PREDICTION MODELS IN HEALTH-CARE SYSTEMS:
APPLICATIONS AND INSIGHTS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF
MANAGEMENT SCIENCE AND ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Leila Zia
December 2013
I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

John Weyant, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Margret Vilborg Bjarnadottir, Co-Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Ashish Goel

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.
Abstract

One of the canonical problems in health-care is to find a methodology for making decisions in the absence of adequate data and in the presence of uncertainty, considering the operational and non-operational costs and benefits of such decisions. Patients, physicians, and policy makers face this problem, sometimes on a daily basis, when making decisions about recommending one form of treatment over others, choosing treatment vs. palliative care, choosing the population to receive preventive diagnostic tests, and so on.

The lack of such methodology is a critical problem for the health-care industry, with real and sometimes severe consequences for individuals or the industry as a whole. For example, a patient might choose to receive chemotherapy with the hope of having her cancer controlled when in fact she has only a few weeks left. A policy maker might recommend a screening policy for men older than 50 to detect early stages of prostate cancer without fully taking into account the relation between aggressive and costly interventions in the early stages of the disease and the outcome.

An accessible and reliable methodology for making health-related decisions can benefit every constituency in the health-care system, however, finding such methodology is by no means an easy task for multiple reasons. First, data collection, storage, and access to the data has not been a simple and economical task until recently. As a result, the research into the relation between data and outcomes has mainly focused on small data-sets with the disadvantage that the insights from each study have been specific to the data-set used in that study. Therefore, generalizability has been an issue. Second, the uncertainty in medical outcomes is not well understood, i.e., even in the presence of patient data, two patients with seemingly equivalent characteristics can have different outcomes. There are two reasons for such paradoxical observation. First, the data-set might not be rich enough, either because of missing values or missing important variables. Second, the models chosen to characterize the outcome using the data might not be the right models for explaining the relation between the data and the outcome.

This dissertation contributes to the literature on medical decision-making by proposing a framework that can enable patients, physicians, and policy makers to make more informed
decisions in the absence of sufficient data and the presence of uncertainty. As examples, two health-care applications, namely, predicting non-attendance at medical appointments and colorectal cancer mortality are studied.

With the first application, the underlying challenges are to find patients at risk of non-attendance at their next medical appointment and apply an optimal intervention to increase their chance of attending their appointment. While the application area is specific, the methodology developed to find optimal intervention policies is applicable to other health-care settings. Moreover, the approach to building prediction models for non-attendance to medical appointments used in this chapter enhances our understanding of the factors that can predict patient non-attendance.

The second application was chosen with the goal of increasing our understanding of one of the major causes of death from cancer, namely, colorectal cancer, the factors that can predict it, and how these factors can be used by patients, physicians, and policy makers in making informed decisions.

The methodology and approaches this dissertation uses and proposes can also be used to address other major problems in health-care such as predicting rehospitalization (30-day hospital readmission), chronic disorder variations, and hospital acquired infections.
Acknowledgment

No doctoral dissertation can be completed without those who support a student throughout the process of researching and writing it; and this dissertation is no exception. It is my pleasure to thank those who have supported me.

First and foremost, I am very thankful to my principal advisor, Professor Margrét Bjarnadóttir. I feel privileged that I have had the opportunity to work with her. Margrét has always been there for me with all the time and energy an advisor can possibly spend with a student. Her optimism and drive, very neat way of looking at applied problems, persistence, and intelligence have been invaluable assets for me. I am truly grateful for what she has done for me.

I also would like to thank my co-advisor, Professor John P. Weyant. Since I worked with John during my first year tutorial, he has always supported me during the different stages of my doctoral studies. I had many insightful conversations with John in these years, which have helped me navigate my way though the years of research. I am thankful for his support and his generosity when it came to spending time with me.

I thank Professor Ashish Goel for his continuing support. I have learned a great deal from Ashish by taking a first-year tutorial, assisting him with teaching, attending seminars that he organized, meeting with him to discuss my research, and being in several department committees with him.

I would like to thank the chair of my dissertation defense committee, Professor Victor R. Fuchs. Vic has been an exceptionally helpful guide for me. He has always been open to discuss research ideas and available. When I first contacted Vic, I never expected to hear back from him, given how busy he always is. I have been pleasantly surprised with his generosity in the past years.

I also would like to thank Dr. Kim F. Rhoads from the Stanford School of Medicine who was the principal investigator for the colorectal cancer research project. Kim has been very supportive of me and my work and has helped me immensely with the second half of my dissertation research. My gratitudes to Harold S. Luft from Palo Alto Medical Foundation who helped me in navigating to my second project on colorectal cancer mortality
and sharing his insights on possible interesting problems. I would also like to thank Dr. J. Sairamesh of 360Fresh Inc. for providing the data for the first part of this dissertation and making time to meet and discuss my work.

There are several faculty and staff at Stanford, Rutgers, and Sharif Universities that I am particularly thankful to. At Stanford, I would like to thank Professors Jerome H. Friedman and Trevor Hastie. I first learned data mining from Jerry’s excellent class, and gained many important insights through our meetings. Trevor generously shared his expertise in tailoring and implementing statistical algorithms for building my prediction models. I am thankful to Professor Ben Van Roy for the classes I took with him and for the opportunity to teach at Stanford myself. I would like to thank the chair of the Department of Management Science and Engineering at Stanford, Professor Peter W. Glynn, and the associate chair, Professor Stephan Barley. Peter and Steve have consistently worked on making the department a better place for students, and I deeply appreciate their efforts. From the staff at Stanford, my special thanks to Lori Cottle, the student services manager in Management Science and Engineering. Lori deeply cares about the students in the department and she goes beyond her responsibilities to make sure all the students receive the best experience she can offer while in the students services. At Rutgers University, I would like to thank my Masters studies advisor, Professor David W. Coit. I published my first research papers with David and thereby learned a lot from him. He also greatly supported me during my application process to Stanford. At Sharif University, I would like to thank Professors Rasoul Haji and Hashem Mahlooji, who have laid the groundwork for my later research work and thus opened many doors in my academic career.

The experience during my doctoral studies has been greatly enhanced by my colleagues and friends who were always there for sharing ideas and research brainstorming. I am especially thankful to Naveed Chehrazi, Yichuan Ding, Behdad Esfahbod, Krishnamurthy Iyer, Tim C. Kraft, Ehsan Mousavi, Sechan Oh, Shayan Oveis Gharan, Farnaz Ronaghi, Yanchong (Karen) Zheng, and Yu Wu. I am also thankful to my friends for making my life at Stanford a very enjoyable experience.

Lastly, I would like to thank my family. I cannot imagine writing these lines here today without the unconditional support that I have received from my parents, Jila Mansouridara and Mir Mohammad Hossein Zia. During my entire life, and specially in the past years
when I lived far from them, they have always been role models of sacrifice, generosity, and love for me. I deeply appreciate all they have done. I also would like to thank my brother, Alireza Zia, and his son, Barsam. Alireza has been a great support for me throughout my life and Barsam has brightened up many days of my life with his presence. Last but not least, I would like to thank my husband, Bernd Bandemer. Bernd and I met first in our second year at Stanford and since then, my appreciation for who he is has only grown. I thank him for his support, and for all the great things I have learned from him.
# Contents

Abstract \hspace{1cm} v  

Acknowledgment \hspace{1cm} vii  

1 Introduction \hspace{1cm} 1  

2 Non-attendance at Medical Appointments \hspace{1cm} 7  
   2.1 Introduction \hspace{1cm} 7  
   2.2 Optimizing Intervention/Prevention Policies \hspace{1cm} 10  
      2.2.1 Single Intervention Option \hspace{1cm} 11  
      2.2.2 Multiple Intervention Options \hspace{1cm} 16  
   2.3 Case Study: Optimal Intervention Policy in an Outpatient Clinic \hspace{1cm} 20  
      2.3.1 The Data and Prediction Models \hspace{1cm} 21  
      2.3.2 Telephone Reminders \hspace{1cm} 22  
      2.3.3 Optimal Intervention Policy and Sensitivity analysis \hspace{1cm} 24  
      2.3.4 Telephone and Text Reminders \hspace{1cm} 28  
      2.3.5 Variable Importance \hspace{1cm} 29  
   2.4 Discussion and Conclusions \hspace{1cm} 33  

