-hoc Analysis

We solved TTP*-Q and TTP*-DR using the training data and evaluated their optimal solutions against the validation data. For various choices of γ^{fp} and γ^{fn}, we found that TTP*-Q generally remains feasible. In some cases (e.g,. $\gamma^{fp} = 0.01$ and $\gamma^{fn} = 0.03$), TTP*-Q violated one of its constraints — though the magnitude of this violation was generally small (i.e., less than 0.006). In contrast, TTP*-DR was feasible against the validation data in every parameter combination tested. Figure 4.1 shows an example optimal solution from TTP*-Q obtained by setting $\gamma^{fp} = 0.028$ and $\gamma^{fn} = 0.02$. In this example, u^* captures many of the true-positives (sensitivity= 0.91) while maintaining a low false-positive rate (false-positive= 0.016). On the other hand, l^* captures fewer true-negatives (specificity= 0.74) but also maintains a low false-negative rate (false-negative= 0.015). Only 16.7% of the validation data fell between u^* and l^*. That is, only 16.7% of those undergoing diagnosis would result in deferred diagnosis.

In practice, it may be of interest to perform post-hoc analysis on athletes who were ultimately deferred. For example, in Table 4.2, we characterize the deferred athletes based on their risk estimates and clinical assessment variables. We also compare them to athletes who were correctly diagnosed. It can be seen that risk estimates were lower for deferred athletes with concussion compared to those that were correctly diagnosed. Likewise, risk estimates were higher for deferred athletes without concussion compared to those who were correctly diagnosed. We also identified statistically significant differences in the SAC total score, SCAT symptom severity, and SCAT total number of symptoms between deferred athletes and their correctly diagnosed counterparts. Practically speaking, these results confirm just how different the "easy" and the "hard" cases are. Furthermore, these characterizations can be used to inform future clinical decisions. For example, deferred athletes with higher

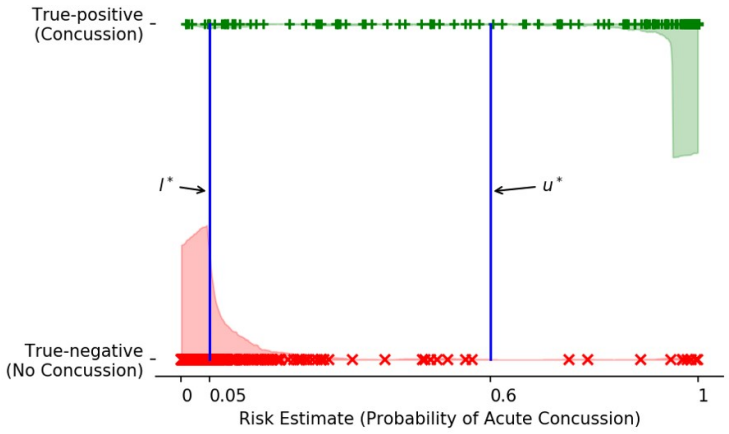


Figure 4.1: Optimal upper threshold (u^*) and lower threshold (l^*) when solving TTP*-Q with $\gamma^{fp} = 0.028$ and $\gamma^{fn} = 0.020$. Risk estimates $> u^*$ are classified as true-positives while risk estimates $< l^*$ are classified as true-negatives. Diagnosis decisions are deferred for risk estimates between u^* and l^*. True positives (+) and true-negatives (x) from validation data are shown with estimated kernel densities.

symptom presentation are more likely to have concussion than those with lower symptom presentation. However, as these clinical variables were already included in our risk estimation model, it may be of greater value to consider additional assessments (e.g., vestibular/ocular-motor screening) for the group of deferred athletes.

4.6.4 Analyzing The Efficiency of Two-threshold Solutions

In practice, decision thresholds which defer too many diagnosis decisions can be problematic. For example, if the necessary treatment following a diagnosis decision must be performed in a timely manner, then deferring too many decisions can cause a delay in receiving that treatment. Additionally, if the action following a deferred decision requires valuable resources (e.g., performing additional diagnostic tests), then deferring too many decisions could put potentially strain those resources. Thus, we evaluated TTP*-Q for acute concussion assessment based on three efficiency measures:

$$e_1 = \frac{p^+(u,l)}{p^D(u,l)}, e_2 = \frac{p^-(u,l)}{p^D(u,l)}, e_3 = \frac{p^+(u,l)}{p^-(u,l)},$$

Table 4.2: Comparison of athletes who were correctly diagnosed and deferred under example TTP*-Q solution

True Outcome	**Concussion**				**No Concussion**			
Diagnosis Decision	*Positive*		*Defer*		*Negative*		*Defer*	
	n=490		n=41		n=465		n=154	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Risk Estimate[†,‡]	0.98	0.06	0.28	0.16	0.01	0.01	0.14	0.10
SAC Total Score [†,‡]	25.71	3.36	27.24	2.25	28.26	1.73	27.97	1.85
BESS Total Score[†,‡]	18.02	8.92	13.00	6.19	10.60	5.10	11.88	5.86
SCAT Total Symptom Severity[†,‡]	31.04	20.83	3.85	3.35	0.14	0.57	0.88	1.68
SCAT Total Number of Symptoms[†,‡]	11.64	5.01	2.37	1.84	0.11	0.42	0.69	1.27

[†]Mean value is significantly different ($p < 0.05$) between concussions with positive and deferred diagnosis decisions using two-sample Student's t-test; [‡]Mean value is significantly different ($p < 0.05$) between non-concussions with negative and deferred diagnosis decisions using two-sample Student's t-test; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; SD, standard deviation

where the $p^+(u,l)$, $p^-(u,l)$ and $p^D(u,l)$ are defined in Section 4.3.5. The measure e_1 is the rate of correct classifications per deferred decision, e_2 is the rate of misclassifications per deferred decision, and e_3 is the rate of correct classifications per misclassification. We computed each of these measures for various parameter combinations. We excluded any results with single-threshold solutions, since we would have $p^D(u,l) = 0$ and thus, e_1 and e_2 cannot be computed. We also varied the proportion of true-positives in the population undergoing acute concussion assessment, q.

From this efficiency analysis, we found that when there is a high proportion of true-positives in the testing population (i.e., q is high), parameter combinations which allow for higher false-positive rates and ultimately end up with high sensitivity result in greater efficiency in terms of correct classifications per misclassification (i.e., $e1$) and deferred decision (i.e., $e3$). Similarly, when most of the testing population is composed of true-negatives (i.e., q is low), parameter combinations which allow for higher false-negative rates result in greater specificity. These parameter combinations are also more efficient in terms of $e1$ and $e3$. Likewise, parameter combinations which have low false-positive rates tend to have lower misclassification rates per deferred decision (i.e., $e2$) when there are fewer true-negatives in the population (i.e., q is lower). Parameter combinations with low false-negative rates are more efficient in terms of $e2$ when there are more true-positives in the population (i.e., q is

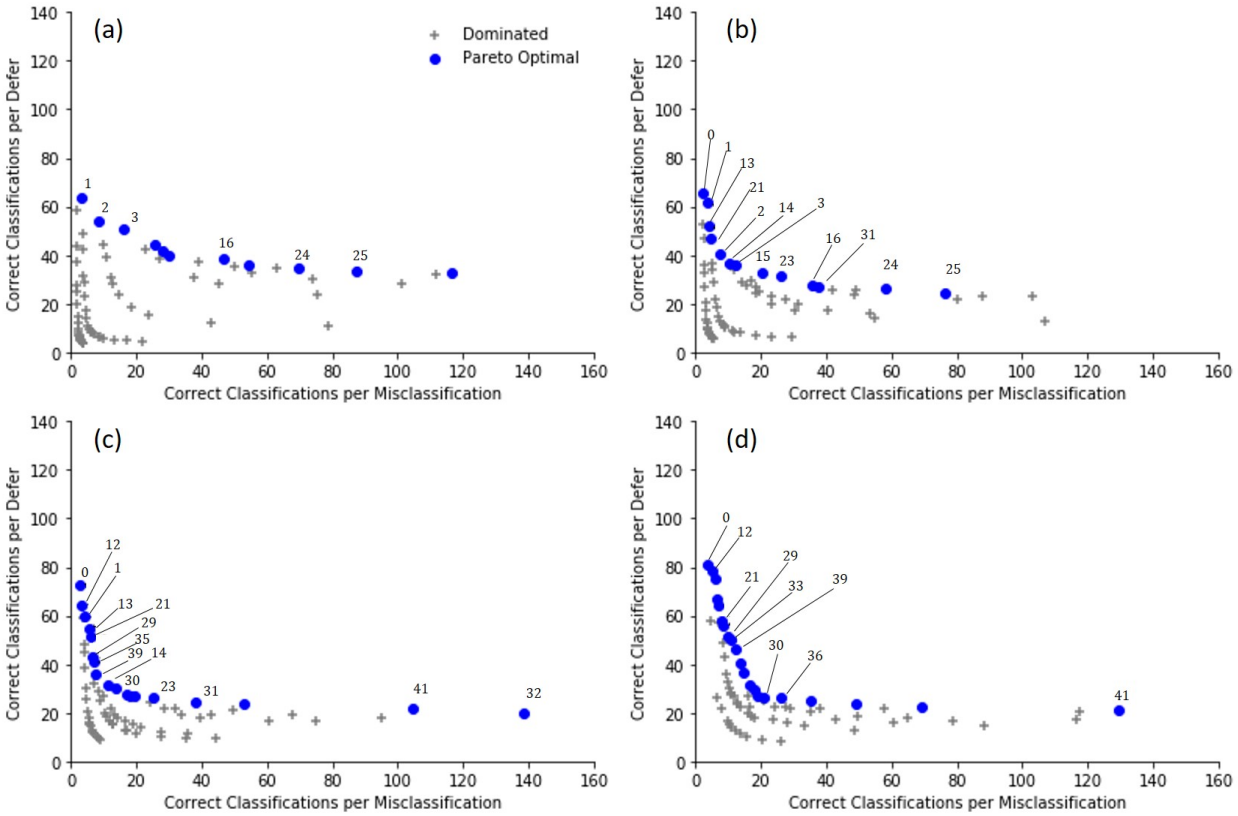


Figure 4.2: (a) $q = 0.2$. (b) $q = 0.4$. (c) $q = 0.6$. (d) $q = 0.8$. Rate of correct classifications per deferred decision (e_1) vs. Rate of correct classifications per misclassification (e_3) for TTP*-Q under various parameter combination and proportion of true-positives, q. Pareto optimal (circles) and dominated (crosses) parameter combinations are shown. Labels are shown only for parameter combinations which are Pareto optimal for at least two values of q. These labels correspond to parameter combinations in Table 4.3.

higher).

We also consider the tradeoff between e_1 and e_3 since decision makers may aim to correctly classify as many patients as possible, while minimizing deferred and incorrect classifications. We plot the values of e_1 and e_3 at different values of q for various combinations of γ^{fp} and γ^{fn} in Figure 4.2.

Certain parameter combinations are only Pareto optimal for one value of q, suggesting the importance of choosing the parameters γ^{fp} and γ^{fn} based on the proportion of true-positives in the testing population rather than the underlying population. In contrast, a few parameter combinations were relatively unaffected by changes in q, i.e., parameter combinations which

Table 4.3: Selected Pareto Optimal parameter combinations for efficiency analysis

	Parameters		*Opt Solution*		*Out-of-sample Performance*			
Label	γ^{fp}	γ^{fn}	u^*	l^*	**Sensitivity**	**Specificity**	**False-positive**	**False-negative**
0	0.01	0.01	0.94	0.02	0.83	0.57	0.01	0.01
1	0.01	0.02	0.94	0.05	0.83	0.74	0.01	0.01
2	0.01	0.03	0.94	0.12	0.83	0.89	0.01	0.04
3	0.01	0.04	0.94	0.22	0.83	0.95	0.01	0.05
12	0.02	0.01	0.73	0.02	0.89	0.57	0.02	0.01
13	0.02	0.02	0.73	0.05	0.89	0.74	0.02	0.01
14	0.02	0.03	0.73	0.12	0.89	0.89	0.02	0.04
15	0.02	0.04	0.73	0.22	0.89	0.95	0.02	0.05
16	0.02	0.05	0.73	0.31	0.89	0.97	0.02	0.06
21	0.03	0.02	0.54	0.05	0.91	0.74	0.02	0.01
23	0.03	0.04	0.54	0.22	0.91	0.95	0.02	0.05
24	0.03	0.05	0.54	0.31	0.91	0.97	0.02	0.06
25	0.03	0.06	0.54	0.38	0.91	0.97	0.02	0.07
29	0.04	0.02	0.40	0.05	0.93	0.74	0.03	0.01
30	0.04	0.03	0.40	0.12	0.93	0.89	0.03	0.04
31	0.04	0.04	0.40	0.22	0.93	0.95	0.03	0.05
32	0.04	0.05	0.40	0.31	0.93	0.97	0.03	0.06
33	0.04	0.06	0.40	0.38	0.93	0.97	0.03	0.07
35	0.05	0.02	0.32	0.05	0.94	0.74	0.03	0.01
36	0.05	0.03	0.32	0.12	0.94	0.89	0.03	0.04
39	0.06	0.02	0.27	0.05	0.95	0.74	0.04	0.01
41	0.06	0.04	0.27	0.22	0.95	0.95	0.04	0.05

remained on the Pareto optimal frontier for at least 2 different values of q. These parameter combinations are summarized in Table 4.3. From this analysis, we find that parameter combinations with low values of γ^{fp} (e.g., labels 1, 2, 3) tend to do well when the proportion of true-positives is low, i.e., $q = 0.2$ and $q = 0.4$. The opposite seems to hold as well — parameter combinations with low γ^{fn} (e.g., labels 12, 30, 39) tend to do well when the proportion of true-positives is high, i.e., $q = 0.6$ and $q = 0.8$. Some parameter combinations performed well for at least 3 values of q (e.g., labels 0, 1, 21). These parameter combinations strike a balance between γ^{fp} and γ^{fn}, though this pattern may not always hold (e.g., label 13). We also find that more conservative parameter combinations, i.e., low γ^{fp} and γ^{fn} such as labels 0 and 1, will achieve more correct classifications for each misclassification. In contrast, less conservative parameter combinations, i.e., higher values of γ^{fp} and γ^{fn} such as labels 33 and 41, will have more correct classifications for each deferred decision.

4.6.5 Comparing TTP* Performance To Existing Methods

We evaluated the model performance for TTP*-Q and TTP*-DR under different parameter combinations, i.e., ϕ, γ^{fp}, and γ^{fn}. We considered the following performance criteria:

(i) False-positive vs. false-negative rate satisfying minimum levels of sensitivity,

(ii) Sensitivity vs. false-positive rate satisfying maximum levels of false-negative rate, and

(iii) Probability of correct classification vs. probability of misclassification,

where the probabilities in criterion (iii) correspond to the definitions of $p^+(u,l)$ and $p^-(u,l)$ in Section 4.3.5. We computed $p^+(u,l)$ and $p^-(u,l)$ at different values of q to evaluate criterion (iii). Furthermore, we compared the performance of these two-threshold solutions against two baseline cases: an optimal one-threshold solution (1T*) and normative value comparison.

Optimal One-Threshold Solution: Applying sample average approximation on the training set of data (i.e., $N^+ = 560$ and $N^- = 706$), we estimated the solution to 1T* as defined in (4.23). We compare TTP to this threshold since the resulting threshold is akin to one chosen from the ROC curve as is commonly done in determining decision thresholds (see Section 4.2).

Normative Value Comparison: While the clinical examination remains to be the golden standard for concussion diagnosis, clinicians commonly use Normative Value Comparison (NC) as an initial acute concussion screening tool (Broglio et al., 2008; Chin et al., 2016; Hänninen et al., 2016; Zimmer et al., 2015). In NC, the clinician compares the performance of an athlete suspected of concussion on various standard assessments, e.g., the SAC, SCAT symptom checklist, and BESS. If the athlete's performance is too many standard deviations (SD) away from a given normative (i.e., mean) value, then the athlete is treated as concussed and otherwise non-concussed. Using the normative values for the training data in Table 4.1, we analyzed a number of one-threshold and two-threshold NC schemes by varying the number of SD away from the normative value (from 0.5 to 3 in increments of 0.25) on the SAC, SCAT symptom checklist, and the BESS. For example, we considered the case where an athlete is assessed as non-concussed if his or her performance on at least one assessment is within 1 SD of the normative value, concussed if beyond 2 SD, or deferred otherwise.

The results of our analysis for criteria (i) and (ii) are presented in Figure 4.3. Only Pareto optimal parameter combinations for TTP*-Q and TTP*-DR are shown. From Figure 4.3(a), we see that, for TTP*-Q and TTP*-DR, at least one of the Pareto optimal parameter combinations achieves a lower false-positive and false-negative rate than 1T* and NC while maintaining the same minimum level of sensitivity. However, when the sensitivity must be ≥ 0.96, TTP*-DR is too conservative and cannot produce a solution which satisfies this requirement. Furthermore, no solution from 1T* or NC is able to achieve at least 0.96 sensitivity while maintaining false-positive and false-negative rates below 0.15.

In Figure 4.3(b), we find that different parameter combinations for TTP*-Q and TTP*-DR are able to achieve similar or better levels of discrimination (i.e., high sensitivity and low false-positive rates) compared to 1T* and NC while keeping lower false-negative rates. Specifically, we point out that when the false-negative rate is constrained to be ≤ 0.01, no 1T* solution is able to achieve a similar performance to TTP*-Q and TTP*-DR. However, several TTP*-Q and TTP*-DR solutions still satisfy the maximum false-negative limit. We also point out that while NC is able to produce a solution with good discrimination at low false-negative rates, it is dominated by the TTP* solutions. That is, TTP*-Q and TTP*-DR are able to achieve similar levels of sensitivity and false-negative rates at lower false-positive rates.

The results for criterion (iii) are presented in Figure 4.4. We find that 1T* is able to match the performance achieved by TTP*-Q and TTP*-DR for one parameter combination. However, TTP* offers solutions with lower probability of misclassification compared to 1T*. Similarly to the analysis of criterion (ii), we find that the TTP* solutions dominate NC in terms of performance. That is, for the same probability of misclassification, each of the TTP* methods achieve a higher probability of correct classification. This finding demonstrates the utility of two-threshold solutions in combination with risk estimation models compared to traditional methods.

Overall, these results suggest that TTP* outperforms both 1T* and NC when comparing across multiple criteria. In particular, the two-threshold solutions are able to maintain high diagnostic accuracy while maintaining low levels of false-positive and false-negative rates.

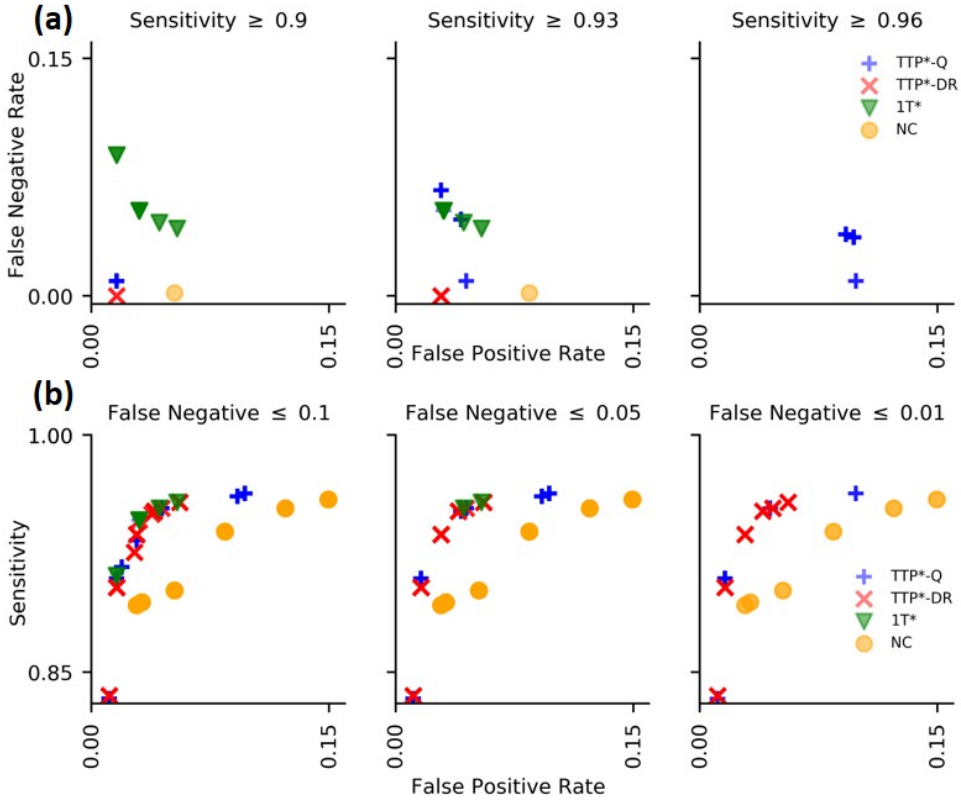


Figure 4.3: (a) False-positive vs. false-negative rate satisfying minimum levels of sensitivity. (b) Sensitivity vs. false-positive rate satisfying maximum levels of false-negative rate. Comparison of Pareto optimal parameter combinations for TTP*-Q, TTP*-DR, optimal single-threshold solutions (1T*), and normative value comparison (NC).

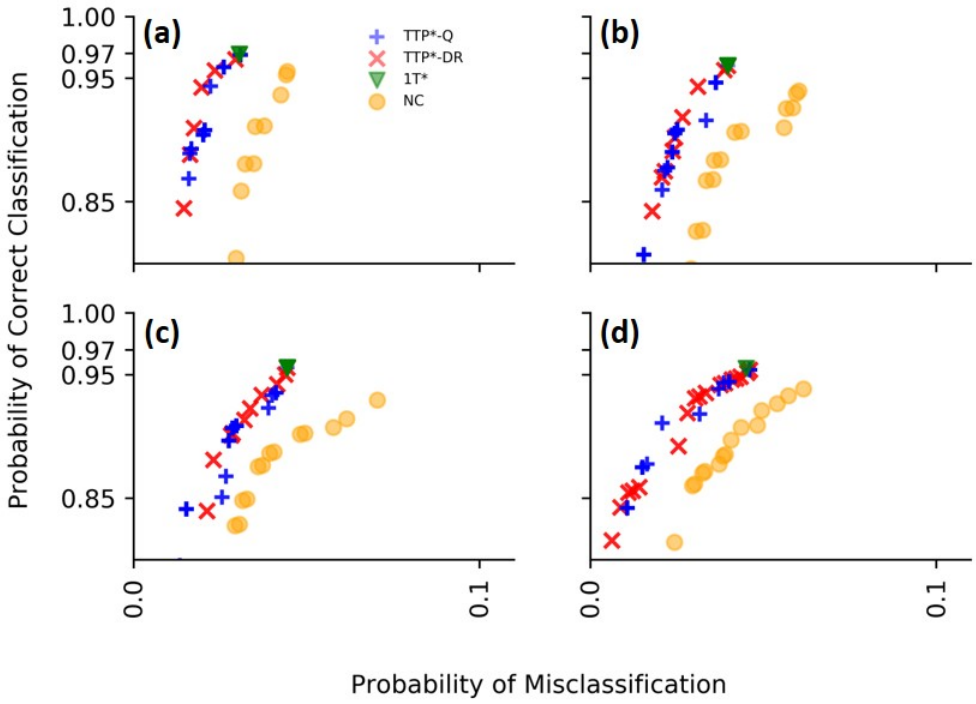


Figure 4.4: (a) $q = 0.2$. (b) $q = 0.4$. (c) $q = 0.6$. (d) $q = 0.8$. **Probability of correct classification vs. probability of misclassification for varying proportions of true-positives, q, and Pareto optimal parameter combinations of TTP*-Q , TTP*-DR, optimal single-threshold solutions (1T*), and normative value comparison (NC).**

4.7 Conclusion

In conclusion, we have designed and analyzed a data-driven method for optimizing an upper and lower threshold for diagnosis decisions. This method (1) reflects the decision maker's risk attitude, (2) determines data-driven thresholds based on risk estimates specific to a certain risk estimation model and target population, and (3) identifies a range of risk estimates for which the risk estimation model faces elevated false-positive and false-negative rates. Through data-driven solution approaches, we avoid needing assumptions on the distribution of risk estimates.

4.7.1 Managerial Insights

Our analysis provides insight into the nature of our modeling framework and its application to acute concussion assessment.

When should two diagnosis decision thresholds be used instead of one? Since the advantage of using two decision thresholds is in its ability to defer patients, the benefit of using two thresholds (instead of one) depends on the risk estimation model's accuracy and the accuracy of the assessments which follow a deferred diagnosis decision. Specifically, if a risk estimation model is inaccurate and patients who are deferred can eventually be diagnosed with sufficient accuracy, then two decision thresholds should be used. We also find that two diagnosis decision thresholds are needed (e.g., possible, probable, or definite concussion) when it is important to restrict both the false-positive and false-negative rates. However, if it is only important to limit either the false-positive or false-negative rate, then a single decision threshold (e.g., concussion or no concussion) suffices. Fortunately, for acute concussion assessment, the added conservatism by restricting misclassification rates does not drastically reduce sensitivity and specificity. Furthermore, two thresholds should be used if the cost associated with misdiagnosis is sufficiently high compared to a deferred diagnosis decision. Alternatively, if the majority of deferred decisions consists of patients who ultimately would have been misdiagnosed or if very few deferred decisions are ultimately misdiagnosed, then two decision thresholds should be used instead of one.

How does one choose parameters which generate good thresholds? For TTP, the decision-maker only needs to consider the constraints on false-positive and false-negative rates, since the relative importance of sensitivity to specificity is trumped by these constraints

in two-threshold solutions. To this end, the degree to which false-positive and false-negative rates are constrained should reflect the relative "cost" associated with each type of misdiagnosis. Furthermore, the false-negative rate should be further constrained if the proportion of true-positives in the testing population is high. However, if the proportion of true-positives is low, then the false-positive rate should be constrained instead. In the case of acute concussion assessment, our efficiency analysis demonstrated that balanced constraints (i.e., γ^{fp} and γ^{fn} are both low and nearly equal) can be robust for different proportions of true-positives in the testing population.

Given the similarities between TTP and MTTP, similar principles can be applied in choosing parameters for MTTP. However, choosing parameters for OTP can be challenging. In our analysis of OTP, we found that the parameters $\gamma^{fp}_{k,j}$ and $\gamma^{fn}_{k,j}$ must satisfy certain conditions in order for OTP to be feasible. To this end, when using the sensitivity-based formulation of OTP (i.e., the objective function is given by (4.14a)), one can satisfy these conditions by setting $\gamma^{fn}_{k,j} = 1$ for all $j > k$ and focus on choosing appropriate values for $\gamma^{fp}_{k,j}$ for all $j \leq k$. In our analysis, we found that the performance of different parameter choices were sensitive to the underlying quality of the risk estimation model. For example, when the risk estimation model was poor, choosing high values of $\gamma^{fp}_{k,j}$ (e.g., 0.11-0.13) was imperative to attaining high accuracy and low mean squared error — regardless of sample size. On the other hand, parameter sets which had low values for $\gamma^{fp}_{k,j}$ (e.g., 0.01-0.05) achieved high accuracy at low mean squared error when the risk estimation models were high quality. Given these guiding principles, one can construct a set of viable parameters and then use methods such as cross-validation to evaluate the performance of different parameter choices.

Which data-driven solution method should be used? In our simulation analysis, we analyzed the performance of each solution method under different sample sizes and the quality of the underlying risk estimation model. If the available sample size is small and the underlying risk estimation model is poor, then TTP*-DR should be applied if ensuring feasibility is important to the application. Additionally, TTP*-DR can be useful when the available sample of data is suspected to be quite different from the distribution generated by the population. However, as sample sizes get larger, the computational burden of TTP*-DR increases to the point where it may not be practical to implement. To this end, TTP*-Q can be executed much more quickly at large sample sizes and performs similarly to TTP*-DR. Hence, TTP*-Q should be used once the sample size is large enough and the underlying

quality of the risk estimation model is reasonably accurate.

How does this methodology compare to existing methods? For acute concussion assessment, combining two-threshold solutions with risk estimation models allows for both greater diagnostic accuracy and lower likelihood of misdiagnoses. In particular, given the same risk estimation model, one-threshold solutions are adept at providing higher sensitivity and specificity but at the cost of greater misclassification rates compared to two-threshold solutions. Compared to common clinical practice such as normative value comparison, combining risk estimation models with data-driven thresholds can drastically improve diagnostic accuracy. Therefore, this modeling framework should be transformed into a clinical decision aid to facilitate its implementation in practice.

In our extensions of TTP to multi-class classification, we found that OTP outperforms a common method for determining decision thresholds in ordinal classification. Specifically, OTP attained higher accuracy and lower mean squared error than a commonly used equidistant threshold method. This improvement in performance can be owed to the fact that OTP has greater flexibility in optimizing thresholds given that it does not require thresholds to be equidistant. However, this improved performance comes at a cost of requiring more parameters. To this end, we have identified some principles to guide the choice of parameters for OTP.

How does this work impact diagnosis decisions in clinical practice? It is natural for clinicians to have varying levels of certainty in their diagnosis decisions based on their initial evaluation of a patient. For instance, consider concussion diagnosis, for which no "gold standard" objective test currently exists, common concussion symptoms are sometimes slow to evolve and not necessarily specific to concussion, and clinical presentation of concussion may vary largely from one patient to the next (Kutcher and Giza, 2014). While the clinical exam remains the standard in this field, our research can provide valuable decision support to clinical judgment by quantifying this uncertainty with real data. Our modeling framework provides a more objective and data-driven way to guide risk-based diagnosis decisions by identifying "easy" cases which can be diagnosed immediately and "hard" cases which may require further evaluation before diagnosing. We believe that our modeling framework permits clinicians to blend their expertise with quantitative evidence; while it does not recommend a next step for patients who are deferred, it has the flexibility to incorporate clinical judgment when necessary. Directing the expertise of clinicians to those patients who most

need it can have a great impact in improving patient care while reducing unnecessary workload for clinicians. To facilitate the use of TTP and its extensions in clinical practice, the models developed in this research may initially undergo a pilot in clinic. After sufficient validation, it may eventually be incorporated into a mobile application or integrated with electronic health records. For example, MTTP can be paired with electronic health records to alert clinicians when patients may be at relatively high risk for developing any number of new conditions. By supplementing clinical decision-making, this research has the potential to improve diagnosis decisions, patient care, and health outcomes.

4.7.2 Limitations and Future Work

This research can be extended to address limitations in our current modeling framework. First, this work only determines two thresholds but some risk classifications may fall into many more categories. Consider the diagnosis of pulmonary hypertension, where patients may be classified into 1 of 5 groups (Galie et al., 2009). In this case, 2 thresholds may not be enough. While we have shown that TTP can be extended or modified to fit multi-label and ordinal classification frameworks, the extension to other types of multi-class classification (e.g., one-vs-all classification) is not so straightforward due to the potential for conflicting results.

The impact of time has been absent in our model formulations and analyses. However, in time-critical applications, it is important to consider whether deferring decisions is still a plausible action to take. Therefore, our modeling framework can be extended to consider the impact that this additional factor would have on our results.

Throughout this work, we assume a fixed risk estimation model. However, one may wish to jointly optimize a risk estimation model and its accompanying decision thresholds. Particularly, when the same data is used to formulate the risk estimation model and the thresholds, one may need to optimize the proportion of data used for the risk estimation model and the threshold problem.

In practice, differences between sample data and the true population distribution can act as barriers to implementing risk estimation models in practice. While our distributionally robust formulation is useful for dealing with such situations, one may obtain decision thresholds which are too conservative to be practically useful. Future research can investigate the

utility of different ambiguity sets for such scenarios.

Finally, this model was motivated by medical diagnosis problems, but may be applied to other application domains where the determination of a dichotomous outcome is critical. Examples of such applications include bankruptcy prediction or natural disaster forecasting.

As medical data continues to become more readily available, methods which can incorporate these data to supplement medical decisions become increasingly relevant. While several extensions can be made to our work, the framework we have developed in this research provides a promising baseline for applying and understanding data-driven risk estimates in diagnosis decisions. Through future implementation and validation in clinical practice, it is our hope that this modeling framework ultimately leads to an improved quality of healthcare delivery.

4.A Quantile Estimation Representations of TTP

We first show that we can rewrite the constraints (4.1b) and (4.1c) as chance constraints. That is,

$$\begin{aligned}\mathbb{E}[fp(u,\xi^-)] \le \gamma^{fp} &\iff \mathbb{P}(u \ge \xi^-) \ge 1-\gamma^{fp} \text{ and}\\ \mathbb{E}[fn(l,\xi^+)] \le \gamma^{fn} &\iff \mathbb{P}(l \le \xi^+) \ge 1-\gamma^{fn}.\end{aligned}$$

Now, we show that the values $u(\gamma^{fp})$ and $l(\gamma^{fn})$ are quantiles. To show this result, we require a preliminary definition. For any $\alpha \in (0,1)$, define the α-quantile $\theta_{\mathcal{X}}(\alpha)$ for the random variable X with distribution $\mathbb{P}_{\mathcal{X}}$ as

$$\theta_{\mathcal{X}}(\alpha) := \inf\left\{x \in \mathbb{R} : \alpha \le \mathbb{P}_{\mathcal{X}}(X \le x)\right\}. \tag{4.26}$$

From the definition of $u(\gamma^{fp})$ and the chance constraint representation of TTP*, we have

$$\begin{aligned}u(\gamma^{fp}) &= \inf\left\{u \in [0,1] : \mathbb{E}[fp(u,\xi^-)] \le \gamma^{fp}\right\}\\ &= \inf\left\{u \in [0,1] : \mathbb{P}(\xi^- \ge u) \le \gamma^{fp}\right\}\\ &= \inf\left\{u \in [0,1] : 1-\gamma^{fp} \le \mathbb{P}(\xi^- \le u)\right\}\\ &= \theta_{\mathcal{P}^-}(1-\gamma^{fp}).\end{aligned}$$

Thus, $u(\gamma^{fp})$ is the $(1-\gamma^{fp})$-quantile of $\mathcal{P}^-$. Similarly, we can show that

$$\begin{aligned}l(\gamma^{fn}) &= \sup\left\{l \in [0,1] : \mathbb{E}[fn(l,\xi^+)] \le \gamma^{fn}\right\}\\ &= \sup\left\{l \in [0,1] : 1-\gamma^{fn} \le \mathbb{P}(\xi^+ \ge l)\right\}\\ &= -\inf\left\{-l \in [-1,0] : 1-\gamma^{fn} \le \mathbb{P}(-l \ge -\xi^+)\right\}\\ &= -\theta_{\tilde{\mathcal{P}}^+}(1-\gamma^{fn}),\end{aligned}$$

where $\tilde{\mathcal{P}}^+$ represents the probability distribution of $-\xi^+$. Thus, $l(\gamma^{fn})$ can be taken to be the negative value of the $(1-\gamma^{fn})$-quantile of $\tilde{\mathcal{P}}^+$.

4.B Reformulating TTP*-DR Constraints

In this section, we show the tractable reformulations for the constraints (4.20a) and (4.20b) based on the reformulations derived in Mohajerin Esfahani and Kuhn, 2018. First, suppose that the thresholds u and l are fixed and let ϵ be a chosen Wasserstein radius. Then,

(a) the false-positive constraint (4.20a) can be reformulated as the following set of constraints:

$$\lambda\epsilon + \frac{1}{N^-}\sum_{n=1}^{N^-} s_n \leq \gamma^{fp} \tag{4.27a}$$

$$(z_{n,1} - v_{n,1})u + 1 + v_{n,1} - z_{n,1}\hat{\xi}_n^- \leq s_n, \quad n = 1, ..., N^- \tag{4.27b}$$

$$(z_{n,1} - v_{n,1})u + 1 - z_{n,1}\hat{\xi}_n^- \leq s_n, \quad n = 1, ..., N^- \tag{4.27c}$$

$$v_{n,2} - z_{n,2}\hat{\xi}_n^- \leq s_n, \quad n = 1, ..., N^- \tag{4.27d}$$

$$-z_{n,2}\hat{\xi}_n^- \leq s_n, \quad n = 1, ..., N^- \tag{4.27e}$$

$$z_{n,1} - v_{n,1} \leq 0, \quad n = 1, ..., N^- \tag{4.27f}$$

$$z_{n,2} - v_{n,2} = 0 \quad n = 1, ..., N^- \tag{4.27g}$$

$$z_{n,k} \leq \lambda, \quad n = 1, ..., N^-,\ k = 1, 2 \tag{4.27h}$$

$$-z_{n,k} \leq \lambda, \quad n = 1, ..., N^-,\ k = 1, 2. \tag{4.27i}$$

(b) the false negative-constraint (4.20b) can be reformulated as the following set of con-

straints:

$$\begin{aligned}
\lambda\epsilon + \frac{1}{N^+}\sum_{n=1}^{N^+} s_n &\leq \gamma^{fn} \\
(z_{n,1} - v_{n,1})l + 1 + v_{n,1} - z_{n,1}\hat{\xi}_n^+ &\leq s_n, \quad n = 1, ..., N^+ \\
(z_{n,1} - v_{n,1})l + 1 - z_{n,1}\hat{\xi}_n^+ &\leq s_n, \quad n = 1, ..., N^+ \\
v_{n,2} - z_{n,2}\hat{\xi}_n^+ &\leq s_n, \quad n = 1, ..., N^+ \\
-z_{n,2}\hat{\xi}_n^+ &\leq s_n, \quad n = 1, ..., N^+ \\
z_{n,1} - v_{n,1} &\geq 0, \quad n = 1, ..., N^+ \\
z_{n,2} - v_{n,2} &= 0 \quad n = 1, ..., N^+ \\
z_{n,k} &\leq \lambda, \quad n = 1, ..., N^+, \; k = 1, 2 \\
-z_{n,k} &\leq \lambda, \quad n = 1, ..., N^+, \; k = 1, 2.
\end{aligned}$$

We now derive the reformulation for (4.20a). Fix the choice of upper threshold u and let

$$\begin{aligned}
\ell(\xi^-) &= \max\{\ell_1(\xi^-), \ell_2(\xi^-)\}, \text{ where} \\
\ell_1(\xi^-) &= \begin{cases} 1 & u \leq \xi^-, \\ -\infty & \text{otherwise}, \end{cases} \\
\ell_2(\xi^-) &= 0.
\end{aligned} \tag{4.29}$$

Over the support $\Xi = [0, 1]$, it can be seen that $\ell(\xi^-) = fp(u, \xi^-)$ for $u \in [0, 1)$ and $u \neq \xi^-$. However, under assumption A1, the case where $u = \xi^-$ is no difficulty. In addition, we can always set (4.20a) to be equal to 0 when $u = 1$. Now, notice that the set Ξ is convex and closed. Furthermore, the functions $-\ell_1$ and $-\ell_2$ are proper, convex, and lower

semicontinuous. Therefore, by Theorem 4.2 in Mohajerin Esfahani and Kuhn, 2018, we have

$$\begin{aligned}
\sup_{\mathcal{Q}\in D^-} \quad & \mathbb{E}_{\mathcal{Q}}[fp(u,\xi^-)] = \\
\inf_{\lambda,s_n,z_{n,k},v_{n,k}} \quad & \lambda\epsilon + \frac{1}{N^-}\sum_{n=1}^{N^-} s_n \\
\text{s.t.} \quad & [-\ell_k]^*(z_{n,k}-v_{n,k}) + \sigma_\Xi(v_{n,k}) - z_{n,k}\hat{\xi}_n^- \le s_n, \qquad n=1,...,N^-,\ k=1,2 \quad (4.30)\\
& ||z_{n,k}||_* \le \lambda \ n=1,...,N^-,\ k=1,2, \quad (4.31)
\end{aligned}$$

where the functions

$$\begin{aligned}
[-\ell_k]^*(z) &= \sup_{\xi\in\mathbb{R}} z\xi + \ell_k(\xi),\\
\sigma_\Xi(v) &= \sup_{\xi\in\Xi} v\xi,\\
||z||_* &= \sup_{|\xi|\le 1} z\xi = |z|,
\end{aligned}$$

are defined in Mohajerin Esfahani and Kuhn, 2018.

Based on our choices for ℓ_1 and ℓ_2, we have

$$[-\ell_1]^*(z) = \sup_{\xi\ge u} z\xi + 1 = \begin{cases} +\infty & z>0 \\ zu+1 & z\le 0, \end{cases} \quad (4.32)$$

$$[-\ell_2]^*(z) = \begin{cases} 0 & z=0 \\ +\infty & \text{otherwise} \end{cases}. \quad (4.33)$$

Additionally, since $\Xi=[0,1]$, we have

$$\sigma_\Xi(v) = \sup_{\xi\in[0,1]} v\xi = \max\{v,0\}. \quad (4.34)$$

Substituting (4.32) - (4.34) into (4.30) and using the following equivalence for (4.34)

$$v_{n,1}^+ \le s_n \iff \begin{cases} v_{n,1}\le s_n \\ 0\le s_n \end{cases}$$

gives us (4.27b)-(4.27g). Furthermore, reformulating (4.31) as

$$||z_{n,k}||_* = |z| \leq \lambda \iff \begin{cases} z_{n,k} \leq \lambda \\ -z_{n,k} \leq \lambda \end{cases}$$

gives us (4.27h)-(4.27i). Thus we have shown the reformulation for (4.20a). To show the reformulation for (4.20b), let

$$\ell_1(\xi^+) = \begin{cases} 1 & l \geq \xi^+, \\ -\infty & \text{otherwise}, \end{cases} \tag{4.35}$$

and define $\ell(\xi^+)$ to be the same as in (4.29). The remainder of the derivation follows similarly as was shown for (4.20a).

4.C Simulation Analysis of Section 4.5 With Trapezoidal Distribution

In this section, we replicate our simulation analysis in Section 4.5 under different distributional assumptions. Specifically, we performed our simulation analysis under the following settings.

Underlying risk estimation distributions: Let

$$\mathcal{P}^+ = \text{Trapezoidal}(a^+, b^+, c^+, d^+) \text{ and } \mathcal{P}^- = \text{Trapezoidal}(a^-, b^-, c^-, d^-),$$

where the parameters $a^+ = a^- = 0$ and $b^+ = b^- = 1$ to ensure that the distributions $\mathcal{P}^+, \mathcal{P}^-$ have support $[0, 1]$. We varied the parameters $(c^+, d^+) \in \{(0.6, 0.7), (0.7, 0.85), (0.85, 0.95)\}$. For each pair (c^+, d^+), we set $c^- = 1 - d^+$ and $d^- = c^- + (d^+ - c^+)$. Each of these distributional set-ups resulted in an area under the receiver operating characteristic curve (AUROC) of $0.649, 0.748$, and 0.810, respectively.

As was done in Section 4.5, we solved TTP*-Q and TTP*-DR 100 times with $\gamma^{fp} = 0.10$,$\gamma^{fn} = 0.05$ under varying sample sizes $N^+, N^- \in \{100, 500, 1000\}$. Both models were evaluated based on optimality gap and maximum constraint violation, as defined in (4.21)

and (4.22), respectively.

Our results are illustrated in Figure 4.C.1.

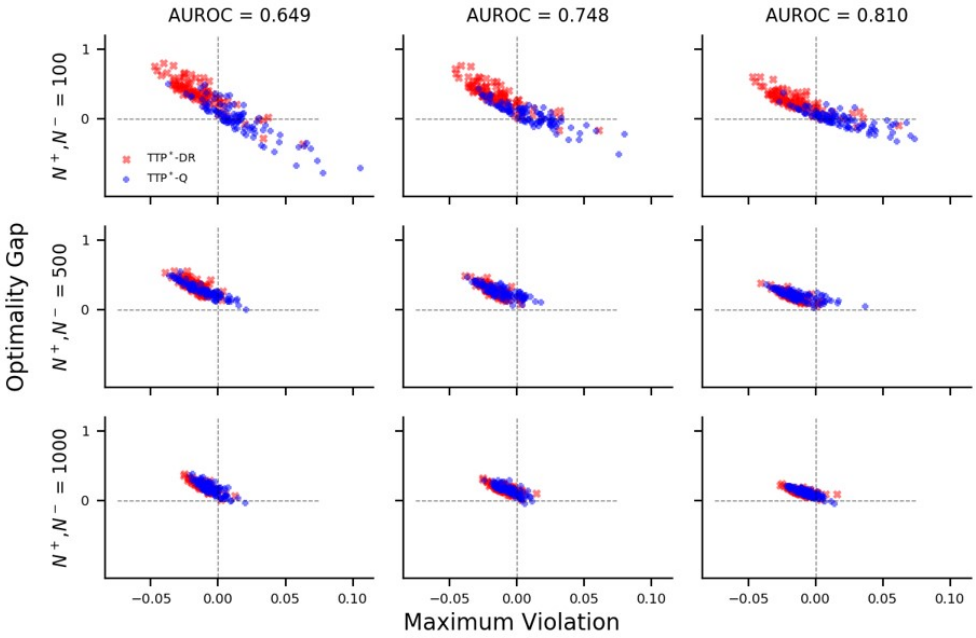


Figure 4.C.1: Distribution of optimality gap and maximum constraint violation under varying quality of risk estimation model (AUROC) and sample size (N^+, N^-). AUROC, area under the receiver operating characteristic curve.

In this analysis, we find that sample size and AUROC play an important role in both feasibility and optimality compared to the true optimal solution. For TTP*-Q, feasibility is low (18%-24%) at a sample size of 100. Once the sample size increases to 500, feasibility increases above 80%, though we find that increasing the sample size to 1000 does not further increase feasibility. Furthermore, increasing AUROC does not correspond to an increase in feasibility until the sample size is 1000. However, increasing AUROC does correspond to a decrease in optimality gap once the sample size is at least 500.

TTP*-DR is largely feasible (83%-86%) at a sample size of 100 and this feasibility increases above 94% for larger sample sizes. Much like TTP*-Q, increasing AUROC does not correspond with an increase in feasibility for TTP*-DR. Unlike TTP*-Q, both increases in AUROC and sample size correspond with decreases in optimality gap for TTP*-DR. When the sample size is at least 500 and AUROC is at least 0.748, we find that TTP*-Q and

TTP*-DR have similar optimality gaps, though TTP*-DR has much greater feasibility.

Comparing this analysis with the one performed in Section 4.5, we find that the proportion of TTP*-Q instances which are feasible is generally higher when the risk estimates follow a Beta distribution instead of a Trapezoidal distribution, even if both distributions have a similar AUROC. For example, with a sample size of 1000, when the risk estimates follow a Beta distribution with an AUROC of 0.680, 95% of the instances were feasible. In contrast, when the risk estimates follow a Trapezoidal distribution with an AUROC of 0.649, only 79% of instances are feasible. We suspect that these differences may be caused by the variance of the distributions used. That is, the Trapezoidal distribution has a higher variance than the Beta distribution with a similar AUROC. Larger sample sizes are needed to accurately estimate optimal thresholds when the risk estimates have distributions with high variance.

Chapter 5

Data-driven Approach to Unlikely, Possible, Probable, and Definite Concussion

5.1 Introduction

In this chapter, we apply the framework from Chapter 4 and risk estimation model from Chapter 2 to develop a certainty-based framework for acute concussion diagnosis. In particular, we extend the framework proposed by Kutcher and Giza, where they recommended incorporating diagnostic certainty to the assessment of concussion (Kutcher and Giza, 2014). That is, rather than a binary diagnosis paradigm (i.e., concussion or no concussion), Kutcher and Giza suggested that concussion diagnosis should be relayed across a spectrum of risk categories (e.g., Possible, Probable, and Definite concussion), with each category reflecting the degree to which a concussion diagnosis is certain. Similar risk-based categories have been used for classifying diagnosis decisions for other diseases, including multiple sclerosis (McDonald et al., 2001), Alzheimer's disease (McKhann et al., 2011), and diabetes (American Diabetes Society, 2016). Compared to traditional binary diagnosis, risk-based diagnosis frameworks account for the evolution of the injury over time and allows for more flexibility in the post-injury management of concussion. Specifically, incorporating certainty in the assessment of concussion can help to determine whether an athlete should be managed as if he or she has a concussion, ultimately improving the quality of patient care. However,

the guidelines developed by Kutcher and Giza were based on clinical experience rather than objective data.