3 Colorectal Cancer Mortality \hspace{1cm} 37  
   3.1 Colorectal Cancer Mortality Literature \hspace{1cm} 38  
      3.1.1 Descriptive Modeling \hspace{1cm} 39  
      3.1.2 Predictive Modeling \hspace{1cm} 40  
      3.1.3 Challenges of the Literature \hspace{1cm} 40  
   3.2 The Data \hspace{1cm} 41  
      3.2.1 Censoring \hspace{1cm} 43  
      3.2.2 Randomization \hspace{1cm} 43  
      3.2.3 Missing Values \hspace{1cm} 44  
   3.3 Prediction Models for Colorectal Cancer Mortality \hspace{1cm} 45
3.3.1 Mortality from the Time of First Hospital Discharge ............... 46
3.4 Variable Importance .................................................. 60

4 Conclusion ................................................................. 71

A ROC Curves and Generating the Family of Prediction Models ....... 73
B Non-attendance Study Variables ........................................ 77
C Colorectal Cancer Study Variables ...................................... 81

Bibliography ................................................................. 89
# List of Tables

2.1 Costs and benefits of a phone system intervention program. .................. 23

3.1 Training set, test set, and the rate of mortality ................................. 44
3.2 Top twenty variables from GUIDE; 30-day and 1-year mortality models. .. 52
3.3 Top twenty variables in GUIDE; 5-year mortality model ....................... 53
3.4 Top twenty important variables from relaxed lasso. ......................... 59
3.5 Ranked relaxed lasso variables as a function of penalty ...................... 65

C.1 Patient demographics and General Information ................................. 81
C.2 Socio-economic status variables .................................................. 82
C.3 Comorbidities ............................................................................. 83
C.4 Hospital discharge data ............................................................... 84
C.5 Hospital-specific information ...................................................... 85
C.6 Tumor Diagnosis ....................................................................... 86
C.7 Treatment .................................................................................. 87
C.8 Days Between Major Events ....................................................... 88
## List of Figures

2.1 Road map for finding optimal intervention policies ........................................ 9  
2.2 Cost function and ROC break-points ................................................................. 14  
2.3 Optimal mixed interventions without set-up costs ............................................. 18  
2.4 Optimal mixed interventions with set-up costs .................................................. 19  
2.5 ROC curves for predicting patient appointment cancellation ............................... 22  
2.6 Optimal classification tree for the case study ..................................................... 25  
2.7 Sensitivity analysis for intervention benefit ....................................................... 26  
2.8 Sensitivity analysis for intervention effectiveness ............................................... 27  
2.9 Sensitivity analysis for employee overhead costs .............................................. 28  
2.10 Optimal mixed intervention policy for the case study ......................................... 30  
2.11 Relative variable importance in predicting patient appointment cancellations; CART ......................................................... 31  
2.12 Relative variable importance in predicting patient appointment cancellations; MART .......................................................... 32  
3.1 The sequence of treatment for CRC patients ..................................................... 43  
3.2 ROC curves for 30-day, 1-year, and 5-year mortality using GUIDE ..................... 47  
3.3 AUC as a function of time horizon for classification tree models .......................... 48  
3.4 Classification trees for the case of equal mis-classification costs .......................... 51  
3.5 ROC curves for 30-day, 1-year, and 5-year mortality with relaxed lasso .............. 56  
3.6 AUC as a function of time horizon for lasso and relaxed lasso models ............... 56  
3.7 Relaxed lasso coefficient size as a function of time horizon ............................... 62  
3.8 Coefficient profiles for relaxed lasso ............................................................... 64  
3.9 Improvement and adjusted improvement in reclassification measures ............... 67  
3.10 Adjusted percentage of improvement in reclassification for relaxed lasso .......... 68  
3.11 Adjusted percentage of improvement in reclassification for type of admission and race ................................................................. 69  
A.1 Two ROC curves generated using two different approaches ............................ 75
In the past few decades, decision making in the context of health-care has become increasingly complex. The decision maker, whether a patient, physician, or policy maker, needs to usually make decisions in a resource constrained system where the outcome of decisions are not known deterministically. For example, a cancer patient should decide whether to receive an aggressive treatment over palliative care; an oncologist should decide what treatment options to offer to a patient; a policy maker should choose a subset of the population to apply a vaccination to, and so on.

There are multiple factors that contribute to the complexity of decision making in health-care settings. These include the structural complexity of the health-care system, interference of roles and responsibilities, and the availability of, sometimes numerous options. We explain each of these here.

**Complex Structure**

Prior to the introduction of employer-based health insurance, the health-care system was fairly simple in structure. Out-of-pocket payments were the main resource for covering the costs of treatment or diagnostic procedures by patients. Since the costs of diagnosing, treating, and controlling the disease had to be paid by physician and covered by patient, the two parties had personal incentives to keep the costs to a minimum while providing or receiving care.
The aforementioned health-care system changed significantly after World War II with the introduction of employer-based health insurance \cite{17} which covered a large number of members of the society. This is the health-care system as we know it today. There are two main characteristics that differentiate the current health-care system from the earlier one. First, in the new system, many patients and physicians enter decision settings in which the costs of care are covered by one or more health insurers. Access to health insurance means that the costs are shared among the patients in the pool of the health insurer. In this new setting, the effect of decisions does not stay local. An ineffective treatment procedure with high cost to a subset of patient population means higher premiums, taxes, etc. for the other patients in the health-care system. Given that the pool of resources, financial or otherwise, is often limited, offering ineffective treatment to a subset of patients could mean denying access to treatment to other patients who could benefit from the shared resources. For example, the diffusion of new innovations into oncology practice that are not clear to provide better patient outcomes might result in settings where the government should decide whether to pay for an experimental treatment for leukemia patients or programs to reduce premature birth \cite{99}.

The second characteristic which differentiates the two health-care systems is the introduction of many advanced treatments and diagnostic procedures. These procedures became available to patients and physicians with the rise of customized medicine and treatments \cite{67}. At first glance, having more options to choose from strikes as a positive development. However, the reality is that more options improve the decision making process only when the trade-offs introduced by choosing each one of them is fully understood. This is not the case for some of the treatments available to patients with chronic diseases today. For example, when it comes to cancer treatments, the new treatment options are more expensive and can potentially prolong a patient’s life time. However, in many cases, it is not clear how long the life-span of the patient can be increased and the trade-offs between longer life-span, quality of life, and costs are unknown.

**Introduction of New Roles**

Since today’s health-care system relies on cost sharing, there is a need for a more centralized decision making system which *chooses* some diagnostic or treatment procedures over
others. The need has led to the introduction of a new entity, namely, a policy maker. The policy maker sets guidelines for treatment and diagnosis, and ideally, the physicians follow these guidelines and provide feedback to the policy maker. One can observe that the role of physicians have changed in this new system. In the past, a physician’s role was to be the advocate of a single patient at the time of decision making [99]. Today, for high cost chronic diseases such as cancer, the physician’s role is moved toward making cost-based decisions while considering a balanced view of the needed care by a patient and the limited resources of the society. A national survey of medical oncologists shows that 56% of oncologists surveyed strongly or somewhat agreed with the statement that costs influence their decisions regarding which cancer treatments to recommend for their patients. 84% stated that patients’ out-of-pocket expenses have influence on their decisions. However, only 42% of the physicians felt that they are prepared to take cost-effectiveness information into their decisions [76].

**Introduction of New Options**

The past three decades have seen a significant change in terms of chronic disease prognosis, diagnosis, control and treatment. For example, cancer treatments used to be more aggressive, painful and disfiguring [23]. The treatments required blunt attack to malignant and healthy cells, aggressive surgical procedures which removed essential parts of the body organs, and more. Today, treatments have become more personalized and less aggressive. However, policy makers, physicians, and patients have to make decisions about which treatment to recommend, prescribe, or receive and each treatment comes with its own set of trade-offs. Many treatments, specially in the case of cancer care, are costly; They are sometimes ineffective or slightly effective and might threaten the patients with prolong life of painful side effects [10, 23, 93].

**High Costs and Questionable Quality of Care**

The United States operates the most expensive health-care system in the world [80]. The high cost of care on its own is not a case for concern given that the society might decide to
invest more on health-care to achieve better health outcomes [3]. However, the unsustainable growth in health-care costs [38] and the short-comings of the system on essential quality metrics have been a source of concern in the health-care community [97]. Patient safety has been a major issue in the health-care system. Medical errors happen routinely [33]. For example, two in every seven Medicare patients suffered from permanent or temporary harm caused during hospitalization while 44% of these errors could be prevented [61]. The rate of adverse events in hospitals is shown to be 33% of all admissions [25]. Many have called for urgent improvement in health-care quality [24, 81, 97].

Lack of (access) to reliable prediction models

Reliable data and prediction models are the basis for many decision making processes in health-care. A patient deciding on whether to choose aggressive cancer treatment over palliative care should know how much time she has left and whether the outcome can prolong her life span and by how much. A policy maker deciding on the subset of population to receive vaccination should find a series of characteristics which define a high risk member. Building prediction models for decision support in these settings require individual level data and robust models that reliably predict which members of the the population are at risk.

The use of prediction models in health-care applications has been limited partly due to lack of access to large data-sets until recently. Even in the recent years that access to large data-sets have become possible, research to build reliable prediction models has been fairly limited. Some applications such as 30-day hospital readmission and breast cancer outcomes have received a lot of attention while many other health-care applications are left out.

How are decisions made in the health-care system today? The answer to this question depends on who the decision maker is. In the past decade, some physicians have used cost-effectiveness research in their active clinical environments [99]. Policy makers have used cost-effectiveness analysis as well as more qualitative analysis, while balancing their decisions with political and social pressures [16]. Patients, however, have been left on their own in their decision making. They usually do not have access to the wealth of information physicians and policy makers have access to. Furthermore, they suffer from the lack of
adequate communication between patients and physicians [62]. A study of 1,057 patient encounters with 3,552 clinical decisions shows that only 9% of the clinical decisions could be considered informed medical decisions [18]. As a result, their decisions are mainly based on financial factors, quality of life, and emotions [103].

This dissertation addresses how access to data can help the three constituencies of patients, physicians, providers, and policy makers to make more informed decisions.

The main body of the dissertation consists of two chapters. Chapter 2 develops a methodology for making decisions about the optimal level and combination of intervention or prevention policies. To this end, chapter 2 uses the costs and benefits of an intervention, optimization methodology and prediction models to find an optimal intervention policy. The methodology developed is applied to find optimal intervention policies for reducing medical appointment non-attendance. Chapter 3 discusses prediction models for colorectal cancer mortality. Three modern statistical methods along with a large state-wide database of patients was used to build the prediction models. A detailed analysis of the models, how they can be used, and what they mean to patients, physicians, and policy makers are provided. Chapter 4 concludes the dissertation. There are three appendices appendices. Appendix A describes how a family of prediction models can be generated and represented by a receiver operating characteristic curve. Appendix B contains independent variables used in building prediction models for non-attendance at medical appointments case study in Chapter 2. Appendix C contains the list of independent variables used in building prediction models in Chapter 3.
Chapter 2

Non-attendance at Medical Appointments

2.1 Introduction

Healthcare spending in the United States reached $2.7 trillion in 2011, which corresponds to a health expenditure per capita of $8,508. Among OECD countries, Norway is second to the United States when it comes to health expenditure per capita with 50% less spending [80]. In the United States, the biggest categories of expenditure are hospital care and physician-and-clinic services comprising more than 50% of the overall costs, followed by prescription drugs and dental services. These high, and continuously rising, healthcare costs have been the focus of politicians, policy makers and researchers alike, resulting in wealth of research [35, 43, 44, 111], public debate and legislative efforts including the Patient Protection and Affordable Care Act.

With mounting pressure to reduce healthcare costs, risk modeling and cost-effectiveness analyses have become active areas of healthcare research. For example, risk modeling

I gratefully acknowledge Dr. J. Sairamesh from 360Fresh, Inc., Ravi Nemana from 360Fresh, Inc. and Dr. Keith Argenbright from Moncrief Cancer Resources, UTSW, and their support for accessing de-identified data for the NCI SBIR project Grant Number: 1 R43 CA141899-01 and Project Title: Early Warning Method and Technologies for Improving Cancer Care and Targeted Interventions. The project was supported by an active IRB (CR0000216/STU 082010-188) sponsored by UTSW for the NCI SBIR project. The content in this chapter is solely the responsibility of the author and does not necessarily represent the official views of the National Cancer Institute or the NIH.
and overutilization studies have focused on hospital readmissions [1, 14, 52, 52, 68, 105], medical errors [5, 30, 53, 69, 88] and overutilization [35, 54]. Cost-effectiveness studies have been applied to healthcare problems such as methadone maintenance treatment and HIV transmission [112], colorectal cancer screening and surveillance [60, 87], and the treatment of depression [107]. In his study, Tengs and his coauthors list five-hundred life saving interventions and their cost-effectiveness [101].

Examining the cost-effectiveness literature, one can verify that, by in large, the studies are focused on large population wide policy implementations without taking into account the heterogeneity of the population. These studies are, however, elaborate on considering the costs and benefits of the policies in making decisions. On the other hand, risk prediction and modeling studies are elaborate on considering the heterogeneity of the population but lack the cost and benefit considerations which can help these studies become useful in practice.

This chapter merges risk predictions and cost-effectiveness analysis, addressing how one can optimally translate risk predictions into intervention or prevention policies by merging risk prediction modeling with operational optimization. This approach is in contrast with earlier cost-effectiveness work that applies interventions, preventions, or treatments to the whole population, or large specific sub-populations. The framework applies to situation where risk models can be built based on individual characteristics, whether that is to predict 30 day hospital readmission, appointment non-attendance, or the onset of disease complications. In short, the proposed methodology will be effective in any setting where it is not cost-effective to apply intervention or prevention programs to the entire population, but it might be optimal to apply it to a targeted higher risk subset of the population.

Figure 2.1 depicts the high-level picture of the methodology developed in this chapter. Given a data-set of patient records and a prediction algorithm, a family of prediction models is generated to predict an outcome, for example, 30-day hospital readmission. The family of prediction models is generated by varying the trade-off between the cost of false-positive and false-negative errors and is represented by a ROC curve, commonly used in classification. The goal is to maximize the total benefit function for the intervention program or combination of programs available. The optimization formulation takes into account the
prediction characteristics through the ROC curve and the cost and benefit functions associated with each available intervention program. The optimal solution corresponds to an operating point on the ROC curve, and thereby the right scale of the program(s).

![Diagram](image)

**Figure 2.1.** Merging prediction models and cost-effectiveness studies with operational considerations.

This is the first study which merges prediction models and optimization beyond the linear cost and linear benefit of diagnostic testing studied in [98]. This study develops a methodology to address imperfect interventions, multi-layered interventions and combinations of interventions, all important factors in real world settings. The case study considers non-attendance to medical appointments, but the general results, managerial insights and the methodology developed can be extended to any actionable risk prediction setting.
2.2 Optimizing Intervention/Prevention Policies

This Section develops the methodology for finding an optimal intervention policy given an ROC curve and program’s operational costs and benefits. The notation and general set-up are introduced first, followed up by building an optimization model that maximizes the overall benefit when a single intervention option is available. The results are extended for the cases of multiple intervention options as well as multi-stage intervention options at the end of this section.

Let $p$ be the prevalence of the outcome (an underlying condition in the population or the event of interest), which the intervention or prevention program is aimed at limiting. Let $p_{ij}$ be the joint probability of predicting the outcome as $j$ when the true value of the outcome is $i$. As the outcome is a binary variable, $i$ and $j$ are in $\{0, 1\}$. Without loss of generality we label the “condition of interest” as 1, and the absence of it as 0. Let $f$ be an ROC curve that summarizes the classification performance of a family of prediction models, by displaying the trade-off between sensitivity and specificity. An ROC curve can be parameterized by $p_{01}$, the false positive probability, in which case the true positive probability can be written as $p_{11} = pf\left(\frac{p_{01}}{1-p}\right)$. Once a $p_{01}$ is selected, the scale of the intervention and the population selected for intervention has been determined. As $p_{01}$ increases from 0 (no one is selected and the intervention program is not applied to anyone) to $1-p$ (everyone is selected and the program is applied to everyone), more patients are selected for intervention, and therefore, the benefit gained increases. However, since the performance of most prediction models are far from the ideal ROC curve, the fraction of correctly identified members goes down and toward $\frac{p}{1-p}$ as the scale of the program is expanded.

Given an intervention, let $B$ be the benefit from intervention where $B$ is a function of the true-positive probability, $p_{11}$, i.e., there is reward for correctly identifying the population with the condition of interest. Furthermore, let $\mu$ represent the effectiveness of the intervention, measured as the fraction of the population with the condition of interest that the intervention is effective on. Let $C$ be the cost from the intervention, which applies to the entire population the intervention is applied to, both the false positives $p_{01}$ as well as the true positives $p_{11}$. Let $G$ be the total benefit function incorporating all costs and benefits of

---

Please refer to Appendix A for details on ROC curves and generation of family of prediction models.
an intervention as well as the ROC curve.

Using the above notation, the total benefit function can be written as

\[ G(p_{01}) = B(\mu p_{11}) - C(p_{01} + pf(\frac{p_{01}}{1-p})). \]  

(2.1)

The first order optimality condition for \( G(p_{01}) \) is

\[ B'(\mu pf(\frac{p_{01}}{1-p}))\mu pf'(\frac{p_{01}}{1-p}) \frac{1}{1-p} - (1 + \frac{p}{1-p}f'(\frac{p_{01}}{1-p}))C'(p_{01} + pf(\frac{p_{01}}{1-p})) = 0 \]

where the left hand side of the above equation is the derivative of \( G \) with respect to \( p_{01} \).