Therefore, the goal of our study is to create a data-driven modeling framework to identify concussed and non-concussed athletes who are Unlikely to have concussion and classify the remaining athletes as having Possible, Probable, or Definite concussion, with each category reflecting increasing diagnostic certainty. While experienced clinicians may be able to quickly synthesize the likelihood of concussion and ultimately identify a post-injury management plan for athletes, our data-driven framework provides a more objective approach which can ultimately benefit those clinicians who may be inexperienced in managing concussion. We then aim to validate our framework to identify how athletes are distributed across each risk classification and identify differences in demographics, time-of-injury characteristics, and standard assessment scores between athletes under each risk classification.

5.2 Materials and Methods

5.2.1 Study Population and Design

To develop our methodology for classifying athletes as Unlikely, Possible, Probable, or Definite concussion, we used data from the CARE Consortium (see Section 1.1.1).

5.2.2 Sample Selection

In our analysis, we focused on the timepoints at baseline, <6h, 24-48h, and unrestricted RTP. We only included baseline data which could be matched with post-injury data. The assessments at <6h and 24-48h were denoted "acute concussion" and those from baseline and unrestricted RTP were denoted "normal performance". We consider those from the unrestricted RTP timepoint to demonstrate a normal performance because they have been cleared for RTP by each institution's local medical staff. We analyzed <6h and 24-48h separately.

5.2.3 Study Variables

For each participant in the study data, we obtained demographic information along with raw scores on the Standard Assessment of Concussion (SAC), Standard Concussion Assessment Tool (SCAT) symptom survey, and the Balance Error Scoring System (BESS) at baseline. For those diagnosed with concussion, we obtained time-of-injury characteristics along with raw scores for SAC, SCAT symptoms, and BESS scores at each post-injury evaluation timepoint. We computed the change score for these athletes by subtracting the raw score at baseline from the raw score at each post-injury timepoint. A positive change score indicates an increase in the measure compared to baseline, whereas a negative change score indicates a decrease compared to baseline. We also filled missing data elements using multiple imputation by chained equations (Royston, 2004). Our modeling variables include:

- **Demographic variables:** age, sex, and number of previous concussions
- **Time of injury characteristics:** whether the athlete experienced loss of consciousness (LOC), post-traumatic amnesia (PTA), retrograde amnesia (RGA), whether the athlete was removed from play immediately, and whether the injury was reported immediately
- **Standard concussion assessments:** SAC total score, SCAT total number of symptoms, SCAT total symptom severity, and BESS total score.

These variables are previously described in Chapter 2.

5.2.4 Data Analysis

Our overall framework for classifying Unlikely, Possible, Probable, or Definite concussion is summarized in Figure 5.1. To create and evaluate our models, we divided our post-injury data into a training set and a validation set. The training set consisted of all data collected between January 23, 2014 and November 29, 2016 while the validation data consisted of all data collected after that date (i.e., November 30, 2016 to October 2, 2017). The CARE Consortium protocol for concussion diagnosis along with assessments performed at baseline and post-injury remained unchanged during this period and thus, rater drift was minimal, if any. We used our training set to develop the models to determine which athletes should

be classified under each risk category. Then, we applied our models to the validation data to evaluate and analyze our framework. We describe each of the steps in our methodology in more detail below.

5.2.5 Model Calibration

Using a randomly chosen 40% of the training data, we created a logistic regression model for estimating risk scores. For any athlete, risk scores are a scalar between 0 and 1, where greater risk scores indicate higher likelihood of acute concussion. We used the multivariate logistic regression model (i.e., the raw score model) from Chapter 2 to estimate risk scores associated with athletes at <6h and 24-48h. We used the raw score models since change scores could not be computed for baseline data and may not always be available for acute concussion assessments in clinical settings. Since time-of-injury characteristics were not available for baselines but were part of these previous models, we assumed in this logistic regression analysis only that for baseline data, injuries were reported immediately and that participants were removed from play immediately.

With the remaining 60% of the training data, we determined risk score thresholds to identify Unlikely and Definite concussions. We first applied our logistic regression models to this subset of training data to obtain risk scores for each athlete. Then, we used these risk scores as the input for the data-driven optimization algorithm of Chapter 4 to determine risk score thresholds. Athletes with risk scores below the lower threshold represent those whose concussion probability was low and thus would be identified as Unlikely concussion. Similarly, athletes with risk scores above the upper threshold are most likely to have concussion and would be classified as Definite concussion. In designing these thresholds, we favored higher sensitivity over lower false-positive rates and lower false-negative rates over higher specificity.

After determining the upper and lower risk score thresholds, we identified athletes in the training set with risk scores between Unlikely and Definite thresholds and used them to determine how athletes should be classified as Possible or Probable, as these cases could not be easily distinguished by our logistic regression model. Categorization of these cases was approached using a classification and regression tree (CART) (Breiman et al., 1984) analysis. CART is a non-parametric statistical modeling technique which produces a decision tree for prediction and is capable of handling categorical variables and continuous variables.

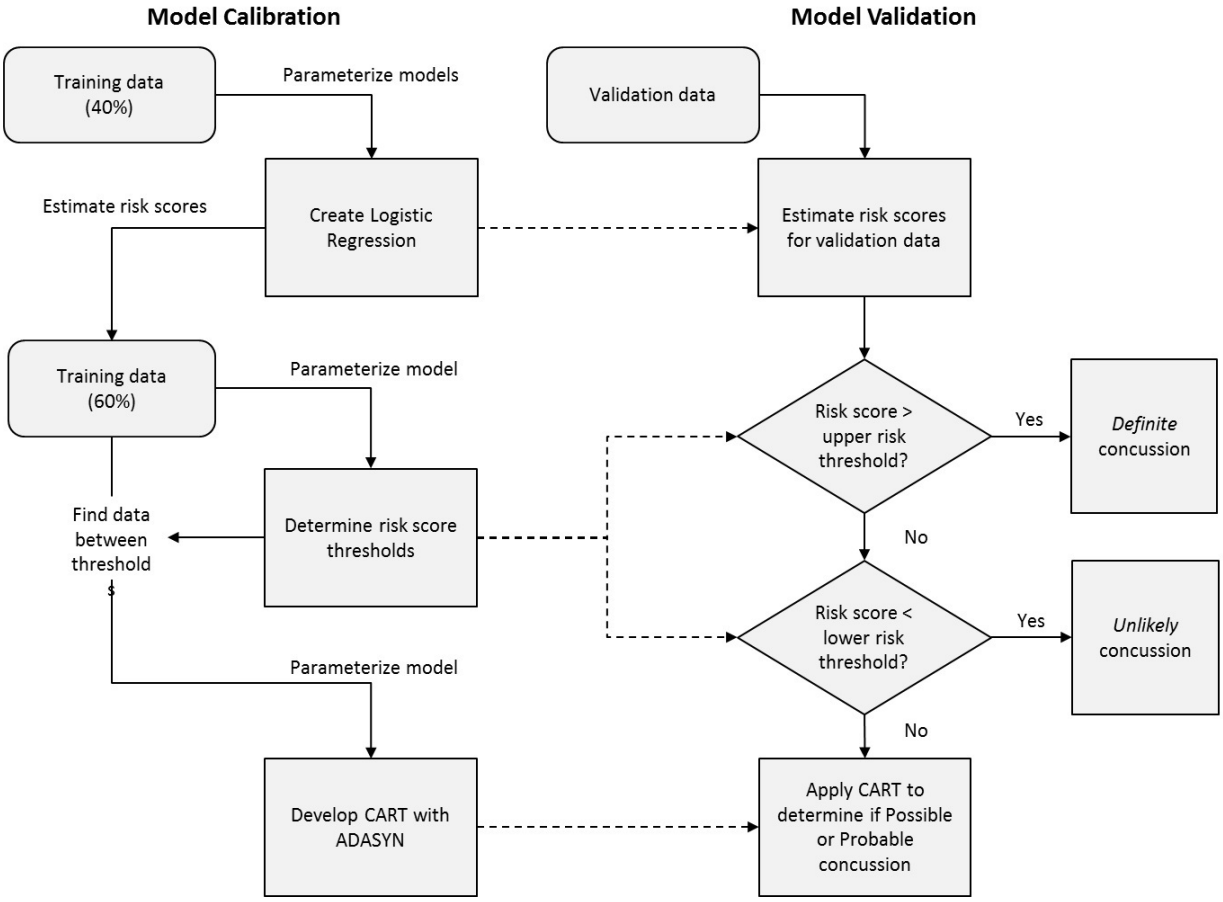


Figure 5.1: Illustration of methodological framework for developing data-driven models which were used to classify athletes as Unlikely, Possible, Probable, or Definite concussion based on certainty of acute concussion. CART, Classification Tree; ADASYN, Adaptive Synthetic Sampling

Compared to other predictive modeling methods (e.g., generalized linear models), CART is advantageous in its interpretability and ability to model highly non-linear relationships between variables. Due to the higher proportion of normal performances to acute concussions in our data, we applied Adaptive Synthetic Sampling (ADASYN) to mitigate data imbalance issues before creating a CART (He et al., 2008). Additionally, for this CART, we restricted the resulting decision tree to include only variables which were available for all timepoints. That is, the resulting decision tree did not include time-of-injury characteristics and change scores for the SAC, SCAT symptom assessments, and the BESS since they were not available for baseline data. Athletes who were predicted to be acute concussions by this CART were classified as Probable concussions while those who were predicted to be normal performances were classified as Possible concussions.

5.2.6 Model Validation

To implement our models, we applied our logistic regression models to the validation data to obtain risk scores for each athlete. Then, we compared these risk scores to the upper and lower thresholds we generated using our optimization algorithm in the model calibration phase. Athletes with risk scores below the lower threshold were classified as Unlikely concussions, while athletes with risk scores above the upper threshold were classified as Definite concussions. We then applied our CART to any athlete with a risk score between these thresholds to classify them as Possible or Probable concussion.

5.2.7 Model Evaluation

After implementing our models on the validation data, we performed additional analysis to evaluate the performance of our classification framework. The goals of this analysis were to (1) analyze how our models classified acute concussions and normal performance throughout each risk category and (2) identify interclass differences (i.e., across different risk classifications) and intraclass differences (i.e., within the same risk classification) in demographics, time-of-injury characteristics, and standard assessment scores for acute concussions and normal performances among the risk classifications.

To achieve the first goal, we determined the percentage of acute concussions and normal performances within each risk classification at both <6h and 24-48h. Ideally, data captured

in the acute post-injury state should place the athlete in greater risk classifications (i.e., Definite or Probable), while data captured at baseline should place the athlete into lower risk classifications (i.e., Unlikely or Possible). We compared the distribution of acute concussions and normal performances using the Kolmogorov-Smirnov test. A significant p-value for this test implies that the distribution of acute concussions and normal performances among the risk classifications is dissimilar.

Since our diagnosis scheme consisted of 4 risk categories instead of 2, we also computed a modified sensitivity and specificity. Our modified computation was founded in recommendations by Kutcher and Giza who indicated that Probable and Definite concussions should be managed as concussions, while Possible concussions should be managed based on clinical judgment. Furthermore, we assume that Unlikely concussions are managed as non-concussions. Therefore, we provide a sensitivity range where the lower bound reflects the proportion of acute concussions that are correctly classified as Probable and Definite and the upper bound reflects the proportion of acute concussions correctly classified as Possible, Probable and Definite. We also provide a range for specificity, where the lower bound reflects the case where no Possible concussions are treated as non-concussed and an upper bound which reflects the case where all Possible concussions are treated as non-concussed. In practice, the true sensitivity and specificity should fall between these bounds depending on how Possible concussions are managed.

To achieve the second goal, we first identified interclass differences in the study variables across each risk classification for acute concussions and normal performances using analysis of variance (ANOVA) tests with Tukey's post-hoc comparisons. For example, we determined if athletes with acute concussion who were classified as having a Probable concussion had any differences in SAC, SCAT symptoms, or BESS compared to acute concussions who were classified as Definite concussions. Next, using Student's t-test, we identified intraclass differences in the study variables between acute concussions and normal performances within each risk classification. All models were created and analyzed using Python 3.5.2 (Python Software Foundation, Beaverton, Oregon, USA).

5.3 Results

In Table 5.1, we summarize the study data at each timepoint with respect to the study variables. Across all timepoints, there were significant differences between training and validation data in height (P=0.0082-0.047), weight (P=0.012-0.047), and number of previous concussions (P<0.001 for all). There were also significant differences in age at baseline (P=0.0012) and 24-48h (P=0.021) and the proportion of males at unrestricted RTP (P=0.013). Among post-injury assessments, there were significant differences between SAC raw scores at baseline (P=0.00085), <6h (P=0.038), and at unrestricted RTP (P=0.0038), SCAT total symptoms raw score at unrestricted RTP (P=0.027), and BESS raw score at <6h (P=0.048) and 24-48h (P=0.00098).

5.3.1 Multivariate Logistic Regression

The model variables and corresponding coefficient values for the multivariate logistic regression models at <6h and 24-48h are shown in Table 5.2. At <6h, all variables were significant except for whether the injury was reported immediately (P=0.16), SAC raw score (P=0.13), and BESS raw score (P=0.080). At 24-48h, all variables were significant except for SAC raw score (P=0.23) and BESS raw score (P=0.94).

Table 5.2: Multivariate logistic regression coefficients at <6h and 24-48h post-injury

	<6h			**24-48h**		
Study Variable	*Coefficient*	*SE*	*p-value*	*Coefficient*	*SE*	*p-value*
Intercept	-0.57	1.71	0.74	-0.37	1.46	0.80
Male Sex	0.71	0.27	0.01	0.45	0.22	0.04
Report injury immediately?	-0.46	0.33	0.16	-1.37	0.22	<0.01
Removed from play immediately?	-0.91	0.35	0.01	NA	NA	NA
SAC raw score	-0.09	0.06	0.13	-0.06	0.05	0.23
SCAT symptom severity raw score	0.06	0.03	0.02	-0.05	0.02	0.02
SCAT total symptoms raw score	0.28	0.07	<0.01	0.47	0.05	<0.01
BESS raw score	0.03	0.02	0.08	0.00	0.02	0.94

SE, standard error; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; NA, variable not included in this model

Table 5.1: Data characteristics of training and validation data with respect to each timepoint

Normal Performance				
	Baseline		**Unrestricted RTP**	
	Training	*Validation*	*Training*	*Validation*
n	751	884	716	629
Height in meters (SD)	1.79 (0.11)**	1.78 (0.12)	1.79 (0.12)*	1.78 (0.12)
Weight in kg (SD)	83.75 (21.65)**	81.95 (21.25)	83.33 (21.4)**	80.74 (21.01)
Age in years (SD)	19.43 (1.30)*	19.24 (1.32)	19.37 (1.30)	19.27 (1.32)
Male Sex (% yes)	60.80%	57.58%	60.06%**	54.05%
Number of previous concussions (SD)	0.73 (1.0)*	0.59 (0.83)	0.70 (0.93)*	0.52 (0.77)
Report injury immediately? (% yes)[1]			41.2%**	47.22%
Removed from play immediately? (% yes)[1]			48.60%	45.31%
LOC? (% yes)[1]			5.45%	4.77%
PTA? (% yes)[1]			11.45%*	7.63%
RGA? (% yes)[1]			5.59%	4.13%
SAC change score (SD)[1]			0.97 (2.13)*	0.51 (1.99)
SAC raw score (SD)	27.05 (2.01)*	27.38 (1.96)	27.93 (1.75)*	28.19 (1.76)
SCAT symptom severity change score (SD)[1]			-4.87 (8.8)	-5.12 (9.71)
SCAT symptom severity raw score (SD)	5.16 (8.54)	5.25 (9.71)	0.63 (1.97)	0.46 (1.67)
SCAT total symptoms change score (SD)[1]			-2.48 (4)	-2.55 (3.83)
SCAT total symptoms raw score (SD)	2.81 (3.85)	2.77 (3.89)	0.47 (1.39)**	0.34 (1.09)
BESS change score (SD)[1]			-2.35 (6.32)	-2.39 (5.9)
BESS raw score (SD)	12.7 (6.32)	12.78 (6.21)	10.46 (5.81)	10.84 (5.36)
Acute Concussion				
	<6h		**24-48h**	
	Training	*Validation*	*Training*	*Validation*
n	546	539	719	694
Height in meters (SD)	1.80 (0.12)**	1.79 (0.11)	1.79 (0.12)*	1.78 (0.12)
Weight in kg (SD)	86.01 (22.45)**	83.32 (21.39)	83.94 (22.08)**	81.33 (20.85)
Age in years (SD)	19.36 (1.32)	19.23 (1.36)	19.34 (1.27)**	19.2 (1.33)
Male Sex (% yes)	64.47%	61.60%	60.08%	56.34%
Number of previous concussions (SD)	0.78 (1.03)*	0.57 (0.78)	0.74 (1.03)*	0.58 (0.82)
Report injury immediately? (% yes)	55.31%**	60.67%	40.47%**	46.54%
Removed from play immediately? (% yes)	57.51%	56.77%	46.18%	47.26%
LOC? (% yes)	5.86%	5.01%	4.45%	4.76%
PTA? (% yes)	12.27%**	9.09%	11.4%**	8.65%
RGA? (% yes)	5.86%	5.57%	5.84%	5.33%
SAC change score (SD)	-0.82 (3.19)*	-1.76 (3.58)	-0.4 (2.61)*	-1.11 (2.72)
SAC raw score (SD)	26.18 (2.94)**	25.83 (3.33)	26.63 (2.41)	26.48 (2.77)
SCAT symptom severity change score (SD)	23.22 (20.75)	23.45 (22.76)	19.48 (21.68)	20.62 (23.27)
SCAT symptom severity raw score (SD)	28.64 (20.83)	28.58 (21.52)	25.21 (21.54)	26.04 (21.93)
SCAT total symptoms change score (SD)	8.01 (5.93)	8.03 (6.43)	7.49 (6.58)	8.05 (6.88)
SCAT total symptoms raw score (SD)	10.86 (5.42)	10.77 (5.57)	10.53 (6.03)	10.91 (6.18)
BESS change score (SD)	3.58 (8.64)	4.02 (8.52)	1.5 (7.48)**	2.58 (8.19)
BESS raw score (SD)	16.35 (8.78)**	17.36 (8.38)	14.42 (7.87)*	15.82 (8.1)

[1]variable not available for baseline data; Change score at a time point is computed as: raw score at timepoint - raw score at baseline; *P<0.01 **P<0.05 Significantly different from validation data at same time point based on Student's t-test; n, number of data points; SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System

5.3.2 Classifying Unlikely, Possible, Probable, and Definite Concussion

We obtained risk score thresholds after applying the training data to our optimization algorithm. At <6h, the lower threshold was 0.047 and the upper threshold was 0.33. At 24-48h, the lower threshold was 0.07 and the upper threshold was 0.46. The CART we developed for <6h and 24-48h are shown in Figure 5.1.

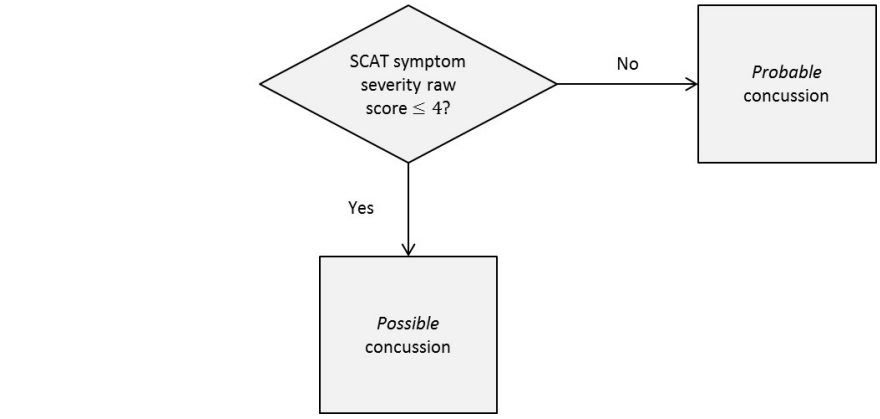
We now provide an example to illustrate how these risk thresholds and CART can be used to determine whether an athlete should be classified as an Unlikely, Possible, Probable, or Definite concussion.

Consider a 19 year-old female athlete who is being assessed for acute concussion 24-48h after injury. She did not report the injury immediately and in her post-injury assessments, obtained total scores of 30 and 12 on the SAC and BESS, respectively. On the SCAT symptom assessment, she reported 4 total symptoms with a total severity of 6. Using the logistic regression model for 24-48h, her risk score is equal to 0.36. Since her risk score is less than the upper threshold of 0.46 and greater than the lower threshold of 0.07 at 24-48h, she is not classified as a Definite or Unlikely concussion. To determine if she is a Possible or Probable concussion, one would refer to the CART for 24-48h. Since her SCAT symptom severity raw score is not 0, her SAC raw score is greater than 25, and her SCAT total symptoms raw score is greater than 1, she would be classified as a Probable concussion.

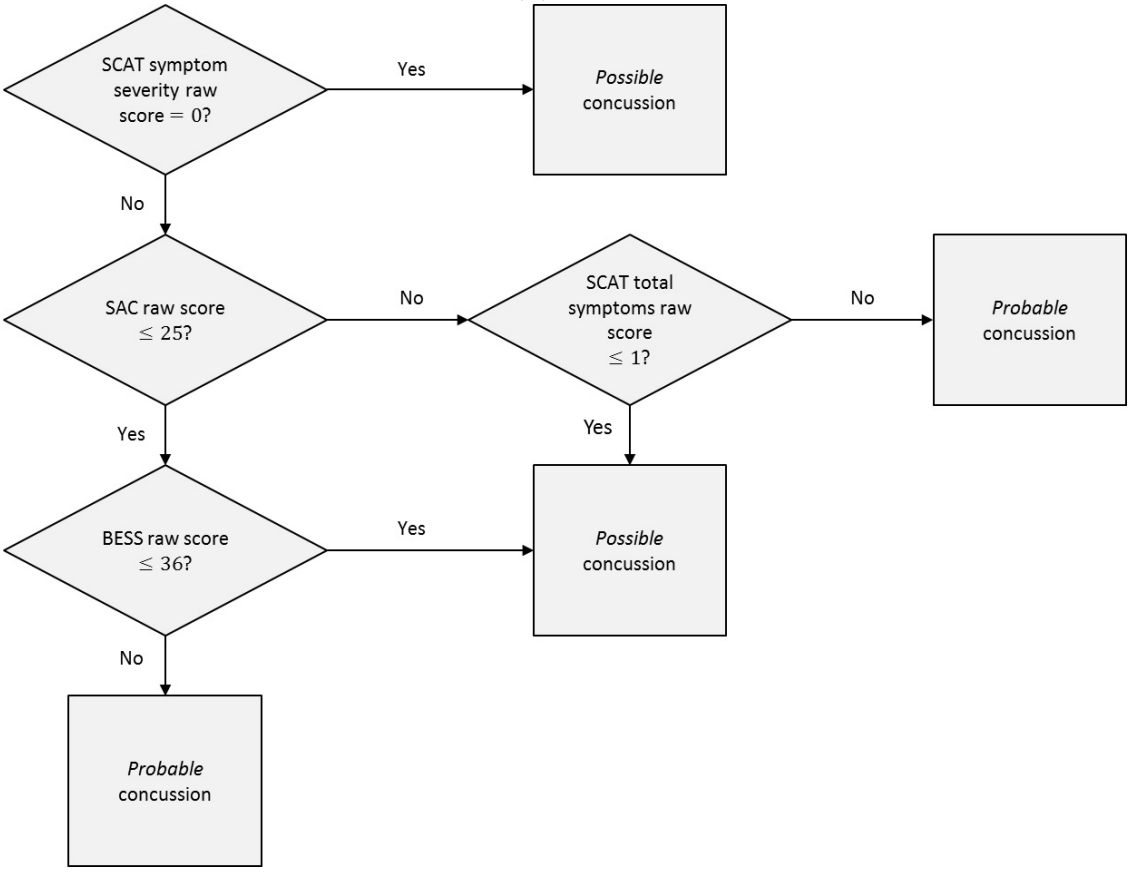
To provide an additional example, consider a 21 year-old male athlete who was assessed for concussion within 6 hours of a suspected injury. His injury was not reported immediately and he was not removed from play immediately. His SAC raw score and BESS raw score were 24 and 12, respectively. He also reported 1 symptom with a severity of 1. Based on these values, his risk score is equal to 0.22. Since his risk estimate is between the lower and upper thresholds of 0.047 and 0.33 at <6h, respectively, then he must either be a Possible or Probable concussion. Since his SCAT symptom severity raw score is ≤ 4, the CART analysis at <6h would classify him as a Possible concussion.

5.3.3 Distribution of Acute Concussions and Normal Performances

The distribution of acute concussions and normal performances within each risk classification is shown in Table 5.3. At <6h, 434 (80.52%) of acute concussions were classified as Definite concussion while only 14 (2.60%) were classified as Unlikely concussion. Among



(a) <6h



(b) 24-48h

Figure 5.1: Classification trees for determining Possible and Probable concussions at <6h and 24-48h post-injury

Table 5.3: Distribution of acute concussions and normal performances among risk classifications at <6h and 24-48h post-injury

		Unlikely		Possible		Probable		Definite	
		n	%	*n*	%	*n*	%	*n*	%
<6h	Acute concussion*	14	2.60%	31	5.75%	60	11.13%	434	80.52%
	Normal performance	696	46.00%	526	34.77%	189	12.49%	102	6.74%
	(Unrestricted RTP)**	282	44.83%	329	52.31%	15	2.38%	3	0.48%
	(Baseline)	414	46.83%	197	22.29%	174	19.68%	99	11.20%
	Total	710	34.60%	557	27.14%	249	12.13%	536	26.12%
24-48h	Acute concussion*	21	3.03%	41	5.91%	110	15.85%	522	75.22%
	Normal performance	714	47.19%	397	26.24%	309	20.42%	93	6.15%
	(Unrestricted RTP)**	275	43.72%	312	49.60%	36	5.72%	6	0.95%
	(Baseline)	439	49.66%	85	9.62%	273	30.88%	87	9.84%
	Total	735	33.30%	438	19.85%	419	18.99%	615	27.87%

*Distributions of acute concussions and normal performances within risk classifications are significantly different at P<0.001 using Kolmogorov-Smirnov test **Distributions of Unrestricted RTP and Baseline timepoints within risk classifications are significantly different at P<0.01 using Kolmogorov-Smirnov test

the remaining acute concussions, 31 (5.75%) were classified as Possible concussion and 60 (11.13%) were classified as Probable concussion. When the <6h algorithm was applied to normal performance data (i.e., baseline and unrestricted RTPs), 696 (46.00%), 526 (34.77%), 189 (12.49%), and 102 (6.74%) were classified as Unlikely, Possible, Probable, and Definite concussion respectively. With the 24-48h algorithm, 522 (75.22%) acute concussions were classified as Definite concussion while 21 (3.03%) were classified as Unlikely concussion. There were 41 (5.91%) and 110 (15.85%) acute concussions classified as Possible and Probable concussion, respectively. Among the normal performances, 714 (47.19%), 397 (26.24%), 309 (20.42%), and 95 (6.15%) were classified as Unlikely, Possible, Probable, and Definite concussion, respectively. With both <6h and 24-48h algorithms, the distributions among risk classifications were different between acute concussions and normal performances based on the Kolmogorov-Smirnov test (P<0.001). Additionally, the distribution of baselines and unrestricted RTPs across the risk classifications were also significantly different at both <6h and 24-48h (P<0.001).

Using our modified calculation for sensitivity and specificity, we obtained a sensitivity range of 91.65-97.40% with the <6h algorithm and 91.06-97.00% with 24-48h algorithm. We also obtained a specificity range of 46.00-80.77% with the <6h algorithm and 47.19-73.43% with the 24-48h algorithm, respectively.

As an ancillary analysis, we performed our analysis without the unrestricted RTP data (data not shown). The resulting logistic regression model, risk score thresholds, and CART models led to a distribution with a sensitivity of 89.42%-98.52% and a specificity of 23.21%-71.60% at <6h. At 24-48h, the sensitivity and specificity ranged from 85.73%-95.67% and 41.52%-71.60%, respectively.

5.3.4 Interclass Differences

The interclass differences for acute concussions and normal performances are shown in Table 5.4 and Table 5.5, respectively. Among acute concussions, all mean raw and change scores for SCAT symptom assessments at Unlikely and Possible concussions are significantly different from Definite concussion at both <6h and 24-48h (P<0.001 for all). Among the SAC and BESS at <6h, only the SAC change score is not significantly different between Definite and Unlikely concussions. In contrast, at 24-48h, only BESS raw score is significantly different between Definite and Unlikely concussions (P=0.021). Possible and Probable concussions are significantly different in SCAT total symptoms raw score at <6h and 24-48h (P=0.0027-0.0082). They are also significantly different in SAC change score (P=0.013), SAC raw score (P<0.001), and SCAT total symptoms change score (P=0.012) at 24-48h.

For normal performances, the mean raw scores for the SAC, SCAT symptom severity, SCAT total symptoms, and the BESS among Definite and Probable concussions are significantly different from Unlikely concussion (P<0.001 for all), except for SAC raw score at 24-48h. At <6h and 24-48h, Possible and Probable concussions were significantly different in SCAT symptom severity raw score and SCAT total symptoms raw score (P<0.001 for all). At 24-48h, Possible and Probable concussions are also significantly different in SAC raw score and BESS raw score (P<0.001 for all).

5.3.5 Intraclass Differences

The intraclass differences are highlighted in Table 5.4 and Table 5.5. Among those classified as Possible concussions, acute concussions and normal performances are significantly different in SCAT symptom severity (P<0.001 at <6h, P=0.016 at 24-48h) and SCAT total symptoms raw score (P<0.001 at <6h, P=0.0019 at 24-48h). There are also significant differences in SAC raw change scores (P=0.0026) and raw scores (P=0.046) for acute concussions and nor-

Table 5.4: Comparison of study variables for acute concussions classified as Unlikely, Possible, Probable, and Definite concussion at <6h and 24-48h post-injury

	<6h			
	Unlikely	*Possible*	*Probable*	*Definite*
n	14	31	60	434
Age in years (SD)	19.93 (0.80)*	19.42 (1.48)	19.23 (1.28)	19.19 (1.37)*
Male Sex (% yes)	78.57%*	74.19%	43.33%	62.67%
Number of previous concussions (SD)	0.43 (0.62)	0.55 (0.87)	0.60 (0.82)	0.56 (0.77)
Report injury immediately? (% yes)	100.00%*[1]	67.74%*[1]	83.33%*	55.76%
Removed from play immediately? (% yes)	100.00%*[1]	64.52%*[1]	76.67%*	52.07%
LOC? (% yes)	14.29%	6.45%*	5.00%	4.61%*
PTA? (% yes)	7.14%	3.23%*	3.33%	10.37%*
RGA? (% yes)	21.43%	6.45%*	0.00%	5.76%*
SAC change score (SD)	0.14 (1.46)	-0.97 (2.86)[1]	-0.55 (2.10)	-1.76 (3.51)
SAC raw score (SD)	28.57 (1.18)[1]	26.61 (2.88)[1]	27.48 (1.80)	25.49 (3.42)*
SCAT symptom severity change score (SD)	-2.79 (5.83)[1]	-0.48 (9.08)[1]	5.18 (7.17)*[1]	28.49 (21.94)*
SCAT symptom severity raw score (SD)	0.71 (1.33)[1]	2.74 (1.24)*[1]	8.70 (3.00)*[1]	34.01 (20.29)*
SCAT total symptoms change score (SD)	-1.21 (3.41)[1]	0.42 (3.98)[1]	2.60 (3.51)*[1]	9.63 (5.85)*
SCAT total symptoms raw score (SD)	0.57 (0.90)[1,2]	2.06 (1.08)*[1,2]	4.73 (1.42)[1]	12.57 (4.53)*
BESS change score (SD)	-1.93 (8.92)[1]	3.97 (7.45)	2.87 (6.39)*	4.94 (9.41)
BESS raw score (SD)	11.64 (6.00)[1]	15.00 (6.51)[1]	13.83 (7.34)	18.40 (9.02)*
	24-48h			
	Unlikely	*Possible*	*Probable*	*Definite*
n	21	41	110	522
Age in years (SD)	19.38 (1.29)	19.20 (1.40)	19.15 (1.29)	19.20 (1.33)*
Male Sex (% yes)	61.90%	73.17%*2	49.09%	56.32%
Number of previous concussions (SD)	0.52 (0.73)	0.66 (0.95)	0.53 (0.72)	0.58 (0.82)*
Report injury immediately? (% yes)	100.00%*[1]	51.22%*[1,2]	76.36%*	37.74%
Removed from play immediately? (% yes)	90.48%*[1]	51.22%*[1]	62.73%*	41.95%
LOC? (% yes)	9.52%	4.88%*	3.64%	4.79%*
PTA? (% yes)	9.52%	7.32%	5.45%	9.39%
RGA? (% yes)	9.52%*	2.44%	7.27%	4.98%*
SAC change score (SD)	-0.52 (2.32)	-1.59 (2.56)*[1,2]	-0.15 (1.63)	-1.28 (2.76)*
SAC raw score (SD)	27.29 (1.72)	25.66 (2.69)*[1,2]	27.76 (1.33)	26.24 (2.90)*
SCAT symptom severity change score (SD)	-3.81 (7.40)[1]	-1.00 (10.92)[1]	4.38 (8.40)*[1]	26.42 (23.33)*
SCAT symptom severity raw score (SD)	0.90 (1.72)*[1]	3.24 (4.11)*[1]	7.15 (5.47)[1]	32.88 (20.95)*
SCAT total symptoms change score (SD)	-2.05 (2.87)[1,2]	-0.54 (3.79)[1,2]	2.63 (3.47)*[1]	10.21 (6.14)*
SCAT total symptoms raw score (SD)	0.33 (0.56)*[1,2]	1.59 (1.56)*[1,2]	4.28 (1.79)*[1]	13.47 (4.74)*
BESS change score (SD)	0.62 (5.60)	0.83 (5.92)	1.46 (6.73)*	3.52 (8.40)*
BESS raw score (SD)	11.62 (5.62)[1]	13.17 (6.29)[1]	14.07 (7.18)[1]	16.89 (8.61)

Change score at a timepoint is computed as: raw score at timepoint - raw score at baseline; *Significantly different (P<0.05) from normal performances in the same risk classification and timepoint based on Student's t-test; [1]Significantly different (P<0.05) from Definite concussion at the same timepoint based on Tukey's post-hoc pairwise comparisons; [2]Significantly different (P<0.05) from Probable concussion at the same timepoint based on Tukey's post-hoc pairwise comparisons; n, number of data points; SD, standard deviation; LOC, loss of consciousness; PTA, post-traumatic amnesia; RGA, retrograde amnesia; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System

Table 5.5: Comparison of study variables for normal performances classified as Unlikely, Possible, Probable, and Definite concussion at <6h and 24-48h post-injury

	<6h			
	Unlikely	*Possible*	*Probable*	*Definite*
n (% baseline)	696 (59.48%)	526 (37.45%)	189 (92.06%)	102 (97.06%)
Age in years (SD)	19.28 (1.27)*	19.33 (1.37)	19.13 (1.32)	18.86 (1.30)*1
Male Sex (% yes)	48.13%*	72.24%1	41.80%2	53.92%
Number of previous concussions (SD)	0.55 (0.81)	0.53 (0.76)	0.58 (0.86)	0.70 (0.84)
Report injury immediately? (% yes)**	84.75%*	14.89%*	53.33%*	33.33%
Removed from play immediately? (% yes)**	84.40%*	11.85%*	46.67%*	33.33%
LOC? (% yes)**	7.45%	2.43%*	6.67%	0.00%*
PTA? (% yes)**	7.09%	7.90%*	13.33%	0.00%*
RGA? (% yes)**	6.03%	2.43%*	6.67%	0.00%*
SAC change score (SD)	0.85 (1.93)	0.54 (2.13)	0.13 (1.82)	-0.33 (1.89)
SAC raw score (SD)	28.11 (1.60)	27.35 (2.25)1	27.40 (1.90)1	27.27 (1.93)1
SCAT symptom severity change score (SD)	-4.98 (8.37)	-5.13 (10.12)	-8.33 (18.17)*	-8.67 (14.27)*
SCAT symptom severity raw score (SD)	0.41 (1.02)	1.02 (1.41)*	8.22 (3.23)1,2	25.65 (15.51)*1,2
SCAT total symptoms change score (SD)	-2.63 (3.51)	-2.54 (3.90)	-1.73 (5.96)*	-1.67 (4.64)*
SCAT total symptoms raw score (SD)	0.28 (0.64)	0.81 (1.16)*1	4.84 (1.59)1,2	11.41 (3.99)*1,2
BESS change score (SD)	-2.72 (6.03)	-1.64 (5.68)	-1.60 (5.55)*	-0.33 (3.68)*
BESS raw score (SD)	10.74 (5.14)	12.96 (5.97)1	13.20 (7.05)1	16.41 (7.77)1,2
	24-48h			
	Unlikely	*Possible*	*Probable*	*Definite*
n (% baseline)	714 (61.48%)	397 (21.41%)	309 (88.35%)	93 (93.55%)
Age in years (SD)	19.36 (1.31)	19.16 (1.30)	19.24 (1.35)	18.86 (1.29)*1
Male Sex (% yes)	56.16%	59.19%*	54.05%	49.46%
Number of previous concussions (SD)	0.54 (0.78)	0.51 (0.80)	0.61 (0.82)	0.76 (0.88)*
Report injury immediately? (% yes)**	100.00%*	2.88%*	30.56%*	33.33%
Removed from play immediately? (% yes)**	77.09%*	18.59%*	33.33%*	50.00%
LOC? (% yes)**	8.73%	0.96%*	8.33%	0.00%*
PTA? (% yes)**	8.73%	6.41%	8.33%	16.67%
RGA? (% yes)**	5.09%*	3.21%	5.56%	0.00%*
SAC change score (SD)	0.85 (2.03)	0.54 (2.07)*	0.28 (1.74)	1.17 (2.19)*
SAC raw score (SD)	27.88 (1.82)	27.30 (2.39)*1	27.92 (1.40)2	27.34 (2.02)*
SCAT symptom severity change score (SD)	-5.09 (8.96)	-4.99 (8.71)	-5.75 (17.90)*	-12.83 (16.08)*
SCAT symptom severity raw score (SD)	0.34 (0.96)*	0.99 (2.29)*	6.64 (5.50)1,2	24.75 (16.6)*1,2
SCAT total symptoms change score (SD)	-2.61 (3.64)	-2.68 (3.73)	-1.06 (4.52)*	-2.50 (6.85)*
SCAT total symptoms raw score (SD)	0.20 (0.49)*	0.66 (1.37)*1	3.85 (1.82)*1,2	11.87 (3.85)*1,2
BESS change score (SD)	-2.06 (6.19)	-2.08 (5.50)	-3.06 (6.42)*	-0.83 (3.02)*
BESS raw score (SD)	11.76 (5.68)	11.28 (5.46)	13.38 (6.76)1,2	15.56 (7.67)*1,2

Change score at a timepoint is computed as: raw score at timepoint - raw score at baseline; *Significantly different (P<0.05) from acute concussions in the same risk classification and timepoint based on Student's t-test **Variable not available for baseline data; [1]Significantly different (P<0.05) from No concussion at the same timepoint based on Tukey's post-hoc pairwise comparisons; [2]Significantly different (p<0.05) from Possible concussion at the same timepoint based on Tukey's post-hoc pairwise comparisons; n, number of data points; SD, standard deviation; LOC, loss of consciousness; PTA, post-traumatic amnesia; RGA, retrograde amnesia; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System

mal performances classified as Possible concussion at 24-48h. Among probable concussions at <6h and 24-48h, acute concussions and normal performances are significantly different in change scores for SCAT symptom severity (P=0.0012-0.0077), SCAT total symptoms (P=0.0093 at <6h, P<0.001 at 24-48h), and BESS (P=0.0074 at <6h, P<0.001 at 24-48h). They are also significantly different in SCAT symptom severity raw score at <6h (P<0.001) and SCAT total symptoms raw score at 24-48h (P<0.001).

To illustrate how these intraclass differences can be used to inform clinical decision-making, we revisit the examples from the Classifying Unlikely, Possible, Probable, and Definite Concussion subsection. Consider the first athlete (19 year-old female) and suppose that her change scores for the SAC, SCAT symptom severity, SCAT total symptoms, and BESS are 0, 6, 4, and -5, respectively. Based on intraclass differences identified in this study for 24-48h, there were significant differences between acute concussions and normal performances for the SCAT symptom severity change score, SCAT total symptoms change score, SCAT total symptoms raw score, and the BESS change score. Comparing this athlete's assessments with the mean values for Probable concussions presented in Table 4 and Table 5, we find that the athlete is more comparable to acute concussion in terms of change scores for the SCAT symptom severity and total number of symptoms. Conversely, she is more comparable to the normal performances in terms of the BESS change score. Following the conservative decision-making approaches that are recommended for concussion management, one could treat this athlete as if she has an acute concussion.

Now, consider the 21 year-old male athlete and suppose that his change scores for the SAC and BESS were 0 and 5 respectively. Additionally, his SCAT symptom severity and total symptoms both decreased by 4 compared to baseline. Based on intraclass differences identified in this study for the Possible concussion group at <6h, there were significant differences between acute concussions and normal performances in the SCAT symptom severity and SCAT total symptom raw scores. Comparing this athlete's values in these measures (of 1 symptom reported with a severity of 1) to the mean values obtained in our analysis, we find that this athlete more closely resembles the normal performances within the Possible concussions – despite the low SAC raw score and high BESS change score. These results could potentially indicate that additional assessments should be performed on this athlete to confirm the possibility that this athlete is not concussed.

5.4 Discussion

Kutcher and Giza proposed a risk-based classification framework for diagnosing acute concussion developed from clinical experience (Kutcher and Giza, 2014). Compared to traditional binary diagnosis, this framework allows the assessment of acute concussion to reflect the physician's diagnostic certainty. Furthermore, taking this approach allows the injury diagnosis to evolve as the injury evolves and more information becomes available. However, while they provided clinical guidelines for each risk classification, they did not provide specific criteria with respect to commonly recommended and implemented concussion assessment tools. In this research, we designed and evaluated a novel data-driven method for classifying athletes evaluated for acute concussion as either Unlikely, Possible, Probable, and Definite concussion. The major contributions of our research are as follows:

- We develop an objective and data-driven framework which stratifies acute concussion assessment by diagnostic certainty. These risk categories lay the foundation for guiding post-injury management decisions.
- We identify key characteristic which can be used to differentiate between acute concussions and normal performances in each risk category.
- We provide additional, quantitative support for the value of a multidimensional battery, the use of change scores in acute concussion assessment, and the potential implications for several demographic factors and time-of-injury characteristics in acute concussion assessment.

The variables used in our logistic regression and CART models are parts of standard concussion assessment batteries, giving foundation for our framework to be used in sporting environments. To our knowledge, we are the first to combine predictive modeling techniques (i.e., logistic regression and CART) and optimization algorithms to classify athletes into concussion risk categories. Erring in the direction of minimizing false negatives, our framework classified most acute concussions (91.07-91.65%) into the higher risk categories (i.e., Probable and Definite concussion) and most normal performances (73.43-80.77%) into the lower risk categories (i.e., Unlikely and Possible concussion). Additionally, few acute concussions were classified as Unlikely concussion (2.60-3.03%) and few normal performances were classified as Definite concussion (6.15-6.74%).

Our most important finding was that athletes classified as Definite concussion had lower SAC, higher SCAT symptom, and higher BESS scores compared to the other risk categories. In comparing these risk groups, Definite concussions exhibited noticeably more symptoms and greater symptom severity compared to the other risk categories while the Unlikely concussions exhibited mean symptom severities and mean total symptoms close to 0. Definite concussions also had much higher BESS raw scores and lower SAC scores compared to Unlikely concussions. These findings demonstrate the ability of our framework to separate the "easy" cases from the "hard" cases and are consistent with previous research demonstrating symptoms are typically the most sensitive to acute concussion (Chin et al., 2016; Garcia et al., 2018; McCrea et al., 2005; Register-Mihalik et al., 2013b; Resch et al., 2016). Our findings also provide support for the utility of using neurocognitive assessments and postural control measures for acute concussion assessment, as demonstrated by previous research (Barr and McCrea, 2001; Broglio, Macciocchi, and Ferrara, 2007; Buckley, Munkasy, and Clouse, 2017; Covassin, Schatz, and Swanik, 2007; Guskiewicz, 2001; Guskiewicz, Ross, and Marshall, 2001; McCrea et al., 2005; Riemann and Guskiewicz, 2000; Sufrinko et al., 2017a; Valovich McLeod et al., 2004)

However, among those classified as Possible or Probable concussions, raw scores for SCAT symptom severity and total symptoms are significantly less for acute concussions in the Possible concussion group compared to all baselines ($P<0.01$ for both measures at <6h and 24-48h using Student's t-test). Additionally, there are no significant differences in the SCAT symptom severity or total symptoms between acute concussions and normal performances in the Probable risk category. These findings demonstrate the difficulty in identifying all acute concussions using symptom raw scores alone. Fortunately, there were some significant differences between acute concussions and normal performances in the Possible and Probable risk categories for change scores in the SAC, SCAT symptom severity, SCAT total symptoms, and the BESS. This result suggests that change scores, which require baseline assessments, have added value when evaluating Possible and Probable concussions and is an important finding regarding the utility of the baseline assessment.

In our analysis, we also sought out to identify differences in athlete demographics and time-of-injury characteristics across and within risk classifications. There were statistically significant differences in age, sex, and number of previous concussions between acute concussions and normal performances within some risk categories. For example, among those

classified as Definite concussions, the athletes were, on average, older than those providing normal performances at both <6h and 24-48 hour. Outside of age, there were no other consistent demographic differences and risk categories. For time-of-injury variables among acute concussions, a larger proportion of those classified as Unlikely concussion reported the injury immediately and were removed from play immediately compared to those who were classified as Definite concussions. This result suggests that those athletes who were removed from play immediately or assessed immediately after injury may be in the earliest stages of an evolving injury whereby neurocognitive declines, increased symptoms, worsening postural control emerge over time (Guskiewicz et al., 2003; Makdissi et al., 2010; McCrea et al., 2003). However, we note that very few acute concussions were classified as Unlikely concussion, and due to this small sample size, this point may require further investigation.

We also found that baselines comprised most normal performances within the Probable and Definite risk categories. Despite varying parameter settings to balance the sensitivity and specificity of our results, we were unable to drastically improve on the proportion of baselines in these upper risk categories. This finding may be due to performance differences between baseline and unrestricted RTP timepoints. Specifically, the baseline data showed lower SAC scores, higher SCAT symptom assessment scores, and higher BESS scores compared to unrestricted RTP data ($P<0.001$ for all measures and in both training and validation sets). As a result, our logistic regression model categorized the baseline performance of some athletes into the higher risk categories. The performance discrepancy between baseline and unrestricted RTP timepoints is consistent with previous studies (Garcia et al., 2018; McCrea et al., 2013; McCrea et al., 2003; Piland et al., 2010; Shehata et al., 2009) and may be attributed to comorbidities (Lovell et al., 2006) or learning effects from multiple assessments prior to return to play (Moreau, Langdon, and Buckley, 2014; Valovich McLeod et al., 2004). Future works may be able to address this shortcoming by incorporating individual items from the SAC, SCAT symptom assessments, and the BESS instead of using total scores. Regardless, this finding highlights the need for clinicians to interpret the administered assessments in the context of the injury, such as an observed mechanism, and differentiate from other injuries and conditions with similar signs and symptoms (Bruce et al., 2017; Kutcher and Giza, 2014; McCrory et al., 2017; Zuckerman et al., 2016)

To account for clinical judgment in our methodology, we used a modified range-based computation for sensitivity and specificity, which provided a 6% sensitivity increase and

¿34% specificity increase in Possible concussion management. Furthermore, the sensitivity of our algorithm mirrors those reported in previous studies (80.0-100.0%) evaluating concussion testing batteries for acute concussion assessment (Broglio, Macciocchi, and Ferrara, 2007; McCrea et al., 2005; Putukian et al., 2015; Resch et al., 2016). Methodological differences between our study and these aforementioned studies account for some differences, including the test battery assessments. For example, both Broglio, Macciocchi, and Ferrara, 2007 and Resch et al., 2016 used the Sensory Organization Test (SOT) for balance assessment instead of the BESS. While both the SOT and BESS reveal similar post-concussion trends in postural control deficits (Guskiewicz, 2001), the SOT has less clinical applicability given its size and cost. Additionally, the diagnosis criteria differed greatly across each study. Broglio, Macciocchi, and Ferrara, 2007; McCrea et al., 2005 and Putukian et al., 2015 used different measures of significant change to indicate concussion while Resch et al., 2016 used both predictive discriminant analyses and clinical interpretation guidelines. In comparison, we paired a data-driven optimization framework with predictive modeling methods (i.e., logistic regression and CART) to classify athletes into risk categories. By using predictive modeling methods, we were able to simultaneously incorporate demographic information and time-of-injury characteristics, along with SAC, SCAT symptoms, and BESS results. Finally, the concussed sample used in the present study (n=1085 for <6h and n=1413 for 24-48h) is much larger than those in the aforementioned studies (n=32-166).