Solving the above for \( p_{01} \): \n
\[ f'(\frac{p_{01}}{1-p}) = \frac{1-p}{p} \frac{C'(p_{01} + pf(\frac{p_{01}}{1-p}))}{\mu B'(\mu pf(\frac{p_{01}}{1-p})) - C'(p_{01} + pf(\frac{p_{01}}{1-p}))}. \]  

(2.2)

Note that the left hand side of Equation (2.2) is the slope of the ROC curve. Whether the local maxima found through Equation (2.2) are a global maxima depends on the cost and benefit functions \( C \) and \( B \) which is explored in the following section.

### 2.2.1 Single Intervention Option

Consider a single intervention with corresponding cost and benefit per intervention. The goal is to find the optimal subset of the population to apply the intervention to unless the intervention does not result in non-negative total benefit in which case the intervention program will not be used at all. The case of linear cost and benefit functions are considered first.

**Linear Cost and Benefit**

Consider an intervention with linear cost and benefit, i.e., the cost is directly proportional to the number of patients the intervention is applied to, and the benefit is directly proportional to the number of correctly identified members with the condition. Let the cost of intervention per patient be \( c_0 \), independent of whether the intervention is applied to the
false-positive or true-positive population. For every correctly identified patient at risk, let the benefit be \( b \). Let \( \mu = 1 \), representing a perfect intervention.

Using equation (2.1) and the above assumptions, the objective function can be written as

\[
\max_{p_{01}} G(p_{01}) = N(b p_{01} \frac{p_{01}}{1-p} - c_0 (p_{01} + p f(\frac{p_{01}}{1-p}))),
\]

where \( N \) is the total population size. Following equation (2.2), the first order optimality condition is

\[
f'(\frac{p_{01}}{1-p}) = \frac{c_0}{b - c_0} \frac{1-p}{p}.
\]

It is easy to show that the second order optimality condition guarantees a concave \( G \) when \( b > c_0 \). Using this fact and equation (2.4), the optimal operating point corresponds to the point on the ROC curve where the slope equals \( \frac{c_0}{b - c_0} \frac{1-p}{p} \), if a such point exists. We call this ratio the critical ratio.

Observe that as the cost, \( c_0 \), increases, the critical ratio corresponding to the slope of the ROC increases. This translates to smaller values of \( p_{01} \), and hence the intervention is applied to a smaller subset of the population. For a fixed cost \( c_0 \), as the benefit \( b \) increases, \( f' \) in equation (2.4) decreases and hence, intervention is applied to an increasing fraction of the population. Moreover, as the prevalence \( p \) increases, \( f' \) in equation (2.4) decreases, resulting in increased intervention rate.

The result presented above was first introduced in the context of diagnostic tests by Sox et al. [98]. The goal of their study was to find an optimal threshold or cut-off point for a diagnostic test result by maximizing the total expected utility of a diagnostic test. A physician would then use the optimal threshold to label a patient as “at-risk” or not depending on whether a patient’s test result is beyond that threshold or not. But what if there is no point on the ROC curve for which the slope of the point is equal to the critical ratio?

Sox et al. did not address this question. However, as we will show, this is a critical question to address since whether the critical ratio matches the slope of the ROC curve at some point \( p_{01} \) depends on the unit cost and benefit of intervention as well as the quality
of the ROC curve itself. Therefore, I suggest the following update to the optimal operating point of the ROC curve, $p_{01}^*$,

$$p_{01}^* = \begin{cases} 
1 - p, & b > c_0, \frac{c_0}{b-c_0} \frac{1-p}{p} < d_{\text{min}} \\
\{ p_{01} | f'(\frac{p_{01}}{1-p}) = \frac{c_0}{b-c_0} \frac{1-p}{p} \}, & b > c_0, d_{\text{min}} < \frac{c_0}{b-c_0} \frac{1-p}{p} < d_{\text{max}} \\
0, & b > c_0, \frac{c_0}{b-c_0} \frac{1-p}{p} > d_{\text{max}} \\
0, & b < c_0 
\end{cases}$$

where $d_{\text{min}}$ and $d_{\text{max}}$ are the minimum and maximum slope of the ROC curve at point (1,1) and (0,0), respectively. Case (i) in the above equation holds when the critical ratio is smaller than $d_{\text{min}}$ which is equivalent to $b > c_0 (1 + \frac{1-p}{pd_{\text{min}}})$. In this case, it is optimal to apply the intervention to everyone given that the benefit of the intervention significantly outweighs the cost. Note that $d_{\text{min}}$, the slope of the ROC curve at point (1,1), is generally a small number close to zero. In case (ii), the critical ratio lies between $d_{\text{min}}$ and $d_{\text{max}}$, and the optimal policy is to apply the intervention to the sub-population corresponding to the point on the ROC curve where the slope is the critical ratio. Cases (iii) and (iv) correspond to when it is not cost efficient to implement the intervention, which is the case when $b \leq c_0 (1 + \frac{1-p}{pd_{\text{max}}})$ or $b < c_0$.

The above set of equations is the building block of the further discussions in this chapter, when we move to more complicated cost structures.

**Piecewise Discontinuous Linear Cost and Linear Benefit Functions**

In most applications, the cost function is not linear. There are two sources of non-linearity in the cost function. First, many intervention programs require some set-up costs. Second, economy of scale or dis-economy of scale might be applicable to the costs, i.e., The per patient intervention cost may increase or decrease as the intervention coverage increases depending on the application. For example, increasing the number of children immunized is generally associated with a decrease in cost per child, however in rural areas (for example in Tanzania) the increasing costs of transportation and administrations cause diseconomies of scale, leading to increased costs [21]. This section extends the methodology to find the
optimal operating point of an ROC curve for a piecewise discontinuous linear cost function and a continuous linear benefit function, capturing these operational realities.

Assume the benefit function, $B$, is linear in $p_{11}$ with the coefficient $\mu b$ where $\mu$ is the effectiveness of the intervention and $b$ is the per patient benefit. Following the notation in equation (2.1), let the cost function be $C(p_{01} + pf(p_{01} / (1-p)))$ where $C$ is a piecewise linear function of the sum of $p_{01}$ and $p_{11}$ with $m$ break-points. Note that if there are additional costs associated with correct identification, they could be incorporated in a linear manner. Figure 2.2 shows a sample cost function and a ROC curve.

![Figure 2.2. Discontinuous piecewise linear cost function and monotone increasing concave ROC function. Every break-point in the cost function corresponds to a point on the ROC curve.](image)

The cost function in Figure 2.2 is parameterized by the variable costs $c_1, c_2, \ldots, c_m$ and the set-up costs $k_1, k_2, \ldots, k_m$. Let $N_i$ be the total number of patients who received intervention by the end of interval $i$. If the intervention is initiated, there is a one-time set-up cost of $k_1$ and for the first $N_1$ patients there is a cost of $c_1$ per patient intervened. When the intervention is applied to more than $N_1$ patients, the system incurs a variable cost of $c_2$ for each of the next $N_2 - N_1$ patients who receive the intervention along with a scale-up cost of $k_2$. For example, this occurs when additional personnel need to be trained as the capacity of the intervention program is extended.

Corresponding to each break-point of the cost function in Figure 2.2 is a point on the ROC curve. For example, $N_1$ corresponds to $p_{01} / (1-p)$ where $p_{01} / (1-p)$ is the false-positive rate corresponding to the end point of interval 1. In general, the end point of interval $i$ on the ROC
curve is found by solving for $p_{01}^i$ in

$$p_{01}^i + p f\left(\frac{p_{01}^i}{1 - p}\right) = \frac{N_i}{N}.$$  

It is easy to see that there are $m + 1$ intervals on the ROC curve corresponding to the $m + 1$ intervals in the cost function. To find the global optimal operating point on the ROC curve, each interval is treated independently and a local optimum is found, applying the methodology of section 2.2.1. The global optimal operating point, $p_{01}^*$, is the maxima over the local optima. Recall the total benefit function

$$G(p_{01}) = B(\mu pf(\frac{p_{01}}{1 - p})) - C(p_{01} + p f(\frac{p_{01}}{1 - p})), $$

and let

$$B(\mu pf(\frac{p_{01}}{1 - p})) = b \mu N p f(\frac{p_{01}}{1 - p}), $$

and

$$C(p_{01} + p f(\frac{p_{01}}{1 - p})) = \begin{cases} 
  c_1 N(p_{01} + p f(\frac{p_{01}}{1 - p})) + k_1 \mathbf{1}(p_{01} > 0), & p_{01} \leq p_{01}^1 \\
  c_i N(p_{01} + p f(\frac{p_{01}}{1 - p})) + C C_i + K_i, & p_{01}^{i-1} < p_{01} \leq p_{01}^i, \\
  & \forall i \in \{2, \ldots, m\}. 
\end{cases}$$

where $K_i$ is the cumulative set-up cost

$$K_i = \sum_{j=1}^{i} k_j,$$

and $CC_i$, a cost correction term, is

$$CC_i = \sum_{j=1}^{i-1} (c_j - c_i)(N_j - N_{j-1}).$$

For each interval $i$, define $\hat{p}_{01}^i$ to be the value of $p_{01}$ satisfying the critical ratio of that
interval.