From a clinical perspective, previous studies have discussed the importance and value of taking a heterogeneous and targeted approach to concussion management (Collins et al., 2016; Collins et al., 2014; Ellis et al., 2016). However, since the focus of this study was on identifying acute concussion, it does not address injury heterogeneity by accounting for potential concussion subtypes or clinical profiles. However, our work lays the foundation to do so using clustering or clinically determined approaches.

Our study is not without limitations. First, we acknowledge that our framework does not provide a recommendation for post-injury management for athletes classified in each risk category. These post-injury decisions are beyond the scope of our study and are an important topic for future research. To this end, clinicians can still benefit from knowing the degree of certainty in a diagnosis decision before determining the next course of action. Second, our study treats all concussions in the CARE data as truly concussed, regardless of the medical staff certainty. Thus, there is the possibility that our models were trained

and validated on athletes who were not actually concussed but were labeled so. Third, the differences between our training and validation data in important clinical measures such as PTA or LOC may have caused differences in the presentation of concussion between those two groups, potentially explaining some of the prediction errors in our models. Determining training and validation data using random subset selection instead of by a time-based cut-off could lead to a more homogeneous division in data and ultimately, improved modeling results. Fourth, since our study data only included athletes aged 18-22, we cannot directly apply our results to populations beyond this group. Therefore, future studies should focus on other population groups, such as youth sports and professional athletes, to determine the generalizability of our results beyond our study population. Fifth, we were limited in our ability to include change scores and time-of-injury characteristics in our models, as these measures were not available for baseline data. Finding ways to incorporate such variables in future analysis may improve our results. Furthermore, our analysis focused on the SAC, SCAT symptom assessments, and the BESS. Data limitations precluded our ability to include assessments such as the Sensory Organization Test, computer based neurocognitive testing, the King-Devick test, and/or the Vestibular/Ocular Motor Screening Assessment that have shown promise in other investigations (Anzalone et al., 2017; Kontos et al., 2016; Mucha et al., 2014; Pearce et al., 2015). Finally, as there is no gold standard for concussion diagnosis, we did not have a comparative mechanism for our results.

The objective, algorithmic approach we proposed and developed for risk-based classification of athletes undergoing acute concussion assessment extends the original framework proposed by Kutcher and Giza, 2014. By applying predictive modeling and optimization methods, our work provides a promising first-step in taking an evidence-based approach to acute concussion assessment stratification. While the clinical examination remains the gold standard for concussion diagnosis, the models we have designed and analyzed have the potential to provide valuable decision support for clinicians.

Chapter 6

Cluster Analysis of Possible and Probable Concussions

6.1 Introduction

In Chapter 5, we applied the methods developed in Chapters 2 and 4 to create a data-driven framework for acute concussion diagnosis which incorporates the degree of uncertainty in that diagnosis (i.e., Unlikely, Possible, Probable, or Definite concussion). This prior analysis focused primarily on comparisons based on composite scores for the Standardized Assessment of Concussion (SAC), full Balance Error Scoring System (BESS), and the Sport Concussion Assessment Tool (SCAT) symptom checklist. However, providing a more granular characterization of athletes with Possible and Probable concussions would be of additional use for clinicians.

Hence, the goal of this research is to characterize performance on the SAC, full BESS, and SCAT symptom checklist for athletes identified as having Possible and Probable concussions. We achieve this goal by first using our previously developed framework to classify athletes as having Possible and Probable concussion and then applying extensive cluster analysis to identify variables which can best separate the concussions and normal performances within the Possible and Probable concussions. The results and insights generated in this research provide a reference for clinicians to use when assessing athletes for whom a diagnosis decision is unclear.

6.2 Materials and Methods

6.2.1 Study Population and Design

To perform our cluster analysis, we used data from the CARE Consortium, as described in Section 1.1.1.

6.2.2 Sample Selection

Since our analysis aimed to characterize Possible and Probable concussions at the acute concussion stage, we focused on assessments performed at <6 hours (n=1456) and 24-48 hours (n=2394). As a point of comparison, we also included assessments at baseline (n=2587) and unrestricted RTP (n=2178) in our analysis, although baseline data were used only to compute change scores (defined below) for the SAC, BESS, and symptom checklist. Therefore, we only included baseline data which could be matched to an athlete's post-injury assessment(s). We denoted the assessments at <6h and 24-48h as "acute concussion" and those from the unrestricted RTP timepoint as "normal performance." We performed separate analyses for the <6h and 24-48h timepoints.

6.2.3 Study Variables

Within the study data, we obtained the athlete's sex along with baseline scores on the SAC, BESS, and SCAT symptom checklist. The SAC (McCrea et al., 1998), BESS (Riemann, Guskiewicz, and Shields, 1999), and SCAT symptom checklist (Concussion in Sport Group, 2013) are considered to be among the most useful tools for evaluating acute concussion (Echemendia et al., 2017). To facilitate our analysis, we first aggregated symptoms into symptom groups based on previous work by Kontos et al., 2019 (see Table 6.1). These symptom groups have shown strong reliability and validity for concussion screening (Kontos et al., 2020; Kontos et al., 2019). Note that although "neck pain" is included in the SCAT symptom checklist, we excluded this symptom from our analysis since it is regarded as a concurrent symptom with concussion and not necessarily a specific symptom of concussion (King, McCrea, and Nelson, 2020). We obtained the score for each symptom group by taking the sum of the severity of the symptoms within each group. In our analysis, we included

the raw score and change score for each symptom group along with the total number of symptoms reported.

Table 6.1: Symptom groups

Symptom Group	Symptoms
Ocular-Vestibular	blurred vision, dizziness, balance problems, nausea or vomiting
Cognitive-Fatigue	feeling like "in a fog", difficulty concentrating, difficulty remembering, feeling slowed down, fatigue, drowsiness, confusion, "don't feel right"
Post-traumatic Migraine	headache, trouble faling asleep, "pressure in the head", sensitivity to light, sensitivity to noise
Anxiety-Mood	nervous or anxious, sadness, more emotional, irritability

Since previous research has identified sex differences in symptom burden and neurocognitive measures post-injury (Covassin et al., 2012; Covassin et al., 2018; Covassin, Schatz, and Swanik, 2007; Dick, 2009; Moran et al., 2020), separate analyses were carried out for male athletes and female athletes. For athletes who sustained concussions, we obtained raw scores for the SAC, BESS, and SCAT symptom checklist at <6h, 24-48h, and unrestricted RTP. For each of the post-injury assessments, we computed a change score by taking the difference between the raw score obtained at the time of post-injury assessment and the score obtained during the baseline assessment for that athlete. A positive change score indicates an increase in that measure compared to baseline and vice versa. Therefore, positive change for the SAC indicates "better" performance for the SAC, whereas a positive change score for symptoms and the BESS indicate "worse" performance. For assessments at each timepoint, all missing data elements were filled using multiple imputation by chained equations with imputation performed within each timepoint (e.g., only data from <6h was used to impute missing data at <6h) (Royston, 2004). Most variables were missing at less than 5% except for BESS total score at baseline (5.4%), <6h (24.5%), and 24-48h (10.3%). Imputation was performed using the software R, Version 3.2.2.

6.2.4 Data Analysis

The purpose of our analysis was to identify characteristics which could be useful in determining which athletes have a true diagnosis of concussion among those athletes who are most difficult to assess. Our analysis was carried out in three stages, as illustrated in Figure 6.1.

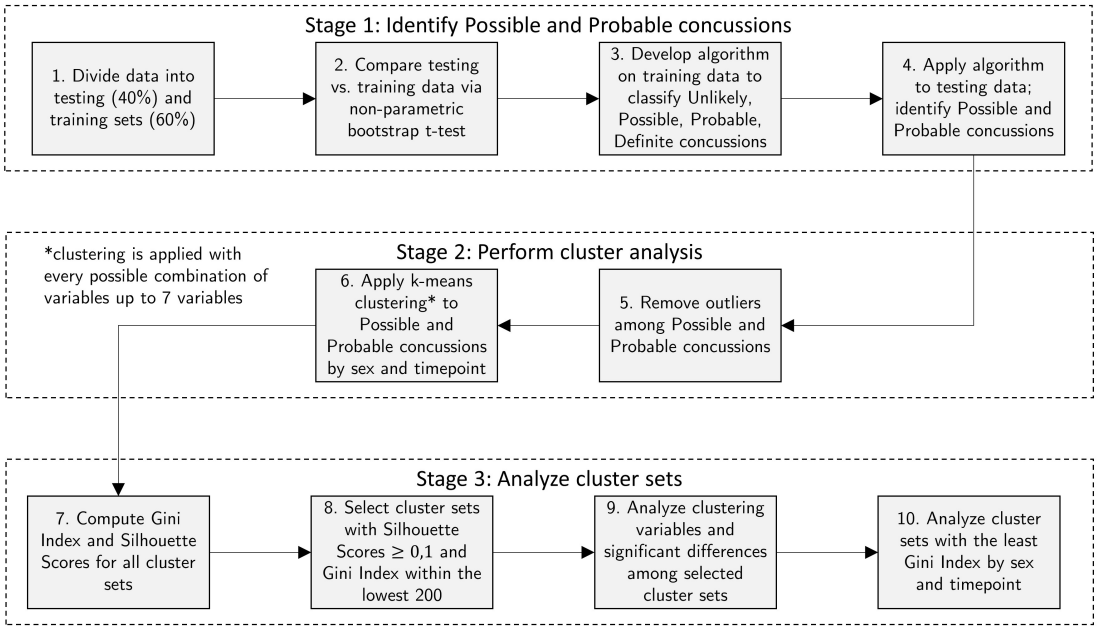


Figure 6.1: Illustration of data analysis procedure

In the first stage of our analysis, we randomly divided the data into a training set (60% of data) and testing set (40% of data). Given the non-normality of our modeling variables, we used two-sample non-parametric bootstrap t-tests (Efron and Tibshirani, 1993) to compare training and testing sets, with a significance level of $\alpha = 0.05$ indicating significant differences. We also computed Cohen's d to quantify the effect size of these differences. Then, we applied our methodology from Chapter 5 on the training set to classify athletes suspected of concussion as Unlikely, Possible, Probable, or Definite concussions. We then applied this algorithm to our testing set to identify Possible and Probable concussions.

In the second stage of our analysis, we performed k-means clustering on this subset of Possible and Probable concussions to identify potential clusters which could describe the difference between those who were likely to have concussion and those who were unlikely to have concussion. Briefly, k-means clustering will place each athlete into one of k different clusters based on which cluster the athlete is "closest" to. To perform this analysis, we first separated the Possible and Probable concussions into four subsets by sex and timepoint: male <6h, female <6h, male 24-48h, and female 24-48h. In early stages of cluster analysis,

we found that extreme values from change scores in SCAT total symptom severity and total number of symptoms were largely influencing our result. Since k-means clustering is known to be sensitive to the presence of outliers (Jain, 2010), we removed a small subset of athletes (n=14-21 athletes at each sex and timepoint) from our analysis whose change score in SCAT total symptom severity or total number of symptoms was greater than 2.5 standard deviations from the mean value within the same sex and timepoint. We remark that closer investigation of these outliers indicate that these athletes scored significantly better post-injury than at baseline (i.e., total symptom severity is much higher at baseline). Using a two-sample non-parametric bootstrap t-test, we identified significant differences for all modeling variables across each subset of Possible and Probable concussions by sex and timepoint. For all variables, we used Cohen's d to quantify effect sizes for all differences. To perform the cluster analysis, we specified the number of clusters and the variables on which to cluster (i.e., a subset of the modeling variables in Section 6.2.3). To maximize the interpretability of our results, we focused on cluster analysis which divided the data into two cluster groups. Additionally, work by Formann, 1984 recommends between 2^m to 5×2^m data samples when clustering on m variables. Hence, due to our sample sizes, we could only test clustering groups which were clustered on at most 7 variables. Therefore, in our analysis, we tested every combinations of variables with up to 7 variables resulting in 63,003 clusters at both <6h and 24-48h. Hereafter, we refer to each group of clusters formed by a specified set of variables as a cluster set.

In the third stage of our analysis, we evaluated the cluster sets for each subset of Possible and Probable concussions (by sex and timepoint) by computing their silhouette scores (Rousseeuw, 1987) and sample-weighted Gini indices (Cowell, 2000). Silhouette scores range from -1 to 1 and measure the dissimilarity between clusters in a cluster set, where greater Silhouette scores indicate greater dissimilarity. Gini indices range from 0 to 1 and measure the purity of clusters within a cluster set. Cluster sets with lower Gini indices are purer, implying that one cluster will contain mostly acute concussions and the other will contain mostly normal performances. We restricted our attention to the cluster sets which had a Silhouette score of at least 0.1 and a Gini index within the lowest 200 of all Gini indices (approximately best 0.32% of all clusters). With these 200 sets of clusters, we analyzed which variables were used in clustering and which of these variables were significantly different (based on a non-parametric bootstrap t-test) between the two clusters in each cluster set.

Finally, we analyzed the cluster sets with the lowest Gini index in each of the four cluster subsets to illustrate the differences between the two purest clusters.

6.3 Results

6.3.1 Characteristics of Study Data

In Table 6.1 and Table 6.2, we summarize the characteristics of normal performances and acute concussions, respectively, with respect to modeling variables. There were very few significant differences between the training and testing data. For assessments performed at the unrestricted RTP timepoint, the mean number of previous concussions (P=0.0085, d=0.11), total number of symptoms raw score (P=0.038, d=0.09), and the Anxiety-Mood symptom group raw score (P=0.031, d=0.09) were significant.

6.3.2 Characteristics of Possible and Probable Concussions

We used the data-driven framework of Chapter 5 to identify Possible and Probable concussions in the testing data. We describe the Possible and Probable concussions at <6h in Table 6.3 and at 24-48h in Table 6.4. For males at <6h, mean values were significantly different between the concussed and normal performance groups for most modeling variables (P<0.05 for all, d=0.51-3.09) except for the SAC raw score (P=0.10, d=0.21), BESS raw score (P=0.20, d=0.20), and Anxiety-Mood raw score (P=0.27, d=0.14) and change score (P=0.37, d=0.14). For females at <6h, there were significant differences between the concussed and normal performance groups in mean values for the BESS change score (P=0.001, d=0.54), total number of symptoms raw score (P=0.0005, d=3.05) and change score (P=0.002, d=1.04), Ocular-Vestibular raw score (P=0.001, d=2.00) and change score (P=0.0005, d=1.34), Cognitive-Fatigue raw score (P=0.0005, d=2.25), and Post-traumatic Migraine raw score (P=0.0005, d=3.29) and change score (P=0.0005, d=1.63). At 24-48h, all variables were significantly different between males with acute concussion and normal performance except for Ocular-Vestibular change score (P=0.59, d=0.10), and Anxiety-Mood raw score (P=0.27, d=0.23) and change score (P=0.95, d=0.01). For females at 24-48h, mean values were significantly different between the concussed and normal performance groups for

Table 6.1: Characteristics of normal performances by timepoint

	Baseline		Unrestricted RTP	
Variable	*Training*	*Testing*	*Training*	*Testing*
n	1623	964	1346	832
Male Sex (% yes)	58.1%	60.79%	56.98%	58.89%
Age, years (SD)	19.10 (1.25)	19.16 (1.30)	19.12 (1.25)	19.12 (1.29)
Height, m (SD)	1.79 (0.11)	1.79 (0.12)	1.78 (0.11)	1.79 (0.12)
Weight, kg (SD)	82.33 (21.55)	83.12 (21.97)	81.78 (21.24)	82.62 (21.97)
Number of previous concussions (SD)	0.60 (0.88)	0.62 (0.88)	0.64 (0.88)	0.54 (0.81)**
SAC RS (SD)	27.30 (1.95)	27.22 (1.98)	28.11 (1.68)	28.11 (1.75)
SAC CS (SD)			0.83 (2.03)	0.78 (2.04)
BESS RS (SD)	13.31 (6.28)	13.74 (6.64)	10.81 (5.65)	10.70 (5.73)
BESS CS (SD)			-2.91 (6.23)	-2.59 (6.51)
SCAT total symptoms RS (SD)	2.79 (3.84)	2.82 (3.95)	0.40 (1.29)	0.30 (0.98)*
SCAT total symptoms CS (SD)			-2.47 (3.94)	-2.51 (3.85)
Vestibular RS (SD)	0.28 (0.99)	0.27 (1.04)	0.01 (0.12)	0.01 (0.15)
Vestibular CS (SD)			-0.25 (0.96)	-0.25 (1.09)
Ocular RS (SD)	0.08 (0.42)	0.07 (0.41)	0.01 (0.14)	0.00 (0.05)
Ocular CS (SD)			-0.06 (0.41)	-0.04 (0.37)
Cognitive-Fatigue RS (SD)	2.52 (4.83)	2.41 (4.54)	0.22 (1.06)	0.15 (0.70)
Cognitive-Fatigue CS (SD)			-2.35 (4.80)	-2.60 (5.41)
Post-traumatic Migraine RS (SD)	1.03 (2.13)	1.13 (2.18)	0.18 (0.74)	0.13 (0.55)
Post-traumatic Migraine CS (SD)			-0.90 (2.14)	-0.92 (2.15)
Anxiety-Mood RS (SD)	1.00 (2.48)	0.98 (2.48)	0.07 (0.52)	0.03 (0.38)*
Anxiety-Mood CS (SD)			-0.97 (2.48)	-0.96 (2.62)
Cervical RS (SD)	0.25 (0.70)	0.27 (0.76)	0.05 (0.27)	0.05 (0.31)
Cervical CS (SD)			-0.23 (0.78)	-0.22 (0.81)

n, sample size; RS, raw score; CS, change score; SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; *P<0.05 **P<0.01 ***P<0.001 Significantly different from training data at same timepoint using 2-sample non-parametric bootstrap t-test; Effect sizes (Cohen's d) for significant differences: Number of previous concussions at Unrestricted RTP, d=0.11; SCAT total symptoms RS, d=0.09; Anxiety-Mood RS, d=0.09

SAC raw score (P=0.0125, d=0.42) and change score (P=0.024, d=0.35), SCAT total symptoms raw score (P=0.0005, d=2.68) and change score (P=0.0005, d=1.22), Ocular-Vestibular raw score (P=0.013, d=0.87), Cognitive-Fatigue raw score (P=0.0005, d=1.54) and change score (P=0.0015, d=0.48), and Post-traumatic raw score (P=0.0005, d=2.13) and change score (P=0.0005, d=1.36).

Table 6.2: Characteristics of acute concussions by timepoint

	<6h		24-48h	
Variable	*Training*	*Testing*	*Training*	*Testing*
n	831	625	1432	962
Male Sex (% yes)	62.33%	63.84%	57.96%	59.98%
Age, years (SD)	19.23 (1.34)	19.23 (1.31)	19.06 (1.25)	19.07 (1.24)
Height, m (SD)	1.80 (0.11)	1.79 (0.12)	1.79 (0.11)	1.79 (0.12)
Weight, kg (SD)	84.67 (22.23)	83.84 (21.89)	82.50 (21.77)	82.86 (22.09)
Number of previous concussions (SD)	0.65 (0.91)	0.68 (0.90)	0.59 (0.86)	0.66 (0.93)
SAC RS (SD)	26.01 (2.96)	25.87 (3.21)	26.56 (2.67)	26.52 (2.61)
SAC CS (SD)	-1.10 (3.05)	-1.35 (3.33)	-0.66 (2.73)	-0.77 (2.64)
BESS RS (SD)	17.11 (8.47)	16.69 (8.38)	15.13 (7.80)	15.13 (8.15)
BESS CS (SD)	3.58 (8.57)	3.15 (8.52)	1.59 (8.12)	1.64 (8.08)
SCAT total symptoms RS (SD)	10.52 (5.64)	11.07 (5.45)	11.01 (6.03)	10.83 (5.78)
SCAT total symptoms CS (SD)	7.95 (6.42)	8.01 (6.61)	8.06 (6.66)	7.85 (6.80)
Vestibular RS (SD)	3.26 (3.37)	3.38 (3.37)	2.51 (3.10)	2.29 (2.85)
Vestibular CS (SD)	3.00 (3.52)	3.01 (3.44)	2.20 (3.17)	1.97 (3.12)
Ocular RS (SD)	0.84 (1.31)	0.96 (1.41)	0.67 (1.16)	0.60 (1.05)
Ocular CS (SD)	0.77 (1.34)	0.87 (1.37)	0.58 (1.29)	0.50 (1.19)
Cognitive-Fatigue RS (SD)	12.56 (10.66)	13.18 (10.54)	11.77 (10.26)	11.32 (9.78)
Cognitive-Fatigue CS (SD)	10.10 (11.80)	10.25 (11.77)	9.01 (10.91)	8.55 (10.83)
Post-traumatic Migraine RS (SD)	7.66 (5.20)	8.05 (5.43)	8.17 (5.99)	7.74 (5.49)
Post-traumatic Migraine CS (SD)	6.70 (5.31)	6.94 (5.55)	7.10 (6.06)	6.63 (5.67)
Anxiety-Mood RS (SD)	2.44 (4.07)	2.57 (4.16)	2.58 (4.12)	2.27 (3.73)
Anxiety-Mood CS (SD)	1.68 (4.57)	1.38 (4.85)	1.56 (4.56)	1.26 (4.25)
Cervical RS (SD)	0.99 (1.42)	1.09 (1.55)	1.21 (1.57)	1.15 (1.49)
Cervical CS (SD)	0.75 (1.52)	0.79 (1.64)	0.95 (1.65)	0.86 (1.62)

n, sample size; RS, raw score; CS, change score; SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; *P<0.05 **P<0.01 ***P<0.001 Significantly different from training data at same timepoint using 2-sample non-parametric bootstrap t-test

6.3.3 Clustering Variables

In Figure 6.1, we illustrate the frequency by which each variable was clustered on among the 200 sets of clusters with the lowest Gini Indices by sex and timepoint. At <6h, the modeling variables most frequently clustered on for males were Post-traumatic Migraine raw score (136/200) and change score (162/200), and the total number of symptoms raw score (190/200) and change score (145/200). Similarly, the most frequently clustered on variables for females at <6h were Post-traumatic Migraine raw score (200/200) and change score (149/200), as well as total number of symptoms (163/200) and Cognitive-Fatigue raw score

Table 6.3: Characteristics of Possible and Probable concussions at <6h

	Male		Female	
Variable	*Normal Performance*	*Concussed*	*Normal Performance*	*Concussed*
n	322	71	226	39
SAC RS (SD)	27.67 (1.90)	27.28 (1.77)	28.17 (1.65)	28.13 (1.69)
SAC CS (SD)	-0.76 (2.15)	0.35 (2.31)***	-0.72 (2.06)	-0.23 (2.34)
BESS RS (SD)	11.55 (5.23)	12.62 (6.35)	10.49 (5.49)	11.97 (4.83)
BESS CS (SD)	2.20 (6.19)	-1.06 (7.32)**	2.51 (6.74)	-0.97 (4.84)***
SCAT total symptoms RS (SD)	0.26 (0.85)	4.13 (2.34)***	0.39 (1.07)	4.72 (2.67)***
SCAT total symptoms CS (SD)	1.71 (2.65)	-2.69 (3.32)***	2.19 (2.70)	-0.95 (4.49)**
Ocular-Vestibular RS (SD)	0.01 (0.08)	0.82 (1.30)**	0.01 (0.11)	1.62 (2.09)***
Ocular-Vestibular CS (SD)	0.11 (0.71)	-0.52 (1.52)**	0.20 (0.79)	-1.23 (2.06)***
Cognitive-Fatigue RS (SD)	0.15 (0.67)	2.52 (2.89)***	0.18 (0.66)	2.46 (2.14)***
Cognitive-Fatigue CS (SD)	1.59 (2.86)	-1.13 (4.40)***	1.85 (2.98)	0.74 (4.48)
Post-traumatic Migraine RS (SD)	0.11 (0.52)	3.37 (2.91)***	0.19 (0.61)	3.79 (2.47)***
Post-traumatic Migraine CS (SD)	0.52 (1.38)	-2.94 (3.22)***	0.83 (1.82)	-2.64 (3.43)***
Anxiety-Mood RS (SD)	0.02 (0.25)	0.06 (0.23)	0.06 (0.64)	0.41 (1.23)
Anxiety-Mood CS (SD)	0.48 (1.31)	0.30 (1.55)	0.86 (2.11)	1.13 (2.90)

n, sample size; RS, raw score; CS, change score; SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; *P<0.05 **P<0.01 ***P<0.001 Significantly different from normal performance of the same sex using 2-sample non-parametric bootstrap t-test; Effect sizes (Cohen's d) for significant differences for males at <6h: SAC CS, d=0.51;BESS CS, d=0.51; SCAT total symptoms RS, d=3.09; SCAT total symptoms CS, d=1.58; Ocular-Vestibular RS, d=1.46; Ocular-Vestibular CS, d=0.70; Cognitive-Fatigue RS, d=1.74; Cognitive-Fatigue CS, d=0.85; Post-traumatic Migraine RS, d=2.47; Post-traumatic Migraine CS, d=1.88; Effect sizes (Cohen's d) for significant differences for females at <6h: BESS CS, d=0.54; SCAT total symptoms RS, d=3.05; SCAT total symptoms CS, d=1.04; Ocular-Vestibular RS, d=2.00; Ocular-Vestibular CS, d=1.34; Cognitive-Fatigue RS, d=2.25; Post-traumatic Migraine RS, d=3.29; Post-traumatic Migraine CS, d=1.63

(103/200).

For males at 24-48h, we found similar results to males at <6h. The variables most frequently clustered on were the total number of symptoms raw score (200/200) and change score (99/200), along with the Post-traumatic Migraine change score (145/200) and BESS change score (101/200). Among females at 24-48h, the most frequently clustered variables were the total number of symptoms raw score (200/200) and change score (93/200), Cognitive-Fatigue change score (103/200), and Anxiety-Mood change score (102/200).

6.3.4 Analysis of Significant Differences

In Figure 6.2, we illustrate the frequency by which each variable was significantly different between the two cluster groups among the 200 sets of clusters with the lowest Gini Indices by sex and timepoint. For males at <6h, the variables which were most often sig-

Table 6.4: Characteristics of Possible and Probable concussions at 24-48h

	Male		Female	
Variable	*Normal Performance*	*Concussed*	*Normal Performance*	*Concussed*
n	236	87	182	54
SAC RS (SD)	27.86 (1.93)	27.03 (2.23)**	28.34 (1.63)	27.65 (1.71)*
SAC CS (SD)	-0.74 (2.13)	0.38 (2.14)***	-0.66 (2.00)	0.02 (1.84)*
BESS RS (SD)	11.24 (6.44)	13.10 (7.50)*	10.69 (6.24)	11.67 (5.09)
BESS CS (SD)	2.17 (6.29)	-1.71 (8.67)***	2.38 (7.05)	1.44 (5.95)
SCAT total symptoms RS (SD)	0.31 (0.88)	2.95 (1.74)***	0.42 (1.04)	3.85 (1.88)***
SCAT total symptoms CS (SD)	1.55 (2.59)	-1.26 (3.29)***	2.24 (2.77)	-1.31 (3.41)***
Ocular-Vestibular RS (SD)	0.01 (0.09)	0.16 (0.48)*	0.01 (0.10)	0.30 (0.66)*
Ocular-Vestibular CS (SD)	0.10 (0.63)	0.18 (1.38)	0.19 (0.83)	0.00 (1.17)
Cognitive-Fatigue RS (SD)	0.18 (0.75)	1.48 (2.02)***	0.18 (0.62)	1.85 (1.98)***
Cognitive-Fatigue CS (SD)	1.44 (2.80)	0.11 (4.04)**	2.11 (3.67)	0.35 (3.49)**
Post-traumatic Migraine RS (SD)	0.13 (0.57)	2.29 (2.38)***	0.21 (0.67)	2.94 (2.41)***
Post-traumatic Migraine CS (SD)	0.52 (1.44)	-1.71 (2.75)***	0.77 (1.69)	-2.06 (3.06)***
Anxiety-Mood RS (SD)	0.03 (0.30)	0.14 (0.73)	0.08 (0.72)	0.46 (1.55)
Anxiety-Mood CS (SD)	0.44 (1.25)	0.43 (1.65)	0.91 (2.25)	0.26 (2.36)

n, sample size; RS, raw score; CS, change score; SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; *P<0.05 **P<0.01 ***P<0.001 Significantly different from normal performance of the same sex using 2-sample non-parametric bootstrap t-test; Effect sizes (Cohen's d) for significant differences for males at 24-48h: SAC RS, d=0.41;SAC CS, d=0.52; BESS RS, d=0.28; BESS CS, d=0.55; SCAT total symptoms RS, d=2.25; SCAT total symptoms CS, d=1.01; Ocular-Vestibular RS, d=0.59; Cognitive-Fatigue RS, d=1.06; Cognitive-Fatigue CS, d=0.42; Post-traumatic Migraine RS, d=1.63; Post-traumatic Migraine CS, d=1.18 Effect sizes (Cohen's d) for significant differences for females at 24-48h: SAC RS, d=0.42; SAC CS, d=0.35; SCAT total symptoms RS, d=2.68; SCAT total symptoms CS, d=1.22; Ocular-Vestibular RS, d=0.87; Cognitive-Fatigue RS, d=1.54; Cognitive-Fatigue CS, d=0.48; Post-traumatic Migraine RS, d=2.13; Post-traumatic Migraine CS, d=1.36

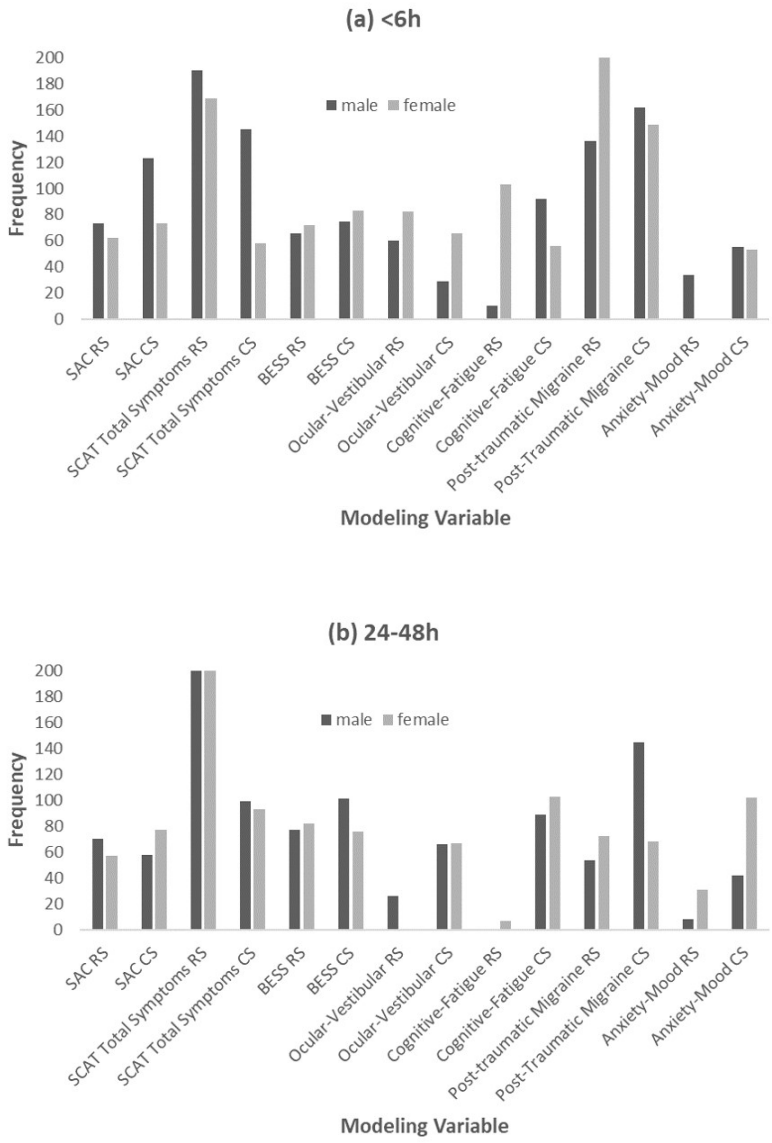


Figure 6.1: Frequency of clustering variables among cluster sets in the lowest 200 Gini Indices by sex at (a) <6h post-injury and (b) 24-48h post-injury

nificantly different between the clusters in each cluster set were total number of symptoms raw score (150/200) and change score (145/200), as well as Post-traumatic Migraine change score (148/200) and Cognitive-Fatigue change score (141/200). Among females at <6h, the most frequently significantly different variables were total number of symptoms change score (148/200), Post-traumatic Migraine raw score (137/200) and change score (145/200), and Ocular-Vestibular change score (145/200). For males at 24-48h, total number of symptoms raw score (153/200) and change score (153/200) as well as the Post-traumatic Migraine raw score (150/200) and change score (150/200) were most often significantly different between clusters in each cluster set. Across females at 24-48h, the variables that were most often significantly different between clusters were total number of symptoms raw score (152/200) and change score (155/200), and Post-traumatic Migraine raw score (147/200) and change score (149/200).

6.3.5 Analysis of Clusters with the Lowest Gini Index

We show the cluster sets which had the lowest Gini Index for males and females at <6h in Table 6.5 and at 24-48h in Table 6.6. For males at <6h (Gini Index=0.224), the first cluster contained 334 athletes (4.49% with concussion) and the second cluster contained 59 athletes (94.92% with concussion). Mean values for all variables were significantly different between the two clusters except for SAC raw score (P=0.42, d=0.10), BESS raw score (P=0.81, d=0.04), and Anxiety-Mood raw score (P=0.21, d=0.19). For females at <6h (Gini Index=0.176), the first cluster had 228 athletes (2.19% with concussion) and the second cluster had 37 athletes (91.89% with concussion). Between the two clusters, mean values for all variables were significantly different except for SAC raw score (P=0.83, d=0.04) and change score (P=0.50, d=0.13), BESS raw score (P=0.055, d=0.34), Cognitive-Fatigue change score (P=0.053, d=0.50), and Anxiety-Mood raw score (P=0.10, d=0.67) and change score (P=0.97, d=0.01). For males at 24-48h (Gini Index=0.362), the first cluster had 81 athletes (82.72% with concussion) and the second cluster had 242 athletes (8.26% with concussion). Mean values for all variables were significantly different between the two clusters except for SAC raw score (P=0.09, d=0.21), BESS raw score (P=0.083, d=0.04), Ocular-Vestibular change score (P=0.84, d=0.04), and Anxiety-Mood raw score (P=0.16, d=0.36) and change score (P=0.93, d=0.01). Finally, for females at 24-48h (Gini Index=0.299), the

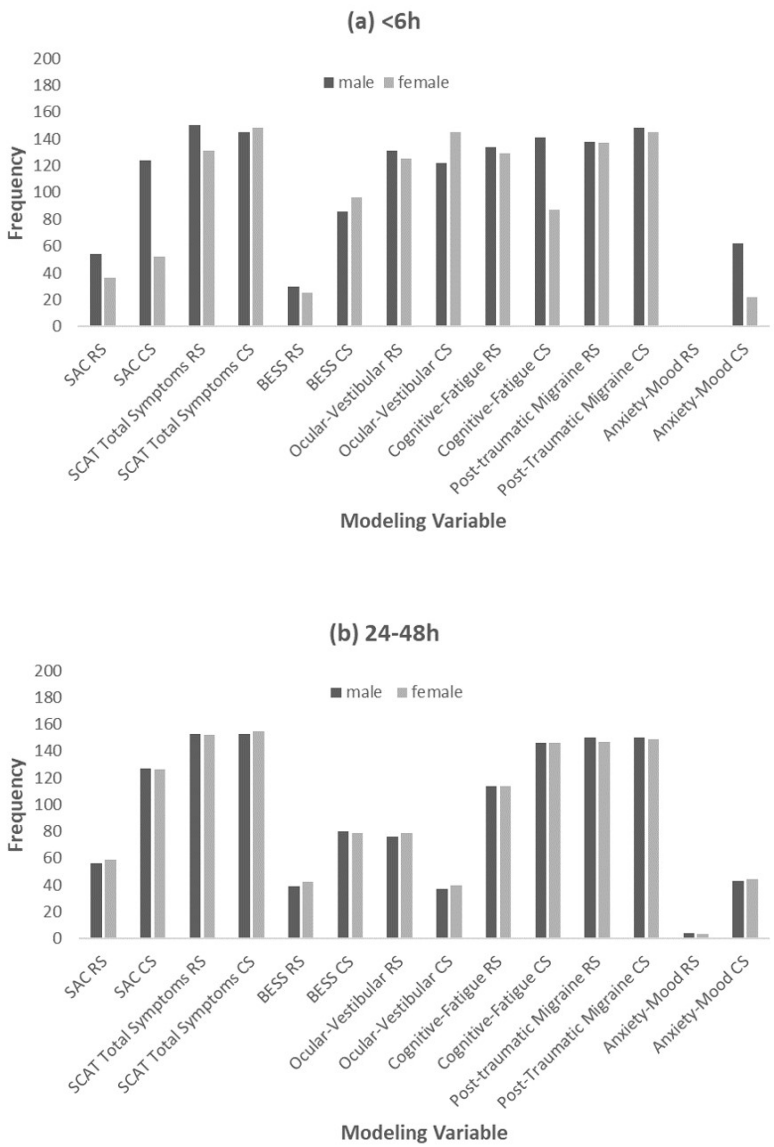


Figure 6.2: Frequency of significant differences in study variables among cluster sets in the lowest 200 Gini Indices by sex at (a) <6h post-injury and (b) 24-48h post-injury

first cluster contained 179 athletes (4.47% with concussion) and the second cluster contained 57 athletes (80.70%) with concussion. Between the two clusters, mean values for all variables were significantly different except for SAC change score (P=0.09, d=0.25), BESS raw score (P=0.57, d=0.08) and change score (P=0.57, d=0.08), Ocular-Vestibular change score (P=0.49, d=0.13), and Anxiety-Mood raw score (P=0.056, d=0.75).

Table 6.5: Two-cluster cluster sets with the lowest Gini index for males and females at <6h

	Male		**Female**	
Variable	*Cluster 1*	*Cluster 2*	*Cluster 1*	*Cluster 2*
n	334	59	228	37
% concussed	4.49%	94.92%	2.19%	91.89%
SAC RS (SD)	27.63 (1.93)	27.44 (1.55)	28.18 (1.64)	28.11 (1.78)
SAC CS (SD)[1]	0.71 (2.19)	-0.27 (2.22)**	0.68 (2.09)	0.41 (2.19)
SCAT total symptoms RS (SD)[1,2]	0.27 (0.81)	4.85 (1.93)***	0.33 (0.84)	5.35 (2.28)***
SCAT total symptoms CS (SD)[1]	-1.72 (2.59)	3.64 (2.88)***	-2.25 (2.76)	1.51 (3.92)***
BESS RS (SD)	11.71 (5.29)	11.92 (6.37)	10.45 (5.44)	12.30 (5.07)
BESS CS (SD)	-1.96 (6.27)	0.37 (7.58)*	-2.43 (6.74)	0.68 (5.00)***
Ocular-Vestibular RS (SD)[1]	0.01 (0.09)	0.97 (1.38)***	0.01 (0.09)	1.73 (2.09)**
Ocular-Vestibular CS (SD)[2]	-0.11 (0.70)	0.64 (1.63)***	-0.22 (0.80)	1.43 (1.94)***
Cognitive-Fatigue RS (SD)[2]	0.15 (0.64)	2.97 (3.01)***	0.13 (0.49)	2.89 (2.01)***
Cognitive-Fatigue CS (SD)[1]	-1.58 (2.81)	1.61 (4.69)***	-1.92 (2.97)	-0.30 (4.46)
Post-traumatic Migraine RS (SD)[1,2]	0.10 (0.42)	4.10 (2.73)***	0.19 (0.59)	4.03 (2.35)***
Post-traumatic Migraine CS (SD)[1,2]	-0.56 (1.33)	3.83 (2.87)***	-0.80 (1.84)	2.65 (3.52)***
Anxiety-Mood RS (SD)	0.02 (0.25)	0.07 (0.25)	0.04 (0.60)	0.54 (1.32)
Anxiety-Mood CS (SD)	-0.51 (1.38)	-0.12 (1.19)*	-0.90 (2.11)	-0.92 (2.97)

n, sample size; RS, Raw Score; CS, Change Score; SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; [1]Cluster variable for males [2]Cluster variable for females *P<0.05 **P<0.01 ***P<0.001 Significantly different from Cluster 1 of the same sex and timepoint using two-sample non-parametric bootstrap t-test; Effect sizes for significant differences for males at <6h: SAC CS, d=0.45; SCAT total symptoms RS, d=4.34; SCAT total symptoms CS, d=2.04; BESS CS, d=0.36; Ocular-Vestibular RS, d=1.78; Ocular-Vestibular CS, d=0.84; Cognitive-Fatigue RS, d=2.16; Cognitive-Fatigue CS, d=1.01; Post-traumatic Migraine RS, d=3.57; Post-traumatic Migraine CS, d=2.66; Anxiety-Mood CS, d=0.29; Effect sizes for significant differences for females at <6h: SCAT total symptoms RS, d=4.37; SCAT total symptoms CS, d=1.28; BESS CS, d=0.48; Ocular-Vestibular RS, d=2.21; Ocular-Vestibular CS, d=1.59; Cognitive-Fatigue RS, d=3.16; Post-traumatic Migraine RS, d=3.73; Post-traumatic Migraine CS, d=1.60

Table 6.6: Two-cluster cluster sets with the lowest Gini index for males and females at 24-48h

	Male		Female	
Variable	*Cluster 1*	*Cluster 2*	*Cluster 1*	*Cluster 2*
n	81	242	179	57
% concussed	82.72%	8.26%	4.47%	80.70%
SAC RS (SD)	27.32 (1.95)	27.75 (2.07)	28.35 (1.65)	27.67 (1.66)**
SAC CS (SD)[1]	-0.09 (2.04)	0.61 (2.21)*	0.63 (2.03)	0.14 (1.80)
SCAT total symptoms RS (SD)[1,2]	3.56 (1.34)	0.17 (0.46)***	0.23 (0.60)	4.25 (1.43)***
SCAT total symptoms CS (SD)[1]	1.60 (3.34)	-1.60 (2.49)***	-2.37 (2.74)	1.53 (3.09)***
BESS RS (SD)	13.02 (7.53)	11.31 (6.47)	10.80 (6.17)	11.26 (5.50)
BESS CS (SD)	1.88 (8.89)	-2.13 (6.26)**	-2.03 (7.03)	-2.60 (6.16)
Ocular-Vestibular RS (SD)[1]	0.19 (0.50)	0.00 (0.06)*	0.00 (0.00)	0.32 (0.66)**
Ocular-Vestibular CS (SD)[2]	-0.15 (1.42)	-0.11 (0.62)	-0.17 (0.80)	-0.05 (1.22)
Cognitive-Fatigue RS (SD)[2]	1.96 (2.10)	0.05 (0.28)***	0.09 (0.36)	2.04 (1.92)***
Cognitive-Fatigue CS (SD)[1]	0.23 (3.97)	-1.52 (2.82)**	-2.18 (3.68)	-0.21 (3.35)***
Post-traumatic Migraine RS (SD)[1,2]	2.64 (2.33)	0.07 (0.32)***	0.15 (0.53)	3.00 (2.30)***
Post-traumatic Migraine CS (SD)[1,2]	2.07 (2.60)	-0.58 (1.42)***	-0.74 (1.69)	1.81 (3.16)***
Anxiety-Mood RS (SD)	0.19 (0.79)	0.02 (0.26)	0.00 (0.00)	0.70 (1.92)
Anxiety-Mood CS (SD)	-0.42 (1.66)	-0.44 (1.25)	-0.96 (2.16)	-0.14 (2.57)*

n, sample size; RS, Raw Score; CS, Change Score; SD, standard deviation; SAC, Standard Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; BESS, Balance Error Scoring System; [1]Cluster variable for males [2]Cluster variable for females *P<0.05 **P<0.01 ***P<0.001 Significantly different from Cluster 1 of the same sex and timepoint using two-sample non-parametric bootstrap t-test; Effect sizes for significant differences for males at 24-48h: SAC CS, d=0.32; SCAT total symptoms RS, d=4.33; SCAT total symptoms CS, d=1.17; BESS CS, d=0.57; Ocular-Vestibular RS, d=0.70; Cognitive-Fatigue RS, d=1.78; Cognitive-Fatigue CS, d=0.56; Post-traumatic Migraine RS, d=2.16; Post-traumatic Migraine CS, d=1.49; Effect sizes for significant differences for females at 24-48h: SAC RS, d=0.41; SCAT total symptoms RS, d=4.59; SCAT total symptoms CS, d=1.38; Ocular-Vestibular RS, d=0.98; Cognitive-Fatigue RS, d=1.96; Cognitive-Fatigue CS, d=0.55; Post-traumatic Migraine RS, d=2.35; Post-traumatic Migraine CS, d=1.19; Anxiety-Mood CS, d=0.36

6.4 Discussion

In this research, we extend our prior work from Chapter 5 using cluster analysis to characterize the Possible and Probable concussions. To our knowledge, this analysis is one of the few which focuses specifically on the subgroup of athletes whose concussions are most difficult to categorize both clinically and based on the algorithm we developed. The findings in this research provide valuable guidance for clinicians by identifying key components of the SAC, BESS, and SCAT symptom checklist which best separate concussions and normal performances among the Possible and Probable concussion groups.

Similarly to our initial analysis, we found that the total number of symptoms was an important differentiator between concussions and normal performances among the algorithmically classified Possible and Probable concussions. To this end, the present study performs a more detailed analysis of symptoms among these subpopulations by incorporating symptom groups which have been previously identified as being useful for the assessment and treatment of concussion (Kontos et al., 2019). For males and females at both <6h and 24-48h, our analysis identified the Post-traumatic Migraine symptom group (both raw and change scores) as an important differentiator between athletes with acute concussion and those without. To a lesser degree, both raw scores and change scores for the Cognitive-Fatigue symptom group as well as raw scores for the Cervical group (i.e., neck pain) were also frequently clustered on and significantly different between the best-performing cluster groups for males and females at 24-48h. However, the same trend does not hold for these symptom groups at <6h. This result aligns with the findings by Kontos et al., 2019, who quantified the diagnostic utility of a clinical profile-based approach and found that, compared to all symptom groups, the Post-traumatic Migraine symptom group could most accurately discriminate between concussed athletes and controls. In contrast, they found that the Ocular and Vestibular symptom groups had the second and third highest discriminative ability, respectively, among all symptom groups. In our analysis, the Ocular-Vestibular group played a small role in our analysis. This difference may be attributed to differences in sample selection, where they consider a sample of athletes aged 12-19 years whereas we select a subset of collegiate athletes who were algorithmically difficult to identify as having concussion. Finally, we did note that the Anxiety-Mood symptom group was rarely clustered on or significantly different among the cluster sets which best separated between normal performances and acute concussions. This result mimics the findings of Kontos et al., 2019, who found that the Anxiety-Mood symptom group was the worst-performing symptom group among those which we studied in our analysis.