\[ \hat{p}_{01}^i = \left\{ p_{01} \mid f'(\frac{p_{01}}{1 - p}) = \frac{1 - p}{p} \frac{c_i}{\mu b - c_i} \right\}, \quad (2.5) \]

and define \( p_{01}^{*i} \) for interval \( i \) as

\[
 p_{01}^{*i} = \begin{cases} 
 p_{01}^i, & \mu b - c_i \geq 0, \frac{1-p}{p} \frac{c_i}{\mu b - c_i} < d_i \\
 \hat{p}_{01}^i, & \mu b - c_i \geq 0, d_i \leq \frac{1-p}{p} \frac{c_i}{\mu b - c_i} \leq d_{i-1} \\
 p_{01}^{i-1}, & \mu b - c_i \geq 0, \frac{1-p}{p} \frac{c_i}{\mu b - c_i} > d_{i-1} \\
 p_{01}^{i-1}, & \mu b - c_i < 0,
\end{cases} \quad (2.6)
\]

where \( d_{i-1} \) and \( d_i \) are the slopes of the ROC curve at the break-points defining interval \( i \).

Note that the \( p_{01}^{*i} \) are found independent of start and scale up costs. The global optimal operating point of the ROC curve is found by

\[
 p_{01}^* \in \{ p_{01} \mid G(p_{01}) = \max \{0, G(p_{01}^{*1}), G(p_{01}^{*2}), \ldots, G(p_{01}^{*m+1}) \} \}. \quad (2.7)
\]

In the next section, this result is extended to the case where there are multiple intervention options available and the decision maker should choose what percentage of the population should receive what type of intervention, if at all.

### 2.2.2 Multiple Intervention Options

In reality, a healthcare organization may be considering a menu of intervention options, two-stage interventions, or combinations of different intervention options. For a simple choice between mutually exclusive intervention options, a comparison of the optimal values of each intervention benefit function will result in an optimal policy selection. However, when mixing of interventions or multi-stage interventions are considered, more analysis is needed to find the optimal policy. This section presents the results for these two cases. For presentation simplicity, \( B \) and \( C \) are assumed to be linear in this section. However, the results presented here are easily extendable to nonlinear cost cases.
Multi-Stage Interventions

Multi-stage intervention is an intervention in which every stage of intervention identifies potential candidates for the next stage. The most common multi-stage intervention is the two-stage intervention. For example, in the context of adherence to medical appointments, the first stage of an intervention may involve a phone call reminder. In some cases, the phone call reveals that further action is needed for the patient to attend her appointment. For example, transportation to and from the medical center may have to be arranged in the secondary intervention stage.

Let \( \eta \) be the additional effectiveness of the intervention if an extra cost \( c_\eta \) is spent on top of the initial intervention cost for a subset of the true positive population, and let \( c_1 = c_0 + c_\eta \) be the total cost of the intervention for those true positives that need the two step intervention. Let \( \mu \) be the effectiveness of the intervention at the first stage. The benefit function \( B \) is updated to

\[
B((\mu + \eta)p f\left(\frac{p_{01}}{1-p}\right)) = N(\mu + \eta) b p f\left(\frac{p_{01}}{1-p}\right)
\]

Since the cost \( c_\eta \) is only applied to a subset of the true-positive population, the cost function is

\[
C = N(c_0 p_{01} + (\mu c_0 + \eta c_1) p f\left(\frac{p_{01}}{1-p}\right)).
\]

The first-order optimality condition for maximizing the total benefit function \( G \) results in a critical ratio of

\[
f'(\frac{p_{01}}{1-p}) = \frac{1-p}{p} \frac{c_0}{(\mu + \eta) b - (\mu c_0 + \eta c_1) - c_0}.
\]

The critical ratio can again be used to find the optimal intervention policy. More generally, for a \( k \) step intervention policy where \( \eta_k \) is the additional effectiveness gained by step \( k \) and \( c_k \) is the total cost for the intervention up to and including step \( k \), the critical ratio is

\[
f'\left(\frac{p_{01}}{1-p}\right) = \frac{1-p}{p} \frac{c_0}{(\mu + \sum_{i=1}^{k} \eta_i) b - (\mu c_0 + \sum_{i=1}^{k} \eta_i c_i) - c_0}.
\]
Mixed Interventions

If multiple intervention options are available, it may not be optimal to limit the intervention policy to a single strategy. Instead one can apply different interventions to different sub-sets of the population. As an example, it may not be cost efficient to have a nurse call everyone that has an appointment; however, it may be cost efficient to call the patients identified at high risk of non-attendance, and send a text message to those at a lower risk of missing their appointments.

Consider $P_1$ and $P_2$ as two intervention options. Let $P_1$ be a high-cost, high reward policy compared to $P_2$, i.e., $c_1 > c_2$ and $\mu_1 > \mu_2$ where $c_i$ and $\mu_i$ are the cost and effectiveness of policy $i$, respectively. Let $G_1(p_{01})$ and $G_2(p_{01})$ be the net benefit function for $P_1$ and $P_2$, respectively. Figure 2.3 shows an example of possible benefit functions for the two interventions, labeled by $G_1$ and $G_2$. For this example, the optimal intervention policy is to apply $P_1$ up to $p_{s01}$, the false-positive probability at the switching point, and $P_2$ up to $p_{01}^2$. For any $p_{01} > p_{01}^2$, no intervention is applied. The optimal net benefit, $G^*$, is shown in Figure 2.3(a).

![Figure 2.3](image)

**Figure 2.3.** Figure (a) shows $G_1$, $G_2$, and $G^*$. Figure (b) shows the derivative of $G_1$ and $G_2$.

In general, for a continuous intervention cost function for $p_{01} > 0$, if a switching point, $p_{01}^s$, exists where $G'_1(p_{01}) > G'_2(p_{01})$ for all $p_{01} < p_{01}^s$ and $G'_1(p_{01}) < G'_2(p_{01})$ for all $p_{01} > p_{01}^s$, it is optimal to apply $P_1$ up to $p_{01}^s$ and switch to $P_2$ for all $p_{01}$ between the switch point and the optimal level of intervention under $P_2$, $p_{01}^{2s}$. Such switching point is the $p_{01}$
2.2. Optimizing Intervention/Prevention Policies

satisfying

\[ G'_1(p_{01}) = G'_2(p_{01}) \Rightarrow f'(\frac{p_{01}}{1-p}) = \frac{1-p}{b(\mu_1 - \mu_2) - (c_1 - c_2)} \]  

(2.8)

In this case, the switch point, \( p_{01}^s \), is the point beyond which the marginal rate of net benefit from the cheaper intervention, \( P_2 \), outweighs the marginal benefit from the more expensive intervention, \( P_1 \).

Note that the existence of start-up costs does not directly affect the switch point. This is shown in Figure 2.4. Rather, they will impose conditions on the overall benefit gain. In other words, in the presence of a set-up cost for \( P_2 \), if a switching point \( p_{01}^s \) exists, the switch to \( P_2 \) is optimal if the following two conditions hold:

\[
\begin{align*}
G_1(p_{01}^s) + G_2(p_{01}^{2*}) - G_2(p_{01}^s) - k_2 &> G_1(p_{01}^{1*}) \\
G_1(p_{01}^s) + G_2(p_{01}^{2*}) - G_2(p_{01}^s) - k_2 &> G_2(p_{01}^{2*})
\end{align*}
\]  

(2.9)

The above condition assures that the benefit from the mixed policy is greater than the benefit from the single policy at optimum. Together, the two conditions ensure that using a mixed policy is better than using any of the individual policies \( P_1 \) and \( P_2 \).

In the next section, we apply the methodology developed in this section to find optimal intervention policies for the problem of non-attendance at medical appointments.