In previous studies analyzing concussion assessment batteries, the SAC and BESS have been found to perform similarly for identifying concussion in collegiate athletes (Chin et al., 2016; Downey, Hutchison, and Comper, 2018; Garcia et al., 2018; McCrea et al., 2005). Our analysis extends this literature by characterizing the clinical utility of the SAC and BESS among Possible and Probable concussions. Among males at <6h and 24-48h, we found that SAC change score seemed to play a more important role than BESS raw score or change

score in differentiating between acute concussions and normal performances. However, for females, this trend is reversed. That is, the BESS raw score and change score seemed to be more important than the SAC raw and change scores. Overall, the SAC and BESS appear to play a moderately important role in identifying concussion among Possible and Probable concussions, although they remain overshadowed by the total number of symptoms and the Post-traumatic Migraine symptoms. Given that the SAC and BESS were included in the initial models (see Table 5.2) but not found to be significant, our findings in the current study might suggest that the SAC and BESS could have more utility across the smaller, albeit important subset of athletes whose concussions are difficult to assess. This finding bears similarity to the results by Broglio et al., 2019, whose analysis of the SCAT using a classification and regression tree approach found that increased symptoms typically indicate concussion. However, the SAC and/or BESS would be beneficial to ascertain the concussion assessment for athletes who report low symptom levels or no symptoms — suggesting the importance of a stepwise approach to concussion assessment and diagnosis. Overall, these findings also provide additional support for a multidimensional approach to concussion assessment.

Previous studies have attempted to quantify the value of change scores for the clinical assessment of acute concussion, generally finding that they only have a small marginal benefit over raw scores or a standard normative reference Chin et al., 2016; Echemendia et al., 2012; Garcia et al., 2018; Putukian et al., 2015; Randolph, 2011; Schmidt et al., 2012. However, none of these studies specifically examined the utility of change scores within the subgroup of athletes who are most difficult to assess for concussion. This analysis provides support regarding the clinical value of change scores among Possible and Probable concussions. Specifically, change scores for the total number of symptoms and Post-traumatic Migraine symptom group were frequently used as clustering variables across males and females at <6h and 24-48h. Similarly, change scores for the total number of symptoms, Post-traumatic Migraine symptom group, and Cognitive-Fatigue symptom group were frequently found to be significantly different between concussed and normal performance groups for males and females at both <6h and 24-48h. These results suggest that among athletes whose assessments were not immediately indicative of a concussion, those who displayed elevated cognitive and migraine symptoms compared to baseline were more likely to have concussion.

Differences in concussion presentation by sex have been identified in previous research

(Covassin et al., 2012; Covassin et al., 2018; Covassin, Schatz, and Swanik, 2007; Dick, 2009; Moran et al., 2020). While our analysis presented some sex-related differences, we found that there were more similarities within timepoints than within sex with regard to the variables which were most frequently clustered on and significantly different. For example, the total number of symptoms, the Post-traumatic Migraine symptom group, and Cognitive-Fatigue symptom group were consistently important between both males and females. While further investigation is warranted to support our findings, these initial results suggest that, within the same timepoint, similar injury assessment approaches can potentially be taken for male and female athletes with Possible/Probable concussion.

This study is not without limitations. First, this study focused on a collegiate athlete population and additional studies would be needed to support whether our findings translate to other patient and athlete populations. Second, this study only included analysis of the SAC, BESS, and SCAT symptom checklist. Future studies could investigate the utility of other assessment methods such as the Vestibular/Ocular-Motor Screening (VOMS), King-Devick or Tandem Gait tests. Third, our cluster analysis only considered two-cluster groups. There could potentially be clusters of three or more variables which can better describe differences between Possible and Probable concussions. Fourth, this work served to characterize Possible and Probable concussions and our methodology was not explicitly designed to develop clinical guidelines. Future work can extend our research by operationalizing our findings, which would facilitate its implementation in clinical settings. Finally, there is no perfect diagnostic marker for concussion and hence, our results relied on clinical judgment to determine which patients had concussion. To this end, this limitation cannot be improved upon until an objective marker for concussion is developed.

To facilitate the assessment and post-injury management of athletes, previous researchers have proposed a certainty-based diagnosis framework (i.e., Possible, Probable, and Definite) for concussion based on clinical experience (Kutcher and Giza, 2014). The present study builds on our analysis in Chapter 5, which aimed to develop a data-driven approach to quantifying these injury designations, paying particular attention to the subset of athletes whose injuries are most difficult to assess, i.e., Possible and Probable concussions. From our previous work, we found that those athletes with high symptom loads are likely to have concussion (i.e., Definite concussion). However, when overall symptoms are low, clinicians should specifically direct their attention to the presence of Post-traumatic Migraine or Cognitive-Fatigue

symptoms, as well as the SAC and BESS evaluations. While these findings can facilitate the assessment of athletes with acute concussion, the clinical exam remains the gold standard for concussion diagnosis.

Chapter 7

Estimating the Value of Incorporating Patient Behavior in Return-to-play Decisions

7.1 Introduction

Chapters 2-6 focused on the assessment and diagnosis of concussion. Once athletes are diagnosed with concussions, they typically progress through a return-to-play protocol until a clinician determines it is safe for them to return-to-play. In this chapter, we develop a general modeling framework for optimizing treatment cessation decisions which accounts for uncertainty in patient behavior and then apply this model to the problem of optimizing the timing of return-to-play from concussion.

In both clinical research and practice, experts have been increasingly advocating for patient-centered care (Epstein and Street, 2011). Patient-reported outcomes (PROs), such as symptom presentation, are a critical component of patient-centered care. PROs help physicians assess a patient's health status and make treatment decisions (Deshpande et al., 2011). One major challenge in incorporating PROs is the potential for patients to purposely present misleading information (Lohr and Zebrack, 2009). For example, in a recent study, between 60% and 80% of participants reported that they had intentionally withheld clinically relevant information from their physicians (Levy et al., 2018). For physicians, inaccurate PROs can complicate the process of accurately assessing a patient's health, leading

to erroneous treatment decisions or wasted clinical resources.

Patients may inaccurately report PROs for several reasons including poor memory, poor questionnaire or survey design, lack of knowledge, or the desire to respond in socially acceptable ways (Levy et al., 2018; Newell et al., 1999). However, in some cases, patients may be *strategic*. That is, strategic patients have different objectives than their doctors and consequently, they deliberately misrepresent PROs to influence the doctor's health assessment or treatment decisions. For example, strategic patients may malinger, i.e., exaggerate their symptoms, for the purpose of "avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs" (Bass and Halligan, 2014). Strategic patients might also intentionally under-report symptoms. For example, high school and collegiate athletes who are assessed for sports-related concussion may under-report symptoms in an effort to return-to-play (RTP) more quickly (Conway et al., 2018; Kerr et al., 2014b; Register-Mihalik et al., 2013a). While the impact of strategically reported PROs has been studied in different epidemiological and psychological contexts, research on its impact in medical practice has been limited (Barsky, 2002).

In spite of the potential issues associated with strategically reported PROs, there is also a danger in assuming all patients exhibit strategic reporting behavior; for example, physician biases about strategic or deceptive behavior can hamper open communication between the doctor and patient, leading to a damaged doctor-patient relationship and negative health outcomes. Furthermore, these issues disproportionately affect patients from marginalized groups (Dehon et al., 2017; Perloff et al., 2006), potentially due to biased expectations or negative stereotypes associated with these patients. Such consequences can be potentially mitigated when treatment planning occurs over several visits, wherein physicians have an opportunity to dynamically adjust their beliefs about the patient's PRO-reporting behavior as they continue to interact over time. To this end, the operations research community has explored sequential treatment planning in many settings including chronic disease management (Denton et al., 2009; Helm et al., 2015; Kazemian et al., 2019; Mason et al., 2014; Mason et al., 2012; Schell et al., 2019), cancer treatment (Ayer, Alagoz, and Stout, 2012; Ayer et al., 2016; Ayvaci et al., 2017; Cevik et al., 2018; Lavieri et al., 2012; Lee, Lavieri, and Volk, 2018; Zhang et al., 2012), and weight-loss management (Aswani et al., 2018), among others. Yet, few of these modeling approaches consider the impact and uncertainty in a patient's PRO-reporting behavior. To ensure the highest quality of patient care and

health outcomes, it is critical to develop medical decision-making models which can learn and incorporate patient behavior while optimizing treatment decisions.

With these challenges in mind, the goals of this chapter are two-fold. Our first goal is to **formulate and solve an optimal sequential decision-making model which jointly captures uncertainty in a patient's health and PRO-reporting behavior**. To address this goal, we formulate a novel Behavior-Learning Multi-agent Partially Observable Markov Decision Process (BLM-POMDP). In this framework, the doctor dynamically optimizes the timing of treatment cessation while learning the patient's behavior type and evolving health state through a combination of (potentially strategically) reported symptoms and objective clinical measures. We find that key theoretical properties for Partially Observable Markov Decision Processes (POMDPs) do not hold for the BLM-POMDP, making its theoretical analysis and numerical solution challenging. Nevertheless, we characterize the BLM-POMDP in terms of Behavior-Aware Multi-agent POMDPs (BAM-POMDP) in which the patient's behavior is known exactly. Then, we leverage the relationship between BAM-POMDPs and POMDPs to provide intuition about the nature of optimal treatment decisions for strategic patients and derive a tractable approximate solution algorithm. Our second goal is to **estimate and quantify the Value of Incorporating Patient Behavior (VoIPB)**. To address this goal, we apply the BLM-POMDP to a case study on RTP from sports-related concussion. In this case study, we parameterize and validate the BLM-POMDP using a large multi-site dataset on concussion among collegiate athletes as well as values from sports medicine literature. Drawing on recent estimates for symptom-reporting behavior across collegiate athletes, we estimate lower and upper bounds on the VoIPB in terms of total discounted health utility by comparing our BLM-POMDP with a POMDP and two practice-based approaches which do not account for patient behavior. We further contextualize the VoIPB by considering its impact in terms of (1) reducing the likelihood of premature RTP and (2) increasing each athlete's total health-adjusted athletic exposures after RTP.

From a theoretical perspective, we make the following contributions.

1. **We formulate a novel BLM-POMDP framework.** Overall, the BLM-POMDP builds on existing modeling approaches for multi-agent POMDPs. In comparison to existing approaches, we consider the case in which agents perform actions sequentially in each decision period (i.e., they do not perform actions simultaneously) and may not

be in cooperation (i.e., the patient may be strategic). Compared to previous modeling frameworks in the medical decision-making literature, the BLM-POMDP explicitly models the interaction between the patient and doctor, including their anticipation of each other's actions. This framework extends POMDPs in a way that naturally fits patient-doctor interactions in many healthcare settings.

2. **We characterize the structure of the BLM-POMDP's optimal value function and optimal policy.** Solving the BLM-POMDP is a formidable task since it generalizes POMDPs which are generally intractable (i.e., PSPACE-Complete) and its state space is expressed as the product space of several probability simplexes. To this end, we prove that the BLM-POMDP can be decomposed into several Behavior-Aware Multi-agent POMDPs (BAM-POMDPs) — a special case of the BLM-POMDP in which the patient's behavior type is known perfectly. However, because the BAM-POMDP models interactions between the doctor and patient, classic results for POMDPs (i.e., convexity of the value function) do not hold, in general, for the BAM-POMDP. Nevertheless, we leverage Blackwell dominance and characterization of the patient's behavior to establish connections between the POMDP and BAM-POMDP. These structural results lead to insights for managing strategic patients and approximating the optimal policy.

We also make the following practical contributions to the management of sports-related concussion.

3. **We are one of the first groups to approach RTP from concussion with a decision-theoretic and data-driven approach.** With concussion identified as a major public health issue, management of sports-related concussion has undergone significant changes in the last decade. Both national and international guidelines are shifting from primarily expert consensus-based best practices to data-driven guidelines grounded in the most recent research. To this end, the criteria for determining RTP continues to be left to clinical judgment. We extend the development of data-driven RTP guidelines by applying and evaluating the BLM-POMDP to RTP from concussion among collegiate athletes.

4. **We provide a data-driven framework to tailor athlete-specific RTP criteria and demonstrate its improvement over current practice.** By leveraging

multi-center data from the CARE Consortium and injury rates derived from the literature, we tailor optimal RTP criteria for male collegiate football players with different concussion histories. This framework, which can be generalized to across a broader range of athlete types, accounts for differences in health-related utilities, symptom-reporting behavior, injury presentation over time, pre-RTP recovery trajectories, and sport-specific injury risks. On a variety of clinically relevant performance measures, our numerical analysis demonstrates a marked improvement of these tailored RTP criteria over existing approaches.

The remainder of this chapter is organized as follows. In Section 7.2, we review the related literature. In Section 7.3, we present our problem setting and BLM-POMDP modeling framework to optimize the timing of treatment cessation for potentially strategic patients. In Section 7.4, we derive structural results for the BLM-POMDP and related managerial implications. These structural results allow us to derive a grid-based approximation to the BLM-POMDP in Section 7.5. In Section 7.6, we apply the BLM-POMDP to optimize the timing of RTP from sports-related concussion. Finally, in Section 7.7, we provide a discussion of our findings and concluding remarks.

7.2 Literature Review

Overall, this research builds upon the medical decision-making literature in which modeling frameworks incorporate patient behavior. Previous work in this area primarily focused on patients' adherence to optimized treatment decisions. For example, research by Shechter et al., 2008 and Barnett et al., 2017 assumed that patient adherences were known problem parameters whereas Mason et al., 2012 and Ayer et al., 2016 assumed that adherence was unknown but unaffected by a patient's health state or treatment decisions. Patient adherence has also been modeled as a function of medication cost (Schell et al., 2019) or the use of adherence-improving interventions (Lobo et al., 2017). While adherence affects treatment decisions, adherence does not modify a doctor's interpretation of health assessments. For example, in the work by Ayer et al., 2016, the likelihood of seeing a positive or negative mammogram result depends only on the patient's underlying health and not their adherence type. In contrast, a patient's PRO-reporting behavior changes the likelihood that they would report different symptom levels. Hence, a patient who reports "high" symptom levels

could signal different health statuses depending on their behavior type. This aspect of PRO-reporting behavior couples the patient's behavior type with their health, thereby prohibiting the application of well-established modeling frameworks such as POMDPs and requiring the development of new modeling frameworks such as the BLM-POMDP.

In contrast to the literature focusing on adherence to medication, others have developed modeling frameworks where the patient's behavior is derived from his or her own objectives. Schottmüller, 2013 models a patient-doctor interaction in which the patient knows that the doctor has financial incentives to recommend certain treatment decisions. In this paper, the author develops and analyzes a single-period cheap talk model to derive the patient's symptom-reporting behavior. In contrast, our modeling framework incorporates a modified cheap talk framework into a larger, multi-period treatment decision problem. To this end, Aswani et al., 2018; Mintz et al., 2017, and Mintz et al., 2019 develop multi-period frameworks in which the decision-maker must determine appropriate weightloss interventions or incentives for patients with obesity. The patient's behavior (i.e., exercise and diet levels) is determined by the solution to his or her own utility-maximization problem which is solved independently of the doctor's response to their actions. In contrast, we focus on the case in which the patient considers the doctor's potential response in determining a symptom-reporting strategy. Finally, work by Zhang, Wernz, and Hughes, 2018 presents a stochastic game analysis in which a patient and doctor jointly determine chronic disease management activities. While we both model patient-doctor interactions in a multi-period stochastic setting, their model considers a perfectly observable state space and "switching" structure in their actions. That is, only the patient can perform actions in some states and only the doctor can perform actions in other states. In contrast, our model considers a partially observable state space and leader-follower structure in the actions where both parties perform actions in every period.

Our modeling framework also bears similarity to approaches developed in multi-agent POMDP literature. The classic multi-agent POMDP (MPOMDP) considers the problem of coordinating decisions among cooperative and independent agents with imperfect knowledge of the state space (Messias, Spaan, and Lima, 2011; Pynadath and Tambe, 2002). In this setting, the agents share their observations amongst each other without noise (i.e., honestly) or costs. The MPOMDP was generalized into the Decentralized POMDP (DEC-POMDP) in which only subsets of agents communicate (Amato et al., 2013). Finally, the Interac-

tive POMDP (I-POMDP) presents a modeling framework in which autonomous agents (who may not coordinate) account for other autonomous agents in their decision-making processes (Gmytrasiewicz and Doshi, 2005). In contrast to the MPOMDP, DEC-POMDP, and I-POMDP, we model agents with potentially conflicting objectives who perform actions sequentially. Furthermore, we assume that the patient can add noise (i.e., strategically report) to the doctor's observations. To this end, some multi-agent POMDP frameworks take a *leader-follower* approach (Chang, Erera, and White, 2015; Krishnamurthy, 2012; Zhuang, Bier, and Alagoz, 2010), i.e., in each decision period, the leader performs an action and the follower observes that action and chooses their own action accordingly. Among these modeling approaches, the work by Krishnamurthy, 2012 is most similar to ours. They model an optimal stopping time problem in which the global decision-maker observes the actions of a local decision-maker and uses these observations to make an estimate of the system state and determine whether to stop the system or not. In contrast to their modeling approach, the local decision-maker in our model (i.e., the patient) considers the resulting action taken by the global decision-maker (i.e., the doctor) whereas in their model, the local agents act independently of the global agent.

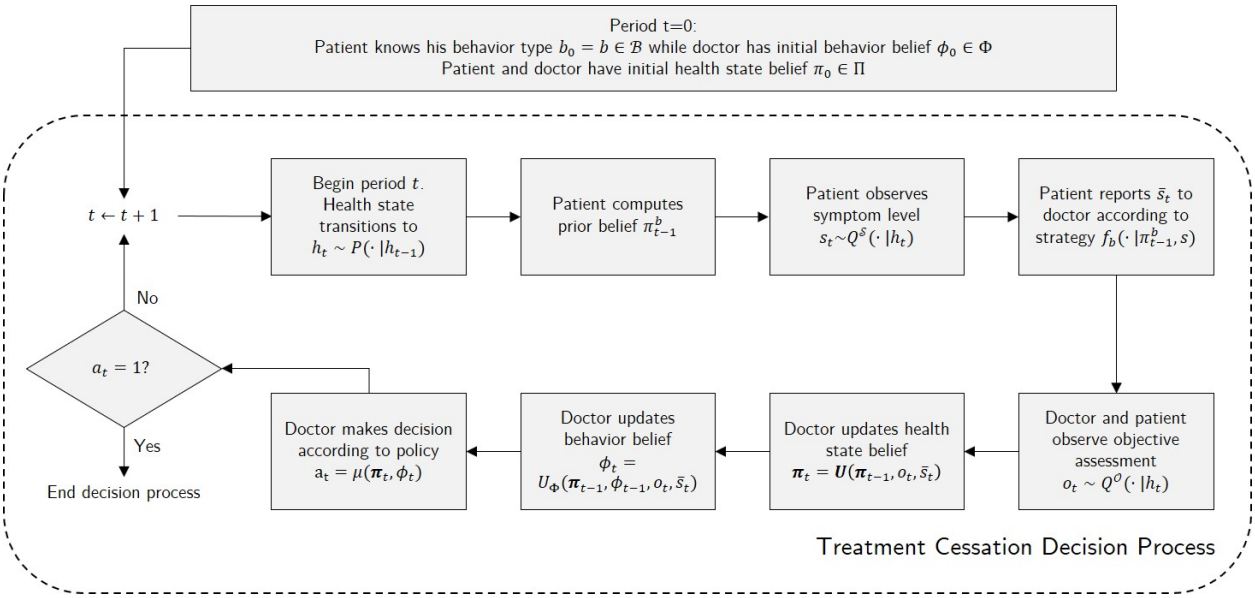
7.3 Modeling Framework

In this section, we describe our problem setting and model formulation for optimizing the timing of treatment cessation with potentially strategic patients. This framework can be modified easily to fit the treatment initiation context.

7.3.1 Problem Description

The problem setting we consider is illustrated in Figure 7.1 and proceeds as follows. A patient (him) is recovering from a diagnosed disease and is undergoing a fixed treatment regimen. He meets periodically with a doctor (her). In each meeting, the patient observes a measure which he later reports to his doctor (i.e., a PRO). To simplify exposition, we refer to this measure and PRO as the observed symptom level and reported symptom level, respectively. The patient, who may be strategic, acts according to his own objectives. That is, he reports a symptom level which could be different from what he has actually observed in an effort

Figure 7.1: Illustration of BLM-POMDP Treatment Cessation Decision Process



to manipulate the doctor's treatment decision. The doctor then performs separate clinical evaluations on the patient (i.e., objective assessments). Based on the reported symptoms and the objective assessments, the doctor assesses the patient's health, updates her belief about the patient's behavior type, and determines whether to cease treatment. If treatment ceases, then the decision process ends. Otherwise, the patient's health evolves stochastically and the decision process repeats in the next meeting between the two. This decision process continues indefinitely until treatment ceases or the patient dies.

In this setting, the patient's recovery depends on receiving treatment and thus, premature treatment cessation could result in adverse health outcomes for the patient. Conversely, delayed treatment cessation may also have consequences, e.g., side effects and costs related to unnecessary treatment. We now describe the components of this model in more detail.

7.3.2 Patient's Health Dynamics

For each decision period $t = 1, 2, ...$ we model the patient's health, its progression over time, and its corresponding observations as a stationary discrete-time Hidden Markov Model (HMM). For clarity, we use a subscript t to denote the status of an HMM component during a specific decision period. The components of this HMM are as follows.

Core Health States: Let $\mathcal{H} := \{0, ..., H\}$ be the ordered set of unobservable core health states. The health state $h = 0$ represents the state in which the patient has fully recovered from the disease while health states $h = 1, ..., H$ represent diseased states in increasing severity. For example, in the case of concussion, the health states can be modeled such that $h = 0$ represents no concussion, $h = 1$ represents asymptomatic concussion, and $h = 2$ represents symptomatic concussion. Without loss of generality, death can be modeled by setting H as an observable absorbing state.

State Transition Probabilities: The initial state distribution is given by the vector $\pi_0 := \begin{bmatrix}\mathbb{P}(h_0 = 0) & \cdots & \mathbb{P}(h_0 = H)\end{bmatrix}^\top$. The patient's health state evolves at the start of each period t according to a transition probability matrix P with components $P(h'|h) := \mathbb{P}(h_{t+1} = h'|h_t = h)$ for all $h, h' \in \mathcal{H}$. Because h is not directly observable by the patient or doctor, it must be inferred from symptoms and objective health assessments (i.e., objective observations).

Symptom observations: After the patient's health state evolves to h at the start of period t, the patient observes a symptom level $s \in \mathcal{S} := \{0, ..., S\}$ where all $s \in \mathcal{S}$ are ordered in increasing severity. Only the patient observes s in each period. The likelihood of observing s_t depends on h_t and is summarized in an observation probability matrix $Q^{\mathcal{S}}$ with entries $Q^{\mathcal{S}}(s|h) := \mathbb{P}(s_t = s|h_t = h)$ for all $s \in \mathcal{S}$ and $h \in \mathcal{H}$. After observing s, the patient reports $\bar{s} \in \mathcal{S}$ to the doctor according to his strategy, which we formally define in Section 7.3.4.

Objective observations: After the patient reports $\bar{s}$ to the doctor, she performs an additional clinical assessment (e.g., a blood biomarker or neurocognitive exam) which cannot be strategically manipulated by the patient. The results of this assessment are denoted by the objective observation $o \in \mathcal{O} := \{0, ..., O\}$, where all $o \in \mathcal{O}$ are ordered in increasing severity. The likelihood of observing o_t depends on h_t. These probabilities are summarized in an observation probability matrix $Q^{\mathcal{O}}$ with components $Q^{\mathcal{O}}(o|h) := \mathbb{P}(o_t = o|h_t = h)$ for all $o \in \mathcal{O}$ and $h \in \mathcal{H}$. We assume that $\mathbb{P}(o_t = o, s_t = s|h_t = h) = Q^{\mathcal{O}}_t(o|h)Q^{\mathcal{S}}_t(s|h)$.

We next describe the doctor's treatment cessation problem (Section 7.3.3) and the patient's symptom-reporting problem (Section 7.3.4).

7.3.3 Doctor's Treatment Cessation Problem

In this section, we formulate the doctor's treatment cessation problem. We denote by $\mathcal{B} := \{0, 1, ..., B\}$ the set of all patient behavior types. Each $b \in \mathcal{B}$ dictates a symptom-reporting strategy f_b which we detail in Section 7.3.4. In addition, each patient knows his own behavior type $b_0 \in \mathcal{B}$. We begin this section by formulating the Behavior-Aware Multi-agent POMDP (BAM-POMDP) which assumes that the patient's behavior type b_0 is also known by the doctor. Then, we generalize the BAM-POMDP by formulating the Behavior-Learning Multi-agent POMDP (BLM-POMDP) in which the doctor learns the patient's behavior type over time.

BAM-POMDP Formulation

When $b_0 = b \in \mathcal{B}$ is known, the doctor's treatment cessation problem is given by a BAM-POMDP with the following components.

Health state beliefs: Prior to beginning the decision process, the doctor has an initial health state belief $\pi_0 \in \Pi$ where π_0 has components $\pi_0(h) = \mathbb{P}(h_0 = h)$ for all $h \in \mathcal{H}$ and $\Pi := \{\pi \in \mathbb{R}_+^{|\mathcal{H}|} : \sum_{h \in \mathcal{H}} \pi(h) = 1\}$ is the $|\mathcal{H}|-1$ probability simplex. Before making a decision in period t, the doctor's information state is given by the tuple $(\pi_0, b_0, \bar{s}_1, ..., \bar{s}_t, o_1, ..., o_t)$. If the patient reported symptoms honestly (i.e., $\bar{s}_t = s_t$ for all t), a sufficient statistic for the doctor's information state is the health belief state $\pi_t^b \in \Pi$ with components $\pi_t^b(h) = \mathbb{P}(h_t = h | \pi_0, b_0 = b, \bar{s}_1, ..., \bar{s}_t, o_1, ..., o_t)$ for all $h \in \mathcal{H}$. Given $\pi_{t-1}^b = \pi, o_t = o$, and $\bar{s}_t = s_t = s$,

$$\pi_t^b = \bar{U}(\pi, o, s) := \frac{D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi}{\bar{C}(\pi, o, s)}, \tag{7.1}$$

where

$$\begin{aligned} \bar{C}(\pi, o, s) &:= 1^\top D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi \\ D_o^{\mathcal{O}} &:= \text{diag}(Q^{\mathcal{O}}(o|0), ..., Q^{\mathcal{O}}(o|H)) \\ D_{\bar{s}}^{\mathcal{S}} &:= \text{diag}(Q^{\mathcal{S}}(\bar{s}|0), ..., Q^{\mathcal{S}}(\bar{s}|H)). \end{aligned}$$

However, if the patient reports symptoms strategically, then a sufficient statistic for her information state can only be constructed if the patient's strategy f_b can be recovered.

Therefore, we assume that the patient has *limited private memory*, i.e., at the beginning of period t, his information state is given by $(\pi_0, b_0, \bar{s}_1, ..., \bar{s}_{t-1}, o_1, ..., o_{t-1}, s_t)$. Limited private memory implies that the patient no longer remembers the *actual* symptoms that he observed in the past but has access to any previous observations which were shared between him and the doctor (e.g., via medical records). This assumption may be reasonable when there is low reliability and accuracy in patient recall of medical history and symptoms (Cohen and Java, 1995; Simon and Gureje, 1999; Van Den Brink, Bandell-Hoekstra, and Huijer Abu-Saad, 2001). Within this class of patient strategies and known behavior type $b_0 = b$, the doctor's belief state is updated according to

$$\pi_t^b = U_b(\pi, o, \bar{s}) := \frac{D_o^{\mathcal{O}} \bar{D}_{\bar{s}}^{\pi,b} P^\top \pi}{C_b(\pi, o, \bar{s})}, \tag{7.2}$$

where $C_b(\pi, o, \bar{s}) := \mathbb{P}(o_t = o, \bar{s}_t = \bar{s} | \pi_t = \pi, b_0 = b) = 1^\top D_o^{\mathcal{O}} \bar{D}_{\bar{s}}^{\pi,b} P^\top \pi$, $\bar{D}_{\bar{s}}^{\pi,b} := \text{diag}(\mathbb{P}(\bar{s}|\pi, b, h = 0), ..., \mathbb{P}(\bar{s}|\pi, b, h = H))$, and $\mathbb{P}(\bar{s}|\pi, b, h) = \sum_{s \in \mathcal{S}} f_b(\bar{s}|\pi, s) Q^{\mathcal{S}}(s|h)$. In (7.2), the likelihood of observing $\bar{s}$ depends on the patient's strategy f_b, which depends on $\pi_{t-1}^b = \pi$. This interdependence between π_t^b, f_b, and π_{t-1}^b links the doctor's decisions with the patient's reported symptoms.

Actions: After observing $\bar{s}_t$ and o_t in decision epoch t, the doctor must choose an action $a \in \mathcal{A} := \{0, 1\}$. If she chooses $a = 0$, the decision process continues in the next period. Otherwise, choosing $a = 1$ ceases treatment and the decision process ends.

Rewards: After performing an action, the doctor immediately receives a reward based on her reward function $r^d : \mathcal{H} \times \mathcal{A} \to \mathbb{R}_+$, where $r^d(\cdot, 0)$ describes the reward received for continuing treatment and $r^d(\cdot, 1)$ describes a large lump sum reward received for ceasing treatment. For convenience we denote $r_a^d := \begin{bmatrix} r^d(0, a) & \cdots & r^d(H, a) \end{bmatrix}^\top$ for all $a \in \mathcal{A}$.

Optimality Equations: The doctor aims to determine a stationary policy $\mu : \Pi \to \mathcal{A}$ which maximizes her expected total discounted reward. The optimal value function V_b can be determined by solving the following Bellman equation for all $\pi \in \Pi$:

$$V_b(\pi) = \max_{a \in \mathcal{A}} \begin{cases} \pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi, o, \bar{s}) V_b(U_b(\pi, o, \bar{s})) & a = 0 \\ \pi^\top r_1^d & a = 1 \end{cases}, \tag{7.3}$$

and the optimal policy $\mu_b^*(\pi)$ corresponds to the maximizing action for $V_b(\pi)$.

BLM-POMDP Formulation

The BAM-POMDP assumes that $b_0 = b$ is known to the doctor. In practice, it is not typically known which type of behavior a patient exhibits. Instead, the doctor forms a belief over the patient's behavior type which she can update dynamically during their interactions. The BLM-POMDP generalizes the BAM-POMDP in this regard. We now describe the main differences between the BLM-POMDP and BAM-POMDP.

Health state and behavior type beliefs: When the patient's behavior type is unknown to the doctor, her information state at time t is given by $(\pi_0, \phi_0, o_1, ..., o_t, \bar{s}_1, ..., \bar{s}_t)$, where $\phi_0 \in \Phi$ is an initial behavior belief vector with components $\phi_0(b) = \mathbb{P}(b_0 = b)$ for all $b \in \mathcal{B}$ and $\Phi := \{\phi \in \mathbb{R}_+^{|\mathcal{B}|} : \sum_{b\in\mathcal{B}} \phi(b) = 1\}$ represents the $|\mathcal{B}| - 1$ probability simplex. We can summarize her information state by the tuple $(\boldsymbol{\pi}_t, \phi_t)$, where $\boldsymbol{\pi}_t = (\pi_t^0, ..., \pi_t^B)$ consists of health beliefs for all $b \in \mathcal{B}$ and the vector ϕ_t is the doctor's behavior belief vector in period t with components $\phi_t(b) = \mathbb{P}(b_0 = b|\pi_0, \phi_0, o_1, ..., o_t, \bar{s}_1, ..., \bar{s}_t)$. Given the previous state $(\boldsymbol{\pi}_{t-1} = \boldsymbol{\pi}, \phi_{t-1} = \phi)$ and current observations $o_t = o, \bar{s}_t = s$, the behavior belief vector is updated according to

$$\phi_t = U_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) := \frac{D^{\boldsymbol{\pi}}_{o,\bar{s}}\phi}{C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})}, \tag{7.4}$$

where

$$C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) := \mathbb{P}(o_t = o, \bar{s}_t = \bar{s}|\boldsymbol{\pi}_t = \boldsymbol{\pi}, \phi_t = \phi) = 1^\top D^{\boldsymbol{\pi}}_{o,\bar{s}}\phi$$
$$D^{\boldsymbol{\pi}}_{o,\bar{s}} = \text{diag}(C_0(\pi^0, o, \bar{s}), ..., C_B(\pi^B, o, \bar{s})).$$

Additionally, the collection of health state belief vectors is updated according to $\boldsymbol{\pi}_t = \mathbf{U}(\boldsymbol{\pi}, o, \bar{s}) = (U_0(\pi^0, o, \bar{s}), ..., U_B(\pi^B, o, \bar{s}))$.

Optimality Equations: In this setting, the optimal value function V is obtained by solving the following Bellman Equations for all $\boldsymbol{\pi}$ and ϕ

$$V(\boldsymbol{\pi}, \phi) = \max_{a\in\mathcal{A}} \begin{cases} \sum_{b\in\mathcal{B}} \phi(b)(\pi^b)^\top r_0^d & \\ \quad +\rho \sum_{o\in\mathcal{O}} \sum_{\bar{s}\in\mathcal{S}} C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) V(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), U_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})) & a = 0 \\ \sum_{b\in\mathcal{B}} \phi(b)(\pi^b)^\top r_1^d & a = 1 \end{cases} \tag{7.5}$$

and the optimal policy $\mu^*(\boldsymbol{\pi}, \phi)$ corresponds to the maximizing action for $V(\boldsymbol{\pi}, \phi)$. When $\phi = e_b$ for some $b \in \mathcal{B}$ (where e_b denotes a vector of 0s with a 1 in the b^{th} component), we have $V(\boldsymbol{\pi}, e_b) = V_b(\pi^b)$. That is, the BAM-POMDP is a special case of the BLM-POMDP. Given that the BLM-POMDP's state space, $\Pi^{|\mathcal{B}|} \times \Phi$, is the Cartesian product of several probability simplexes, it is apparent that the BLM-POMDP suffers from the curse of dimensionality. To this end, our analytical study in Section 7.4 gives way to a numerical solution approach for approximating μ^*.

7.3.4 Patient's Symptom-reporting Problem

We now formulate the patient's symptom-reporting problem (PSRP) for a patient of behavior type $b_0 = b$ through a modified cheap talk framework. In addition to our assumption of limited private memory, we assume that the patient is *myopic*, i.e., he follows a myopic decision rule and assumes that the doctor does so as well. This assumption is reasonable since patients have been shown to heavily discount future health or monetary utilities (Chapman and Elstein, 1995). Furthermore, previous medical decision-making literature has validated myopic patient models in the context of weight-loss intervention (Aswani et al., 2018; Mintz et al., 2017; Mintz et al., 2019). In our model, the patient also assumes that the doctor is *naive*, i.e., the doctor assumes that the patient is honest and her belief update is given by (7.1). This model of symptom-reporting results in greater information transmission compared to the classic cheap talk model, more closely resembling cheap talk behavior identified in empirical studies (Kawagoe and Takizawa, 2009). Hereafter, we refer to the doctor in the PSRP as the *naive doctor* to distinguish her from the actual doctor.

At the start of period t, the patient's belief about his own health is summarized by the tuple (π^b_{t-1}, s_t). His strategy f_b maps his pre-observation belief π^b_{t-1} and observed symptom s_t to the symptom he will report. More precisely, $f_b(\bar{s}|\pi, s) = 1$ if the patient reports $\bar{s}_t = \bar{s}$ given $\pi^b_{t-1} = \pi$ and $s_t = s$. To determine f_b, he maximizes his expected reward, where his reward function is given by $r^p : \mathcal{H} \times \mathcal{A} \times \mathcal{B} \to \mathbb{R}_+$. Thus, for any (π, s), the PSRP is given by

$$\max_{\bar{s} \in \mathcal{S}} \mathbb{E}_{h,o}[r^p(h, \alpha(\pi, o, \bar{s}), b)|\pi, s], \tag{7.6}$$

where α is the naive doctor's action. Based on the patient's assumptions about the naive

doctor,

$$\alpha(\pi, o, \bar{s}) := \arg\max_{a \in \mathcal{A}} \mathbb{E}_h[r_M^d(h, a)|\pi, o, \bar{s}] = \begin{cases} 1 & \bar{U}(\pi, o, \bar{s})^\top \bar{r}_M^d \geq 0 \\ 0 & \bar{U}(\pi, o, \bar{s})^\top \bar{r}_M^d < 0 \end{cases}, \tag{7.7}$$

where the post-observation belief state is given by (7.1), the function $r_M^d : \mathcal{H} \times \mathcal{A} \to \mathbb{R}_+$ is the naive doctor's myopic reward function, and $\bar{r}_M^d := r_M^d(h, 1) - r_M^d(h, 0)$. We now proceed with our derivation of f_b. Let $\mathcal{S}_b^*(\pi, s)$ be the set of symptoms which maximizes (7.6). Notice that $|\mathcal{S}_b^*(\pi, s)|$ may be greater than 1, e.g., if $\alpha(\pi, o, \bar{s}) = 0$ for all $o \in \mathcal{O}$ and $\bar{s} \in \mathcal{S}$, the patient's expected reward constant for every $\bar{s} \in \mathcal{S}$. To this end, we assume that the patient has a tie-breaking function δ such that $\delta(\mathcal{S}_b^*(\pi, s))$ admits a single $\bar{s} \in \mathcal{S}_b^*(\pi, s)$. In our analysis, we take

$$\delta(\mathcal{S}_b^*(\pi, s)) = \arg\min_{s' \in \mathcal{S}_b^*(\pi, s)} |s - s'|. \tag{7.8}$$

We show in Section 7.4.3 that (7.8) is a viable tie-breaking rule (i.e., admits a unique $\bar{s}$) with minimal assumptions. Further, (7.8) has a desirable interpretation; it implies that the patient will only under-report or over-report symptoms as much as is needed to obtain his desired outcome. With this construction, the patient's strategy is given by the function $f_b : \Pi \times \mathcal{S} \times \mathcal{S} \to \{0, 1\}$, where $f_b(\bar{s}|\pi, s) = 1$ if $\delta(\mathcal{S}_b^*(\pi, s)) = \bar{s}$ and $f_b(\bar{s}|\pi, s) = 0$ otherwise.

7.4 Analytical Results

In this section, we provide a theoretical characterization of the BLM-POMDP, BAM-POMDP, and PSRP. In Section 7.4.1, we show that the BLM-POMDP can be decomposed into several BAM-POMDPs. In light of this result, we analyze the structure of the BAM-POMDP in Section 7.4.2. This analysis characterizes the optimal solution to the BAM-POMDP through comparisons between the POMDP and BAM-POMDP and gives way to the solution methodology presented in Section 7.5. Finally, in Section 7.4.3, we characterize the patient's symptom-reporting behavior as modeled by the PSRP. The following definitions and modeling assumptions are used throughout this section.

Definition 7.1 (First-order Stochastic Dominance (FOSD))**.** *Let π_1, π_2 denote two pmfs over a discrete set $\mathcal{X}$. Then π_1 stochastically dominates π_2 in the first-order sense (denoted*

$\pi_1 \succeq_S \pi_2$) *if*

$$\sum_{i \geq x} \pi_1(i) \geq \sum_{i \geq x} \pi_2(i) \text{ for all } x \in \mathcal{X}.$$

Definition 7.2 (Monotone Likelihood Ratio (MLR))**.** *Let π_1, π_2 denote two pmfs over a discrete set $\mathcal{X}$. Then π_1 MLR dominates π_2 (i.e., $\pi_1 \succeq_R \pi_2$) if $\pi_1(j)\pi_2(i) \geq \pi_1(i)\pi_2(j)$ for all $j < i$.*

The notion of FOSD is related to MLR dominance since $\pi_1 \succeq_R \pi_2$ implies $\pi_1 \succeq_S \pi_2$, though the reverse is not generally true. However, when $|\mathcal{X}| = 2$, FOSD and MLR dominance are equivalent.

Definition 7.3 (Supermodular)**.** *A function of two variables $f(x, y)$ is said to be supermodular if for every $x' > x$ and $y' > y$, $f(x, y') + f(x', y) \geq f(x, y) + f(x', y')$.*

Assumption 7.1. *The doctor's reward function $r^d(h, a)$ and patient's $r^p(h, a, b)$ are supermodular in (h, a) and non-increasing in h for all $a \in \mathcal{A}$.*

Assumption 7.1 implies that the total benefit from ceasing treatment on time and continuing treatment when necessary exceeds the utility gained in delayed or premature treatment cessation. In practice, this assumption is easily satisfied since delayed and premature treatment cessation are associated with costs rather than benefits. Additionally, the assumption regarding monotonicity in health states implies that the patient and doctor receive more benefit when the patient is healthier.

Definition 7.4 (Totally Positive of Order 2 (TP2))**.** *A matrix M is TP2 if all of its second-order minors are non-negative.*

MLR dominance provides an equivalent definition for TP2 matrices. Specifically, a transition or observation kernel M is TP2 if the $i + 1^{th}$ row MLR dominates the i^{th} row, i.e., $M_{i+1} \succeq_R M_i$.

Assumption 7.2. *The matrices P, $Q^{\mathcal{O}}$, and $Q^{\mathcal{S}}$ are TP2.*

Assumption 7.2 implies that sicker patients are more likely to remain sick or die than healthier patients. Furthermore, sicker patients are more likely to present more "severe" assessments and symptoms than healthier patients.

Throughout this section, we denote by $F^{\pi,b}$ an $\mathcal{S} \times \mathcal{S}$ matrix with components $F^{\pi,b}(\bar{s}|s) = f_b(\bar{s}|\pi, s)$. Note that we can express the observation probability matrix for reported symptoms as $\bar{Q}^{\mathcal{S}}(\pi, b) = Q^{\mathcal{S}} F^{\pi,b}$ with entries $\mathbb{P}(\bar{s}_t = \bar{s}|h_t = h, \pi_t = \pi, b_0 = b)$. Additionally, $\bar{D}_{\bar{s}}^{\pi,b}$ in (7.2) can be expressed as $\bar{D}_{\bar{s}}^{\pi,b} = \text{diag}(Q^{\mathcal{S}} F_{\bar{s}}^{\pi,b})$, where $F_{\bar{s}}^{\pi,b}$ is the $\bar{s}^{th}$ column of $F^{\pi,b}$.

7.4.1 Analysis of the BLM-POMDP

The BLM-POMDP suffers from the curse of dimensionality since its state space, $\Pi^{|\mathcal{B}|} \times \Phi$, is given by the product of multiple probability simplexes. In the main result of this section, we show that V and μ^* can be constructed from $|\mathcal{B}|$ independently solved BAM-POMDP models. We begin with the following Lemma.

Lemma 7.1. *The BLM-POMDP value function $V(\boldsymbol{\pi}, \phi)$ is linear in ϕ.*

Proof. We prove this result by induction on the value iteration algorithm. The base case holds trivially. Now, suppose that $V_n(\boldsymbol{\pi}, \phi)$ is linear in ϕ for $n = 0, 1, ..., N-1$. For $n = N$, consider each of the following cases.

- **Case 1:** Suppose that $V_N(\boldsymbol{\pi}, \phi) = \sum_{b \in \mathcal{B}} \phi(b)(\pi^b)^\top r_1^d$. Clearly, $V_N(\boldsymbol{\pi}, \phi)$ is linear in ϕ.
- **Case 2:** Suppose that

$$V_N(\boldsymbol{\pi}, \phi) = \sum_{b \in \mathcal{B}} \phi(b)(\pi^b)^\top r_0^d + \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) V_{N-1}(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), U_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}))$$

  Since the immediate reward term $\sum_{b \in \mathcal{B}} \phi(b)(\pi^b)^\top r_0^d$ is linear in ϕ, it suffices to show that the expected value to go term $\sum_{\bar{s} \in \mathcal{S}} C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) V_{N-1}(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), U_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}))$ is linear in ϕ. Take any $o \in \mathcal{O}$ and $\bar{s} \in \mathcal{S}$. Since V_{N-1} is linear in ϕ by the induction

hypothesis, we have

$$
\begin{aligned}
&C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) V_{N-1}(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), U_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})) \\
&= C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) V_{N-1}\left(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), \sum_{b \in \mathcal{B}} \frac{C_b(\pi^b, o, \bar{s})\phi(b)}{C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})} e_b\right) \\
&= C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) \sum_{b \in \mathcal{B}} \frac{C_b(\pi^b, o, \bar{s})\phi(b)}{C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})} V_{N-1}(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), e_b) \\
&= C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) \sum_{b \in \mathcal{B}} \frac{C_b(\pi^b, o, \bar{s})\phi(b)}{C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})} V^b_{N-1}(U_b(\pi^b, o, \bar{s})) \\
&= \sum_{b \in \mathcal{B}} \phi(b) C_b(\pi^b, o, \bar{s}) V^b_{N-1}(U_b(\pi^b, o, \bar{s})),
\end{aligned}
$$

where $V^b_{N-1}(\pi^b) = V_{N-1}(\boldsymbol{\pi}, e_b)$ is the value function estimate on iteration $N-1$ for a BAM-POMDP with a patient whose behavior type is b. Hence, we have

$$
V_N(\boldsymbol{\pi}, \phi) = \sum_{b \in \mathcal{B}} \phi(b) \Big((\pi^b)^\top r^d_0 + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi^b, o, \bar{s}) V^b_{N-1}(U_b(\pi^b, o, \bar{s})) \Big),
$$

which is linear in ϕ.

Since $V_N(\boldsymbol{\pi}, \phi)$ is linear in ϕ, $V_n(\boldsymbol{\pi}, \phi)$ is linear in ϕ for any positive integer n. Furthermore, $V(\boldsymbol{\pi}, \phi)$ is linear in ϕ since $\lim_{n \to \infty} V_n(\boldsymbol{\pi}, \phi) = V(\boldsymbol{\pi}, \phi)$. □

Now we show the main result.

Theorem 7.1. *The value function for the BLM-POMDP can be reformulated as*

$$
V(\boldsymbol{\pi}, \phi) = \max_{a \in \mathcal{A}} \begin{cases} \sum_{b \in \mathcal{B}} \phi(b) \Big((\pi^b)^\top r^d_0 + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi^b, o, \bar{s}) V_b(U_b(\pi^b, o, \bar{s})) \Big) & a = 0 \\ \sum_{b \in \mathcal{B}} \phi(b) (\pi^b)^\top r^d_1 & a = 1 \end{cases}. \tag{7.9}
$$

Proof. From Lemma 7.1, it follows that

$$
C_\Phi(\boldsymbol{\pi}, o, \bar{s}) V(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), U_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})) = \sum_{b \in \mathcal{B}} \phi(b) C_b(\pi^b, o, \bar{s}) V_b(U_b(\pi^b, o, \bar{s})),
$$

for all $o \in \mathcal{O}$ and $\bar{s} \in \mathcal{S}$. By performing this substitution, rearranging sums, and grouping like terms, we have

$$V(\boldsymbol{\pi}, \phi) = \max_{a \in \mathcal{A}} \begin{cases} \sum_{b \in \mathcal{B}} \phi(b)(\pi^b)^\top r_0^d & \\ \quad + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s}) V(\mathbf{U}(\boldsymbol{\pi}, o, \bar{s}), U_\Phi(\boldsymbol{\pi}, \phi, o, \bar{s})) & a = 0 \\ \sum_{b \in \mathcal{B}} \phi(b)(\pi^b)^\top r_1^d & a = 1 \end{cases}$$
$$= \max_{a \in \mathcal{A}} \begin{cases} \sum_{b \in \mathcal{B}} \phi(b) \Big((\pi^b)^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi^b, o, \bar{s}) V_b(U_b(\pi^b, o, \bar{s})) \Big) & a = 0 \\ \sum_{b \in \mathcal{B}} \phi(b)(\pi^b)^\top r_1^d & a = 1 \end{cases}.$$

Hence, $V(\boldsymbol{\pi}, \phi)$ has the desired form in (7.9). □

Theorem 7.1 implies that the expected rewards for each action is a weighted sum of the expected reward for all $|\mathcal{B}|$ BAM-POMDPs. This result suggests that doctors need not consider those behavior types that have been ruled out. For example, if a patient begins to report symptoms that are too low for an over-reporting type of patient, then the doctor can treat the patient as if he were an under-reporting or honest type of patient. Moreover, it follows that if $\mu_b^*(\pi^b) = 1$ (resp., $\mu_b^*(\pi^b) = 0$) for all b such that $\phi(b) > 0$, then $\mu^*(\boldsymbol{\pi}, \phi) = 1$ (resp., $\mu^*(\boldsymbol{\pi}, \phi) = 0$). Therefore, if it is optimal to cease treatment (resp., continue treatment) for all remaining plausible patient behavior types (i.e., $\phi(b) > 0$), then it is optimal to cease treatment (resp., continue treatment) for a patient whose behavioral type is not known with certainty.

7.4.2 Analysis of the BAM-POMDP

Since the BLM-POMDP can be decomposed into $|\mathcal{B}|$ independent BAM-POMDPs, we characterize the BAM-POMDP's optimal policy. We suppress b from our notation throughout this section to simplify exposition. If patients are honest, i.e., $F^\pi = I$ for all $\pi \in \Pi$, the BAM-POMDP reduces to the POMDP. We denote by $\bar{V}$ and $\bar{\mu}^*$ the optimal value function and policy associated with this POMDP. We can immediately establish the following results.

Theorem 7.2. *If the patient is honest, the following properties hold.*

1. *Let* $\bar{\Pi}_1 := \{\pi \in \Pi : \bar{\mu}^*(\pi) = 1\}$ *denote the set in which it is optimal to cease treatment.*

Then, $\bar{\Pi}_1$ is convex and $\bar{\mu}^*$ is characterized by a switching curve on the boundary of $\bar{\Pi}_1$.

2. *Define*

$$r_L^d := r_1^d - (r_0^d + \rho P r_1^d)$$
$$\Pi_L := \Big\{\pi \in \Pi : \pi^\top r_L^d \geq 0\Big\}$$
$$\Pi_L^= := \Big\{\pi \in \Pi : \pi^\top r_L^d = 0\Big\}.$$

For each $h \in \mathcal{H}$, let

$$\nu_h := \{\pi \in \Pi_L^= : \pi(j) = 0 \text{ for } j \neq 0, h\} = \Big\{\pi \in \Pi : \pi(0) = \frac{-r_L^d(h)}{r_L^d(0) - r_L^d(h)}, \pi(h) = 1 - \pi(0)\Big\},$$

define the vertices in which $\Pi_L^=$ intersects the faces of Π. If

$$\Big(D_O^{\mathcal{O}} D_S^{\mathcal{S}} P^\top \nu_i\Big)^\top r_L^d \geq 0 \text{ for all } i \in \mathcal{H}, \tag{7.10}$$

then $\bar{\Pi}_1 = \Pi_L$, i.e., $\mu^*(\pi) = 1$ if $\pi^\top r_L^d \geq 0$ and $\bar{\mu}^*(\pi) = 0$ otherwise.

Proof. We prove these results under the general assumption that F^π is TP2 and constant in π.

1. This result follows from Lovejoy, 1987b.

2. We show this proof in three parts.

   Part 1: First we show that if condition (7.10) holds, then the set Π_L is closed under belief updating. Take any $\pi \in \Pi_L$ and let ℓ_0 denote the line connecting the vertex e_0 with π. Since $\pi \in \Pi_L$, we can extend ℓ_0 past π until it intersects the hyperplane $\Pi_L^=$ at some belief state $\underline{\pi}$. Clearly, $\underline{\pi} \succeq_R \pi$. Furthermore, we can express $\underline{\pi}$ as a convex combination of the vertices $\{\nu_i\}_{i \in \mathcal{H}}$. Hence, we have

$$\Big(D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi\Big)^\top r_L^d \geq \Big(D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \underline{\pi}\Big)^\top r_L^d \geq \Big(D_O^{\mathcal{O}} D_S^{\mathcal{S}} P^\top \underline{\pi}\Big)^\top r_L^d = \sum_{i \in \mathcal{H}} \lambda_i \Big(D_O^{\mathcal{O}} D_S^{\mathcal{S}} P^\top \nu_i\Big)^\top r_L^d \geq 0,$$

where the $\lambda_i \in [0,1]$ satisfy $\sum_i \lambda_i = 1$ and $\sum_{i \in \mathcal{H}} \lambda_i \nu_i = \underline{\pi}$, the inequalities leverage MLR dominance and the fact that the vector r_L^d is nonincreasing in h, and the last inequality follows from condition (7.10). Therefore, $\bar{U}(\pi, o, \bar{s}) \in \Pi_L$.

Part 2: We prove by induction on the value iteration algorithm that if $\pi \in \Pi_L$, then $\bar{V}(\pi) = \pi^\top r_1^d$ which implies that $\mu^*(\pi) = 1$. For the base case, arbitrarily assume that $\bar{V}_0(\pi) = \pi^\top r_1^d$ for all $\pi \in \Pi$. Hence, the base case holds trivially. Now, assume that $\bar{V}_n(\pi) = \pi^\top r_1^d$ for all $\pi \in \Pi_L$, $n = 0, 1, ..., N-1$. For $n = N$, take any $\pi \in \Pi_L$. Since $U(\pi, o, \bar{s}) \in \Pi_L$ for any $\pi \in \Pi_L$, we have

$$\begin{aligned}
\bar{V}_N(\pi) &= \max\left\{\pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} \bar{C}(\pi, o, \bar{s}) \bar{V}_{N-1}(\bar{U}(\pi, o, \bar{s})), \quad \pi^\top r_1^d\right\} \\
&= \max\left\{\pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} \bar{C}(\pi, o, \bar{s}) \bar{U}(\pi, o, \bar{s})^\top r_1^d, \quad \pi^\top r_1^d\right\} \\
&= \max\left\{\pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} \left(D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi\right)^\top r_1^d, \quad \pi^\top r_1^d\right\} \\
&= \max\left\{\pi^\top \left(r_0^d + \rho P r_1^d\right), \quad \pi^\top r_1^d\right\} = \pi^\top r_1^d.
\end{aligned}$$

Hence, $V_N(\pi) = \pi^\top r_1^d$. Since $\lim_{n \to \infty} \bar{V}_n(\pi) = \bar{V}(\pi)$, it follows that $\bar{V}(\pi) = \pi^\top r_1^d$ and $\bar{\mu}^*(\pi) = 1$.

Part 3: We now show that if $\pi \notin \Pi_L$, then $\mu^*(\pi) = 0$. For any $\pi \notin \Pi_L$, we have

$$\begin{aligned}
\pi^\top r_1^d &< \pi^\top \left(r_0^d + \rho P r_1^d\right) = \pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} \bar{C}(\pi, o, \bar{s}) \left(\frac{D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi}{\bar{C}(\pi, o, \bar{s})}\right)^\top r_1^d \\
&\leq \pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} \bar{C}(\pi, o, \bar{s}) \bar{V}(\bar{U}(\pi, o, \bar{s})).
\end{aligned}$$

Hence, $\mu^*(\pi) = 0$.

□

Theorem 7.2 characterizes the BAM-POMDPs optimal policy for honest patients. Namely, there is a (potentially non-linear) threshold such that it is optimal for the doctor to cease treatment for patients who are healthier than that threshold. We note that the interpretable decision policy in Property 2 is better known as the myopic one-step look ahead policy

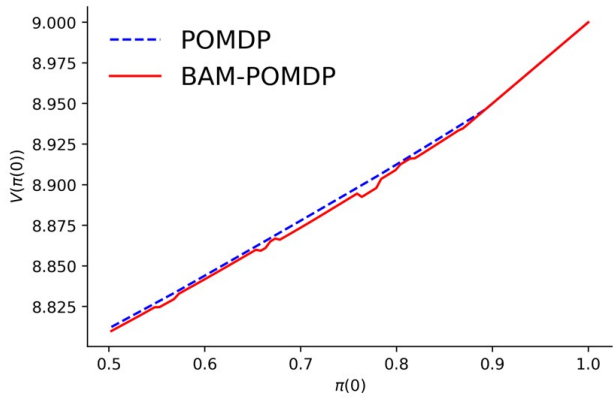


Figure 7.1: The BAM-POMDP's value function is not piecewise linear and convex in π like the POMDP's.

(Yasuda, 1988). This policy states that it is optimal to cease treatment in the current period if treatment cessation yields a greater reward than waiting one period and ceasing treatment in the following period.

Remark 7.1. *If the patient's strategy is independent of π, i.e., $F^\pi = F$ for all $\pi \in \Pi$, then the likelihood of observing any reported symptom $\bar{s}$ is also independent of π and can be summarized by the observation probability matrix $\bar{Q}^{\mathcal{S}} = Q^{\mathcal{S}}F$. In this case, the BAM-POMDP also reduces to a POMDP and if F is TP2, Theorem 7.2 applies.*

When patients are strategic and their strategies depend on π, the BAM-POMDP's value function, in general, is not piecewise linear and convex in π as it for POMDPs (see Figure 7.1). This technical challenge implies that typical approaches to establishing structural results for POMDPs cannot be taken for the BAM-POMDP. To this end, we begin our analysis by establishing the connection between treatment cessation for honest and strategic patients.

Proposition 7.1. *For any patient strategy f, the following inequalities hold:*

$$\sum_{\bar{s}\in\mathcal{S}} C(\pi, o, \bar{s})\bar{V}(U(\pi, o, \bar{s})) \leq \sum_{\bar{s}\in\mathcal{S}} \bar{C}(\pi, o, \bar{s})\bar{V}(\bar{U}(\pi, o, \bar{s})) \text{ for any } o \in \mathcal{O} \text{ and } \pi \in \Pi \quad (7.11)$$

$$V(\pi) \leq \bar{V}(\pi) \text{ for all } \pi \in \Pi. \quad (7.12)$$

Proof. We begin our proof with the following preliminary result.

Lemma 7.2. *The POMDP value function $\bar{V}(\pi)$ is convex in π.*

Proof. Since this result holds for finite horizon POMDPs (Smallwood and Sondik, 1973), induction on the value iteration algorithm implies that the result holds for infinite horizon POMDPs. □

We now show the desired result in two parts.

Part 1: To show that (7.11) holds, we utilize the following relationship between the belief update functions $U(\cdot)$ and $\bar{U}(\cdot)$:

$$\begin{aligned} U(\pi, o, \bar{s}) &= \sum_{s \in \mathcal{S}} \bar{U}(\pi, o, s) \frac{\bar{C}(\pi, o, s)}{C(\pi, o, \bar{s})} f(\bar{s}|\pi, s) \\ C(\pi, o, \bar{s}) &= \sum_{s \in \mathcal{S}} \bar{C}(\pi, o, s) f(\bar{s}|\pi, s). \end{aligned} \tag{7.13}$$

Now, for any $o \in \mathcal{O}$ and $\pi \in \Pi$, we have

$$\sum_{\bar{s} \in \mathcal{S}} C(\pi, o, \bar{s}) \bar{V}(U(\pi, o, \bar{s})) = \sum_{\bar{s} \in \mathcal{S}} C(\pi, o, \bar{s}) \bar{V}\left(\sum_{s \in \mathcal{S}} \bar{U}(\pi, o, s) \frac{\bar{C}(\pi, o, s)}{C(\pi, o, \bar{s})} f(\bar{s}|\pi, s) \right) \tag{7.14}$$

$$\leq \sum_{\bar{s} \in \mathcal{S}} C(\pi, o, \bar{s}) \sum_{s \in \mathcal{S}} \frac{\bar{C}(\pi, o, s)}{C(\pi, o, \bar{s})} f(\bar{s}|\pi, s) \bar{V}(\bar{U}(\pi, o, s)). \tag{7.15}$$

Note that the equality in (7.14) follows from (7.13). The inequality in (7.15) follows from the application of Jensen's Inequality (since $\bar{V}(\pi)$ is convex in π by Lemma 7.2) and the fact that the expression $\frac{\bar{C}(\pi,o,s)}{C(\pi,o,\bar{s})} f(\bar{s}|\pi, s)$ induces a probability distribution over s. Furthermore,

$$\begin{aligned} \sum_{\bar{s} \in \mathcal{S}} C(\pi, o, \bar{s}) \sum_{s \in \mathcal{S}} \frac{\bar{C}(\pi, o, s)}{C(\pi, o, \bar{s})} f(\bar{s}|\pi, s) \bar{V}(\bar{U}(\pi, o, s)) &= \sum_{\bar{s} \in \mathcal{S}} \sum_{s \in \mathcal{S}} f(\bar{s}|\pi, s) \bar{C}(\pi, o, s) \bar{V}(\bar{U}(\pi, o, s)) \\ &\leq \sum_{s \in \mathcal{S}} \bar{C}(\pi, o, s) \bar{V}(\bar{U}(\pi, o, s)), \end{aligned}$$

since $f(\bar{s}|\pi, s) \in \{0, 1\}$ for all $\bar{s} \in \mathcal{S}$. Hence, we have shown that the inequality (7.11) holds.

Part 2: Now, we show that (7.12) holds by induction on the value iteration algorithm. By setting $V_0(\pi) = \pi^\top r_1^d$ for all $\pi \in \Pi$, it is obvious that $V_0(\pi) = \pi^\top r_1^d \leq \bar{V}(\pi)$. Now, assume that for iterations $n = 1, ..., N-1$, we have $V_n(\pi) \leq \bar{V}(\pi)$ for all $\pi \in \Pi$. Take

any arbitrary $\pi \in \Pi$. If $V_N(\pi) = \pi^\top r_1^d$, then we have our desired result. If $V_N(\pi) = \pi^\top r_0^d + \rho \sum_{o\in\mathcal{O}} \sum_{\bar{s}\in\mathcal{S}} C(\pi, o, \bar{s}) V_{N-1}(U(\pi, o, \bar{s}))$, we have

$$\begin{aligned}
\bar{V}(\pi) &\geq \pi^\top r_0^d + \rho \sum_{o\in\mathcal{O}} \sum_{\bar{s}\in\mathcal{S}} \bar{C}(\pi, o, \bar{s}) \bar{V}(\bar{U}(\pi, o, \bar{s})) \\
&\geq \pi^\top r_0^d + \rho \sum_{o\in\mathcal{O}} \sum_{\bar{s}\in\mathcal{S}} \bar{C}(\pi, o, \bar{s}) V_{N-1}(\bar{U}(\pi, o, \bar{s})) \qquad (7.16) \\
&\geq \pi^\top r_0^d + \rho \sum_{o\in\mathcal{O}} \sum_{\bar{s}\in\mathcal{S}} C(\pi, o, \bar{s}) V_{N-1}(\bar{U}(\pi, o, \bar{s})), \qquad (7.17)
\end{aligned}$$

where (7.16) follows from the induction hypothesis and (7.17) follows from (7.11). Therefore, by induction, this inequality $V_n(\pi) \leq \bar{V}(\pi)$ holds for all n. Since $\lim_{n\to\infty} V_n(\pi) = V(\pi)$, it follows that $V(\pi) \leq \bar{V}(\pi)$, which completes the proof. □

Proposition 7.1 shows that the doctor's performance is no better with a strategic patient than it would be with an honest patient. This result can be explained by the notion of Blackwell dominance (Blackwell, 1953). Since $\bar{Q}^{\mathcal{S}}(\pi) = Q^{\mathcal{S}} F^\pi$, symptoms reported by a strategic patient are less informative than those reported by an honest patient, making the doctor's decision process more challenging. Nevertheless, we can leverage (7.11) and (7.12) to characterize μ^*.

Theorem 7.3. *For any f, the doctor's optimal policy μ^* satisfies the following properties.*

1. *It is optimal to wait one more period if $\pi^\top r_L^d < 0$.*
2. *If it is optimal to cease treatment assuming that the patient will be honest in the next period, then it is optimal to cease treatment. That is, $\bar{\mu}^*(\pi) = 1$ implies $\mu^*(\pi) = 1$.*

Proof. The proof for each property follows.

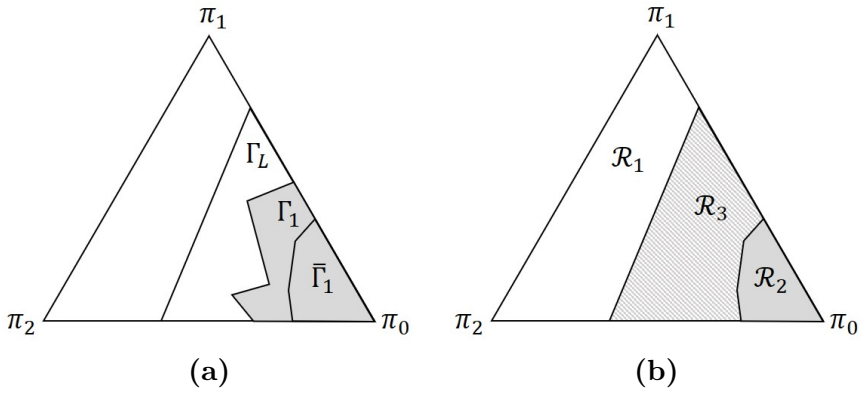
1. This result is identical to the proof of Property 2 in Theorem 7.2.
2. Take any $\pi \in \bar{\Pi}_1$. Then,

$$\begin{aligned}
\pi^\top r_1^d &\geq \pi^\top r_0^d + \rho \sum_{o\in\mathcal{O}} \sum_{\bar{s}\in\mathcal{S}} \bar{C}(\pi, o, \bar{s}) \bar{V}\Big(\bar{U}(\pi, o, \bar{s})\Big) \qquad (7.18) \\
&\geq \pi^\top r_0^d + \rho \sum_{o\in\mathcal{O}} \sum_{\bar{s}\in\mathcal{S}} C(\pi, o, \bar{s}) V\Big(U(\pi, o, \bar{s})\Big), \qquad (7.19)
\end{aligned}$$

where (7.18) follows from the fact that $\pi \in \bar{\Pi}_1$ and (7.19) follows from Proposition 7.1. From (7.19), it follows that $\mu^*(\pi) = 1$, as desired.

□

Figure 7.2: (a) Theorem 7.3 implies that $\bar{\Gamma}_1 \subseteq \Gamma_1 \subseteq \Gamma_L$ with $\mu^*(\pi) = 1$ in all shaded regions and $\mu^*(\pi) = 0$ in all non-shaded regions. (b) Illustration of variable-resolution grid regions in Section 7.5 based on Theorem 7.3. Allocation of grid points should be prioritized in $\mathcal{R}_3$, followed by $\mathcal{R}_1$ and $\mathcal{R}_2$.



Theorem 7.3 implies that $\bar{\Gamma}_1 \subseteq \Gamma_1 \subseteq \Gamma_L$ (see Figure 7.2a), where $\Gamma_1 = \{\pi \in \Pi : \mu^*(\pi) = 1\}$. This characterization guides our allocation of grid points for approximating μ^* as described in Section 7.5 (see Figure 7.2b). Furthermore, μ^* inherits some important structural properties of $\bar{\mu}^*(\pi)$. For example, Π_1 contains a convex subset since $\bar{\Pi}_1$ is convex. Therefore, μ^* contains a (potentially non-linear) threshold for which it is optimal to cease treatment for all patients believed to be healthier than that threshold. We can also establish necessary conditions which guarantee that μ^* is characterized by an interpretable linear decision threshold.

Corollary 7.1. *If* (7.10) *holds, then* $\mu^*(\pi) = 1$ *if* $\pi^\top r_L^d \geq 0$ *and* $\mu^* = 0$ *otherwise.*

Proof. Since (7.10) is satisfied, $\bar{\Pi}_1 = \Pi_L$ by Theorem 7.2. Since Theorem 7.3 implies that $\bar{\Pi}_1 \subseteq \Pi_1 \subseteq \Pi_L$, it follows that $\Pi_1 = \Pi_L$. □

The beauty of Theorem 7.3 and Corollary 7.1 is that these results hold without requiring any knowledge of the patient's strategy f. Furthermore, the conditions required for these results can be verified using only the problem data. It might be tempting, then, to suggest

that the patient's strategy can be ignored with little consequence. This assertion is wrong; it is critical that the patient's strategy is incorporated in health state belief updates. Given the same prior belief π and observations $(o, \bar{s})$, a doctor who anticipates symptom under-reporting (resp., over-reporting) from the patient will update her belief about the patient's health more conservatively (resp., aggressively) than a doctor who believes that the patient is reporting symptoms truthfully.

Proposition 7.2. *Suppose that* $\arg\max_{\bar{s}} f(\bar{s}|\pi, s)$ *is non-decreasing in* s.

1. *If* F^π *is lower triangular, then* F^π *is TP2 and* $U(\pi, o, \bar{s}) \succeq_R \bar{U}(\pi, o, \bar{s})$.
2. *If* F^π *is upper triangular, then* F^π *is TP2 and* $\bar{U}(\pi, o, \bar{s}) \succeq_R U(\pi, o, \bar{s})$.

Proof. We begin with the following lemma.

Lemma 7.3. *Suppose that for every* $\pi \in \Pi$,

$$Q^{\mathcal{S}}(\bar{s}|h)\bar{Q}^{\mathcal{S}}(\bar{s}|h', \pi) \geq Q^{\mathcal{S}}(\bar{s}|h')\bar{Q}^{\mathcal{S}}(\bar{s}|h, \pi), \tag{7.20}$$

for each $\bar{s} \in \mathcal{S}$ *and* $h \geq h' \in \mathcal{H}$. *Then,* $\bar{U}(\pi, o, \bar{s}) \succeq_R U(\pi, o, \bar{s})$ *for any* $o \in \mathcal{O}$, $\bar{s} \in \mathcal{S}$, *and* $\pi \in \Pi$. *Conversely, if*

$$Q^{\mathcal{S}}(\bar{s}|h)\bar{Q}^{\mathcal{S}}(\bar{s}|h', \pi) \leq Q^{\mathcal{S}}(\bar{s}|h')\bar{Q}^{\mathcal{S}}(\bar{s}|h, \pi), \tag{7.21}$$

then $U(\pi, o, \bar{s}) \succeq_R \bar{U}(\pi, o, \bar{s})$ *for any* $o \in \mathcal{O}$, $\bar{s} \in \mathcal{S}$, *and* $\pi \in \Pi$.

Proof. We begin by showing that $\bar{U}(\pi, o, \bar{s}) \succeq_R U(\pi, o, \bar{s})$ holds if (7.20) is satisfied. By definition, $\bar{U}(\pi, o, \bar{s}) \succeq_R U(\pi, o, \bar{s})$ if and only if

$$\begin{aligned} &[\bar{U}(\pi, o, \bar{s})]_h[U(\pi, o, \bar{s})]_{h'} \geq [\bar{U}(\pi, o, \bar{s})]_{h'}[U(\pi, o, \bar{s})]_h \\ \iff &\Big[\frac{D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi}{1^\top D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi}\Big]_h \Big[\frac{D_o^{\mathcal{O}} \bar{D}_{\bar{s}}^{\pi} P^\top \pi}{D_o^{\mathcal{O}} \bar{D}_{\bar{s}}^{\pi} P^\top \pi}\Big]_{h'} \geq \Big[\frac{D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi}{1^\top D_o^{\mathcal{O}} D_{\bar{s}}^{\mathcal{S}} P^\top \pi}\Big]_{h'} \Big[\frac{D_o^{\mathcal{O}} \bar{D}_{\bar{s}}^{\pi} P^\top \pi}{D_o^{\mathcal{O}} \bar{D}_{\bar{s}}^{\pi} P^\top \pi}\Big]_h. \end{aligned}$$

After expanding out the terms in both sides of the expression, canceling out normalization

constants, and canceling out the $Q^{\mathcal{O}}(o|h)$ terms, the inequality holds if

$$\begin{aligned} & Q^{\mathcal{S}}(\bar{s}|h) \sum_{h'' \in \mathcal{H}} P(h|h'')\pi(h'')\bar{Q}^{\mathcal{S}}(\bar{s}|h',\pi) \sum_{h'' \in \mathcal{H}} P(h'|h'')\pi(h'') \\ & \quad \geq Q^{\mathcal{S}}(\bar{s}|h') \sum_{h'' \in \mathcal{H}} P(h'|h'')\pi(h'')\bar{Q}^{\mathcal{S}}(\bar{s}|h',\pi) \sum_{h'' \in \mathcal{H}} P(h|h'') \\ \iff & \; Q^{\mathcal{S}}(\bar{s}|h)\bar{Q}^{\mathcal{S}}(\bar{s}|h',\pi) \geq Q^{\mathcal{S}}(\bar{s}|h')\bar{Q}^{\mathcal{S}}(\bar{s}|h,\pi). \end{aligned}$$

Clearly, this inequality holds if (7.20) is satisfied. Showing that (7.21) implies $U(\pi, o, \bar{s}) \succeq_R \bar{U}(\pi, o, \bar{s})$ is similar and has been omitted. □

We now proceed with the proof for each property in Proposition 7.2 below.

1. Suppose that the conditions hold. We first show that F^π is TP2 for any $\pi \in \Pi$. Take any $\pi \in \Pi$. By definition, F^π is TP2 if $f(\bar{s}|\pi, s)f(\bar{s}'|\pi, s') \geq f(\bar{s}'|\pi, s)f(\bar{s}|\pi, s')$ for any $s > s'$ and $\bar{s} > \bar{s}'$. The only non-trivial case occurs if the right-hand side of this expression is equal to 1, which requires that $f(\bar{s}'|\pi, s)f(\bar{s}|\pi, s') = 1$. However, this case cannot occur since $\arg\max_{\bar{s}} f(\bar{s}|\pi, s)$ is non-decreasing in s. Hence, F^π is TP2.

   Now, we show that $U(\pi, o, \bar{s}) \succeq_R \bar{U}(\pi, o, \bar{s})$. From Lemma 7.3, it suffices to show that (7.21) holds. Take any $\bar{s} \in \mathcal{S}$ and $h \geq h'$. Since $\arg\max_{\bar{s}} f(\bar{s}|\pi, s)$ is non-decreasing in s and F^π is lower triangular, we can rewrite (7.21) as

$$Q^{\mathcal{S}}(\bar{s}|h') \sum_{s \geq \bar{s}} Q^{\mathcal{S}}(s|h) \geq Q^{\mathcal{S}}(\bar{s}|h) \sum_{s \geq \bar{s}} Q^{\mathcal{S}}(s|h'). \tag{7.22}$$

   Since $Q^{\mathcal{S}}$ is TP2, $[Q^{\mathcal{S}}]_h \succeq_R [Q^{\mathcal{S}}]_{h'}$ for any $h \geq h'$. Therefore,

$$Q^{\mathcal{S}}(\bar{s}|h')Q^{\mathcal{S}}(s|h) \geq Q^{\mathcal{S}}(\bar{s}|h)Q^{\mathcal{S}}(s|h') \text{ for any } h \geq h', s \geq \bar{s}. \tag{7.23}$$

   Summing both sides of (7.23) over all $s \geq \bar{s}$ gives (7.22), which completes the proof.

2. The proof is similar to the previous case and has been omitted.

□

Proposition 7.2 implies that if a patient under-reports (resp., over-reports) symptoms, a doctor who does not account for strategic behavior would think that the patient is healthier

(resp., sicker) than he actually is. Consequently, failing to account for strategic behavior can result in premature (resp., delayed) treatment cessation for patients who under-report (resp., over-report) symptoms. In Section 7.4.3, we verify that PSRP satisfies these conditions. Altogether, Theorem 7.3 and Proposition 7.2 illustrate the important role played by the health state belief update for strategic patients. We strengthen these findings with our numerical analysis in Section 7.6.

7.4.3 Analysis of Patient Strategies

We now analyze the structure of f_b as determined by the PSRP. We suppress b from our notation throughout this section to simplify exposition. We begin this section by reformulating the patient's bi-level optimization problem (7.6) into a form which is interpretable and amenable to analytical study. For any $\pi \in \Pi$ and $s \in \mathcal{S}$,

$$\begin{aligned}
&\arg\max_{\bar{s}\in\mathcal{S}} \quad \mathbb{E}_{h,o}[r^p(h,\alpha(\pi,o,\bar{s}))|\pi,s] \\
=&\arg\max_{\bar{s}\in\mathcal{S}} \quad \mathbb{E}_{h,o}[r^p(h,1)\alpha(\pi,o,\bar{s}) + r^p(h,0)(1-\alpha(\pi,o,\bar{s}))|\pi,s] \\
=&\arg\max_{\bar{s}\in\mathcal{S}} \quad \mathbb{E}_{h,o}[\bar{r}^p(h)\alpha(\pi,o,\bar{s})|\pi,s] \\
=&\arg\max_{\bar{s}\in\mathcal{S}} \quad \sum_{h\in\mathcal{H}} \left(\frac{\sum_{h'\in\mathcal{H}} P(h|h')\pi(h')Q^{\mathcal{S}}(s|h)}{\sum_{h''\in\mathcal{H}} Q^{\mathcal{S}}(s|h'') \sum_{h'\in\mathcal{H}} P(h''|h')\pi(h')} \right) \bar{r}^p(h) \sum_{o\in\mathcal{O}} \alpha(\pi,o,\bar{s}) Q^{\mathcal{O}}(o|h) \\
=&\arg\max_{\bar{s}\in\mathcal{S}} \quad U^p(\pi,s)^\top \bar{r}^\pi_{\bar{s}}, \qquad (7.24)
\end{aligned}$$

where $\bar{r}^p = r^p_1 - r^p_0$, $U^p(\pi,s) = D^{\mathcal{S}}_s P^\top \pi / 1^\top D^{\mathcal{S}}_s P^\top \pi$, and $\bar{r}^\pi_{\bar{s}}$ is a vector with components $\bar{r}^\pi_{\bar{s}}(h) = \bar{r}^p(h) \sum_{o\in\mathcal{O}} \alpha(\pi,o,\bar{s}) Q^{\mathcal{O}}(o|h)$ for all $h \in \mathcal{H}$. The form of (7.6) derived in (7.24) has computational implications. For fixed π, $\alpha(\pi,o,\bar{s})$ can be pre-computed for all $o \in \mathcal{O}$ and $\bar{s} \in \mathcal{S}$ since it is independent of the patient's strategy f. Therefore, each vector $\bar{r}^\pi_{\bar{s}}$ can be computed through matrix operations, implying that the computational solution to (7.6) can be determined rather easily. We now proceed with our analysis of f.

In Section 7.3.4, we allude to the interpretability of δ as defined in (7.8). We now provide conditions which guarantee that δ is viable (i.e., admits a unique $\bar{s} \in \mathcal{S}$ for all $\pi \in \Pi, s \in \mathcal{S}$) and identify its implications on the patient's strategy f.

Proposition 7.3. *Let $\bar{r}^p$ be a vector with components $\bar{r}^p(h) = r^p(h,1) - r^p(h,0)$ and suppose that δ has the form specified in (7.8). The following properties hold.*

1. *If $\bar{r}^p \geq \bar{r}^d_M$ or $\bar{r}^p \leq \bar{r}^d_M$, then δ admits a unique $\bar{s}$ for all $\pi \in \Pi$ and $s \in \mathcal{S}$.*
2. *If $\bar{r}^p \geq \bar{r}^d_M$ or $\bar{r}^p \leq \bar{r}^d_M$, then $F^\pi = I$ if either $\bar{U}(\pi, 0, 0)^\top \bar{r}^d_M < 0$ or $\bar{U}(\pi, O, S)^\top \bar{r}^d_M \geq 0$.*
3. *If $\bar{r}^p \geq \bar{r}^d_M$ (resp., $\bar{r}^p \leq \bar{r}^d_M$), then F^π is lower (resp., upper) triangular for all $\pi \in \Pi$.*
4. *If $r^p(h,0) \geq r^p(h,1)$ for all $h \geq 1$ and*

$$\sum_{o \leq o^+} Q^{\mathcal{O}}(o|h^+) + \sum_{o \leq o^-} Q^{\mathcal{O}}(o|h^-) \geq \sum_{o \leq o^+} Q^{\mathcal{O}}(o|h^-) + \sum_{o \leq o^-} Q^{\mathcal{O}}(o|h^+) \ , \tag{7.25}$$

for all $o^+ \geq o^-, h^+ \geq h^-$, then $\delta(\mathcal{S}^*(\pi, s))$ is non-decreasing in s.

Proof. We begin by identifying important properties of the naive doctor's treatment cessation decision, $\alpha(\cdot)$.

Lemma 7.4. *The naive doctor's action $\alpha(\pi, o, \bar{s})$ satisfies the following properties.*

1. *$\alpha(\pi, o, \bar{s})$ is non-increasing in o.*
2. *$\alpha(\pi, o, \bar{s})$ is non-increasing in $\bar{s}$.*
3. *$\alpha(\pi, o, \bar{s})$ is MLR non-decreasing. That is, $\pi' \succeq_R \pi$ implies $\alpha(\pi, o, \bar{s}) \geq \alpha(\pi', o, \bar{s})$.*

Proof. Throughout the proof of each property, we use the fact that $\alpha(\pi, o, \bar{s})$ is non-decreasing in $\bar{U}(\pi, o, \bar{s})^\top \bar{r}^d_M$. We now show the proof for each property.

1. By Lemma 1.2 in (Lovejoy, 1987a), $o' \geq o$ implies that $\bar{U}(\pi, o', \bar{s}) \succeq_R \bar{U}(\pi, o, \bar{s})$ by Assumption 7.2. Since $\bar{r}^d_M$ is non-increasing in h and MLR dominance implies FOSD, we have $\bar{U}(\pi, o', \bar{s})^\top \bar{r}^d_M \leq U(\pi, o, \bar{s})^\top \bar{r}^d_M$, implying that $U(\pi, o, \bar{s})^\top \bar{r}^d_M$ is non-increasing in o. Hence, $\alpha(\pi, o, \bar{s})$ is non-increasing in o, as desired.
2. This proof is similar to the previous and has been omitted.
3. This result follows from the fact that $\bar{U}(\pi, o, \bar{s})$ preserves MLR dominance (see Lemma 1.2 in Lovejoy, 1987a).

□

We also use the properties established in Lemma 7.4 to establish the following monotonicity result.

Lemma 7.5. *If $r^p(h,1) \leq r^p(h,0)$ for all $h \geq 1$ and (7.25) holds, then for any $\bar{s} \leq \bar{s}'$, the vector $\bar{r}^\pi_{\bar{s}} - \bar{r}^\pi_{\bar{s}'}$ is non-increasing in h.*

Proof. Take any $\bar{s} \leq \bar{s}'$ and define the thresholds

$$o^+ = \max\{o \in \mathcal{O} : \alpha(\pi, o, \bar{s}) = 1\} \text{ and } o^- = \max\{o \in \mathcal{O} : \alpha(\pi, o, \bar{s}') = 1\}.$$

From Lemma 7.4, $\alpha(\cdot)$ is non-increasing in o and $\bar{s}$, so $o^+ \geq o^-$, $\alpha(\pi, o, \bar{s}) = 1$ for all $o \leq o^+$, and $\alpha(\pi, o, \bar{s}') = 1$ for all $o \leq o^-$. If $o^+ = o^-$, the desired result is achieved. Otherwise,

$$\left[\bar{r}^\pi_{\bar{s}} - \bar{r}^\pi_{\bar{s}'}\right]_h = \sum_{o \in \mathcal{O}} \bar{r}^p(h) Q^{\mathcal{O}}(o|h)(\alpha(\pi, o, \bar{s}) - \alpha(\pi, o, \bar{s}')) = \bar{r}^p(h) \sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|h).$$

For $h = 0$, we clearly have

$$\bar{r}^p(0) \sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|0) \geq \bar{r}^p(h') \sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|h'),$$

for any $h' \geq 0$ since $\bar{r}^p(0) \geq 0 \geq \bar{r}^p(h')$. Otherwise, take any arbitrary $h \leq h'$. We have

$$\bar{r}^p(h) \sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|h) \geq \bar{r}^p(h') \sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|h'),$$

since $0 \geq \bar{r}(h) \geq \bar{r}^p(h')$ and $\sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|h) \leq \sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|h')$ by (7.25). □

The proof for each property of Proposition 7.3 follows.

1. Suppose that $\bar{r}^p \geq \bar{r}^d_M$ and take any arbitrary $\pi \in \Pi$ and $s \in \mathcal{S}$. If $\bar{s} \leq s$ for all $\bar{s} \in \mathcal{S}^*(\pi, s)$, then clearly, $\delta(\mathcal{S}^*(\pi, s)) = \max\{\bar{s} \in \mathcal{S}^*(\pi, s)\}$. If there exists some $\bar{s} \in \mathcal{S}^*(\pi, s)$ such that $\bar{s} > s$, then $s \in \mathcal{S}^*(\pi, s)$ since $U^p(\pi, s)^\top \bar{r}^\pi_s \geq U^p(\pi, s)^\top \bar{r}^\pi_{\bar{s}}$ for all $\bar{s} \geq s$ (see (7.29) in the proof of Property 3). Therefore, $\delta(\mathcal{S}^*(\pi, s)) = s$. Hence,

$\delta(\mathcal{S}^*(\pi, s))$ admits a unique $\bar{s} \in \mathcal{S}$ for all $\pi \in \Pi$ and $s \in \mathcal{S}$. A similar procedure can be used to show the result in the case that $\bar{r}^p \leq \bar{r}^d_M$.

2. Suppose that $\bar{r}^p \leq \bar{r}^d_M$ or $\bar{r}^p \geq \bar{r}^d_M$ and $\bar{U}(\pi, 0, 0)^\top \bar{r}^d_M < 0$. It follows that $\alpha(\pi, 0, 0) = 0$. From Lemma 7.4, $\alpha(\cdot)$ is non-increasing in o and $\bar{s}$, so $\alpha(\pi, o, \bar{s}) = 0$ for all $o \in \mathcal{O}$ and $\bar{s} \in \mathcal{S}$. Therefore, $s \in \mathcal{S}^*(\pi, s)$ for all $s \in \mathcal{S}$. Now, since $\bar{r}^p \leq \bar{r}^d_M$ or $\bar{r}^p \geq \bar{r}^d_M$, δ is a viable tie-breaking rule and since $s \in \mathcal{S}^*(\pi, s)$, we have $\delta(\mathcal{S}^*(\pi, s)) = s$ for all $s \in \mathcal{S}$. Hence, $F^\pi = I$. Showing that $F^\pi = I$ when $\bar{U}(\pi, O, S)^\top \bar{r}^d_M \geq 0$ follows a similar procedure and has been omitted.

3. Take any $s \in \mathcal{S}$ and $\pi \in \Pi$. First, we show that

$$U^p(\pi, s)^\top \bar{r}^\pi_s \geq U^p(\pi, s)^\top \bar{r}^\pi_{\bar{s}}, \tag{7.26}$$

for all $\bar{s} \geq s$. Subtracting the right-hand side of (7.26) from the left-hand side and writing the terms explicitly gives us

$$\sum_{h \in \mathcal{H}} \mathbb{P}(h|\pi, s) \bar{r}^p(h) \sum_{o \in \mathcal{O}} Q^{\mathcal{O}}(o|h)(\alpha(\pi, o, s) - \alpha(\pi, o, \bar{s})). \tag{7.27}$$

Now, define the thresholds

$$o^+ = \max\{o \in \mathcal{O} : \alpha(\pi, o, s) = 1\} \text{ and } o^- = \max\{o \in \mathcal{O} : \alpha(\pi, o, \bar{s}) = 1\}.$$

From Lemma 7.4, $\alpha(\cdot)$ is non-increasing in o and $\bar{s}$, so $o^+ \geq o^-$, $\alpha(\pi, o, s) = 1$ for all $o \leq o^+$, and $\alpha(\pi, o, \bar{s}) = 1$ for all $o \leq o^-$. If $o^+ = o^-$, (7.27) holds and the proof is complete. Otherwise, we can rewrite (7.27) as

$$\begin{aligned}\sum_{h \in \mathcal{H}} \mathbb{P}(h|\pi, s) \bar{r}^p(h) \sum_{o=o^-+1}^{o^+} Q^{\mathcal{O}}(o|h) &= \sum_{o=o^-+1}^{o^+} \sum_{h \in \mathcal{H}} \mathbb{P}(h|\pi, s) \mathbb{P}(o|h) \bar{r}^p(h) \\ &= \sum_{o=o^-+1}^{o^+} \mathbb{P}(o|s, \pi) \sum_{h \in \mathcal{H}} \mathbb{P}(h|\pi, o, s) \bar{r}^p(h), \end{aligned} \tag{7.28}$$

where the equality in (7.28) follows from the fact that $\mathbb{P}(o|h) = \mathbb{P}(o|h, s, \pi)$ by independence from s, π given h and $\mathbb{P}(h|\pi, s)\mathbb{P}(o|h, \pi, s) = \mathbb{P}(h, o|\pi, s) = \mathbb{P}(h|o, \pi, s)\mathbb{P}(o|\pi, s)$.

Now, since $\alpha(\pi, o, s) = 1$ for all $o \leq o^+$ and $\bar{r}^p \geq \bar{r}^d_M$, it follows that

$$0 \leq \bar{U}(\pi, o, s)^\top \bar{r}^d_M \leq \bar{U}(\pi, o, s)^\top \bar{r}^p = \sum_{h \in \mathcal{H}} \mathbb{P}(h|\pi, o, s)\bar{r}^p(h). \tag{7.29}$$

Therefore, since each term $\mathbb{P}(o|s, \pi)$ is non-negative, (7.28) is non-negative which implies that (7.26) holds. Now, if $s \in \mathcal{S}^*(\pi, s)$, then $f(s|\pi, s) = 1$ since δ follows the form in (7.8). Otherwise, there exists some $\bar{s} < s$ such that $U^p(\pi, s)^\top \bar{r}^\pi_{\bar{s}} > U^p(\pi, s)^\top \bar{r}^\pi_s$, so $\arg\max_{\bar{s}} f(\bar{s}|\pi, s) < s$. Hence, $\arg\max_{\bar{s}} f(\bar{s}|\pi, s) \leq s$ for all $s \in \mathcal{S}$, which implies that F^π is lower triangular. Showing that $\bar{r}^p \leq \bar{r}^d_M$ implies F^π is upper triangular proceeds similarly.

4. We prove Property 4 by contradiction. Suppose that there exist $s > s'$ and $\bar{s} < \bar{s}'$ such that $f(\bar{s}|\pi, s) = 1$ and $f(\bar{s}'|\pi, s') = 1$. Since δ follows the form specified in (7.8), it must be the case that $U^p(\pi, s)^\top \bar{r}^\pi_{\bar{s}} > U^p(\pi, s)^\top \bar{r}^\pi_{\bar{s}'}$. Therefore, we have

$$U^p(\pi, s)^\top (\bar{r}^\pi_{\bar{s}} - \bar{r}^\pi_{\bar{s}'}) > 0 \geq U^p(\pi, s')^\top (\bar{r}^\pi_{\bar{s}} - \bar{r}^\pi_{\bar{s}'}). \tag{7.30}$$

Now, Theorem 7.2 implies that $U^p(\pi, s) \succeq_R U^p(\pi, s')$ and Lemma 7.5 implies that $\bar{r}^\pi_{\bar{s}} - \bar{r}^\pi_{\bar{s}'}$ is non-increasing in h. Therefore, we have

$$U^p(\pi, s')^\top (\bar{r}^\pi_{\bar{s}} - \bar{r}^\pi_{\bar{s}'}) \geq U^p(\pi, s)^\top (\bar{r}^\pi_{\bar{s}} - \bar{r}^\pi_{\bar{s}'}),$$

which contradicts (7.30) and completes the proof.

□

Proposition 7.3 classifies symptom-reporting behavior as either honest, under-reporting, and over-reporting and then characterizes each type. For example, a patient who over-values treatment cessation compared to the naive doctor (i.e., $\bar{r}^p \geq \bar{r}^d_M$) will under-report symptoms unless (1) he agrees with the naive doctor's expected decision or (2) he believes that cannot influence the naive doctor's expected decision. Furthermore, patients who fit the conditions imposed in Proposition 7.3 also satisfy the conditions specified in Proposition 7.2, implying that under-reporting (resp., over-reporting) patients are sicker (resp., healthier) than they report. Overall, these results help clinicians to anticipate the kinds of symptom-reporting

behavior they might see in practice and how they can adjust their interpretation of reported symptoms accordingly.

7.5 Solution Methodology

By Theorem 7.1, the BLM-POMDP's optimal policy is given by a weighted sum of BAM-POMDPs. However, BAM-POMDPs generalize POMDPs, for which finding an optimal policy is undecidable (Madani, Hanks, and Condon, 2003). Hence, we use this section to detail an approximate solution method. Grid-based approximations are commonly used for estimating the optimal value functions and optimal policies for infinite horizon POMDPs (Sandikçi, 2011), i.e., Π is discretized into a finite grid $\mathcal{G} = \{g_1, ..., g_{|\mathcal{G}|}\}$ and the value function is estimated for all $g \in \mathcal{G}$.

We construct $\mathcal{G}$ as a variable-resolution grid. From Theorem 7.3, Π can be divided into regions $\mathcal{R}_1 := \{\pi \in \Pi : \pi^\top r_L^d < 0\}$, $\mathcal{R}_2 := \bar{\Pi}_1$, and $\mathcal{R}_3 := \Pi \setminus (\mathcal{R}_1 \cup \mathcal{R}_2)$. Since the optimal policy μ_b^* is known in $\mathcal{R}_1$ and $\mathcal{R}_2$, we prioritize allocation of grid points in $\mathcal{R}_3$, followed by $\mathcal{R}_1$ and $\mathcal{R}_2$ (see Figure 7.2b). Hence, $\mathcal{G} := \left\{ \bigcup_{i=1}^3 \mathcal{G}_i \right\} \cup \left\{ \bigcup_{h \in \mathcal{H}} e_h \right\}$, where $\mathcal{G}_i$ is the grid approximation to $\mathcal{R}_i$ and e_h is the h^{th} vertex of Π. Now, let $\hat{V}_b$ denote our approximation to V_b and let $\hat{V}$ be the POMDP grid-based approximation evaluated over $\mathcal{G}$. For any $\pi \in \mathcal{R}_2$, we have $\hat{V}_b(\pi) = \pi^\top r_1^d$. Otherwise,

$$\hat{V}_b(\pi) = \begin{cases} \pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi, o, \bar{s}) \sum_{g' \in \mathcal{G}} \lambda_{g'}^{U_b(\pi,o,\bar{s})} \hat{V}(U_b(\pi, o, \bar{s})) & \pi \in \mathcal{R}_1 \\ \max \left\{ \pi^\top r_0^d + \rho \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi, o, \bar{s}) \sum_{g' \in \mathcal{G}} \lambda_{g'}^{U_b(\pi,o,\bar{s})} \hat{V}(U_b(\pi, o, \bar{s})), \pi^\top r_1^d \right\} & \pi \in \mathcal{R}_3 \end{cases},$$

where the weights λ_g^π for any $\pi \in \Pi$ are determined by solving the following linear program:

$$\min_{\lambda_g^\pi} \left\{ \sum_{g \in \mathcal{G}} \lambda_g^\pi \hat{V}(g) : \sum_{g \in \mathcal{G}} \lambda_g^\pi g = \pi, \ \sum_{g \in \mathcal{G}} \lambda_g^\pi = 1, \ 0 \leq \lambda_g^\pi \leq 1 \text{ for all } g \in \mathcal{G} \right\}. \tag{7.31}$$

Since $\{e_0, ..., e_H\} \in \mathcal{G}$, (7.31) is always feasible. Now, let $\tilde{\mu}$ denote an approximate policy for the BLM-POMDP where we set $V_b = \hat{V}_b$ in (7.9) for all $b \in \mathcal{B}$. We complete this section by estimating an upper bound on the optimality gap of this approximate policy. Let $v_b(\boldsymbol{\pi}_0, \phi_0) \approx \mathbb{E}^{\tilde{\mu}}[\sum_{t=1}^\infty \rho^{t-1} r^d(h_t, a_t) | b_0 = b, \boldsymbol{\pi}_0, \phi_0]$ denote the expected total discounted

rewards (e.g., estimated via simulation) of the BLM-POMDP for a patient with behavior type $b_0 = b$ under policy $\tilde{\mu}$ with initial beliefs $\boldsymbol{\pi}_0 = (\pi_0^b = \pi_0)_{b \in \mathcal{B}}$ and ϕ_0. By optimality of V_b and applications of Proposition 7.1 and Jensen's Inequality, we have

$$v_b(\boldsymbol{\pi}_0, \phi_0) \leq \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi_0, o, \bar{s}) V_b(U_b(\pi_0, o, \bar{s})) \leq \sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi_0, o, \bar{s}) \sum_{g \in \mathcal{G}} \lambda_g^{U_b(\pi_0, o, \bar{s})} \hat{V}(g).$$

These inequalities provide an upper bound on the optimality gap given by

$$\text{Optimality Gap UB } (\%) = 100 \times \left(1 - \frac{v_b(\boldsymbol{\pi}_0, \phi_0)}{\sum_{o \in \mathcal{O}} \sum_{\bar{s} \in \mathcal{S}} C_b(\pi_0, o, \bar{s}) \sum_{g \in \mathcal{G}} \lambda_g^{U_b(\pi_0, o, \bar{s})} \hat{V}(g)} \right). \tag{7.32}$$

7.6 RTP From Sports-Related Concussion

In this section, we apply the BLM-POMDP to a case study on optimizing return-to-play (RTP) from sports-related concussion. Our analysis focuses on male collegiate football players since football has one of the greatest risks of injury among all collegiate sports and is a focal point for concussion in popular media (Baugh and Kroshus, 2016). We provide background on RTP from sports-related concussion in Section 7.6.1. We briefly summarize our BLM-POMDP model for optimizing RTP from concussion and in Section 7.6.3, we detail the data sources used to derive model parameters. In Section 7.6.4, we describe benchmark RTP polices against which we compare the BLM-POMDP and the simulation framework we use to evaluate each RTP policy in Section 7.6.5. We then analyze the BAM-POMDP optimal policies in Section 7.6.6, the effect of symptom-reporting behavior on each RTP policy in Section 7.6.7, and estimate the Value of Incorporating Patient Behavior (VoIPB) in Section 7.6.8. This analysis aims to (1) provide insights on how to optimize the timing of RTP for potentially strategic athletes and (2) quantify the benefits in accounting for symptom-reporting behavior in RTP decisions.