Figure 2.4. This figure shows \( G_1 \), \( G_2 \), and \( G^* \) in the presence of set-up costs \( k_1 \) and \( k_2 \). The value of \( G^* \) drops at the break-point, \( p_{01}^s \), equal to the value of \( k_2 \), the set-up cost for \( P_2 \).
2.3 Case Study: Optimal Intervention Policy in an Outpatient Clinic

This section applies the methodology developed earlier in this chapter to find the optimal intervention policy to reduce the patient appointment cancellation rate in an outpatient clinic.

Appointment cancellation and no-shows are a part of a larger non-adherence issue in the healthcare system. In the United States, reported appointment non-attendance (also known as non-compliance or non-adherence) rates differ by the type of practice and disease, but ranges from 6% to over 50% [55, 72]. Non-attendance is undesirable for many reasons. The patient who misses an appointment suffers from discontinuity of care, worsening of medical outcomes, increased morbidity risk, and more frequent emergency room visits and hospitalizations [13, 74]. For the healthcare organization, non-attendance means higher expenses due to rescheduling costs and decreased revenue from empty time slots. The revenue decrease associated with non-attendance ranges from 3% to 14% of overall revenue taking into consideration both the administrative overhead costs and the direct loss of revenue [72].

To the best of my knowledge, there are no cost-effectiveness studies evaluating the different intervention options. How cost effective an intervention is, is highly dependent on the population it is applied to. The current study directly addresses this issue by optimally selecting the right subgroup to maximize the overall benefit to the decision maker. In contrast to cost effectiveness studies, a multitude of personal reasons for and characteristics contributing to a patient missing an appointment are well studied. Characteristics such as young, male, of lower socio-economic status, without insurance and/or divorced are all associated with higher probability of missing appointments [6, 7, 27, 39, 66, 71, 73, 78, 95, 96]. Patients with a history of failed appointments, who have been waiting a long time for their appointments and have a hard time getting time off from work are known to be more likely to miss appointments [6, 7, 36, 39, 73, 95]. Surveys report patients citing numerous reasons for not showing up, including forgetfulness, lack of transportation, financial issues, lack of child care and changes in medical conditions [6, 7, 27, 39, 71, 95].

I apply the optimization methodology to the case of appointment non-attendance at the
University of Texas Southwestern (UTSW) outpatient clinic, and explore the cost effectiveness of intervention programs generally applied to appointment no-shows. Through this process, insights into the most important variable affecting appointment non-attendance are gained.

2.3.1 The Data and Prediction Models

The data-set consists of visit records for 23,647 patients who had appointments between January 3, 2005 and May 1, 2011, a total of 592,470 visit records. Over 66% of the patients have been diagnosed with cancer, and 19% of the patients have at least one cancer treatment at the clinic. On average, there were 315 daily scheduled patient visits and the clinic experienced an appointment non-attendance rate of about 24%, which caused serious scheduling issues, in addition to the negative effects on patients, staff and the clinic as previously discussed. If a patient misses her appointment, the appointment is considered “canceled when the patient informs the clinic ahead of time; it is labeled “no-show” otherwise. Given the high rate of cancellation in the outpatient clinic versus no-show (22% versus 2%), I focus on appointment cancellations for the rest of this section. However, it should be noted that the predictive performance and the insights gained are comparable.

The prediction models are built using a randomly chosen training set consisting of 80% of the data (471,082 patient visit records). The performance of the prediction models is evaluated on the remaining 20% of the data, the test set. A total of 104 independent variables were considered when building the prediction models. The variables can be categorized into five categories: those capturing treatment information, severity of the disease, information about the day of the visit, general and demographic information, and patient’s commitment to past appointments. Appendix B provides a complete list of variables.

Number of interpretable prediction algorithms were employed, including logistic regression (LR), regularized Logistic Regression (LASSO) [102], multiple additive regression trees (MART) [41, 42], multivariate adaptive regression splines (MARS) [40], and Classification trees (CART) [19]. The ROC curves were created by directly minimizing the cost-ratio of false positives and false negatives as opposed to varying a cut-off value.

For details please refer to Appendix A.
The area under the ROC curve (AUC) for CART is 0.812, which compares favorable with the AUC of the LR 0.741, Lasso 0.786 and MARS 0.781. MART outperformed the CART algorithm with AUC of 0.823, however, CART was selected for the study both due to the ease of interpretability of its results which is important in healthcare settings, and, in addition, because of ease of embedding classification tree decision rules in any software. Figure 2.5 shows the ROC curves for the five prediction algorithms applied.

![ROC curves for the five prediction algorithms](image)

**Figure 2.5.** ROC curves corresponding to the five prediction algorithm applied.

### 2.3.2 Telephone Reminders

The health services literature has studied the effectiveness of different interventions for reducing non-attendance. Clinical reminders, whether in the form of computer generated messages, personal telephone reminders or mail have been shown to increase attendance rates [66, 92]. For example, in a study of an academic outpatient practice, computer-generated telephone appointment reminders decreased the no-show rate by 25%, while
staff-operated appointment reminders decreased this rate by up to 41% [83]. The effectiveness of text message appointment reminders is reported to be similar to automated telephone reminders [94]. Providing information to new patients has been shown to be an effective intervention for that patient group [6, 50, 66], reducing the non-attendance rate from 15% down to single digits [50]. Financial disincentives can also reduce non-attendance [8].

For the case study we consider a year long intervention program. To reduce the non-adherence rate, the clinic considers an intervention policy through which a nurse will call patients with an appointment reminder. Using the above approach, we identify the right scale of the program, including which patients to call and how many nurses to train, in order to maximize the total intervention benefit. Table 2.1 summarizes the cost associated with the intervention program as well as the benefit gained by the clinic when the intervention is effective.

<table>
<thead>
<tr>
<th>Cost and Benefit of Phone Call Intervention</th>
<th>Value (in dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One time system set-up cost</td>
<td>$100,000</td>
</tr>
<tr>
<td>One day training cost (per nurse)</td>
<td>$394.56</td>
</tr>
<tr>
<td>Phone call cost (per patient)</td>
<td>$7.54</td>
</tr>
<tr>
<td>Profit (per attended appointment)</td>
<td>$107</td>
</tr>
</tbody>
</table>

Table 2.1. Costs and benefits of a phone system intervention program.

The one time set-up cost includes the cost of setting up a system with embedded decision rules for appointment non-attendance on a daily basis, setting up phone lines, and initializing a training program for the nurses who will be calling the patients. The training program is a one day paid program. The hourly employment compensation and direct costs for a nurse working in a clinic is $49.32 [22]. This cost includes nurse’s wage, benefits, insurance and other costs incurred to the employee. The overhead costs, such as the cost of electricity, office space, etc., are assumed to be 30% of the compensation costs. Therefore the hourly cost of a nurse for the clinic is estimated to be $64.12. The cost per nurse for the one day training program is $394.56, equivalent to eight hours of nurse’s pay. We assume that the average phone call time is 7 minutes including all pre and post processing time. As a result, the cost per phone call is $7.54. If the intervention is effective, the profit gained by an attended appointment is estimated to be $107. The estimation is computed based
on the average outpatient gross revenues of teaching hospitals and clinics in California in 2007 [82] with similar number of outpatient visits as UTSW, i.e., the revenue from hospitals and clinics with 100,000 to 200,000 outpatient appointment per year. A similar data is not available for the UTSW outpatient clinic. A profit margin of 6% was considered for this year [104]. Finally, a conservative effectiveness rate of 25% was chosen for the telephone reminders. This rate is chosen based on the research that shows the effectiveness of telephone reminders on decreasing non-attendance is anywhere from 25% to 93% [50, 83].

### 2.3.3 Optimal Intervention Policy and Sensitivity analysis

A natural cubic spline was fit to the CART ROC curve and the optimal scale of the telephone reminder program was computed using the methodology in section 2.2.1. The optimal operating point (0.22, 0.63) corresponds to applying the intervention to 31% of the patients, which requires 6 part-time nurses. Without the intervention program, 22% of the 315 daily appointments are canceled. With the intervention program at the scale recommended by the optimal solution, 25% of the intervened cancellations are aborted (based on the estimate of $\mu$), or just over 17 appointments, with an estimated net benefit of $16.2 per day. Figure 2.6 shows the first three levels of the classification tree corresponding to the optimal operating point. The complete classification tree contains more than 50 nodes.

Given the level of uncertainty of the estimated profit per appointment, the effectiveness of the intervention, and the overhead cost, we chose to run sensitivity analysis on these parameters.

**Intervention Benefit, $b$**

Keeping all other parameters constant, we vary $b$, the benefit from the intervention, between $100 and $500. The lowest profit for which the program is implemented is $106. The benefit per avoided cancellation required for the intervention to be applied to everyone is $273,881. This high number is reflective of the flat shape of the ROC curve close to (1,1). In this area, we are able to correctly identify 30% of those who will attend at their appointments, with only a few mistakes. The upper bound was set to $500.

Figure 2.7 shows the fraction of the population included in the program, as well as the
2.3. CASE STUDY: OPTIMAL INTERVENTION POLICY IN AN OUTPATIENT CLINIC

Figure 2.6. The first three levels of the optimal classification tree for the appointment cancellation problem. The green node corresponds to the patients predicted to attend their appointment while the red node corresponds to those predicted to cancel. All terminal nodes should be colored as green or red. The white terminal nodes indicate that the branching continues at that node. The numbers next to each terminal node show the number of those who attend and those who cancel their appointments, respectively. The splitting variable along with the split criteria is shown in every middle node.

The preliminary treatment phase is indicated by the absence of a branch. The number of physician visits during the last clinic visit is 1269. The number of cancellations during the last clinic visit is 125205. The fraction of previous reschedules is less than 13%. The number of previous cancellations is less than or equal to 2.5.

Overall benefit as a function of $b$. The dashed lines in (a) mark the fraction of the intervened population beyond which an extra nurse is needed to expand the intervention program. The first six nurses are added as soon as the benefit reaches $106. The seventh nurse is added when the benefit reaches $110. At this point, 33% of the population is included in the program. As expected, the fraction of patients included is a nondecreasing function of the benefit. The flat regions prior to jumps represent thresholds, where a larger increase in the benefit is required for it to be economical to add an additional nurse. We note that as long as the benefit $b$ is at least $195$, it is optimal to contact over 60% of the patients.

The intervention program will be implemented if the optimal net benefit value $G^*$ is non-negative. As shown in Figure 2.7, part (b), $G^*$ exhibits an almost linear relationship with $b$ for larger values of $b$. Since the incremental set-up costs are low, i.e., $1.58 per nurse per day, the discontinuities in the overall benefit function are not visible in the figure. Even
for the initial start-up cost which is large, the effect of the start-up is heavily dominated by the benefit of applying the intervention to 31% of the population. The linear behavior of the net benefit function is expected when looking at the overall benefit function, $G$. As $b$ increases, the critical ratio in Equation 2.5, decreases. We know that the overall benefit function in each interval $i$ is

$$G(p_{01}) = N((\mu b - c_i)p f(\frac{p_{01}}{1-p}) - c_i p_{01} - CC_i - K_i$$

If $b$ is large enough such that $\mu b - c_i$ is large, $G^*$ is almost linear in $b$ given that $f(\frac{p_{01}}{1-p})$ is large. Note that $f(\frac{p_{01}}{1-p})$ is large since the point at which the slope of the ROC is small corresponds to the point with high true-positive value.

![Figure 2.7](image)

**Figure 2.7.** Sensitivity with respect to the intervention benefit $b$. The dashed lines in (a) mark the fraction of the intervened population beyond which an extra nurse is needed to expand the program.

**Intervention Effectiveness, $\mu$**

Keeping all other parameters constant, we vary $\mu$, the effectiveness of the intervention, between 1% and 100%. When $\mu = 1$, 69% of the population are contacted. The lowest
value of $\mu$ for which the program is run is 24.7%, only slightly lower than the conservative 25% effectiveness considered in the base case. The sensitivity of the fraction of the patients included and the optimal intervention benefit is shown in Figure 2.8, as a function of $\mu$. The sensitivity analysis of $\mu$ follows closely the sensitivity analysis of $b$; as the benefit of intervention is the product of these two values, altering one, but not the other has the same effect.

![Figure 2.8.](image)

**Figure 2.8.** Sensitivity with respect to intervention effectiveness $\mu$. The dashed lines in (a) mark the fractions beyond which a new nurse needs to be added to expand the program.

### Employee Overhead Costs

Keeping all other parameters constant, we vary the overhead cost from 0% to 100% which affects the variable cost of the program, i.e., the cost per phone call, $c_i$.

As the overhead cost varies from 0% and 100%, the scope of the program decreases from 49% to 0% as seen in Figure 2.9 (a). When the overhead cost is 0%, 9 nurses call 49% of the population. As the overhead cost increases, fewer patients are called. When the overhead cost reaches 8%, the training cost of the 9th nurse no longer outweighs the benefit gained and therefore, the optimal scope of the program reduces to 8 nurses. The drop in
Figure 2.9 (a) represents the drop in the number of patients contacted, once it is no longer economical to have the 9th nurse. The horizontal parts of the curve, for example, between overhead costs of 8% and 13%, is the range for which the nurses work at full capacity. When the overhead cost is 34% or more, the program is not implemented.

Figure 2.9 (b) shows an almost linear relation between the optimal net benefit function and the overhead cost up to overhead cost of 34%. Since the training cost compared to the optimal net benefit is small, the function appears continuous for the range of overhead cost values plotted. At 34%, since the program is not run anymore, the optimal total benefit is 0.

![Figure 2.9. Sensitivity with respect to employee’s overhead costs.](image)

### 2.3.4 Telephone and Text Reminders

Suppose that the choice of interventions is between telephone reminders, text reminders, or a mixture of the two. Sending text messages to patients is a cheaper but less effective intervention when compared to telephone reminders. As part of the problem assumptions, the effectiveness of text messaging is 19% and its cost is $0.05 per text message, with an initial set-up cost of $10,000. Using Equation (2.8) and checking for the conditions in
Equation (2.9), it becomes clear that telephone reminders should never be used when text messaging is possible. In this setting, it is optimal to send text reminders to 92% of the population.

What should be the per patient benefit from the intervention for mixing of telephone and text reminders to be optimal? To answer this question, a sensitivity analysis on the unit benefit, $b$, was performed where $b$ was varied by a unit at each step. The smallest value of $b$ for which mixing is optimal was identified to be $227. For this value of $b$, the optimal intervention policy is to apply text messaging to the first 3.4% of the population. At this point, the total benefit from telephone reminders becomes positive and since the slope of the total benefit function for telephone reminder is more than text messaging, it is optimal to remind the next 4% of the population by telephone, i.e., up to 7%, which is the switch point for the two benefit functions. At this point, text messaging becomes one more time optimal. The next 85% of the population will receive text messages, up to and including 92.1%. The rest of the population does not receive any reminder. Figure 2.10 shows the two total benefit functions.

### 2.3.5 Variable Importance

By finding the important variables in predicting appointment cancellations, policy, medical, and operational insights can be gained about which variables should be focused on when making decisions. In this section, the reduction in the loss function, i.e., mean squared error, is used as the measure of importance of each independent variable. The corresponding value for each variable is then normalized such that the most important variable has value equal to 100. Figures 2.11 and 2.12 show the relative variable importance ranking for MART and CART when the costs of false-positive and false-negative are equal. For MART, only the variables with relative importance of more than 40 are included while for CART, variables with lower importance measures are also presented.

As shown in the figure, the previous appointments’ history plays an important role in predicting appointment non-attendance. Fraction of previous appointments canceled, fraction of previous appointments canceled during the last visit day, number of cancellations during the last visit day, number of physician appointments the patient had during the
last visit day, and so on are all important in predicting appointment cancellation, i.e., the history of patient behavior can predict their future behavior. Two of the other important variables are minimum and maximum phase of treatment. These variables refer to the minimum and maximum phase of treatment the patient is at (preliminary or advanced) across all treatments the patient has previously received. MART also measures some variables as important that CART does not. For example, distance from the patient’s home to the clinic, age of the patient, appointment day of month, and appointment day of year are also important in predicting cancellations using MART.

Figure 2.10. The benefit functions corresponding to telephone reminders and text messages as two intervention policies. The optimal mixed strategy is to use text reminders upto 3%, telephone reminders upto 7%, text reminders upto 92.1% and none beyond that point.
Maximum phase of treatment
Minimum phase of treatment
Fraction of previous appointments attended
Number of cancellations during the last visit day
Fraction of previous cancellations
Number of physician appointments during the last visit day
Fraction of previous rescheduled appointments; II
Minimum cycle of treatment
Cancelled previous appointment since patient not available
Number of previous cancellations
Fraction of previous rescheduled appointments; I
Number of previous rescheduled appointments attended
Number of days since the last visit day
Cancelled previous visit and rescheduled
Number of treatment appointments during the last visit day
Non-cancer treatment during last visit

Figure 2.11. CART’s relative variable importance


Minimum phase of treatment
No. of days since last visit day
Stage of the last cancer diagnosis
Fraction of previous visits attended
No. of physician appointments during the last visit
No. of days since the last treatment initiation
No. of cancellations during the previous visit day
Age
Fraction of previous rescheduled appointments; I
Appointment day of month
No. of previous appointments attended
Fraction of previous cancellations
Distance from patient home to the clinic
No. of days since the last disease onset
Fraction of previous rescheduled appointments; III
Maximum phase of treatment
No. of previous cancellations
No. of days since the last cancer onset
Fraction of previous rescheduled appointments; II
Appointment week of month (1-4)
Expected start time of the visit
Appointment day (1-365)
Age at the last onset date
No. of disease diagnoses

(a) MART’s relative variable importance

Figure 2.12. Relative variable importance for predicting cancellations with equal costs of misclassification.
2.4 Discussion and Conclusions

This chapter provided a modeling framework for the application of interventions or preven-
tions to heterogeneous populations. Based on a family of prediction models, an optimal
subgroup of patients was selected for inclusion in an intervention program, taking into
account both the costs and the benefits of the intervention(s), a critical consideration for
prediction models being applicable in operational settings.