7.6.1 Background on Sports-Related Concussion

Concussion, the most common type of traumatic brain injury, has been identified as a major public health issue (McCrory et al., 2017). In the short-term, concussion is associated

with the alteration of neurologic function and a wide-ranging set of symptoms which include confusion and memory loss. Furthermore, while the exact relationship is unclear, concussion may be associated with long-term consequences such as cognitive impairment, neurodegenerative disease, depression, and early onset dementia (Guskiewicz et al., 2005; Guskiewicz et al., 2007; Kerr et al., 2014a; Kerr et al., 2012; Kerr et al., 2018a). Improving concussion management is critical to improving patient health outcomes.

For sports-related concussions, a key component of the management protocol is determining when the patient may RTP. During the RTP decision process, current guidelines recommend a multi-faceted approach which combines objective clinical measures (e.g., neurocognitive and balance assessments) and subjective measures (e.g., self-reported symptoms) to estimate an athlete's health status. While self-reported symptoms are typically the most indicative measure of concussion (Chin et al., 2016; Garcia et al., 2018; McCrea et al., 2005; Register-Mihalik et al., 2013b; Resch et al., 2016), athletes may purposely under-report or over-report symptoms to expedite or delay RTP, among other reasons (Conway et al., 2018; Kerr et al., 2014b; Kroshus et al., 2015a; Kroshus et al., 2015b; Register-Mihalik et al., 2013a). Given the consequences associated with premature RTP (e.g., increased risk of injury (McCrea et al., 2020)) and delayed RTP (e.g., reduced health benefits from physical activity and exercise), understanding the role of symptom self-reporting behavior and its impact in the RTP decision process is critical to improving health outcomes.

7.6.2 Modeling RTP From Concussion

We now summarize the modifications to the BLM-POMDP for modeling athlete-specific RTP decisions (see Appendix 7.A for details). First, we model the set of health states as $\mathcal{H} = \{0, 1, 2\}$ where $h = 0, 1$ and 2 represent recovered, asymptomatic concussion, and symptomatic concussion, respectively. The objective measures are given by $\mathcal{O} = \{30+, 29, 28, 27, 0\text{-}26\}$, where each $o \in \mathcal{O}$ represents a range of scores on the Standard Assessment of Concussion (SAC) — a neurocognitive exam used to measure impairment after concussion. The subjective measures are given by $\mathcal{S} = \{0, 1, 2\text{-}4, 5\text{-}13, 14\text{-}32, 33+\}$, where each $s \in \mathcal{S}$ represents total symptom severity scores on the Sport Concussion Assessment Tool (SCAT) graded symptom checklist. The rewards r^d and r^p represent health utilities, where the lump sum rewards $r^d(h, 1)$ are derived from a post-RTP Markov Reward Process

(MRP) with state space $\Omega = \mathcal{H} \cup \{3\}$ and state transition matrix P^{Ω}. The absorbing state $\omega = 3$ represents a time-loss injury. The MRP reward function $r^{\Omega} : \Omega \to \mathbb{R}_+$ accounts for post-RTP injury risk, health benefits for being in play, and health disutilities from playing while injured. The naive doctor's myopic reward function is given by $r^d_M(h, 0) = r^d(h, 0)$ and $r^d_M(h, 1) = r^{\Omega}(h)$ for all $h \in \mathcal{H}$. We model r^p for each behavior type $b \in \mathcal{B}$ by modifying the perceived benefit in RTP relative to r^d_M. Specifically, $b = 0$ corresponds with honest symptom-reporting while $b > 0$ (resp., $b < 0$) corresponds with symptom under-reporting (resp., over-reporting). Finally, we augment $P, Q^{\mathcal{S}}, Q^{\mathcal{O}}, r^d$, and r^p with an athlete specific state $\theta \in \Theta$, where the set Θ contains modifying factors such as sex, concussion history, and sport. Since we focus on male collegiate football players, we set $\Theta = \{0, 1+\}$, where $\theta \in \Theta$ represents the athlete's number of previous concussions.

7.6.3 Model Parameters and Data Sources

For each $\theta \in \Theta$, we jointly derived the HMM parameters P_θ, $Q^{\mathcal{O}}_\theta$, and $Q^{\mathcal{S}}_\theta$ by applying the Baum-Welch algorithm (Rabiner, 1989) on data from the National Collegiate Athletic Association and Department of Defense (NCAA-DoD) Concussion Assessment, Research, and Education (CARE) Consortium (Broglio et al., 2017). This dataset combines concussion assessment data from 29 NCAA universities and military service academies throughout the United States. To our knowledge, this dataset is the largest available on concussion among collegiate athletes. We validated our HMMs using held-out testing data. To parameterize our post-RTP MRP, we utilized injury rates published in the sports medicine literature (Harada et al., 2019; Herman et al., 2017; Kerr et al., 2018b; McCrea et al., 2009). To derive health utilities for our reward functions, we used health-related quality of life estimates from the concussion and sports medicine literature (Cowee and Simon, 2019; McAllister et al., 2001; Weber et al., 2019) combined with a previously published health utility estimating function from the medical decision-making literature (Lawrence and Fleishman, 2004). Given these rewards, we set $\mathcal{B} = \{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$ where $b = -1$ corresponds to a patient who prefers never to RTP and $b = 1$ corresponds to a patient who always prefers to RTP for each $\theta \in \Theta$. We provide additional details on these parameters in Appendix 7.B.

7.6.4 Benchmark Policies

For the BLM-POMDP, we set $\pi_0 = e_2$, i.e., all athletes are assumed to have symptomatic concussions at the start of the RTP decision process, and

$$\phi_0 = \begin{bmatrix} 0.025 & 0.025 & 0.05 & 0.15 & 0.5 & 0.15 & 0.05 & 0.025 & 0.025 \end{bmatrix}^\top,$$

i.e., the doctor assumes that the patient is likely to be honest. The BLM-POMDP is not sensitive to ϕ_0 in this case study (see Appendix 7.C). We compared the BLM-POMDP to a POMDP which does not account for symptom-reporting behavior and two practice-based policies which assume that the athlete reports symptoms honestly. The first practice-based policy, denoted *Myopic*, is a simple myopic policy which permits RTP once SAC and symptom scores are normal (Broglio, Macciocchi, and Ferrara, 2009), i.e., in period t' where $t' = \inf\{t : o_t = \bar{s}_t = 0\}$. The second policy, denoted *CurrPrac*, more closely mimics current practice by permitting RTP 7 days after the athlete presents normal SAC and symptom scores at least once (McCrea et al., 2020), i.e., in period $t' + 7$ where period $t' = \sup\{\inf_t\{t : \bar{s}_t = 0\}, \inf\{t : o_t = o\}\}$.

7.6.5 Simulation Framework

We evaluated the performance by each RTP policy using discrete event simulation with $1,000$ replications for each $\theta \in \Theta$ and $b_0 \in \mathcal{B}$. In each iteration, we assumed that the athlete initially had acute concussion, i.e., $h_0 = 2$. Although we formulated the BLM-POMDP for an infinite horizon, we evaluated the policy over 90 days since the majority of RTP decisions are made well within that timeframe (McCrea et al., 2020). The sequence of events in each decision period follow the illustration in Figure 7.1, with all problem data following the models and parameters defined in Sections 7.A and 7.6.3. Once the RTP policy permits RTP, we simulated the post-RTP MRP (see Section 7.A.2).

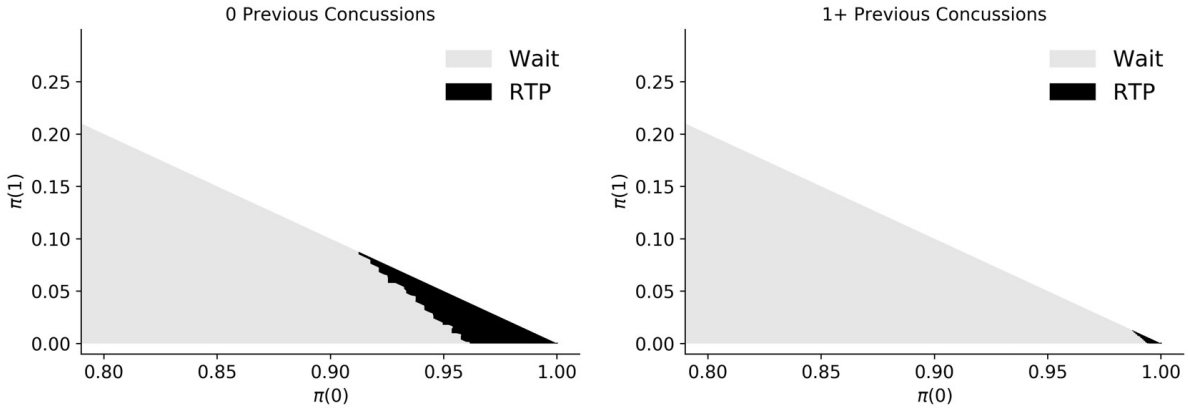
7.6.6 Analysis of BAM-POMDP RTP Policies

In this section, we analyze how concussion history and symptom-reporting behavior modify optimal RTP policies, providing some insights on how optimal RTP policies might change for athletes not included in our study. We also estimate the optimality gaps associated with

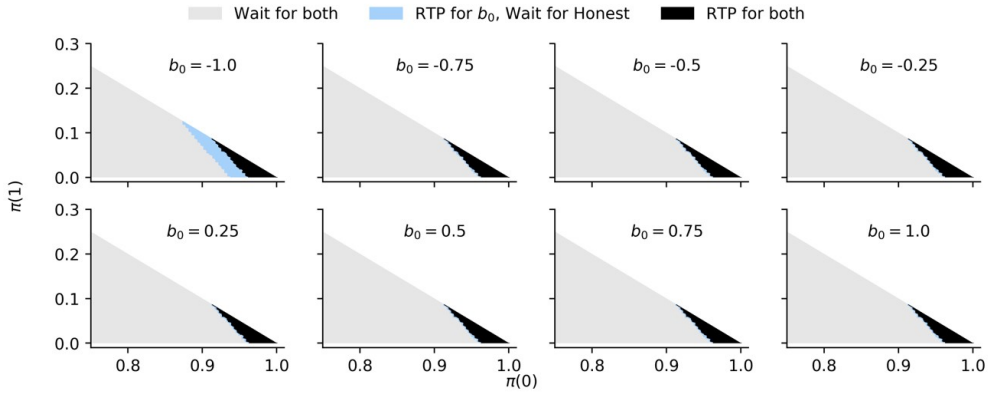
our grid-based approximation. To construct the BLM-POMDP optimal policy, we solved a BAM-POMDP for each $\theta \in \Theta$ and $b_0 \in \mathcal{B}$ using the grid-based approximation of Section 7.5 with 400, 150, and 550 grid points in regions $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$, respectively.

The BAM-POMDP policies for honest athletes are illustrated in Figure 7.1. Recall that these policies are equivalent to the optimal POMDP policies by Theorem 7.2. The optimal policy is more conservative for athletes with 1+ previous concussions compared to 0 previous concussions. This difference is likely due to the increased post-RTP injury risk incurred by athletes with a greater concussion history. Extrapolating beyond our analysis, we expect that athletes at higher risk of injury after RTP require a more conservative RTP policy than those with lower risk (e.g., athletes playing soccer vs. golf). Furthermore, we expect that if the risk of injury post-RTP is sufficiently high, then an optimal policy would never permit RTP.

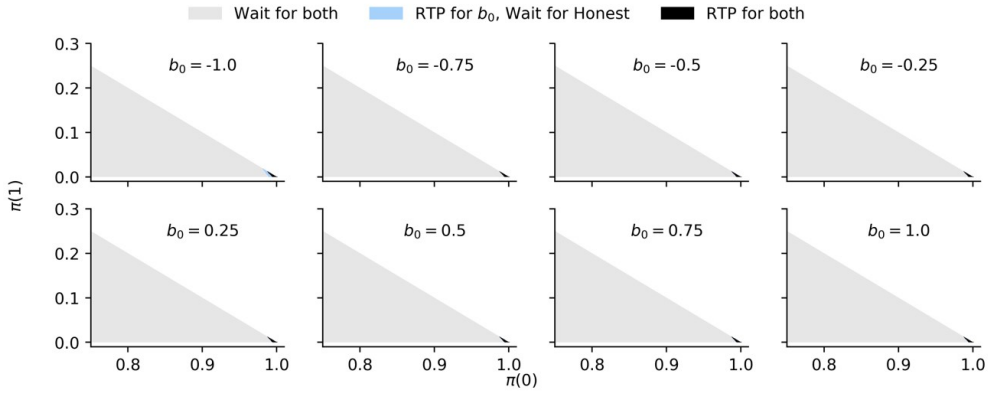
Figure 7.1: BAM-POMDP policies for honest athletes. Black indicates that it is optimal to RTP and gray indicates that it is optimal to wait. For belief states not shown, it is optimal to wait.



For fixed concussion history, RTP policies are more aggressive for athletes who greatly over-report symptoms (i.e., $b_0 = -1$) compared to honest athletes. However, there are only minor differences between the BAM-POMDP policy for honest athletes and all other behavior types (see Figure 7.3). This surprising result might suggest that adaptations to symptom-reporting behavior in the health state belief update sufficiently modify RTP decisions and few compensations are needed within the RTP policy itself. Furthermore, among athletes who under-report symptoms, the RTP policy cannot become any more conservative than it



(a) 0 Previous Concussions



(b) 1+ Previous Concussions

Figure 7.3: Differences in BAM-POMDP policies relative to honest behavior (i.e., $b_0 = 0$). RTP, return-to-play.

already is for honest patients (see Theorem 7.3). Hence, potential differences in performance between the BLM-POMDP and POMDP are likely driven by differences in the health state belief update and not the actual RTP policy. We illustrate this concept for a sample patient in Figure 7.4, where we see that athletes who over-report symptoms will delay RTP and those who under-report symptoms will cause premature RTP compared to the BLM-POMDP.

Figure 7.4: Illustration of BLM-POMDP and POMDP health state belief evolution for sample athletes. Differences in performance between the BLM-POMDP and POMDP are primarily due to differences in health belief updates rather than differences in RTP policy. RTP, return-to-play; $\bar{s}_t$, reported symptom; o_t, objective assessment

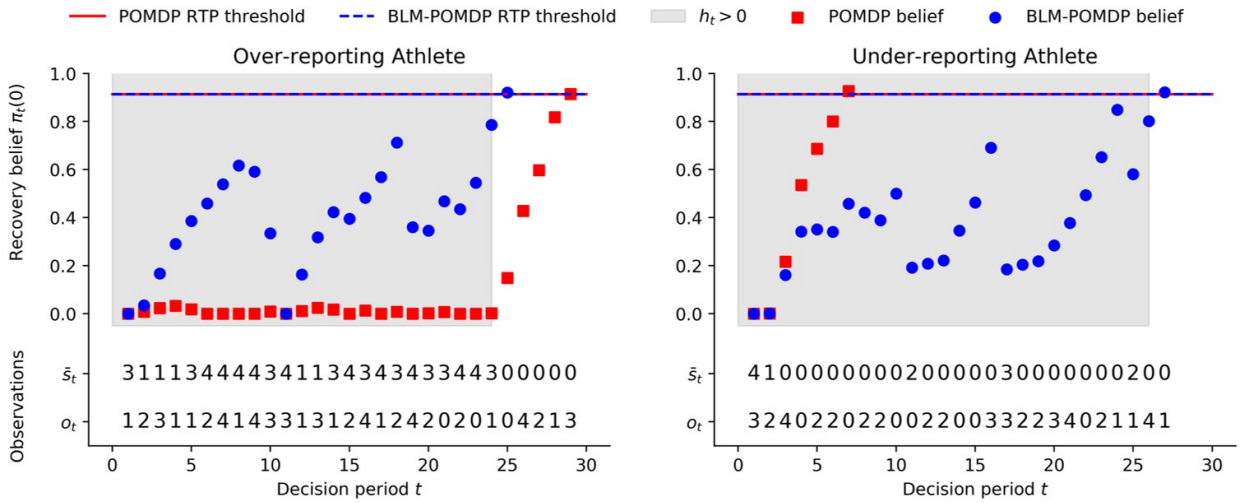


Table 7.1: Optimality Gap UB (%) for BLM-POMDP policies based on simulation estimates

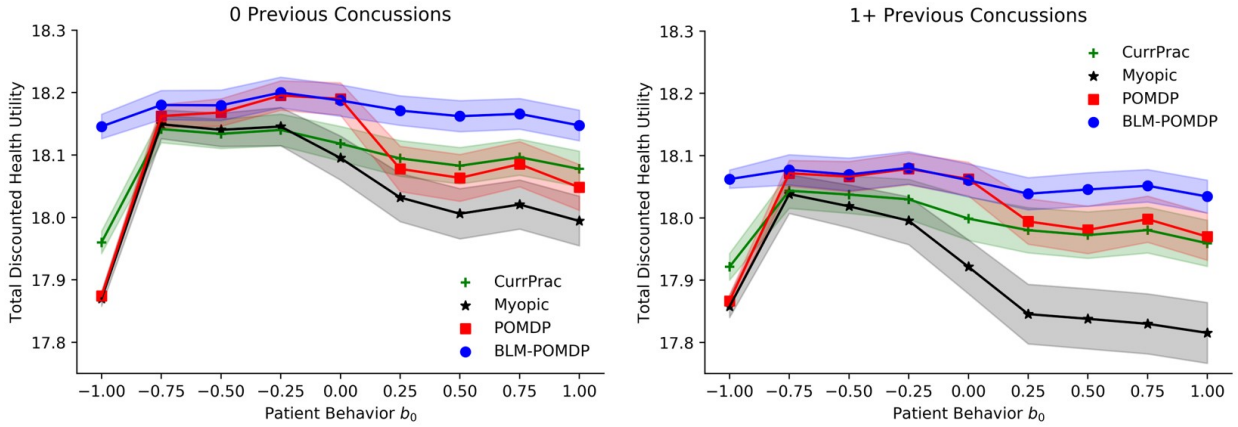
	Symptom-reporting Behavior b_0								
θ	-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
0	0.271%	0.082%	0.085%	-0.028%	0.042%	0.130%	0.180%	0.159%	0.261%
1+	0.217%	0.134%	0.176%	0.117%	0.227%	0.346%	0.308%	0.276%	0.371%

7.6.7 Effect of Symptom-reporting Behavior on RTP Policy Performance

We now analyze how symptom-reporting behavior affects each RTP policy. We begin by estimating the optimality gap associated with the approximate BLM-POMDP policy via (7.32) (see Table 7.1). Overall, these gaps are no larger than 0.371% across all policies, behavior types, and concussion histories. Given the magnitude of these estimates, we suspect that the BLM-POMDP policies are close to optimal.

We present the expected total discounted health utilities achieved by each policy in Figure 7.5. The BLM-POMDP outperforms all benchmark policies although the POMDP perform similarly among honest athletes. Hence, the simpler POMDP suffices for this group. Among

Figure 7.5: Total discounted health utilities for each RTP policy and behavior type $b_0 \in \mathcal{B}$. Markers indicate mean values and shaded areas indicate 95% confidence intervals of the mean.



strategic athletes, the BLM-POMDP most outperforms the benchmark RTP policies when athletes are under-reporting symptoms (i.e., $b_0 > 0$) or when athletes never want to RTP (i.e., $b_0 = -1$). Hence, accounting for symptom-reporting behavior is most beneficial among athletes who under-report symptoms.

We also investigated the timing of RTP decisions relative to the timing of recovery from concussion. For athletes who are permitted to RTP, we compute the *relative RTP delay*, which measures the difference between the time to RTP and the time until the athlete recovers. A value of 0 indicates perfect timing, whereas negative values indicate premature RTP and positive values indicate delayed RTP. We illustrate the relative RTP delay for the BLM-POMDP and POMDP in Figure 7.7. When patients are honest, both the BLM-POMDP and POMDP are more conservative (i.e., take longer to RTP) than the practice-based policies. Across different behavior types, the time to RTP is relatively stable for the BLM-POMDP, with only a small proportion of athletes permitted to RTP before a full recovery is made. In alignment with Proposition 7.2, the POMDP and practice-based policies permit RTP sooner (resp., later) for under-reporting (resp., over-reporting) athletes compared to those who are honest. In fact, when $b_0 = -1$, the POMDP never allows RTP and the practice-based policies permit only 6%-34% of athletes to RTP. In contrast, the BLM-POMDP permits all athletes to RTP. These results suggest that for $b_0 = -1$, differences in performance between the BLM-POMDP and benchmark policies are driven by

the suboptimality of never permitting RTP whereas for under-reporting athletes, the costs associated with premature RTP (e.g., increased injury risk) far outweigh the costs of delayed RTP.

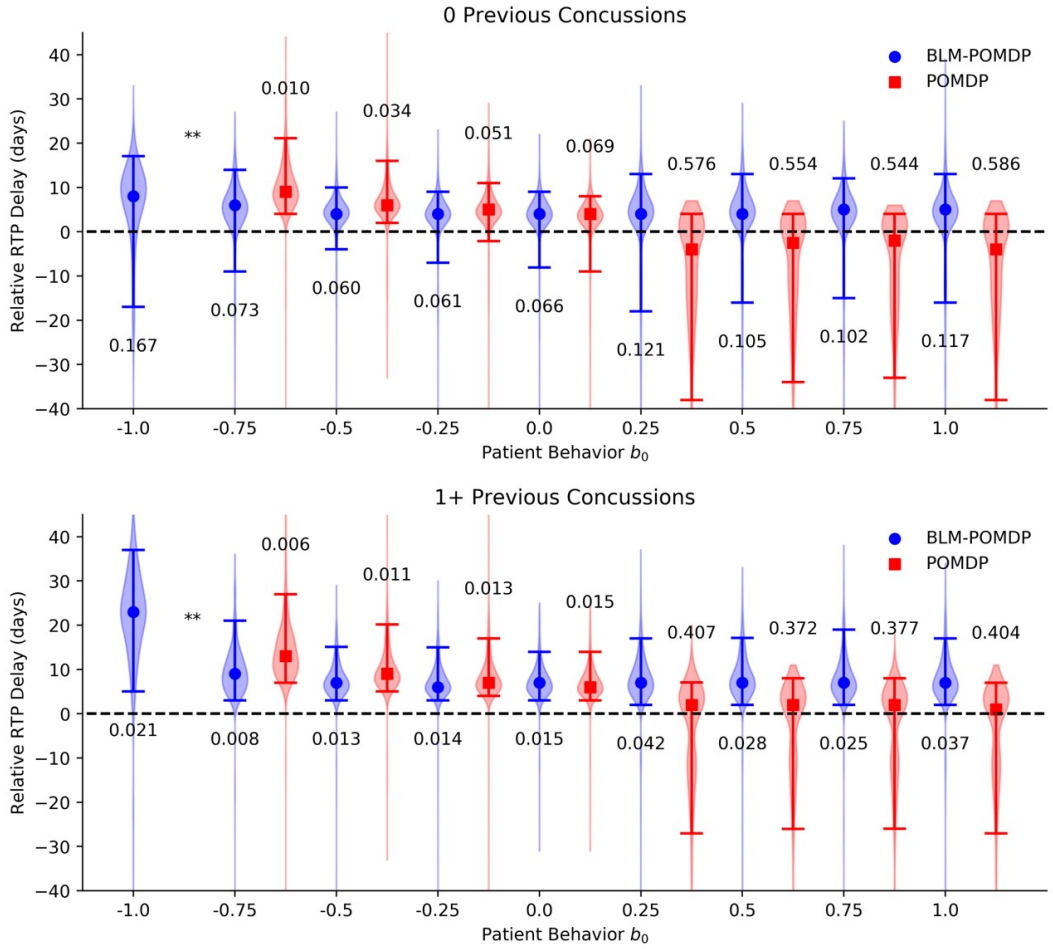


Figure 7.7: Violin plots illustrating distribution of relative RTP delay. Median RTP delay is shown by a circle for BLM-POMDPs and a square for POMDPs in each violin plot, while lower and upper bars indicate 5^{th} and 95^{th} percentiles, respectively. Probability of Premature RTP is shown below BLM-POMDP violin plots and above POMDP violin plots. ** No athletes RTP for POMDP when $b_0 = -1$. RTP, return-to-play.

7.6.8 Estimating the Value of Incorporating Patient Behavior

We now define and estimate the VoIPB. Let $v^p_{b,\theta}$ denote the expected total discounted health utilities for policy $p \in \{\text{BLM-POMDP, POMDP, Myopic, CurrPrac}\}$ and athlete characterized by (b, θ). We define the VoIPB according to

$$\mathbb{V}(\varphi, \theta, p') := \sum_{b \in \mathcal{B}} \varphi(b) \Big(v^{\text{BLM-POMDP}}_{b,\theta} - v^{p'}_{b,\theta} \Big),$$

for each policy $p' \in \{\text{POMDP}, \text{Myopic}, \text{CurrPrac}\}$ where $\varphi \in \Phi$ is a distribution over behavior types. Simply stated, the VoIPB quantifies the benefit in applying the BLM-POMDP over policy p' which does not incorporate patient behavior. Since φ is difficult to estimate, we estimate lower and upper bounds on the VoIPB by

$$\begin{aligned} \mathbb{V}_{LB}(\beta_0, \beta_1, \theta, p') &:= \min_{\varphi \in \Phi'(\beta_0, \beta_1)} \mathbb{V}(\varphi, \theta, p') \\ \mathbb{V}_{UB}(\beta_0, \beta_1, \theta, p') &:= \max_{\varphi \in \Phi'(\beta_0, \beta_1)} \mathbb{V}(\varphi, \theta, p'), \end{aligned}$$

respectively, where the set $\Phi'(\beta_0, \beta_1)$ is given by

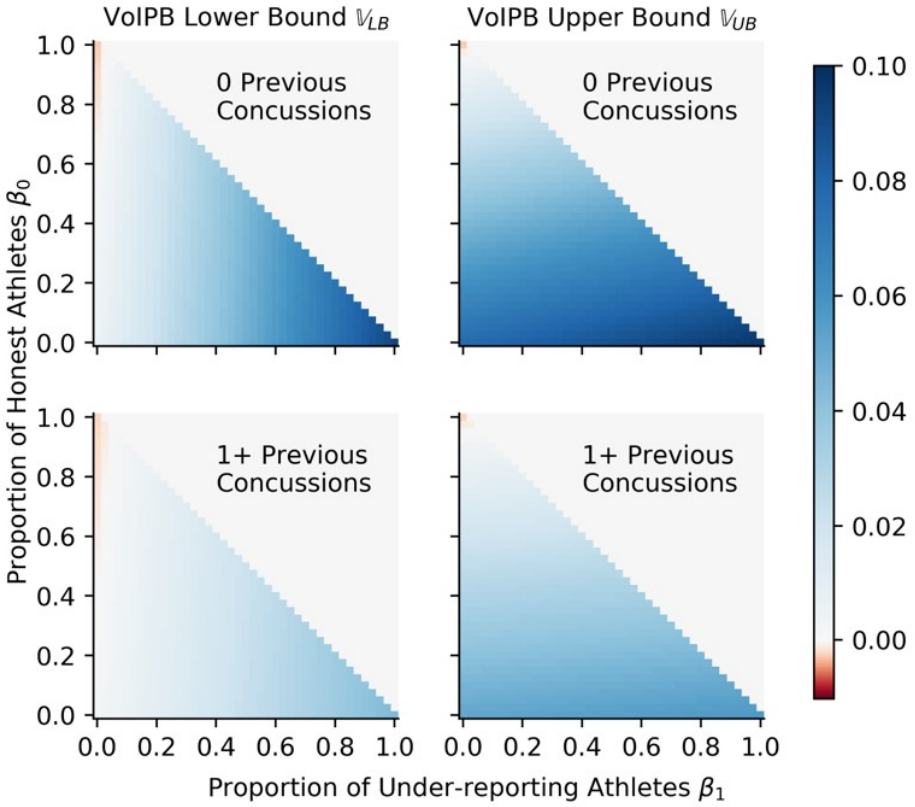
$$\Phi'(\beta_0, \beta_1) := \left\{ \varphi \in \Phi : \begin{array}{cc} \varphi(0) = \beta_0 & \sum_{b>0} \varphi(b) = \beta_1 \\ \varphi(b+1) \geq \varphi(b) \text{ for all } b > 0 & \varphi(b) \geq \varphi(b-1) \text{ for all } b < 0 \end{array} \right\}. \tag{7.33}$$

In (7.33), β_0 and β_1 denote the proportion of honest and under-reporting athletes, respectively. For example, $\Phi'(0.5, 0.4)$ defines the set of all distributions over symptom-reporting behavior when 50% of athletes are honest, 40% under-report symptoms, and 10% over-report symptoms. Finally, notice that $\Phi'(\beta_0, \beta_1)$ contains only distributions with monotone tails, i.e., the proportion of extremely strategic athletes is no greater than the proportion who subtly over- or under-report symptoms.

Bounds on the VoIPB over the POMDP are shown in Figure 7.8. Incorporating patient behavior only fails to be beneficial when the proportion of honest patients in the population is high and there are no under-reporting athletes (e.g., $\beta_0 = 0.75$, $\beta_1 = 0$ for $\theta = 0$). However, this setting is unrealistic given that an estimated 50%-60% of athletes under-report symptoms (Conway et al., 2018; Meier et al., 2015). To this end, both $\mathbb{V}_{LB}(\cdot, \text{POMDP})$

and $\mathbb{V}_{UB}(\cdot, \text{POMDP})$ are increasing as the proportion of under-reporting athletes increases, implying that incorporating patient behavior is more beneficial with athletes who under-report (vs. over-report) symptoms. These trends hold over policies $p' \in \{\text{Myopic}, \text{CurrPrac}\}$. Furthermore, $\mathbb{V}_{LB}(\beta_0, \beta_1, \theta, p') > 0$ for all β_0, β_1 implying the benefit in incorporating patient-behavior and tailoring RTP policies over current practice.

Figure 7.8: Lower ($\mathbb{V}_{LB}$) and upper ($\mathbb{V}_{UB}$) bounds on the Value of Incorporating Patient Behavior (VoIPB) over the POMDP.



To contextualize the VoIPB for clinicians, we quantify the Probability of Premature RTP (PoPRTP) and the Total Health-adjusted Athletic Exposures (THAEs). PoPRTP is defined as $\mathbb{P}(a_t = 1, h_t > 0)$. Minimizing PoPRTP is critical given the association between premature RTP and devastating consequences (e.g., second impact syndrome or late-life neurocognitive impairment (Cantu and Gean, 2010; Guskiewicz et al., 2005)). THAEs are calculated by $\mathbb{E}[\sum_{t=1}^{120} r_\theta^\Omega(\omega_t)\mathbb{1}\{\omega_t \neq 3\}]$ for athletes who RTP and 0 for athletes who do not. They measure

Table 7.2: BLM-POMDP reduction in Probability of Premature RTP (PoPRTP) and gain in Total Health-adjusted Athletic Exposures (THAEs) over benchmark policies

	$\theta = 0$				$\theta = 1+$			
	PoPRTP		**THAEs**		**PoPRTP**		**THAEs**	
p'	*Reduction*	*Relative Risk*	*Gain*	*% Improvement*	*Reduction*	*Relative Risk*	*Gain*	*% Improvement*
POMDP	0.21-0.27	3.26-3.94	1.12-3.49	2.45%-7.67%	0.18-0.22	7.4-9.31	0.85-2.91	2.18%-7.52%
Myopic	0.39-0.47	5.2-6.14	1.92-3.91	4.3%-8.75%	0.51-0.58	18.9-24.56	3.01-4.66	8.22%-12.72%
CurrPrac	0.25-0.29	3.66-4.2	1.31-2.92	2.92%-6.5%	0.32-0.35	12.02-15.4	1.5-2.89	3.92%-7.59%

Ranges are computed over distributions $\varphi_{LB}(\beta_0, \beta_1, \theta, p')$ and $\varphi_{UB}(\beta_0, \beta_1, \theta, p')$ with $\beta_0 \in [0.3, 0.5]$ and $\beta_1 \in [0.5, 0.6]$. Relative Risk = $\text{PoPRTP}_{p'}/\text{PoPRTP}_{\text{BLM-POMDP}}$; % Improvement = $100\times(\text{THAEs}_{\text{BLM-POMDP}}/\text{THAEs}_{p'}$ - 1).

an athlete's health-weighted participation in sport after RTP, providing a surrogate measure for safe RTP.

For each $p' \in \{\text{POMDP, Myopic, CurrPrac}\}$, we evaluated PoPRTP and THAEs under distributions $\varphi_{LB}(\beta_0, \beta_1, \theta, p') = \arg\min_{\varphi \in \Phi'(\beta_0, \beta_1)} \mathbb{V}(\varphi, \theta, p')$ and $\varphi_{UB}(\beta_0, \beta_1, \theta, p') = \arg\max_{\varphi \in \Phi'(\beta_0, \beta_1)} \mathbb{V}(\varphi, \theta, p')$. We assume $\beta_0 \in [0.3, 0.5]$ and $\beta_1 \in [0.5, 0.6]$, reflecting the estimated 50%-60% of athletes who under-report symptoms (Conway et al., 2018; Meier et al., 2015). This analysis is summarized in Table 7.2. Overall, incorporating patient behavior can drastically reduce the PoPRTP and increase the THAEs experienced by athletes after RTP. For example, compared to CurrPrac, incorporating patient behavior reduces the PoPRTP by at least 25% and increases post-RTP participation by up to 4.66 THAEs. A reduction in PoPRTP of this scale has tremendous clinical implications, suggesting drastic reductions in the risk of catastrophic short-term injury (e.g., repeat concussion) and long-term neurodegenerative disease.

7.7 Conclusion, Limitations, and Future Directions

The potential for strategically reported PROs in patient-centered care can make health assessments and treatment decisions challenging. In this chapter, we formulated the BLM-POMDP — a novel multi-agent, multi-period, stochastic dynamic programming framework which incorporates uncertainty around the patient's health and PRO-reporting behavior. Despite its formidable state space representation, we leveraged structural characteristics to

develop an approximation to the optimal policy which achieved a relatively small optimality gap. We then applied the BLM-POMDP to optimize the timing of RTP from sports-related concussion by incorporating CARE Consortium data and published literature values to parameterize and validate our model. The BLM-POMDP outperformed several benchmark RTP policies in terms of increased health utilities, decreased PoPRTP, and increased THAEs — especially in the presence of symptom under-reporting. In particular, the BLM-POMDP's improvement over current practice suggests a drastic reduction in short-term risks of catastrophic injury and long-term risks such as neurodegenerative disease.

Our analysis on RTP from concussion reveals valuable insights for clinicians. First, athletes with a greater risk of injury after RTP require a more conservative approach compared to those at lower risk (e.g., athletes with 1+ vs. 0 previous concussions or football vs. track). That is, the degree of certainty required about the athlete's recovery should increase with the athlete's post-RTP risks. Secondly, there is a greater potential benefit in incorporating symptom-reporting behavior for athletes who are under-reporting symptoms. Specifically, the costs associated with premature RTP are far greater than those associated with delayed RTP. Given the high rate of symptom under-reporting described in the literature, clinicians should cautiously interpret symptoms that signal full recovery early in the RTP decision process, although efforts should be made to learn each athlete's symptom-reporting behavior over time and adjust their interpretation of reported symptoms accordingly. Beyond concussion, our main takeaway is that *incorporating patient behavior can be at least as important as optimizing the treatment decision policy.* Our analysis showed that gap in performance between the VoIPB is largely owed to learning and incorporating patient behavior in the interpretation of PROs, rather than adjusting the treatment policy. Hence, clinicians must aim to understand the drivers of strategic behavior and how this behavior manifests in PROs.

This work can be extended in several ways. First, future research can extend the BLM-POMDP to consider larger action spaces which may be more appropriate for other medical decision-making contexts such as hypertension treatment planning (Schell et al., 2019; Zargoush et al., 2018). Second, future research can consider alternative models of patient behavior, e.g., patients who are not myopic. Finally, the BLM-POMDP can be applied readily to a broad class of stopping time problems. For example, within medical decision-making, the prescription of opioids among potentially malingering patients fits the BLM-POMDP framework. Beyond healthcare, the BLM-POMDP can be extended to applications in finance (e.g.,

sequential fraud detection), military (e.g., optimal search with adversarially placed clues), and energy (e.g., sequential changepoint detection with adversarially determined signals).

As medicine continues its shift towards patient-centered care, the importance of incorporating PROs and understanding patient behavior in medical decision-making will continue rise. Our research provides a novel modeling framework and detailed numerical analysis which illustrate the importance of adaptively learning and accounting for a patient's PRO-reporting behavior throughout a long treatment planning process. By understanding patients' objectives and expectations, doctors can better account for individual differences in patient behavior and tailor their treatment decisions accordingly, ultimately improving each patient's health outcomes.

7.A Modeling RTP From Sports-Related Concussion

In this section, we detail the application of the BLM-POMDP to determine the optimal timing of RTP for athletes diagnosed with sports-related concussion. We begin by describing our HMM model for concussion recovery dynamics before RTP in Section 7.A.1 and after RTP in Section 7.A.2. Then, we derive the doctor's and patient's rewards in Section 7.A.3.

7.A.1 Concussion Recovery Dynamics

We illustrate our model of concussion recovery dynamics in Figure 7.A.1. During the RTP process, the athlete's state space is defined as $\mathcal{H} \times \mathcal{B} \times \Theta$, where $\mathcal{H}$ is the set of unobservable pre-RTP core health states, $\mathcal{B}$ is the set of athlete behavior types, and Θ is the set of demographic information. We describe $\mathcal{H}$ and Θ in this section, while $\mathcal{B}$ is described in Section 7.A.3.

Prior to RTP, the patient's unobservable health states are given by $\mathcal{H} = \{0, 1, 2\}$, where $h = 0$ describes the state where the patient has recovered from concussion, $h = 1$ describes the state where the patient's concussion is asymptomatic (i.e., the athlete is still concussed but shows minimal symptom presentation, neurocognitive deficits, or postural control deficits), and $h = 2$ describes the state where the patient's concussion is symptomatic (i.e., the concussion presents with a high symptom load and high degree of neurocognitive and postural control deficits).

In general, the set of demographic information Θ can include information such as age, sex, number of previous concussions, and sport. These demographic states modify state transition probabilities, observation probabilities, and rewards. In our analysis of Section 7.6, we restrict our attention to male football players with 0 or 1+ previous concussions. That is, we take $\Theta = \{0, 1+\}$, where $\theta \in \Theta$ describes the athlete's number of previous concussions, since concussion history is among the most widely studied factors related to injury presentation and recovery dynamics for concussion (Harada et al., 2019; Tsushima et al., 2019). We remark that the model formulation which follows can be applied readily to a broader set of demographic features such as those we have mentioned previously.

We assume that the injury recovers progressively according to a Bakis model. That is, the

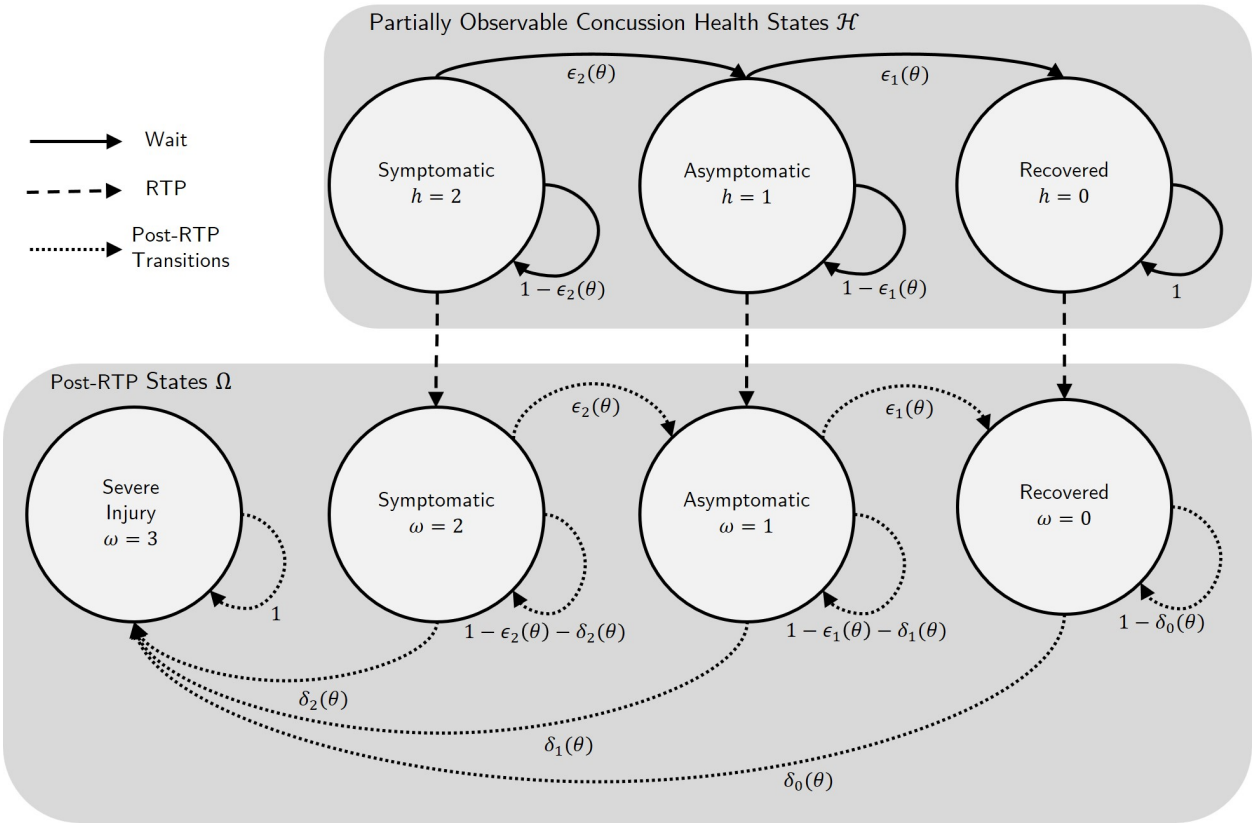


Figure 7.A.1: Illustration of Pre-RTP States, Post-RTP States, and State Transitions.

state transition probabilities between $h \in \mathcal{H}$ take the form

$$P_\theta = \begin{bmatrix} 1 & 0 & 0 \\ \epsilon_1(\theta) & 1-\epsilon_1(\theta) & 0 \\ 0 & \epsilon_2(\theta) & 1-\epsilon_2(\theta) \end{bmatrix}. \tag{7.34}$$

Imposing this structure on P_θ implies that the athlete's cannot directly recover from concussion if it is currently in the symptomatic stage. Furthermore, the athlete's concussion cannot worsen. Instead, it can either improve towards recovery or remain the same.

We model the set of objective observations as $\mathcal{O} = \{0, 1, 2, 3, 4\}$, where an observation of $o = 0, 1, 2, 3$, and 4 corresponds to a total SAC score of 30+, 29, 28, 27, and 0-26, respectively. Lower SAC total scores are correlated with the presence of concussion. Furthermore, these categories represent quintiles of the distribution of total SAC scores in the CARE Consortium Data used to parameterize our models. The set of subjective observations is given by $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$, where an observation of $s = 0, 1, 2, 3$, and 4 corresponds to a total symptom severity score of $0, 1$, 2-4, 5-13, 14-32, and 33+, respectively. Higher symptom severity scores are correlated with the presence of concussion. Since a total symptom severity score 0 represents over 40% of all total symptom severity scores, the categories $1-5$ represent the quintiles of all total symptom severity scores in the CARE Consortium Data not including 0.

7.A.2 Post-RTP Markov Process

Once the doctor permits the athlete to RTP, the athlete enters a post-RTP Markov process in which the doctor can no longer perform actions. We illustrate the post-RTP states and their transitions in Figure 7.A.1. The set of post-RTP states is given by $\Omega = \{0, 1, 2, 3\}$ where $\omega = 0, 1, 2$ is analogous to the pre-RTP health states $h = 0, 1, 2$, respectively, and $\omega = 3$ corresponds to the state in which the athlete has suffered a time-loss injury. We assume that the states $\omega = 3$ results in the athlete being removed from play and is modeled as an absorbing state. For demographic state θ, the state transition probabilities between

$\omega \in \Omega$ are given by

$$P_\theta^\Omega = \begin{bmatrix} 1-\delta_0(\theta) & 0 & 0 & \delta_0(\theta) \\ \epsilon_1(\theta) & 1-\epsilon_1(\theta)-\delta_1(\theta) & 0 & \delta_1(\theta) \\ 0 & \epsilon_2(\theta) & 1-\epsilon_2(\theta)-\delta_2(\theta) & \delta_2(\theta) \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{7.35}$$

7.A.3 Rewards

Since health-related quality of life (HRQoL) have been suggested as important primary outcomes for managing elite athletes with sports-related injuries (Parsons and Snyder, 2011), we model the doctor's and athlete's reward functions using health utilities derived from physical and mental HRQoL measures. Health utilities take on values in $[0, 1]$, where a value of 0 represents death and a value of 1 represents perfect health.

Doctor's Reward Function

The doctor's reward function incorporates physical and mental HRQoL for the patient at each health state along with potential gains and losses in HRQoL from increased exercise and potential injury risks associated with RTP. Specifically, we set the doctor's reward function as

$$r_\theta^d(h, a) = \begin{cases} \zeta\Big(\text{PCS}(h, \theta), \text{MCS}(h, \theta)\Big) & a = 0 \\ W_\theta(h) & a = 1 \end{cases}, \tag{7.36}$$

where $\zeta(\cdot)$ is a function which maps physical and mental HRQoL scores to health utilities, $\text{PCS}(h, \theta)$ and $\text{MCS}(h, \theta)$ are physical and mental HRQoL composite scores, respectively, when the athlete's health state is h and demographic group is θ, and $W_\theta(h)$ is the athlete's expected T-day total discounted reward when they RTP in health state h. In our numerical analysis we take $T = 120$ which is consistent with other studies on injury after RTP from concussion (Brooks et al., 2016; Cross et al., 2016; Herman et al., 2017)

The lump sump reward $r^d(h, 1) = W_\theta(h)$ is derived from a Markov reward process which combines the post-RTP Markov process from Section 7.A.2 with health utilities associated with all post-RTP states $\omega \in \Omega$. In this Markov reward process, the reward for being in

state ω is given by

$$r_\theta^\Omega(\omega) = \begin{cases} \zeta\Big(\text{PCS}(\omega,\theta) + \text{PCS}^+(\omega,\theta) - \text{PCS}^-(\omega,\theta), \\ \qquad \text{MCS}(\omega,\theta) + \text{MCS}^+(\omega,\theta) - \text{MCS}^-(\omega,\theta)\Big) & \omega = 0,1,2 \\ \zeta\Big(\text{PCS}(3,\theta), \text{MCS}(3,\theta)\Big) & \omega = 3 \end{cases}, \tag{7.37}$$

where $\text{PCS}^+(h,\theta)$ and $\text{MCS}^+(h,\theta)$ (resp., $\text{PCS}^-(h,\theta)$ and $\text{MCS}^-(h,\theta)$) are the post-RTP boost (resp., decline) in physical and mental HRQoL composite scores for an athlete in demographic group θ while in health state h due to increased exercise and activity levels. The post-RTP value function is given by

$$\begin{aligned} W_\theta^t(\omega) &= r_\theta^\Omega(\omega) + \rho \sum_{\omega' \in \Omega} P_\theta^\Omega(\omega'|\omega) W_\theta^{t+1}(\omega') \text{ for } t = 1, ..., T-1 \\ W_\theta^T(\omega) &= r_\theta^\Omega(\omega). \end{aligned} \tag{7.38}$$

We set $r^d(h,1) = W_\theta(h) = W_\theta^0(h)$ for all $h \in \mathcal{H}$.

Athlete's Reward Function and Symptom-Reporting Problem

We model the athlete's reward function similarly to the doctor's in the sense that it also incorporates physical and mental HRQoL measures. The main difference in our construction of r_θ^p is that the athlete's gain (or loss) in physical and mental HRQoL post-RTP differs from the doctor's. Specifically, an athlete with θ previous concussions and behavior type $b \in \mathcal{B}$ has the reward function

$$r_\theta^p(h,a,b) = \begin{cases} r^d(h,0) & a = 0 \\ \zeta\Big(\text{PCS}(\omega,\theta) + (1+b)\text{PCS}^+(\omega,\theta) - \text{PCS}^-(\omega,\theta), \\ \text{MCS}(\omega,\theta) + (1+b)\text{MCS}^+(\omega,\theta) - \text{MCS}^-(\omega,\theta)\Big) & a = 1 \end{cases}, \tag{7.39}$$

for all $h \in \mathcal{H}$, $a \in \mathcal{A}$, and $b \in \mathcal{B}$. Furthermore, the athlete is modeled to assume that the naive doctor's myopic reward function is given by $r_{\theta,M}^d(h,0) = r_\theta^d(h,0)$ and $r_{\theta,M}^d(h,1) = r_\theta^\Omega(h)$ for all $h \in \mathcal{H}$. With this parameterization, the athlete will report symptoms honestly if $b = 0$, under-report symptoms if $b > 0$, and over-report symptoms if $b < 0$.

7.B Derivation of Model Inputs

In this section, we detail the derivation of model inputs for our case study in Section 7.6.2.

7.B.1 Parameterizing the HMM

Care Consortium Data

We parameterized the HMM for concussion recovery dynamics (see Section 7.A.1) using multi-center longitudinal data from the National Collegiate Athletic Association and Department of Defense (NCAA-DoD) Concussion Assessment, Research, and Education (CARE) Consortium (Broglio et al., 2017). This dataset combines concussion assessment data across 29 NCAA universities and military service academies throughout the United States.

All study participants are evaluated at the start of each athletic season using several assessments, including the Standard Assessment for Concussion (SAC) and the Sport Concussion Assessment Tool (SCAT) symptom survey. The SAC is a neuropsychological assessment which includes measures of orientation, immediate memory, concentration, and delayed recall. The scores from each of these measures can be summed to obtain a SAC total score, which ranges from 0-30. Lower SAC total scores indicate worse performance. The SCAT symptom survey is a checklist containing of 22 symptoms, each of which are graded from 0-6 with higher scores indicating greater symptom severity. The severity scores can be summed to obtain a SCAT total symptom severity score.

Throughout the course of the athletic season, participants who are diagnosed with concussion are reassessed using the SAC and the SCAT symptom survey at several timepoints: <6 hours post-injury, 24-48 hours post-injury, at the time that the patient becomes asymptomatic, at the time that the patient is cleared to RTP, and 7 days post-RTP.

Baum-Welch Algorithm

We separated male athletes in the CARE Consortium data by concussion history (i.e., 0 and 1+) to parameterize an HMM for each $\theta \in \Theta = \{0, 1+\}$. Since post-injury assessments were not available for each day after the time of injury, we applied linear interpolation to estimate SAC total scores and total symptom severity scores between observed post-injury assessments. We then divided each subset of data into a training set (80%) for parame-

terizing the HMM and a testing set (20%) for validating the HMM. For each training set, we applied an expectation-maximization algorithm known as the Baum-Welch algorithm (Rabiner, 1989) to determine the parameters P_θ, $Q_\theta^{\mathcal{O}}$, and $Q_\theta^{\mathcal{S}}$ for each $\theta \in \Theta$. Since the Baum-Welch algorithm produces varying parameters depending on its initialization, we performed 100 initializations with $\epsilon_1(\theta)$ and $\epsilon_2(\theta)$ generated from a uniform random sampling and initial values of $Q_\theta^{\mathcal{O}}$ and $Q_\theta^{\mathcal{S}}$ estimated using the testing data for every $\theta \in \Theta$. We retained the 10 sets of HMM parameters which produced the greatest log-likelihood values based on the training set. At the recommendation of our clinical collaborators, we chose the set of parameters $(P_\theta, Q_\theta^{\mathcal{O}}, Q_\theta^{\mathcal{S}})$ corresponding to the minimum value of $\epsilon_1(\theta)$ among these 10 candidate parameter sets since it would result in more conservative (i.e., slower) concussion recovery dynamics. The resulting HMM parameters for each $\theta \in \Theta$ are as follows:

$$P_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0.102 & 0.898 & 0 \\ 0 & 0.532 & 0.468 \end{bmatrix} \quad Q_0^{\mathcal{O}} = \begin{bmatrix} 0.220 & 0.229 & 0.183 & 0.156 & 0.212 \\ 0.178 & 0.201 & 0.260 & 0.166 & 0.194 \\ 0.060 & 0.144 & 0.211 & 0.168 & 0.416 \end{bmatrix}$$

$$Q_0^{\mathcal{S}} = \begin{bmatrix} 0.892 & 0.047 & 0.0502 & 0.011 & 0 & 0 \\ 0.401 & 0.222 & 0.261 & 0.100 & 0.015 & 0.001 \\ 0.010 & 0.006 & 0.064 & 0.306 & 0.366 & 0.247 \end{bmatrix}$$

$$P_{1+} = \begin{bmatrix} 1 & 0 & 0 \\ 0.102 & 0.898 & 0 \\ 0 & 0.484 & 0.516 \end{bmatrix} \quad Q_{1+}^{\mathcal{O}} = \begin{bmatrix} 0.240 & 0.222 & 0.217 & 0.145 & 0.176 \\ 0.231 & 0.289 & 0.193 & 0.160 & 0.127 \\ 0.115 & 0.166 & 0.197 & 0.174 & 0.348 \end{bmatrix}$$

$$Q_{1+}^{\mathcal{S}} = \begin{bmatrix} 0.868 & 0.053 & 0.051 & 0.026 & 0.002 & 0 \\ 0.372 & 0.252 & 0.247 & 0.111 & 0.017 & 0.001 \\ 0.011 & 0.014 & 0.092 & 0.347 & 0.367 & 0.170 \end{bmatrix}.$$

Validation

For each $\theta \in \Theta$, we validated the chosen HMM $(P_\theta, Q_\theta^{\mathcal{O}}, Q_\theta^{\mathcal{S}})$ by applying the HMM to the held-out testing data and comparing the distribution of actual SAC total scores and total symptom severity scores to the expected predicted distribution of SAC total scores and total symptom severity scores. Then, we computed the Kullback-Leibler (KL) divergence between

these two distributions, where KL divergence is computed according to

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \right),$$

for a discrete reference distribution P with finite support $\mathcal{X}$ and discrete comparison distribution Q. KL divergence is commonly used for measuring similarity between distributions, including in validation of HMM models (Celeux and Durand, 2008; Wang and Pham, 2011). The value of $D_{KL}(P||Q)$ is bounded below by 0 and is increasing in value as P and Q become more dissimilar. In our validation procedure, we set P as the empirical distribution of SAC total scores or SCAT total symptom severity scores from the testing data. We set Q as the expected predicted observation probabilities when the HMM was applied to testing data. The results of this validation are presented in Figure 7.B.1. Given the small D_{KL} values, we conclude that our HMM models are reasonable approximations for modeling concussion recovery dynamics.

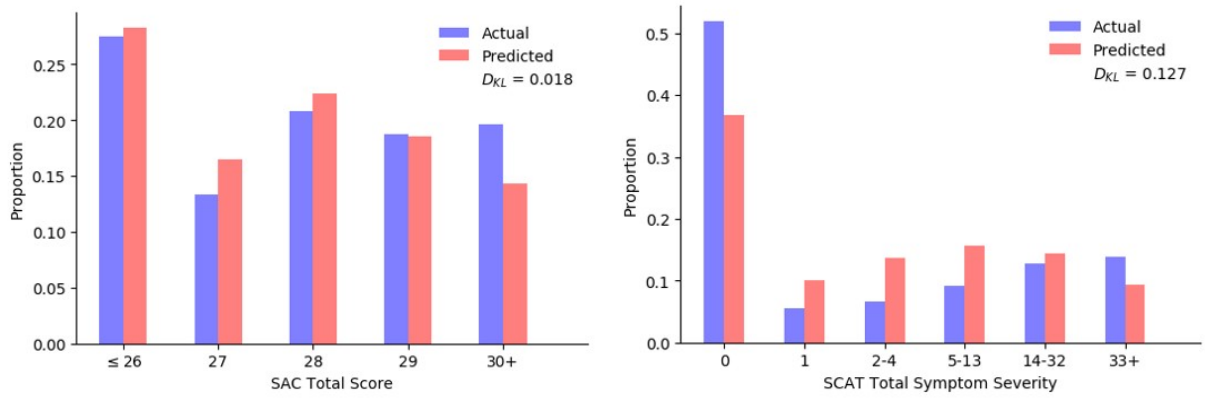
7.B.2 Post-RTP Injury Rates

For each $h \in \mathcal{H}$ and $\theta \in \Theta$, we estimated values of $\delta_h(\theta)$ by applying sex and concussion history-related adjustments to collegiate football injury rates reported in sports medicine literature. Specifically, we estimated

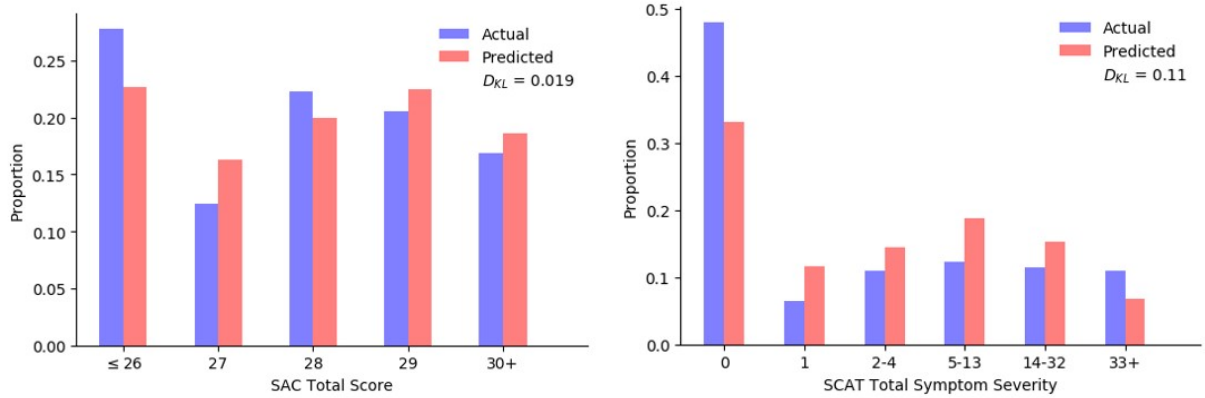
$$\delta_h(\theta) = \frac{\varrho_{\text{inj}}^{h,\theta}}{1 + \varrho_{\text{inj}}^{h,\theta}} \quad \text{and} \quad \varrho_{\text{inj}}^{h,\theta} = \frac{\eta_{\text{inj}}^{\text{sport}}/1000}{1 - \eta_{\text{inj}}^{\text{sport}}/1000} \times \kappa_{\text{sex}} \times \kappa_{\text{history}} \times \kappa_h,$$

where

- $\varrho_{\text{inj}}^{h,\theta}$ are the odds associated with time-loss injuries for athletes with state (h, θ),
- $\eta_{\text{inj}}^{\text{sport}}$ is the sport-specific time-loss injury rate per 1000 athletic exposures,
- κ_{sex} is the sex-specific increased odds in injury rate due to RTP from concussion,
- κ_{history} is the sex-specific increased odds in injury rate due to number of previous concussions, and



(a) 0 previous concussions



(b) 1+ previous concussions

Figure 7.B.1: Distribution of SAC total scores and SCAT total symptom severity scores for testing data and HMM predictions for all $\theta \in \Theta$. SAC = Standard Assessment of Concussion; SCAT = Sport Concussion Assessment Tool; D_{KL} = Kullback-Leibler divergence

- κ_h is the increased odds in injury rate due to athletic activity while in health state h.

We summarize the parameter base values and sources in Table 7.B.1. All values were taken directly from the published literature values except for κ_h, which was estimated as the odds increase in same-season concussion rates from 1999-2001 to 2014-2017, where we assume that more athletes were permitted to RTP prematurely in 1999-2001 since their time to RTP was much shorter than it is now (McCrea et al., 2020).

Table 7.B.1: Injury rate parameters and sources

Parameter	Modifier	Base value	Source
$\eta_{\text{inj}}^{\text{sport}}$	Men's football	7.21	Kerr et al., 2018b
κ_{sex}	Male	2.5	Herman et al., 2017
κ_{history}	Male, 0 previous concussions	1	
	Male, 1+ previous concussions	1.3	Harada et al., 2019
κ_h	$h = 0$	1	
	$h = 1$	1.74	McCrea et al., 2020
	$h = 2$	1.74	McCrea et al., 2020

7.B.3 Rewards

Concussion history-specific values of $\text{PCS}(h, \theta)$ and $\text{MCS}(h, \theta)$ for male athletes were generously provided by Dr. Michelle Weber upon request. These PCS-12 and MCS-12 values are derived from the data in Weber et al., 2019. From McAllister et al., 2001, we set $\text{PCS}^+(h, \theta) = 1.9$ and $\text{MCS}^+(h, \theta) = 2.7$ for all $h \in \mathcal{H}$ and $\theta \in \Theta$. Additionally, we set

$$\text{PCS}^-(h, \theta) = \begin{cases} 0 & h = 0 \\ 4.2 & h = 1 \\ 8.3 & h = 2 \end{cases} \quad \text{and} \quad \text{MCS}^-(h, \theta) = \begin{cases} 0 & h = 0, 1 \\ 3.3 & h = 2 \end{cases} \quad \text{for all } \theta \in \Theta.$$

Finally, we set $\text{PCS}(3, \theta) = 49$ and $\text{MCS}(3, \theta) = 51$ for all $\theta \in \Theta$ (McAllister et al., 2001). A previously published and validated regression equation was used to transform these SF-12 values into health utility values (Lawrence and Fleishman, 2004). The health utility values for the doctor's and athlete's reward functions are presented in Table 7.B.2 and Table 7.B.3, respectively. We parameterized the naive doctor's reward function as $r_{M,\theta}^d(h, a) = r_\theta^p(h, a, 0)$ for all $h \in \mathcal{H}$ and $a \in \mathcal{A}$.

Table 7.B.2: Health utility values for the doctor's reward function for each $\theta \in \Theta$

θ	h	$a=0$	$a=1$
0	0	0.902	18.343
0	1	0.913	17.922
0	2	0.868	17.622
1+	0	0.920	18.444
1+	1	0.889	17.769
1+	2	0.841	17.397

Table 7.B.3: Health utility values for the patient's reward function for each $\theta \in \Theta$ and $b \in \mathcal{B}$.

θ	h	$a=0$	$a=1$								
			$b=-1.0$	$b=-0.75$	$b=-0.5$	$b=-0.25$	$b=0.0$	$b=0.25$	$b=0.5$	$b=0.75$	$b=1.0$
0	0	0.902	0.902	0.915	0.928	0.942	0.955	0.968	0.981	0.995	1.008
0	1	0.913	0.854	0.867	0.880	0.894	0.907	0.920	0.933	0.947	0.960
0	2	0.868	0.719	0.732	0.746	0.759	0.772	0.785	0.799	0.812	0.825
1+	0	0.920	0.920	0.933	0.946	0.960	0.973	0.986	0.999	1.012	1.026
1+	1	0.889	0.830	0.843	0.856	0.870	0.883	0.896	0.909	0.922	0.936
1+	2	0.841	0.692	0.705	0.718	0.731	0.745	0.758	0.771	0.784	0.798

7.C Importance of Initial Beliefs for BLM-POMDP

Here, we estimate the performance of the BLM-POMDP as a function of the initial behavior belief ϕ_0. Specifically, we evaluated the BLM-POMDP under different initial behavior belief states given by

$$\phi_0^1 = \begin{bmatrix} 0.025 & 0.025 & 0.05 & 0.15 & 0.5 & 0.15 & 0.05 & 0.025 & 0.025 \end{bmatrix}^\top$$
$$\phi_0^2 = \begin{bmatrix} 0.025 & 0.02 & 0.025 & 0.05 & 0.05 & 0.075 & 0.10 & 0.15 & 0.5 \end{bmatrix}^\top$$
$$\phi_0^3 = \begin{bmatrix} 0.5 & 0.15 & 0.1 & 0.075 & 0.05 & 0.05 & 0.025 & 0.025 & 0.025 \end{bmatrix}^\top.$$

The initial values of ϕ_0^1, ϕ_0^2, and ϕ_0^3 correspond to initial assumptions of honest, under-reporting, and over-reporting behavior, respectively. These results are presented in Figure 7.C.1. We expected that the BLM-POMDP would achieve greater performance when the initial belief ϕ_0 is "closer" to the actual patient behavior. However, in almost every case, the differences in performance are negligibly small. These results may be attributed to the fact

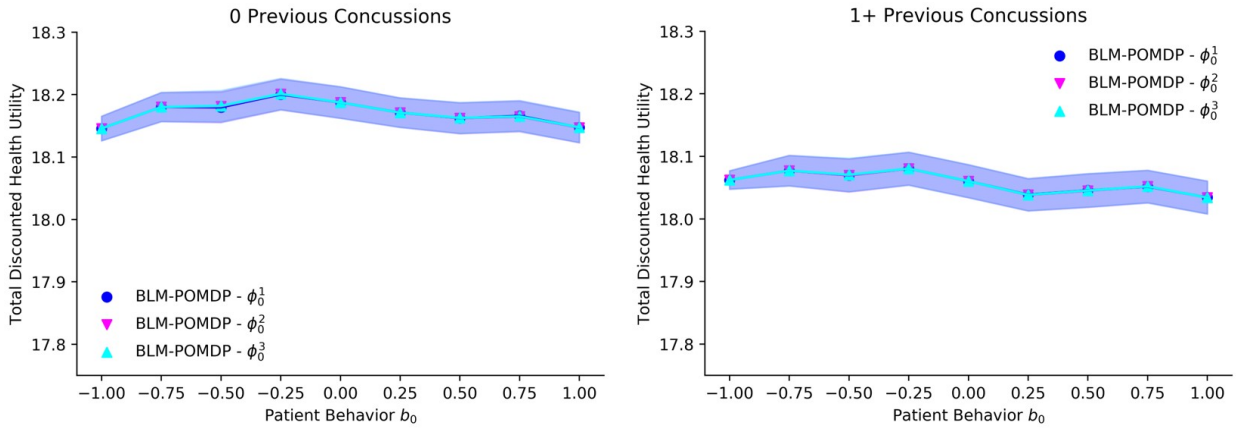


Figure 7.C.1: Total discounted health utilities for BLM-POMDP with different initial behavior beliefs. ϕ_0^1, ϕ_0^2, and ϕ_0^3 correspond to initial beliefs of honest, under-reporting, and over-reporting, respectively. Markers indicate mean values and shaded areas indicate 95% confidence interval of the mean.

that (1) the BLM-POMDP has enough time to learn the patient's behavior and (2) when athletes under-report or over-report symptoms, they rule out the opposite behavior type. We conclude that while it is generally important to "match up" initial behavior beliefs, it is more important to dynamically update one's belief about patient behavior.

Chapter 8

Conclusion and Future Work

Motivated by high-impact problems in the management of sports-related concussion, this thesis combines data-driven analytics and operations research to develop new methodological frameworks in predictive and prescriptive analytics. The analyses carried out within this dissertation reveal both practical insights on the nature of concussion management and technical insights regarding the nature of our modeling approaches.

This dissertation is divided into three parts, each of which coincides with a key aspect of the concussion management protocol. Chapters 2 and 3 comprise the first part of this dissertation, in which we focus on concussion assessment. We developed predictive models which quantify the value of multi-dimensional approaches to concussion assessment, estimate the marginal utility in incorporating athlete-specific baseline information, and identify the most important components of the SCAT — a concussion assessment battery which is commonly used throughout the world. The second part of this dissertation, containing Chapters 4-6, focuses on concussion diagnosis. In these chapters, we formulated the TTP, a novel data-driven stochastic optimization approach to determine diagnostic decision thresholds, and analyze the value of this approach for acute concussion diagnosis by creating a certainty-based concussion diagnosis framework (i.e., Unlikely, Possible, Probable, and Definite concussion). Our analysis provides an important characterization of athletes for whom concussion diagnosis decisions are easily made (i.e., Unlikely or Definite concussion) and those whose diagnoses are not straightforward (i.e., Possible or Probable concussion). The final part of this dissertation, Chapter 7, addresses the optimal timing of RTP from sports-related concussion. In this chapter, we formulated and characterized the BLM-POMDP — multi-agent, multi-period,

stochastic dynamic programming framework which incorporates the patient's and doctor's perspectives while accounting for uncertainty in the patient's health and symptom-reporting behavior. Our application of the BLM-POMDP to RTP from concussion quantifies the benefit (i.e., increased health-related quality of life, decreased likelihood of premature RTP, and increased athletic exposures after RTP) in accounting for symptom-reporting behavior in RTP decisions, while also demonstrating the utility of tailored RTP policies over existing RTP approaches.

The research in this dissertation provides a starting point for data-driven analytics in concussion management. In the following sections, we briefly discuss how future work can extend our analysis and modeling frameworks within concussion and beyond concussion.

8.1 An Integrated Approach to Personalized Concussion Management

In this dissertation, we separately studied concussion diagnosis decisions and RTP decisions. Yet, these components of the concussion management protocol are intimately linked; the "costs" associated with misdiagnoses may be determined by considering outcomes owing to the RTP process such as health-related quality of life, likelihood of premature RTP, or time to RTP. Likewise, the certainty revolving around an athlete's diagnosis might dictate the approach taken in their graded RTP protocol — ultimately affecting their recovery trajectory and RTP criteria. The dependence between all concussion management decisions from the point of diagnosis until the RTP decision suggests a stochastic programming formulation given by

$$\begin{aligned}
\max_{u,l} \quad & \sum_{h\in\mathcal{H}} \Bigg(\mathbb{E}[V_\theta^{RTP}|h_0=h]\mathbb{P}(h_0=h|f(\theta)\leq l) + \mathbb{E}[V_\theta^1(f(\theta))|h_0=h]\mathbb{P}(h_0=h|l<f(\theta)<u) \\
& \qquad + \mathbb{E}[V_\theta^2(f(\theta))|h_0=h]\mathbb{P}(h_0=h|f(\theta)\geq u) \Bigg) \\
\text{s.t.} \quad & \mathbb{E}_\theta[fp(u,f(\theta))] \leq \gamma^{fp} \\
& \mathbb{E}_\theta[fn(l,f(\theta))] \leq \gamma^{fn} \\
& 0 \leq l \leq u \leq 1,
\end{aligned}$$

where $f : \Theta \to [0, 1]$ maps athlete characteristics $\theta \in \Theta$ to a risk score, V_θ^{RTP} describes the expected health outcomes (e.g., health utilities) for an athlete who RTPs, $V_\theta^1(f(\theta))$ describes the expected health outcomes for an athlete classified as Possible/Probable concussion and who enters the injury management protocol with an initial risk score of $f(\theta)$, and $V_\theta^2(f(\theta))$ describes the health outcomes for an athlete classified as Definite concussion and who enters the concussion management protocol with an initial risk score of $f(\theta)$. Here, the functions V_θ^{RTP}, $V_\theta^1(f(\theta))$, and $V_\theta^2(f(\theta))$ may be derived from MRPs and POMDPs, making the solution of this problem a challenging task. Given the intractability of this problem, determining the optimal decision thresholds (u, l) as well as the optimal post-injury protocols associated with V_θ^1 and V_θ^2 can require advances in multi-stage stochastic programming, stochastic control, and simulation optimization. Incorporating symptom-reporting behavior and/or a data-driven robust optimization formulation to this problem (e.g., see Section 4.4.2) would also provide interesting and non-trivial extensions. Key questions which can be addressed by this research include the following.

1. How do diagnosis decisions impact downstream post-injury management decisions and vice versa?

2. How does the quality of the risk prediction model f affect decisions throughout the protocol?

3. How do athlete-specific factors such as age, sex, sport, concussion history, baseline testing performance, symptom-reporting behavior, etc. play a role in RTP and post-injury management decisions?

4. Which athletes most benefit from a personalized approach?

5. How does this integrated approach compare to existing practice? What are the advantages and disadvantages?

6. Are there simple heuristics which perform relatively well (compared to the optimal integrated approach) and are much more interpretable?

8.2 Data-driven Decision-making in the Management of Military Concussions

Military settings are often regarded as another area in which improving concussion management can have substantial impact on the health outcomes of personnel (O'Connor, 2019). Although the management of military concussion can build on many findings related to sports-related concussions (Lew et al., 2007), there are challenges specific to military concussion which warrant special attention. One specific challenge is in the diagnosis of concussion within military settings. In sports, concussion diagnosis can be facilitated by a stoppage of play after a potential concussive event. However, among military personnel in combat environments, such stoppages are not possible given the long, continuous nature of combat missions (Chapman and Diaz-Arrastia, 2014). While surveillance technologies can help to identify potential concussive events (e.g., blast-related trauma), there is a heavy reliance on self-reporting to identify possible concussion. As with sports-related concussion, however, under-reporting concussion symptoms is pervasive and the reasons owing to this behavior are complex and heterogeneous (Rawlins et al., 2019). It is also well-known that the presence of comorbidities (e.g., physical injuries and emotional stress) make it difficult to determine whether concussion-like symptoms are caused by a concussion or a different cause altogether (Rigg and Mooney, 2011). For these reasons, future research can focus on developing models which address the following questions.

1. What person-specific risk factors are associated with concussion? How should these risk factors dictate concussion management decisions?

2. How can post-deployment screening be designed to facilitate the identification of concussion and incentivize truthful symptom under-reporting?

3. How can surveillance technologies help to identify which military personnel have likely sustained a concussion during a combat mission?

4. How can comorbid conditions be ruled out or utilized to help identify the presence of concussion?

Future research can also consider concussion management decisions beyond diagnosis. For example, which personnel should be permitted to return to combat in consideration of risks

associated with doing so? Additionally, how can veterans with potentially undiagnosed traumatic brain injury be identified and provided the care they need? While answering these research questions can result in a profound impact for those impacted by military concussions, this research can also advance the application and theory of predictive modeling and data-driven stochastic optimization.

8.3 Design and Analysis of Shared Decision-Making Frameworks

As mentioned in Chapter 7, healthcare has been transitioning towards patient-centered care. Shared decision-making comprises an important aspect of patient-centered care. Specifically, shared decision-making is defined as "an approach where clinicians and patients share the best available evidence when faced with the task of making decisions, and where patients are supported to consider options, to achieve informed preferences" (Elwyn et al., 2012). In contrast to the decision process modeled in Chapter 7, shared decision-making involves patients and doctors making decisions jointly based on the patient's preferences. Key elements of this decision-making framework include the presentation of treatment alternatives and deliberation over preferences and options, whereby all information exchanges between the doctor and patient hinge on rapport and trust between the two. Hence, important research questions on the design and analysis of shared decision-making models include the following.

1. What role does information transmission and transparency play in shared decision-making framework? For example, how does the order of information revelation affect the patient's choices and decisions?

2. How should shared decisions in previous meetings play a role in future shared decisions when treatment planning occurs over several periods?

3. How should evolving data regarding treatment risk, patient-specific disease risk, patient-specific disease evolution, predicted future health outcomes, etc. be incorporated in shared decision-making processes?

4. How do the doctor's objectives and patient's objectives interact throughout this decision process?

Given the need to model both the patient and doctor, sequential game theoretic frameworks may be appropriate for the design and analysis of shared decision-making. For example, cheap talk frameworks can be useful given the costless information exchanges between the two parties (Crawford and Sobel, 1982). However, cheap talk models in sequential decision-making settings are notoriously difficult to solve in the presence of information asymmetry and partially observable information. Furthermore, refining and selecting cheap talk equilibriums can be burdensome (Chen, Kartik, and Sobel, 2008). Hence, this research can potentially advance the theory of sequential cheap talk and multi-agent stochastic control while also making a valuable contribution to data-driven decision-making in the context of shared decision-making.

8.4 Conclusion

This dissertation develops new methodologies in predictive and prescriptive analytics to determine

1. how large clinical datasets can inform our understanding of personalized disease progression/risk over time, and

2. how this understanding of disease progression/risk can be translated into improved medical decision-making which account for multiple stakeholders' perspectives.

Our findings highlight important insights for the assessment of acute concussion, as well as making data-driven diagnosis and RTP decisions. This research also advances the theory of data-driven analytics and operations research by formulating and characterizing the TTP and BLM-POMDP frameworks. As the Big Data revolution continues to grow, there will continue to be new research questions at the intersection of data analytics, operations research, and healthcare — including the three extensions proposed in this section. While much work remains to be done, it is our hope that this dissertation strengthens the foundation of research in these fields and that future research strives to build on these methods to address new healthcare challenges in the years to come.

Bibliography

Ahmed, S. and A. Shapiro (2008). "Solving Chance-Constrained Stochastic Programs via Sampling and Integer Programming". *State-of-the-Art Decision-Making Tools in the Information-Intensive Age*. June 2015. INFORMS, pp. 261–269.

Ahsen, M. E., M. U. S. Ayvaci, and S. Raghunathan (2019). "When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis". *Information Systems Research* 30.1, pp. 97–116.

Alagoz, O., T. Ayer, and F. S. Erenay (2011). "Operations Research Models for Cancer Screening". *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Amato, C., G. Chowdhary, A. Geramifard, N. K. Üre, and M. J. Kochenderfer (2013). "Decentralized control of partially observable Markov decision processes". *Proceedings of the IEEE Conference on Decision and Control*, pp. 2398–2405.

American Diabetes Society (2016). "2. Classification and Diagnosis of Diabetes". *Diabetes Care* 39.Supplement 1, S13–S22.

Andelic, N., N. Hammergren, E. Bautz-Holter, U. Sveen, C. Brunborg, and C. Røe (2009). "Functional outcome and health-related quality of life 10 years after moderate-to-severe traumatic brain injury". *Acta Neurologica Scandinavica* 120.1, pp. 16–23.

Andersen, E. B. (1977). "Sufficient statistics and latent trait models". *Psychometrika* 42.1, pp. 69–81. arXiv: `arXiv:1011.1669v3`.

Anderson, K. M. (1991). "A nonproportional hazards Weibull accelerated failure time regression model." *Biometrics* 47.1, pp. 281–8.

Anzalone, A. J., D. Blueitt, T. Case, T. McGuffin, K. Pollard, J. C. Garrison, M. T. Jones, R. Pavur, S. Turner, and J. M. Oliver (2017). "A Positive Vestibular/Ocular Motor Screening (VOMS) Is Associated With Increased Recovery Time After Sports-Related Concussion in Youth and Adolescent Athletes". *The American Journal of Sports Medicine* 45.2, pp. 474–479.

Asken, B. M., R. M. Bauer, K. M. Guskiewicz, M. A. McCrea, J. D. Schmidt, C. C. Giza, A. R. Snyder, Z. M. Houck, A. P. Kontos, T. W. McAllister, S. P. Broglio, J. R. Clugston, S. Anderson, J. Bazarian, A. Brooks, T. Buckley, S. Chrisman, M. Collins, J. DiFiori, S. Duma, B. Dykhuizen, J. T. Eckner, L. Feigenbaum, A. Hoy, L. Kelly, T. D. Langford, L. Lintner, G. McGinty, J. Mihalik, C. Miles, J. Ortega, N. Port, M. Putukian, S. Row-

son, and S. Svoboda (2018). "Immediate Removal From Activity After Sport-Related Concussion Is Associated With Shorter Clinical Recovery and Less Severe Symptoms in Collegiate Student-Athletes". *The American Journal of Sports Medicine* 46.6, pp. 1465–1474.

Asken, B. M., M. A. McCrea, J. R. Clugston, A. R. Snyder, Z. M. Houck, and R. M. Bauer (2016). ""Playing Through It": Delayed Reporting and Removal From Athletic Activity After Concussion Predicts Prolonged Recovery." *Journal of athletic training* 51.4, pp. 329–335.

Aswani, A., P. Kaminsky, Y. Mintz, E. Flowers, and Y. Fukuoka (2018). "Behavioral modeling in weight loss interventions". *European Journal of Operational Research* 0, pp. 1–15.

Ayer, T., O. Alagoz, and N. K. Stout (2012). "OR Forum—A POMDP Approach to Personalize Mammography Screening Decisions". *Operations Research* 60.5, pp. 1019–1034.

Ayer, T., O. Alagoz, N. K. Stout, and E. S. Burnside (2016). "Heterogeneity in Women's Adherence and Its Role in Optimal Breast Cancer Screening Policies". *Management Science* 62.5, pp. 1339–1362.

Ayvaci, M. U. S., O. Alagoz, and E. S. Burnside (2012). "The Effect of Budgetary Restrictions on Breast Cancer Diagnostic Decisions". *Manufacturing & Service Operations Management* 14.4, pp. 600–617.

Ayvaci, M. U. S., M. E. Ahsen, S. Raghunathan, and Z. Gharibi (2017). "Timing the Use of Breast Cancer Risk Information in Biopsy Decision-Making". *Production and Operations Management* 26.7, pp. 1333–1358.

Barnett, C. L., S. A. Tomlins, D. J. Underwood, J. T. Wei, T. M. Morgan, J. E. Montie, and B. T. Denton (2017). "Two-Stage Biomarker Protocols for Improving the Precision of Early Detection of Prostate Cancer". *Medical Decision Making* 37.7, pp. 815–826.

Barr, W. B. and M. A. McCrea (2001). "Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion". *Journal of the International Neuropsychological Society* 7.6, S1355617701766052.

Barsky, A. J. (2002). "Forgetting, Fabricating, and Telescoping". *Archives of Internal Medicine* 162.9, p. 981.

Bass, C. and P. Halligan (2014). "Factitious disorders and malingering: Challenges for clinical assessment and management". *The Lancet* 383.9926, pp. 1422–1432.

Bates, D. W., S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar (2014). "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients". *Health Affairs* 33.7, pp. 1123–1131.

Baugh, C. M. and E. Kroshus (2016). "Concussion management in US college football: progress and pitfalls". *Concussion* 1.1, cnc.15.6.

Bayati, M., S. Bhaskar, and A. Montanari (2018). "Statistical analysis of a low cost method for multiple disease prediction". *Statistical Methods in Medical Research* 27.8, pp. 2312–2328.

Bell, D. R., K. M. Guskiewicz, M. a. Clark, and D. a. Padua (2011). "Systematic Review of the Balance Error Scoring System". *Sports Health* 3.3, pp. 287–295.

Bertsimas, D. and A. King (2016). "OR Forum—An Algorithmic Approach to Linear Regression". *Operations Research* 64.1, pp. 2–16.

Bertsimas, D., A. King, and R. Mazumder (2016). "Best subset selection via a modern optimization lens". *The Annals of Statistics* 44.2, pp. 813–852. arXiv: `arXiv:1507.03133v1`.

Bertsimas, D., J. Silberholz, and T. Trikalinos (2018). "Optimal healthcare decision making under multiple mathematical models: application in prostate cancer screening". *Health Care Management Science* 21.1, pp. 105–118.

Blackwell, D. (1953). "Equivalent Comparisons of Experiments". *The Annals of Mathematical Statistics* 24.2, pp. 265–272.

Bozdogan, H. (1987). "Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions". *Psychometrika* 52.3, pp. 345–370.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Broglio, S. P., R. C. Cantu, G. A. Gioia, K. M. Guskiewicz, J. Kutcher, M. Palm, and T. C. V. McLeod (2014). "National Athletic Trainers' Association Position Statement: Management of Sport Concussion". *Journal of Athletic Training* 49.2, pp. 245–265.

Broglio, S. P., M. S. Ferrara, K. Sopiarz, and M. S. Kelly (2008). "Reliable Change of the Sensory Organization Test". *Clinical Journal of Sport Medicine* 18.2, pp. 148–154.

Broglio, S. P., J. Harezlak, B. Katz, S. Zhao, T. McAllister, M. McCrea, J. Hazzard, L. Kelly, D. Campbell, J. Jackson, G. McGinty, P. O'Donnell, K. Cameron, A. Susmarski, J. Goldman, C. Giza, T. Buckley, T. Kaminski, J. Clugston, J. Schmidt, L. Feigenbaum, J. T. Eckner, S. Anderson, C. Master, A. Kontos, S. Chrisman, and A. Brooks (2019). "Acute Sport Concussion Assessment Optimization: A Prospective Assessment from the CARE Consortium". *Sports Medicine* 49.12, pp. 1977–1987.

Broglio, S. P., S. N. Macciocchi, and M. S. Ferrara (2007). "Sensitivity of the concussion assessment battery". *Neurosurgery* 60.6, pp. 1050–1057.

Broglio, S. P., S. N. Macciocchi, and M. S. Ferrara (2009). "Neurocognitive performance of concussed athletes when symptom free." *Journal of athletic training* 42.4, pp. 504–8.

Broglio, S. P., M. McCrea, T. McAllister, J. Harezlak, B. Katz, D. Hack, and B. Hainline (2017). "A National Study on the Effects of Concussion in Collegiate Athletes and US Military Service Academy Members: The NCAA–DoD Concussion Assessment, Research and Education (CARE) Consortium Structure and Methods". *Sports Medicine* 47.7, pp. 1437–1451.

Brooks, M. A., K. Peterson, K. Biese, J. Sanfilippo, B. C. Heiderscheit, and D. R. Bell (2016). "Concussion Increases Odds of Sustaining a Lower Extremity Musculoskeletal Injury After Return to Play Among Collegiate Athletes". *The American Journal of Sports Medicine* 44.3, pp. 742–747.

Broshek, D. K., T. Kaushik, J. R. Freeman, D. Erlanger, F. Webbe, and J. T. Barth (2005). "Sex differences in outcome following sports-related concussion". *Journal of Neurosurgery* 102.5, pp. 856–863.

Bruce, J. M. and R. J. Echemendia (2004). "Concussion history predicts self-reported symptoms before and following a concussive event". *Neurology* 63.8, pp. 1516–1518.

Bruce, J. M., R. J. Echemendia, W. Meeuwisse, M. G. Hutchison, M. Aubry, and P. Comper (2017). "Development of a risk prediction model among professional hockey players with visible signs of concussion". *British Journal of Sports Medicine*, bjsports–2016–097091.

Buckley, T. A., B. A. Munkasy, and B. P. Clouse (2017). "Sensitivity and Specificity of the Modified Balance Error Scoring System in Concussed Collegiate Student Athletes". *Clinical Journal of Sport Medicine*, p. 1.

Cantu, R. C. and A. D. Gean (2010). "Second-Impact Syndrome and a Small Subdural Hematoma: An Uncommon Catastrophic Result of Repetitive Head Injury with a Characteristic Imaging Appearance". *Journal of Neurotrauma* 27.9, pp. 1557–1564.

Carney, N., J. Ghajar, A. Jagoda, S. Bedrick, C. Davis-O'Reilly, H. Du Coudray, D. Hack, N. Helfand, A. Huddleston, T. Nettleton, and S. Riggio (2014). "Concussion guidelines step 1: Systematic review of prevalent indicators". *Neurosurgery* 75.SUPPL. 1, pp. 3–15.

CDC (2016). *Centers for Disease Control and Prevention - Injury Prevention & Control: Traumatic Brain Injury & Concussion*.

Celeux, G. and J.-B. Durand (2008). "Selecting hidden Markov model state number with cross-validated likelihood". *Computational Statistics* 23.4, pp. 541–564.

Cevik, M., T. Ayer, O. Alagoz, and B. L. Sprague (2018). "Analysis of Mammography Screening Policies under Resource Constraints". *Production and Operations Management* 27.5, pp. 949–972.

Chang, Y., A. L. Erera, and C. C. White (2015). "A leader–follower partially observed, multiobjective Markov game". *Annals of Operations Research* 235.1, pp. 103–128.

Chapman, G. B. and A. S. Elstein (1995). "Valuing the Future". *Medical Decision Making* 15.4, pp. 373–386.

Chapman, J. C. and R. Diaz-Arrastia (2014). "Military traumatic brain injury: A review". *Alzheimer's & Dementia* 10.3, S97–S104.

Chen, Y., N. Kartik, and J. Sobel (2008). "Selecting Cheap-Talk Equilibria". *Econometrica* 76.1, pp. 117–136.

Chin, E. Y., L. D. Nelson, W. B. Barr, P. McCrory, and M. A. McCrea (2016). "Reliability and validity of the sport concussion assessment tool-3 (SCAT3) in high school and collegiate athletes". *American Journal of Sports Medicine* 44.9, pp. 2276–2285.

Cohen, G. and R. Java (1995). "Memory for medical history: Accuracy of recall". *Applied Cognitive Psychology* 9.4, pp. 273–288.

Collins, M. W., A. P. Kontos, D. O. Okonkwo, J. Almquist, J. Bailes, M. Barisa, J. Bazarian, O. J. Bloom, D. L. Brody, R. Cantu, J. Cardenas, J. Clugston, R. Cohen, R. Echemendia, R. Elbin, R. Ellenbogen, J. Fonseca, G. Gioia, K. Guskiewicz, R. Heyer, G. Hotz,

G. L. Iverson, B. Jordan, G. Manley, J. Maroon, T. McAllister, M. McCrea, A. Mucha, E. Pieroth, K. Podell, M. Pombo, T. Shetty, A. Sills, G. Solomon, D. G. Thomas, T. C. Valovich McLeod, T. Yates, and R. Zafonte (2016). "Statements of Agreement From the Targeted Evaluation and Active Management (TEAM) Approaches to Treating Concussion Meeting Held in Pittsburgh, October 15-16, 2015". *Neurosurgery* 79.6, pp. 912–929. arXiv: 15334406.

Collins, M. W., A. P. Kontos, E. Reynolds, C. D. Murawski, and F. H. Fu (2014). "A comprehensive, targeted approach to the clinical care of athletes following sport-related concussion". *Knee Surgery, Sports Traumatology, Arthroscopy* 22.2, pp. 235–246.

Concussion in Sport Group (2013). *Sport Concussion Assessment Tool - 3rd Edition*.

Conway, F. N., M. Domingues, R. Monaco, L. M. Lesnewich, A. E. Ray, B. L. Alderman, S. M. Todaro, and J. F. Buckman (2018). "Concussion Symptom Underreporting Among Incoming National Collegiate Athletic Association Division I College Athletes". *Clinical Journal of Sport Medicine* 0.0, p. 1.

Covassin, T., C Buz Swanik, and M. L. Sachs (2003). "Sex Differences and the Incidence of Concussions Among Collegiate Athletes". *Journal of Athletic Training* 38.3, pp. 238–244.

Covassin, T., R. Elbin, W. Harris, T. Parker, and A. Kontos (2012). "The Role of Age and Sex in Symptoms, Neurocognitive Performance, and Postural Stability in Athletes After Concussion". *The American Journal of Sports Medicine* 40.6, pp. 1303–1312.

Covassin, T., J. L. Savage, A. C. Bretzin, and M. E. Fox (2018). "Sex differences in sport-related concussion long-term outcomes". *International Journal of Psychophysiology* 132.August 2017, pp. 9–13.

Covassin, T., P. Schatz, and C. B. Swanik (2007). "Sex differences in neuropsychological function and post-concussion symptoms of concussed collegiate athletes". *Neurosurgery* 61.2, pp. 345–350.

Covassin, T., D. Stearne, and R. Elbin (2008). "Concussion history and postconcussion neurocognitive performance and symptoms in collegiate athletes". *Journal of Athletic Training* 43.2, pp. 119–124.

Cowee, K. and J. E. Simon (2019). "A history of previous severe injury and health-related quality of life among former collegiate athletes". *Journal of Athletic Training* 54.1, pp. 64–69.

Cowell, F. A. (2000). "Measurement of inequality". *Handbook of Income Distribution*. Vol. 1. Chap. 2, pp. 87–166.

Crawford, V. P. and J. Sobel (1982). "Strategic Information Transmission". *Econometrica* 50.6, p. 1431.

Cross, M., S. Kemp, A. Smith, G. Trewartha, and K. Stokes (2016). "Professional Rugby Union players have a 60% greater risk of time loss injury after concussion: a 2-season prospective study of clinical outcomes". *British Journal of Sports Medicine* 50.15, pp. 926–931.

Daneshvar, D. H., D. O. Riley, C. J. Nowinski, A. C. McKee, R. A. Stern, and R. C. Cantu (2011). "Long-Term Consequences: Effects on Normal Development Profile After Concussion". *Physical Medicine and Rehabilitation Clinics of North America* 22.4, pp. 683–700. arXiv: NIHMS150003.

Daníelsson, J. and C. G. de Vries (1997). "Tail index and quantile estimation with very high frequency data". *Journal of Empirical Finance* 4.2-3, pp. 241–257.

De Beaumont, L., M. Lassonde, S. Leclerc, and H. Théoret (2007). "Long-term and cumulative effects of sports concussion on motor cortex inhibition". *Neurosurgery* 61.2, pp. 329–336.

Degeling, K., H. Koffijberg, and M. J. IJzerman (2017). "A systematic review and checklist presenting the main challenges for health economic modeling in personalized medicine: towards implementing patient-level models". *Expert Review of Pharmacoeconomics and Outcomes Research* 17.1, pp. 17–25.

Dehon, E., N. Weiss, J. Jones, W. Faulconer, E. Hinton, and S. Sterling (2017). "A Systematic Review of the Impact of Physician Implicit Racial Bias on Clinical Decision Making". *Academic Emergency Medicine* 24.8, pp. 895–904.

Deneef, P. and D. L. Kent (1993). "Using Treatment-tradeoff Preferences to Select Diagnostic Strategies". *Medical Decision Making* 13.2, pp. 126–132.

Denton, B. T., M. Kurt, N. D. Shah, S. C. Bryant, and S. a. Smith (2009). "Optimizing the Start Time of Statin Therapy for Patients with Diabetes". *Medical Decision Making* 29.3, pp. 351–367.

Deo, S., K. Rajaram, S. Rath, U. S. Karmarkar, and M. B. Goetz (2015). "Planning for HIV Screening, Testing, and Care at the Veterans Health Administration". *Operations Research* 63.2, pp. 287–304.

Deo, S. and M. Sohoni (2015). "Optimal Decentralization of Early Infant Diagnosis of HIV in Resource-Limited Settings". *Manufacturing & Service Operations Management* 17.2, pp. 191–207.

Deshpande, P. R., B. L. Sudeepthi, S. Rajan, and C. P. Abdul Nazir (2011). "Patient-reported outcomes: A new era in clinical research". *Perspectives in Clinical Research* 2.4, p. 137.

Dick, R. W. (2009). "Is there a gender difference in concussion incidence and outcomes?" *British Journal of Sports Medicine* 43.Suppl_1, pp. i46–i50.

Downey, R. I., M. G. Hutchison, and P. Comper (2018). "Determining sensitivity and specificity of the Sport Concussion Assessment Tool 3 (SCAT3) components in university athletes". *Brain Injury* 32.11, pp. 1345–1352.

Ebell, M. (2010). "AHRQ White Paper: Use of Clinical Decision Rules for Point-of-Care Decision Support". *Medical Decision Making* 30.6. Ed. by W. R. Hersh and R. L. Street, pp. 712–721.

Echemendia, R. J., J. M. Bruce, C. M. Bailey, J. F. Sanders, P. Arnett, and G. Vargas (2012). "The utility of post-concussion neuropsychological data in identifying cognitive change

following sports-related MTBI in the absence of baseline data". *Clinical Neuropsychologist* 26.7, pp. 1077–1091.

Echemendia, R. J., W. Meeuwisse, P. McCrory, G. A. Davis, M. Putukian, J. Leddy, M. Makdissi, S. J. Sullivan, S. P. Broglio, M. Raftery, K. Schneider, J. Kissick, M. McCrea, J. Dvorak, A. K. Sills, M. Aubry, L. Engebretsen, M. Loosemore, G. Fuller, J. Kutcher, R. Ellenbogen, K. Guskiewicz, J. Patricios, and S. Herring (2017). "The Sport Concussion Assessment Tool 5th Edition (SCAT5)". *British Journal of Sports Medicine* 5.5, pp. 1–3.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Eisenberg, M. A., W. P. Meehan, and R. Mannix (2014). "Duration and Course of Post-Concussive Symptoms". *Pediatrics* 133.6, pp. 999–1006.

El-Amine, H., E. K. Bish, and D. R. Bish (2018). "Robust Postdonation Blood Screening Under Prevalence Rate Uncertainty". *Operations Research* 66.1, pp. 1–17.

Elbin, R. J., A. M. Sufrinko, P. Schatz, J. French, L. Henry, S. Burkhart, M. W. Collins, and A. P. Kontos (2016). "Removal From Play After Concussion and Recovery Time". *Pediatrics* 138.3, e20160910–e20160910.

Ellis, M. J., L. N. Ryner, O. Sobczyk, J. Fierstra, D. J. Mikulis, J. A. Fisher, J. Duffin, and W. A. C. Mutch (2016). "Neuroimaging Assessment of Cerebrovascular Reactivity in Concussion: Current Concepts, Methodological Considerations, and Review of the Literature". *Frontiers in Neurology* 7.APR, pp. 1–16.

Elwyn, G., D. Frosch, R. Thomson, N. Joseph-Williams, A. Lloyd, P. Kinnersley, E. Cording, D. Tomson, C. Dodd, S. Rollnick, A. Edwards, and M. Barry (2012). "Shared Decision Making: A Model for Clinical Practice". *Journal of General Internal Medicine* 27.10, pp. 1361–1367.

Epstein, R. M. and R. L. Street (2011). "The Values and Value of Patient-Centered Care". *The Annals of Family Medicine* 9.2, pp. 100–103.

Erenay, F. S., O. Alagoz, and A. Said (2014). "Optimizing Colonoscopy Screening for Colorectal Cancer Prevention and Surveillance". *Manufacturing & Service Operations Management* 16.3, pp. 381–400.

Felder, S. and T. Mayrhofer (2014). "Risk Preferences: Consequences for Test and Treatment Thresholds and Optimal Cutoffs". *Medical Decision Making* 34.1, pp. 33–41.

Formann, A. K. (1984). *Die latent-class-analyse: Einführung in Theorie und Anwendung*. Beltz.

Galie, N., M. M. Hoeper, M. Humbert, A. Torbicki, J.-L. Vachiery, J. A. Barbera, M. Beghetti, P. Corris, S. Gaine, J. S. Gibbs, and Others (2009). "Guidelines for the diagnosis and treatment of pulmonary hypertension". *European Heart Journal* 30.20, pp. 2493–2537.

Garcia, G.-G. P., S. P. Broglio, M. S. Lavieri, M. McCrea, and T. McAllister (2018). "Quantifying the Value of Multidimensional Assessment Models for Acute Concussion: An Anal-

ysis of Data from the NCAA-DoD Care Consortium". *Sports Medicine* 48.7, pp. 1739–1749.

Garcia, G.-G. P., M. S. Lavieri, R. Jiang, T. W. McAllister, M. A. McCrea, and S. P. Broglio (2019). "A Data-Driven Approach to Unlikely, Possible, Probable, and Definite Acute Concussion Assessment". *Journal of Neurotrauma* 36.10, pp. 1571–1583.

Garcia, G.-G. P., M. S. Lavieri, R. Jiang, M. A. McCrea, T. W. McAllister, and S. P. Broglio (2020a). "Data-driven stochastic optimization approaches to determine decision thresholds for risk estimation models". *IISE Transactions* 52.10, pp. 1098–1121.

Garcia, G.-G. P., J. Yang, M. S. Lavieri, T. W. McAllister, M. A. McCrea, and S. P. Broglio (2020b). "Optimizing Components of the Sport Concussion Assessment Tool for Acute Concussion Assessment". *Neurosurgery* Accepted.

Gessel, L. M., S. K. Fields, C. L. Collins, R. W. Dick, and R. D. Comstock (2007). "Concussions among United States high school and collegiate athletes." *Journal of Athletic Training (National Athletic Trainers' Association)* 42.4, pp. 495–503.

Giessen, A. van, G. A. de Wit, K. G. Moons, J. A. Dorresteijn, and H. Koffijberg (2018). "An alternative approach identified optimal risk thresholds for treatment indication: an illustration in coronary heart disease". *Journal of Clinical Epidemiology* 94, pp. 122–131.

Giza, C. C., J. S. Kutcher, S. Ashwal, J. Barth, T. S. D. Getchius, G. A. Gioia, G. S. Gronseth, K. Guskiewicz, S. Mandel, G. Manley, D. B. McKeag, D. J. Thurman, and R. Zafonte (2013). "Summary of evidence-based guideline update: evaluation and management of concussion in sports: report of the Guideline Development Subcommittee of the American Academy of Neurology." *Neurology* 80.24, pp. 2250–7.

Glasziou, P. and J. Hilden (1986). "Threshold Analysis of Decision Tables". *Medical Decision Making* 6.3, pp. 161–168.

Gmytrasiewicz, P. J. and P. Doshi (2005). "A Framework for Sequential Planning in Multi-Agent Settings". *Journal of Artificial Intelligence Research* 24, pp. 49–79. arXiv: 1109.2135.

Godbole, S. and S. Sarawagi (2004). "Discriminative Methods for Multi-labeled Classification". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 3056, pp. 22–30.

Greiner, M., D. Pfeiffer, and R. Smith (2000). "Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests". *Preventive Veterinary Medicine* 45.1-2, pp. 23–41.

Greiner, M., D. Sohr, and P. Göbel (1995). "A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests". *Journal of Immunological Methods* 185.1, pp. 123–132.

Güneş, E. D., E. L. Örmeci, and D. Kunduzcu (2015). "Preventing and diagnosing colorectal cancer with a limited colonoscopy resource". *Production and Operations Management* 24.1, pp. 1–20.

Guskiewicz, K. M. (2001). "Postural Stability Assessment Following Concussion: One Piece of the Puzzle". *Clinical Journal of Sport Medicine* 11.3, pp. 182–189.

Guskiewicz, K. M., S. W. Marshall, J. Bailes, M. McCrea, R. C. Cantu, C. Randolph, and B. D. Jordan (2005). "Association between Recurrent Concussion and Late-Life Cognitive Impairment in Retired Professional Football Players". *Neurosurgery* 57.4, pp. 719–726.

Guskiewicz, K. M., S. W. Marshall, J. Bailes, M. A. McCrea, H. P. Harding, A. Matthews, J. K. Register-Mihalik, and R. C. Cantu (2007). "Recurrent Concussion and Risk of Depression in Retired Professional Football Players". *Medicine & Science in Sports & Exercise* 39.6, pp. 903–909.

Guskiewicz, K. M., M. McCrea, S. W. Marshall, R. C. Cantu, C. Randolph, W. Barr, J. A. Onate, and J. P. Kelly (2003). "Cumulative Effects Associated With Recurrent Concussion in Collegiate Football Players". *Journal of the American Medical Association* 290.19, p. 2549.

Guskiewicz, K. M., J. Register-Mihalik, P. McCrory, M. McCrea, K. Johnston, M. Makdissi, J. Dvořák, G. Davis, and W. Meeuwisse (2013). "Evidence-based approach to revising the SCAT2: introducing the SCAT3". *British Journal of Sports Medicine* 47.5, pp. 289–293.

Guskiewicz, K. M., S. E. Ross, and S. W. Marshall (2001). "Postural Stability and Neuropsychological Deficits After Concussion in Collegiate Athletes." *Journal of athletic training* 36.3, pp. 263–273.

Hänninen, T., M. Tuominen, J. Parkkari, M. Vartiainen, J. Öhman, G. L. Iverson, and T. M. Luoto (2016). "Sport concussion assessment tool – 3rd edition – normative reference values for professional ice hockey players". *Journal of Science and Medicine in Sport* 19.8, pp. 636–641.

Harada, G. K., C. M. Rugg, A. Arshi, J. Vail, and S. L. Hame (2019). "Multiple Concussions Increase Odds and Rate of Lower Extremity Injury in National Collegiate Athletic Association Athletes After Return to Play". *The American Journal of Sports Medicine* 47.13, pp. 3256–3262.

Harmon, K. G., J. A. Drezner, M. Gammons, K. M. Guskiewicz, M. Halstead, S. A. Herring, J. S. Kutcher, A. Pana, M. Putukian, and W. O. Roberts (2013). "American Medical Society for Sports Medicine position statement: concussion in sport". *British Journal of Sports Medicine* 47.1, pp. 15–26.

Harrell, F. E. and C. E. Davis (1982). "A new distribution-free quantile estimator". *Biometrika* 69.3, pp. 635–640.

Hartz, A., W. P. McKinney, R. Centor, A. Krieg, G. Simms, and S. Henck (1986). "Stochastic Thresholds". *Medical Decision Making* 6.3, pp. 145–148.

He, H., Y. Bai, E. A. Garcia, and S. Li (2008). "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". *Proceedings of the International Joint Conference on Neural Networks* 3, pp. 1322–1328.

Helm, J. E., M. S. Lavieri, M. P. Van Oyen, J. D. Stein, and D. C. Musch (2015). "Dynamic Forecasting and Control Algorithms of Glaucoma Progression for Clinician Decision Support". *Operations Research* 63.5, pp. 979–999.

Herman, D. C., D. Jones, A. Harrison, M. Moser, S. Tillman, K. Farmer, A. Pass, J. R. Clugston, J. Hernandez, and T. L. Chmielewski (2017). "Concussion May Increase the Risk of Subsequent Lower Extremity Musculoskeletal Injury in Collegiate Athletes". *Sports Medicine* 47.5, pp. 1003–1010.

Humphreys, I., R. L. Wood, C. J. Phillips, and S. Macey (2013). "The costs of traumatic brain injury: a literature review". *ClinicoEconomics and Outcomes Research* 5.1, pp. 281–287.

Jacobson, S. H., G. Yu, and J. A. Jokela (2016). "A double-risk monitoring and movement restriction policy for Ebola entry screening at airports in the United States". *Preventive Medicine* 88, pp. 33–38.

Jain, A. K. (2010). "Data clustering: 50 years beyond K-means". *Pattern Recognition Letters* 31.8, pp. 651–666.

Jónasson, J. O., S. Deo, and J. Gallien (2017). "Improving HIV Early Infant Diagnosis Supply Chains in Sub-Saharan Africa: Models and Application to Mozambique". *Operations Research* 65.6, pp. 1479–1493.

Jund, J., M. Rabilloud, M. Wallon, and R. Ecochard (2005). "Methods to Estimate the Optimal Threshold for Normally or Log-Normally Distributed Biological Tests". *Medical Decision Making* 25.4, pp. 406–415.

Kawagoe, T. and H. Takizawa (2009). "Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information". *Games and Economic Behavior* 66.1, pp. 238–255.

Kazemian, P., J. E. Helm, M. S. Lavieri, J. D. Stein, and M. P. Van Oyen (2019). "Dynamic Monitoring and Control of Irreversible Chronic Diseases with Application to Glaucoma". *Production and Operations Management* 28.5, pp. 1082–1107.

Kerr, Z. Y., K. R. Evenson, W. D. Rosamond, J. P. Mihalik, K. M. Guskiewicz, and S. W. Marshall (2014a). "Association between concussion and mental health in former collegiate athletes". *Injury Epidemiology* 1.1, p. 28.

Kerr, Z. Y., S. W. Marshall, H. P. Harding, and K. M. Guskiewicz (2012). "Nine-Year Risk of Depression Diagnosis Increases With Increasing Self-Reported Concussions in Retired Professional Football Players". *The American Journal of Sports Medicine* 40.10, pp. 2206–2212.

Kerr, Z. Y., J. K. Register-Mihalik, S. W. Marshall, K. R. Evenson, J. P. Mihalik, and K. M. Guskiewicz (2014b). "Disclosure and non-disclosure of concussion and concussion symptoms in athletes: Review and application of the socio-ecological framework." *Brain injury* 28.8, pp. 1009–1021.

Kerr, Z. Y., L. C. Thomas, J. E. Simon, M. McCrea, and K. M. Guskiewicz (2018a). "Association Between History of Multiple Concussions and Health Outcomes Among Former

College Football Players: 15-Year Follow-up From the NCAA Concussion Study (1999-2001)". *The American Journal of Sports Medicine* 46.7, pp. 1733–1741.

Kerr, Z. Y., G. B. Wilkerson, S. V. Caswell, D. W. Currie, L. A. Pierpoint, E. B. Wasserman, S. B. Knowles, T. P. Dompier, R. Dawn Comstock, and S. W. Marshall (2018b). "The first decade of web-based sports injury surveillance: Descriptive epidemiology of injuries in United States high school football (2005-2006 through 2013-2014) and National collegiate athletic association football (2004-2005 through 2013-2014)". *Journal of Athletic Training* 53.8, pp. 738–751.

King, J. A., M. A. McCrea, and L. D. Nelson (2020). "Frequency of Primary Neck Pain in Mild Traumatic Brain Injury/Concussion Patients". *Archives of Physical Medicine and Rehabilitation* 101.1, pp. 89–94.

Kontos, A. P., R. J. Elbin, A. Trbovich, M. Womble, A. Said, V. F. Sumrok, J. French, N. Kegel, A. Puskar, N. Sherry, C. Holland, and M. Collins (2020). "Concussion Clinical Profiles Screening (CP Screen) Tool: Preliminary Evidence to Inform a Multidisciplinary Approach". *Neurosurgery* 0.0, pp. 1–9.

Kontos, A. P., A. Sufrinko, R. Elbin, A. Puskar, and M. W. Collins (2016). "Reliability and Associated Risk Factors for Performance on the Vestibular/Ocular Motor Screening (VOMS) Tool in Healthy Collegiate Athletes". *The American Journal of Sports Medicine* 44.6, pp. 1400–1406.

Kontos, A. P., A. Sufrinko, N. Sandel, K. Emami, and M. W. Collins (2019). "Sport-related Concussion Clinical Profiles". *Current Sports Medicine Reports* 18.3, pp. 82–92.

Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis (2015). "Machine learning applications in cancer prognosis and prediction". *Computational and Structural Biotechnology Journal* 13, pp. 8–17. arXiv: `9781591404590`.

Krishnamurthy, V. (2012). "Quickest detection pomdps with social learning: Interaction of local and global decision makers". *IEEE Transactions on Information Theory* 58.8, pp. 5563–5587. arXiv: `1007.0571`.

Kroshus, E., B. Garnett, M. Hawrilenko, C. M. Baugh, and J. P. Calzo (2015a). "Concussion under-reporting and pressure from coaches, teammates, fans, and parents". *Social Science & Medicine* 134, pp. 66–75.

Kroshus, E., L. D. Kubzansky, R. E. Goldman, and S. B. Austin (2015b). "Norms, Athletic Identity, and Concussion Symptom Under-Reporting Among Male Collegiate Ice Hockey Players: A Prospective Cohort Study". *Annals of Behavioral Medicine* 49.1, pp. 95–103.

Kutcher, J. S. and J. T. Eckner (2010). "At-risk populations in sports-related concussion". *Current Sports Medicine Reports* 9, pp. 16–20.

Kutcher, J. S. and C. C. Giza (2014). "Sports Concussion Diagnosis and Management". *Continuum* 20.6, pp. 1552–1569.

Langlois, J. A., W. Rutland-Brown, and M. M. Wald (2006). "The Epidemiology and Impact of Traumatic Brain Injury: a brief overview." *The Journal of Head Trauma Rehabilitation* 21.5, pp. 375–378. arXiv: `PRINTEDTOREAD`.

Lau, B. C., A. P. Kontos, M. W. Collins, A. Mucha, and M. R. Lovell (2011). "Which On-field Signs/Symptoms Predict Protracted Recovery From Sport-Related Concussion Among High School Football Players?" *The American Journal of Sports Medicine* 39.11, pp. 2311–2318.

Lavelle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz (2011). "Big Data, Analytics and the Path from Insights to Value". *MIT Sloan Management Review* 52.2, pp. 21–31.

Lavieri, M. S., M. L. Puterman, S. Tyldesley, and W. J. Morris (2012). "When to treat prostate cancer patients based on their PSA dynamics". *IIE Transactions on Healthcare Systems Engineering* 2.1, pp. 62–77.

Lawrence, W. F. and J. A. Fleishman (2004). "Predicting EuroQoL EQ-5D Preference Scores from the SF-12 Health Survey in a Nationally Representative Sample". *Medical Decision Making* 24.2, pp. 160–169.

Lebrun, C. M., M. Mrazik, A. S. Prasad, B. J. Tjarks, J. C. Dorman, M. F. Bergeron, T. A. Munce, and V. D. Valentine (2013). "Sport concussion knowledge base, clinical practises and needs for continuing medical education: a survey of family physicians and cross-border comparison". *British Journal of Sports Medicine* 47.1, pp. 54–59.

Lee, E., M. S. Lavieri, and M. Volk (2018). "Optimal Screening for Hepatocellular Carcinoma: A Restless Bandit Model". *Manufacturing & Service Operations Management* July, msom.2017.0697.

Lee, E., M. S. Lavieri, M. L. Volk, and Y. Xu (2015). "Applying reinforcement learning techniques to detect hepatocellular carcinoma under limited screening capacity". *Health Care Management Science* 18.3, pp. 363–375.

Levy, A. G., A. M. Scherer, B. J. Zikmund-Fisher, K. Larkin, G. D. Barnes, and A. Fagerlin (2018). "Prevalence of and Factors Associated With Patient Nondisclosure of Medically Relevant Information to Clinicians". *JAMA Network Open* 1.7, e185293.

Lew, H. L., D. Thomander, K. T. L. Chew, and J. Bleiberg (2007). "Review of sports-related concussion: Potential for application in military settings." *Journal of rehabilitation research and development* 44.7, pp. 963–974.

Li, Y., M. Zhu, R. Klein, and N. Kong (2014). "Using a partially observable Markov chain model to assess colonoscopy screening strategies – A cohort study". *European Journal of Operational Research* 238.1, pp. 313–326.

Lincoln, A. E., S. V. Caswell, J. L. Almquist, R. E. Dunn, J. B. Norris, and R. Y. Hinton (2011). "Trends in Concussion Incidence in High School Sports". *The American Journal of Sports Medicine* 39.5, pp. 958–963.

Lobo, J. M., B. T. Denton, J. R. Wilson, N. D. Shah, and S. A. Smith (2017). "Using claims data linked with electronic health records to monitor and improve adherence to medication". *IISE Transactions on Healthcare Systems Engineering* 7.4, pp. 194–214.

Lohr, K. N. and B. J. Zebrack (2009). "Using patient-reported outcomes in clinical practice: challenges and opportunities". *Quality of Life Research* 18.1, pp. 99–107.

Lovejoy, W. S. (1987a). "Some Monotonicity Results for Partially Observed Markov Decision Processes". *Operations Research* 35.5, pp. 736–743.

Lovejoy, W. S. (1987b). "Technical Note—On the Convexity of Policy Regions in Partially Observed Systems". *Operations Research* 35.4, pp. 619–621.

Lovell, M. R., G. L. Iverson, M. W. Collins, K. Podell, K. M. Johnston, D. Pardini, J. Pardini, J. Norwig, and J. C. Maroon (2006). "Measurement of Symptoms Following Sports-Related Concussion: Reliability and Normative Data for the Post-Concussion Scale". *Applied Neuropsychology* 13.3, pp. 166–174.

Luedtke, J. and S. Ahmed (2008). "A Sample Approximation Approach for Optimization with Probabilistic Constraints". *SIAM Journal on Optimization* 19.2, pp. 674–699.

Madani, O., S. Hanks, and A. Condon (2003). "On the undecidability of probabilistic planning and related stochastic optimization problems". *Artificial Intelligence* 147.1-2, pp. 5–34.

Maillart, L. M., J. S. Ivy, S. Ransom, and K. Diehl (2008). "Assessing Dynamic Breast Cancer Screening Policies". *Operations Research* 56.6, pp. 1411–1427.

Makdissi, M., D. Darby, P. Maruff, A. Ugoni, P. Brukner, and P. R. McCrory (2010). "Natural History of Concussion in Sport". *The American Journal of Sports Medicine* 38.3, pp. 464–471.

Mangasarian, O. L., W. N. Street, and W. H. Wolberg (1995). "Breast Cancer Diagnosis and Prognosis Via Linear Programming". *Operations Research* 43.4, pp. 570–577.

Maruta, J., A. Lumba-Brown, and J. Ghajar (2018). "Concussion Subtype Identification With the Rivermead Post-concussion Symptoms Questionnaire". *Frontiers in Neurology* 9.December, pp. 1–7.

Mason, J. E., B. T. Denton, N. D. Shah, and S. A. Smith (2014). "Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients". *European Journal of Operational Research* 233.3, pp. 727–738.

Mason, J. E., D. A. England, B. T. Denton, S. A. Smith, M. Kurt, and N. D. Shah (2012). "Optimizing Statin Treatment Decisions for Diabetes Patients in the Presence of Uncertain Future Adherence". *Medical Decision Making* 32.1, pp. 154–166.

Mcafee, A. and E. Brynjolfsson (2012). "Big Data: The Management Revolution". *Harvard Business Review* 90.10, pp. 60–68.

McAllister, D. R., A. R. Motamedi, S. L. Hame, M. S. Shapiro, and F. J. Dorey (2001). "Quality of Life Assessment in Elite Collegiate Athletes". *The American Journal of Sports Medicine* 29.6, pp. 806–810.

McCrea, M, J. P. Kelly, C Randolph, J Kluge, E Bartolic, G Finn, and B Baxter (1998). *Standardized assessment of concussion (SAC): on-site mental status evaluation of the athlete.*

McCrea, M., S. Broglio, T. McAllister, W. Zhou, S. Zhao, B. Katz, M. Kudela, J. Harezlak, L. Nelson, T. Meier, S. W. Marshall, and K. M. Guskiewicz (2020). "Return to play and risk of repeat concussion in collegiate football players: comparative analysis from

the NCAA Concussion Study (1999–2001) and CARE Consortium (2014–2017)". *British Journal of Sports Medicine* 54.2, pp. 102–109.

McCrea, M., K. Guskiewicz, C. Randolph, W. B. Barr, T. a. Hammeke, S. W. Marshall, and J. P. Kelly (2009). "Effects of a symptom-free waiting period on clinical outcome and risk of reinjury after sport-related concussion". *Neurosurgery* 65.5, pp. 876–882.

McCrea, M., K. Guskiewicz, C. Randolph, W. B. Barr, T. a. Hammeke, S. W. Marshall, M. R. Powell, K. Woo Ahn, Y. Wang, and J. P. Kelly (2013). "Incidence, Clinical Course, and Predictors of Prolonged Recovery Time Following Sport-Related Concussion in High School and College Athletes". *Journal of the International Neuropsychological Society* 19.01, pp. 22–33.

McCrea, M., K. M. Guskiewicz, S. W. Marshall, W. Barr, C. Randolph, R. C. Cantu, J. a. Onate, J. Yang, and J. P. Kelly (2003). "Acute Effects and Recovery Time Following Concussion in Collegiate Football Players". *JAMA* 290.19, p. 2556.

McCrea, M. A., W. B. Barr, K. M. Guskiewicz, C. Randolph, S. W. Marshall, R. C. Cantu, J. A. Onate, and J. P. Kelly (2005). "Standard regression-based methods for measuring recovery after sport-related concussion". *Journal of the International Neuropsychological Society* 11.01, pp. 58–69.

McCrory, P., W. Meeuwisse, J. Dvorak, M. Aubry, J. Bailes, S. Broglio, R. C. Cantu, D. Cassidy, R. J. Echemendia, R. J. Castellani, G. A. Davis, R. Ellenbogen, C. Emery, L. Engebretsen, N. Feddermann-Demont, C. C. Giza, K. M. Guskiewicz, S. Herring, G. L. Iverson, K. M. Johnston, J. Kissick, J. Kutcher, J. J. Leddy, D. Maddocks, M. Makdissi, G. T. Manley, M. McCrea, W. P. Meehan, S. Nagahiro, J. Patricios, M. Putukian, K. J. Schneider, A. Sills, C. H. Tator, M. Turner, and P. E. Vos (2017). "Consensus statement on concussion in sport—the 5 th international conference on concussion in sport held in Berlin, October 2016". *British Journal of Sports Medicine* 51, pp. 838–847.

McCrory, P., W. H. Meeuwisse, J. S. Kutcher, B. D. Jordan, and A. Gardner (2013). "What is the evidence for chronic concussion-related changes in retired athletes: behavioural, pathological and clinical outcomes?" *British Journal of Sports Medicine* 47.5, pp. 327–330.

McDonald, W. I., A. Compston, G. Edan, D. Goodkin, H.-P. Hartung, F. D. Lublin, H. F. McFarland, D. W. Paty, C. H. Polman, S. C. Reingold, M. Sandberg-Wollheim, W. Sibley, A. Thompson, S. Van Den Noort, B. Y. Weinshenker, and J. S. Wolinsky (2001). "Recommended diagnostic criteria for multiple sclerosis: Guidelines from the international panel on the diagnosis of multiple sclerosis". *Annals of Neurology* 50.1, pp. 121–127.

McGrath, N. (2010). "Supporting the student-athlete's return to the classroom after a sport-related concussion". *Journal of Athletic Training* 45.5, pp. 492–498.

McGregor, M. and J. J. Caro (2006). "QALYs". *PharmacoEconomics* 24.10, pp. 947–952.

McKhann, G. M., D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, and Others (2011). "The diagno-

sis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". *Alzheimer's & Dementia* 7.3, pp. 263–269. arXiv: NIHMS150003.

McLay, L. A., C. Foufoulides, and J. R. W. Merrick (2010). "Using simulation-optimization to construct screening strategies for cervical cancer". *Health Care Management Science* 13.4, pp. 294–318.

Meehan, W. P., R. C. Mannix, A. Stracciolini, R. Elbin, and M. W. Collins (2013). "Symptom Severity Predicts Prolonged Recovery after Sport-Related Concussion, but Age and Amnesia Do Not". *The Journal of Pediatrics* 163.3, pp. 721–725. arXiv: NIHMS150003.

Meier, T. B., B. J. Brummel, R. Singh, C. J. Nerio, D. W. Polanski, and P. S. Bellgowan (2015). "The underreporting of self-reported symptoms following sports-related concussion". *Journal of Science and Medicine in Sport* 18.5, pp. 507–511.

Messias, J. V., M. T. Spaan, and P. U. Lima (2011). "Efficient offline communication policies for factored Multiagent POMDPs". *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pp. 1–9.

Mintz, Y., A. Aswani, P. Kaminsky, E. Flowers, and Y. Fukuoka (2017). "Behavioral Analytics for Myopic Agents". arXiv: 1702.05496.

— (2019). "Non-Stationary Bandits with Habituation and Recovery Dynamics". *Operations Research* Accepted, pp. 1–46. arXiv: 1707.08423.

Mohajerin Esfahani, P. and D. Kuhn (2018). "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations". *Mathematical Programming* 171.1-2, pp. 115–166.

Moons, K. G. M., D. G. Altman, Y. Vergouwe, and P. Royston (2009). "Prognosis and prognostic research: application and impact of prognostic models in clinical practice". *British Medical Journal* 338.7709, pp. 1487–1490.

Moons, K. G., T. Stijnen, B. C. Michel, H. R. Büller, G.-A. Van Es, D. E. Grobbee, and J. D. F. Habbema (1997). "Application of Treatment Thresholds to Diagnostic-test Evaluation". *Medical Decision Making* 17.4, pp. 447–454.

Moran, R. N., J. Meek, J. Allen, and J. Robinson (2020). "Sex differences and normative data for the m-CTSIB and sensory integration on baseline concussion assessment in collegiate athletes". *Brain Injury* 34.1, pp. 20–25.

Moreau, M. S., J. Langdon, and T. a. Buckley (2014). "The lived experience of an in-season concussion amongst NCAA Division I student-athletes". *International Journal of Exercise Science* 7.1, pp. 62–74.

Mucha, A., M. W. Collins, R. Elbin, J. M. Furman, C. Troutman-Enseki, R. M. DeWolf, G. Marchetti, and A. P. Kontos (2014). "A Brief Vestibular/Ocular Motor Screening (VOMS) Assessment to Evaluate Concussions". *The American Journal of Sports Medicine* 42.10, pp. 2479–2486. arXiv: NIHMS150003.

Nease, R. F., D. K. Owens, and H. C. Sox (1989). "Threshold Analysis Using Diagnostic Tests with Multiple Results". *Medical Decision Making* 9.2, pp. 91–103.

Newell, S. A., A. Girgis, R. W. Sanson-Fisher, and N. J. Savolainen (1999). "The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population". *American Journal of Preventive Medicine* 17.3, pp. 211–229.

Nord, E., N. Daniels, and M. Kamlet (2009). "QALYs: Some Challenges". *Value in Health* 12.SUPPL. 1, S10–S15.

O'Connor, K. L. (2019). "Concussion among military service academy members: identifying risk factors, recovery trajectories, and the role of mental health". *British Journal of Sports Medicine* 53.6, pp. 368–369.

Odetola, F. O., L. Bruski, G. Zayas-Caban, and M. Lavieri (2016). "An innovative framework to improve efficiency of interhospital transfer of children in respiratory failure". *Annals of the American Thoracic Society* 13.5, pp. 671–677.

Oldham, J. R., M. S. DiFabio, T. W. Kaminski, R. M. DeWolf, D. R. Howell, and T. A. Buckley (2018). "Efficacy of Tandem Gait to Identify Impaired Postural Control after Concussion". *Medicine & Science in Sports & Exercise* 50.6, pp. 1162–1168.

Pagnoncelli, B. K., S. Ahmed, and A. Shapiro (2009). "Sample average approximation method for chance constrained programming: Theory and applications". *Journal of Optimization Theory and Applications* 142.2, pp. 399–416.

Parsons, J. T. and A. R. Snyder (2011). "Health-related quality of life as a primary clinical outcome in sport rehabilitation". *Journal of Sport Rehabilitation* 20.1, pp. 17–36.

Patricios, J., G. W. Fuller, R. Ellenbogen, S. Herring, J. S. Kutcher, M. Loosemore, M. Makdissi, M. McCrea, M. Putukian, and K. J. Schneider (2017). "What are the critical elements of sideline screening that can be used to establish the diagnosis of concussion? A systematic review". *British Journal of Sports Medicine* 2000.April 2016, bjsports–2016–097441.

Pauker, S. G. and J. P. Kassirer (1975). "Therapeutic Decision Making: A Cost-Benefit Analysis". *New England Journal of Medicine* 293.5, pp. 229–234.

Pauker, S. G. and J. P. Kassirer (1980). "The Threshold Approach to Clinical Decision Making". *New England Journal of Medicine* 302.20, pp. 1109–1117.

Pearce, K. L., A. Sufrinko, B. C. Lau, L. Henry, M. W. Collins, and A. P. Kontos (2015). "Near Point of Convergence After a Sport-Related Concussion". *The American Journal of Sports Medicine* 43.12, pp. 3055–3061. arXiv: 15334406.

Peck, J. S., J. C. Benneyan, D. J. Nightingale, and S. A. Gaehde (2012). "Predicting Emergency Department Inpatient Admissions to Improve Same-day Patient Flow". *Academic Emergency Medicine* 19.9, E1045–E1054.

Pepe, M. S., G. Longton, and H. Janes (2009). "Estimation and comparison of receiver operating characteristic curves". *Stata Journal* 9.1, pp. 1–16.

Perk, J., G. De Backer, H. Gohlke, I. Graham, Ž. Reiner, W. M. M. Verschuren, C. Albus, P. Benlian, G. Boysen, R. Cifkova, and Others (2012). "European Guidelines on Cardiovascular Disease Prevention in Clinical Practice (Version 2012)". *International Journal of Behavioral Medicine* 19.4, pp. 403–488.

Perloff, R. M., B. Bonder, G. B. Ray, E. B. Ray, and L. A. Siminoff (2006). "Doctor-Patient Communication, Cultural Competence, and Minority Health". *American Behavioral Scientist* 49.6, pp. 835–852.

Pierskalla, W. P. and D. J. Brailer (1994). "Chapter 13 Applications of operations research in health care delivery". *Handbooks in Operations Research and Management Science*. Vol. 6. Amsterdam, Netherlands: Elsevier, pp. 469–505.

Piland, S. G., M. S. Ferrara, S. N. Macciocchi, S. P. Broglio, and T. E. Gould (2010). "Investigation of baseline self-report concussion symptom scores". *Journal of Athletic Training* 45.3, pp. 273–278.

Putukian, M. (2011). "The Acute Symptoms of Sport-Related Concussion: Diagnosis and On-field Management". *Clinics in Sports Medicine* 30.1, pp. 49–61.

Putukian, M., R. Echemendia, A. Dettwiler-Danspeckgruber, T. Duliba, J. Bruce, J. L. Furtado, and M. Murugavel (2015). "Prospective clinical assessment using sideline concussion assessment tool-2 testing in the evaluation of sport-related concussion in college athletes". *Clinical Journal of Sport Medicine* 25.1, pp. 36–42.

Putukian, M., M. Raftery, K. Guskiewicz, S. Herring, M. Aubry, R. C. Cantu, and M. Molloy (2013). "Onfield assessment of concussion in the adult athlete". *British Journal of Sports Medicine* 47.5, pp. 285–288.

Pynadath, D. and M. Tambe (2002). "The Communicative Multiagent Team Decision Problem: Analyzing Teamwork Theories and Models". *Journal of Artifical Intelligence Research* 16.1, pp. 389–423.

Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE* 77.2, pp. 257–286. arXiv: `arXiv:1011.1669v3`.

Randolph, C. (2011). "Baseline Neuropsychological Testing in Managing Sport-Related Concussion". *Current Sports Medicine Reports* 10.1, pp. 21–26.

Randolph, C., S. Millis, W. B. Barr, M. McCrea, K. M. Guskiewicz, T. A. Hammeke, and J. P. Kelly (2009). "Concussion Symptom Inventory: An Empirically Derived Scale for Monitoring Resolution of Symptoms Following Sport-Related Concussion". *Archives of Clinical Neuropsychology* 24.3, pp. 219–229.

Rawlins, M. L. W., B. R. Johnson, J. K. Register-Mihalik, K. DeAngelis, J. D. Schmidt, and C. J. D'Lauro (2019). "United States Air Force Academy Cadets' Perceived Costs of Concussion Disclosure". *Military Medicine* 185.1-2, E269–E275.

Read, J., B. Pfahringer, G. Holmes, and E. Frank (2011). "Classifier chains for multi-label classification". *Machine Learning* 85.3, pp. 333–359.

Register-Mihalik, J. K., K. M. Guskiewicz, T. C. V. McLeod, L. A. Linnan, F. O. Mueller, and S. W. Marshall (2013a). "Knowledge, attitude, and concussion-reporting behaviors

among high school athletes: A preliminary study". *Journal of Athletic Training* 48.5, pp. 645–653.

Register-Mihalik, J. K., K. M. Guskiewicz, J. P. Mihalik, J. D. Schmidt, Z. Y. Kerr, and M. a. McCrea (2013b). "Reliable Change, Sensitivity, and Specificity of a Multidimensional Concussion Assessment Battery". *Journal of Head Trauma Rehabilitation* 28.4, pp. 274–283.

Resch, J. E., C. N. Brown, J. Schmidt, S. N. Macciocchi, D. Blueitt, C. M. Cullum, and M. S. Ferrara (2016). "The sensitivity and specificity of clinical measures of sport concussion: three tests are better than one". *BMJ Open Sport & Exercise Medicine* 2.1, e000012.

Riemann, B. L. and K. M. Guskiewicz (2000). "Effects of mild head injury on postural stability as measured through clinical balance testing." *Journal of athletic training* 35.1, pp. 19–25.

Riemann, B. L., K. M. Guskiewicz, and E. W. Shields (1999). "Relationship between Clinical and Forceplate Measures of Postural Stability". *Journal of Sport Rehabilitation* 8.2, pp. 71–82. arXiv: `10566716`.

Rigg, J. L. and S. R. Mooney (2011). "Concussions and the Military: Issues Specific to Service Members". *PM and R* 3.10 SUPPL. 2, S380–S386.

Roozenbeek, B., H. F. Lingsma, P. Perel, P. Edwards, I. Roberts, G. D. Murray, A. I. Maas, and E. W. Steyerberg (2011). "The added value of ordinal analysis in clinical trials: an example in traumatic brain injury". *Critical Care* 15.3, R127.

Rousseeuw, P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics* 20.C, pp. 53–65. arXiv: `z0024`.

Royston, P. (2004). "Multiple imputation of missing values". *The Stata Journal* 4.3, pp. 224–241.

Royston, P., K. G. M. Moons, D. G. Altman, and Y. Vergouwe (2009). "Prognosis and prognostic research: Developing a prognostic model". *BMJ* 338.mar31 1, b604–b604.

Sandikçi, B. (2011). "Reduction of a POMDP to an MDP". *Wiley Encyclopedia of Operations Research and Management Science*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 1–9.

Sato, T., Y. Takano, R. Miyashiro, and A. Yoshise (2016). "Feature subset selection for logistic regression via mixed integer optimization". *Computational Optimization and Applications* 64.3, pp. 865–880.

Schell, G. J., G.-G. P. Garcia, M. S. Lavieri, J. B. Sussman, and R. A. Hayward (2019). "Optimal Coinsurance Rates for a Heterogeneous Population under Inequality and Resource Constraints". *IISE Transactions* 51.1, pp. 74–91.

Schell, G. J., M. S. Lavieri, J. E. Helm, X. Liu, D. C. Musch, M. P. Van Oyen, and J. D. Stein (2014). "Using filtered forecasting techniques to determine personalized monitoring schedules for patients with open-angle glaucoma". *Ophthalmology* 121.8, pp. 1539–1546.

Schmidt, J. D., J. K. Register-Mihalik, J. P. Mihalik, Z. Y. Kerr, and K. M. Guskiewicz (2012). "Identifying Impairments after Concussion". *Medicine & Science in Sports & Exercise* 44.9, pp. 1621–1628.

Schneider, K. J., G. L. Iverson, C. A. Emery, P. McCrory, S. A. Herring, and W. H. Meeuwisse (2013). "The effects of rest and treatment following sport-related concussion: a systematic review of the literature". *British Journal of Sports Medicine* 47.5, pp. 304–307. arXiv: `arXiv:1011.1669v3`.

Schottmüller, C. (2013). "Cost incentives for doctors: A double-edged sword". *European Economic Review* 61, pp. 43–58.

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2009). *Lectures on Stochastic Programming*. Society for Industrial and Applied Mathematics.

Shapiro, D. E. (1999). "The interpretation of diagnostic tests". *Statistical Methods In Medical Research* 8.2, pp. 113–34.

Shechter, S. M., M. D. Bailey, A. J. Schaefer, and M. S. Roberts (2008). "The Optimal Time to Initiate HIV Therapy Under Ordered Health States". *Operations Research* 56.1, pp. 20–33.

Shehata, N., J. P. Wiley, S. Richea, B. W. Benson, L. Duits, and W. H. Meeuwisse (2009). "Sport concussion assessment tool: Baseline values for varsity collision sport athletes". *British Journal of Sports Medicine* 43.10, pp. 730–734.

Sheppard, J. W. and M. A. Kaufman (2005). "A Bayesian approach to diagnosis and prognosis using built-in test". *IEEE Transactions on Instrumentation and Measurement* 54.3, pp. 1003–1018.

Si, B., I. Yakushev, and J. Li (2017). "A sequential tree-based classifier for personalized biomarker testing of Alzheimer's disease risk". *IISE Transactions on Healthcare Systems Engineering* 7.4, pp. 248–260.

Simon, G. E. and O. Gureje (1999). "Stability of Somatization Disorder and Somatization Symptoms Among Primary Care Patients". *Archives of General Psychiatry* 56.1, p. 90.

Smallwood, R. D. and E. J. Sondik (1973). "The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon". *Operations Research* 21.5, pp. 1071–1088.

Somoza, E. and D. Mossman (1992). "Comparing and Optimizing Diagnostic Tests". *Medical Decision Making* 12.3, pp. 179–188.

Sufrinko, A., J. McAllister-Deitrick, M. Womble, and A. Kontos (2017a). "Do Sideline Concussion Assessments Predict Subsequent Neurocognitive Impairment After Sport-Related Concussion?" *Journal of Athletic Training* 52.4, pp. 1062–6050–52.4.01.

Sufrinko, A. M., G. F. Marchetti, P. E. Cohen, R. Elbin, V. Re, and A. P. Kontos (2017b). "Using Acute Performance on a Comprehensive Neurocognitive, Vestibular, and Ocular Motor Assessment Battery to Predict Recovery Duration After Sport-Related Concussions". *The American Journal of Sports Medicine* 45.5, pp. 1187–1194.

Tejada, J. J., J. S. Ivy, J. R. Wilson, M. J. Ballan, K. M. Diehl, and B. C. Yankaskas (2015). "Combined DES/SD model of breast cancer screening for older women, I: Natural-history simulation". *IIE Transactions* 47.6, pp. 600–619.

Teng, Y., N. Kong, and W. Tu (2015). "Optimizing strategies for population-based chlamydia infection screening among young women: an age-structured system dynamics approach". *BMC Public Health* 15.639, pp. 1–11.

Tschandl, P., N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, A. Lallas, J. Lapins, C. Longo, J. Malvehy, M. A. Marchetti, A. Marghoob, S. Menzies, A. Oakley, J. Paoli, S. Puig, C. Rinner, C. Rosendahl, A. Scope, C. Sinz, H. P. Soyer, L. Thomas, I. Zalaudek, and H. Kittler (2019). "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study". *The Lancet Oncology* 20.7, pp. 938–947.

Tsushima, W. T., A. M. Siu, H. J. Ahn, B. L. Chang, and N. M. Murata (2019). "Incidence and Risk of Concussions in Youth Athletes: Comparisons of Age, Sex, Concussion History, Sport, and Football Position". *Archives of Clinical Neuropsychology* 34.1, pp. 60–69.

Valovich McLeod, T. C., R. C. Bay, K. C. Lam, and A. Chhabra (2012). "Representative baseline values on the Sport Concussion Assessment Tool 2 (SCAT2) in adolescent athletes vary by gender, grade, and concussion history". *American Journal of Sports Medicine* 40.4, pp. 927–933.

Valovich McLeod, T. C., D. H. Perrin, K. M. Guskiewicz, S. J. Shultz, R. Diamond, and B. M. Gansneder (2004). "Serial administration of clinical concussion assessments and learning effects in healthy young athletes." *Clinical Journal of Sport Medicine* 14.5, pp. 287–295.

Van Buuren, S, H Boshuizen, and D Knook (1999). "Multiple imputation of missing blood pressure covariates in survival analysis". *Statistic in Medicine* 18.6, pp. 681–694.

Van Den Brink, M., E. N. Bandell-Hoekstra, and H. Huijer Abu-Saad (2001). "The occurrence of recall bias in pediatric headache: A comparison of questionnaire and diary data". *Headache* 41.1, pp. 11–20.

Vermont, J., J. Bosson, P. François, C. Robert, A. Rueff, and J. Demongeot (1991). "Strategies for graphical threshold determination". *Computer Methods and Programs in Biomedicine* 35.2, pp. 141–150.

Voss, J. D., J. Connolly, K. A. Schwab, and A. I. Scher (2015). "Update on the Epidemiology of Concussion/Mild Traumatic Brain Injury." *Current pain and headache reports* 19.7, p. 506.

Wang, B. and T. D. Pham (2011). "MRI-based age prediction using hidden Markov models". *Journal of Neuroscience Methods* 199.1, pp. 140–145.

Wang, T. J., J. M. Massaro, D. Levy, R. S. Vasan, P. A. Wolf, R. B. D'Agostino, M. G. Larson, W. B. Kannel, and E. J. Benjamin (2003). "A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study." *Journal of the American Medical Association* 290.8, pp. 1049–56.

Weber, M. L., R. C. Lynall, N. L. Hoffman, E. H. Miller, T. W. Kaminski, T. A. Buckley, H. J. Benjamin, C. M. Miles, C. T. Whitlow, L. Lintner, S. P. Broglio, M. McCrea, T. McAllister, and J. D. Schmidt (2019). "Health-Related Quality of Life Following Concussion in Collegiate Student-Athletes With and Without Concussion History". *Annals of Biomedical Engineering* 47.10, pp. 2136–2146.

Weise, K., K. Hübel, E. Rose, M. Schläger, D. Schrammel, M. Täschner, and R. Michel (2006). "Bayesian decision threshold, detection limit and confidence limits in ionising-radiation measurement". *Radiation Protection Dosimetry* 121.1, pp. 52–63.

Williamson, I. J. S. and D Goodman (2006). "Converging evidence for the under-reporting of concussions in youth ice hockey". *British Journal of Sports Medicine* 40.2, pp. 128–132.

Xue, Y., D. Klabjan, and Y. Luo (2019). "Predicting ICU readmission using grouped physiological and medication trends". *Artificial Intelligence in Medicine* 95.August, pp. 27–37.

Yang, Y., J. D. Goldhaber-Fiebert, and L. M. Wein (2013). "Analyzing Screening Policies for Childhood Obesity". *Management Science* 59.4, pp. 782–795.

Yao, Y. (2010). "Three-way decisions with probabilistic rough sets". *Information Sciences* 180.3, pp. 341–353.

Yao, Y. and B. Zhou (2016). "Two Bayesian approaches to rough sets". *European Journal of Operational Research* 251.3, pp. 904–917.

Yasuda, M. (1988). "The optimal value of markov stopping problems with one-step look ahead policy". *Journal of Applied Probability* 25.3, pp. 544–552.

Yengo-Kahn, A. M., A. T. Hale, B. H. Zalneraitis, S. L. Zuckerman, A. K. Sills, and G. S. Solomon (2016). "The Sport Concussion Assessment Tool: a systematic review". *Neurosurgical Focus* 40.4, E6.

Zargoush, M., M. Gümüş, V. Verter, and S. S. Daskalopoulou (2018). "Designing Risk-Adjusted Therapy for Patients with Hypertension". *Production and Operations Management* 27.12, pp. 2291–2312.

Zhang, H., C. Wernz, and D. R. Hughes (2018). "A Stochastic Game Analysis of Incentives and Behavioral Barriers in Chronic Disease Management". *Service Science* 10.3, pp. 302–319.

Zhang, J., B. T. Denton, H. Balasubramanian, N. D. Shah, and B. A. Inman (2012). "Optimization of Prostate Biopsy Referral Decisions". *Manufacturing & Service Operations Management* 14.4, pp. 529–547.

Zhang, M. L., Y. K. Li, X. Y. Liu, and X. Geng (2018). "Binary relevance for multi-label learning: an overview". *Frontiers of Computer Science* 12.2, pp. 191–202.

Zhang, M.-L. and K. Zhang (2010). "Multi-label learning by exploiting label dependency". *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10*. New York, New York, USA: ACM Press, pp. 999–1007.

Zhu, Y. and J. Fang (2016). "Logistic Regression–Based Trichotomous Classification Tree and Its Application in Medical Diagnosis". *Medical Decision Making* 36.8, pp. 973–989.

Zhuang, J., V. M. Bier, and O. Alagoz (2010). "Modeling secrecy and deception in a multiple-period attacker–defender signaling game". *European Journal of Operational Research* 203.2, pp. 409–418.

Zimmer, A., J. Marcinak, S. Hibyan, and F. Webbe (2015). "Normative Values of Major SCAT2 and SCAT3 Components for a College Athlete Population". *Applied Neuropsychology: Adult* 22.2, pp. 132–140.

Zuckerman, S. L., D. J. Totten, K. E. Rubel, A. W. Kuhn, A. M. Yengo-Kahn, and G. S. Solomon (2016). "Mechanisms of Injury as a Diagnostic Predictor of Sport-Related Concussion Severity in Football, Basketball, and Soccer". *Neurosurgery* 63.1, pp. 102–112.

Zufferey, D., T. Hofer, J. Hennebert, M. Schumacher, R. Ingold, and S. Bromuri (2015). "Performance comparison of multi-label learning algorithms on clinical data for chronic diseases". *Computers in Biology and Medicine* 65, pp. 34–43.