A product of the modeling approach taken in this chapter is the ability to determine the
necessary scale for efficiency, i.e., the necessary number of patients for the intervention
program to not only help individual patients, but also improve the bottom line. If the size
of a program is not large enough to overcome set-up cost, or does not have a large enough
at-risk population, the model can answer questions about the size needed for an efficient
program. In that case, there may be the possibility to identify other healthcare units, which
may be able to join the program for economy of scale. For example, a text message system
can most likely be shared across departments and even health-care organizations.

One of the key parameters in any intervention evaluation is the intervention effectiv-
ness, and as previously discussed, there are conflicting effectiveness results reported in the
literature. In addition, each healthcare unit, may have a different population with different
characteristics, and therefore the exact value of the effectiveness may be unknown at the
time of implementation. As a result, it is important for any health organization to monitor
the effectiveness of the intervention program on an on-going basis, not only to learn about
the effectiveness of its programs, but also to respond to (gradual) changes in its operation
and patient population characteristics. Periodically updating prediction models should be
a part of this effort. For the case study used in this chapter, the program was implemented
when the effectiveness of intervention is between 24.7% and 1. If for a specific program,
the estimated effectiveness is close to the lower bound of this range, a trial run or evaluation
of the effectiveness on the specific population may be called for.

Through the analysis of medical appointment prediction models, valuable insights were
gained into appointment cancellation risk factors. One hundred and four independent vari-
dables were considered in building the prediction models of section 2.3. Even though the
variables considered by the models covered a wide range of components such as patient
demographic, treatment, and diagnosis, patient attendance history played a key role in predicting non-attendance. The phase of treatment was also found to be a strong predictor of non-attendance; the patients in more advanced phases of treatments are more likely to attend at their appointments. Considering the risk factors identified by the prediction models will help the decision makers with creative and targeted interventions that can reduce appointment cancellation.

The results presented in this chapter can be extended in the following directions.

**Multi-Level Outcome Variables:** The optimization models provided in this chapter are based on the assumption that the outcome of interest is a qualitative binary random variable such as the patient “survived” or “died”, or the patient “attended” the appointment or “canceled” it. Although binary outcomes are the outcome of choice for many healthcare applications, one can think of applications where more than two outcome levels are required to model the problem. For example, the outcome variable can have three levels corresponding to appointment cancellation, appointment no-show, and attended appointment.

**The Benefit and Cost Structure:** The models presented in section 2.2 are for the case where the per patient benefit from the intervention is constant across all patients and the cost changes based on the total number of patients intervened. An important question to address is how these models should be updated to capture varying costs, benefits, and effectiveness. Such models will be capturing the reality of many health-care problems more closely since in practical applications, the costs and benefits of interventions of riskier patients may differ from healthier patients.

**Quality of Prediction:** The quality of the prediction models have a direct impact on the optimal intervention level and the total benefit from intervention. In the example discussed in section 2.3, an optimal total benefit of $1819 per day was achieved when the intervention was applied to 31% of the population. Depending on the quality of the ROC curve, these numbers can change significantly. For example, for the random guess model, the optimal intervention is to not run the program. However, in the case of perfect prediction, the optimal action is to apply the telephone reminder intervention to the at-risk or true-positive population. This corresponds to 22% of the population and the optimal total benefit in this case is $923 per day.
The case illustrated that there is a substantial difference between constant and perfect prediction. Therefore, increasing the quality of the predictions can result in more targeted interventions and more savings.

Although the application discussed in this chapter was non-attendance in an outpatient clinic, there are numerous other healthcare scenarios that could benefit from this approach, for example 30-day emergency room readmission prevention programs. With the call for continuous process improvement and cost cutting in the healthcare system, careful application of beneficial interventions and preventions, to the right patients at the right time will contribute to more efficient healthcare system.

such that the improvement in the classification error is maximized. To control the size of the tree, a particular partitioning step is only accepted if the improvement exceeds a prespecified threshold. The interested reader is referred to (Hastie et al., 2009, Chapter 9) for additional details.
Colorectal cancer (CRC) is the third most common cancer in both men and women. In 2001-2006, more than 60,000 patients died every year from CRC in the United States [34]. In 2013, an estimated 142,820 patients will be diagnosed with CRC and 50,830 patients will lose their life to the disease. CRC mortality is estimated to be the second most common cause of cancer death in 2013 after lung and bronchus cancer [57].

Recent years have seen dramatic changes in screenings, diagnostic procedures, surgical treatments, chemotherapy, and radiation treatments for CRC patients. The aforementioned procedures have become less invasive and more precise, and they result in longer survival times in many cases [75]. Nonetheless, despite all the advances, many cancer patients still need to go through long periods of therapy and treatment which include, in many cases, exposure of body cells to toxins, and incur years of financial, physical, and psychological burdens if they survive the disease [23]. Estimating patient survival is a delicate task and according to the literature, physicians consistently overestimate it [46, 51].

The inaccurate estimation of patient survival burdens the physician, patient, and healthcare system as a whole. In the absence of accurate survival information, an honest conversation between the physician and patient is not possible. Physicians may recommend aggressive treatments to the patient with the hope of controlling or curing the disease. Faced with the uncertainty in prognosis and lack of sufficient conversation about the different aspects of the options available, the patient is left with few options. She could have chosen palliative care or hospice services, had she known the time left is limited, or, with
the hope of prolonging life or curing the disease, she might have chosen aggressive treat-
ment procedures. The lack of accurate information might also disrupt planning for end of
life which could potentially put a tremendous amount of pressure on the patient and her
family toward the end of her life and afterwards. Finally, in a health-care system already
struggling with high costs, aggressive treatment of a patient close to the end of life with no
hope for an actual cure results in wasted resources that could be used for other patients.

This chapter builds prediction models for CRC mortality that can help patients, physi-
cians, and policy makers estimate the patient’s survival (or time to mortality). The predic-
tion models were built for eleven time horizons, i.e., 30-day, 6-month, 1-year, 1.5-year, ...,5-year. For brevity of the presentation, and when detailed analysis of classification models
were desirable, results corresponding to 30-day, 1-year, and 5-year prediction models were
considered. The performance of the models and variable importance measures are studied
using all the eleven time horizons. Classification trees [20], regularized logistic regression
with lasso [102], and relaxed lasso [70] were used to build the classification models. The
performances of the algorithms were measured by area under the ROC curve (AUC). Dif-
ferent measures of variable importance were used and a new measure was developed to
compare the importance of variables in predicting mortality in each time horizon. Furth-
ermore, the importance of variables was studied as a function of time horizon.

The remainder of this chapter is organized as follows. Section 3.1 reviews the literature
on CRC mortality and summarizes its short-comings. Section 3.2 introduces the data-
set and discusses missing values and randomization. Section 3.3 describes the algorithms
used and Section 3.4 discusses important variables in predicting CRC mortality considering
different importance measures.

3.1 Colorectal Cancer Mortality Literature

The literature of CRC mortality is focused on describing the relation between one or more
independent variables and the outcome, i.e., whether the patient survives a certain period of
time or how long the patient survives. The goal of this section is not to present a compre-
hensive review of the literature but rather use some of the published results to demonstrate
the challenges patients, physicians, and policy makers face in drawing conclusions and
making decisions based on the current literature. To this end, the literature can be divided into research on descriptive modeling and predictive modeling.

### 3.1.1 Descriptive Modeling

Descriptive modeling emphasizes the association between dependent and independent variables with little or no focus on causal relations or predictive models. When carefully done, descriptive modeling can be illuminating. For instance, in the early nineties, a descriptive study showed that screening sigmoidoscopy is associated with reduced rectum and distal colon cancer mortality [77]. The results of this study changed the CRC treatment protocol and to date, screening sigmoidoscopy is considered a screening procedure, which reduces mortality. However, as I will demonstrate below, not all the findings in the literature are as commonly agreed on as screening sigmoidoscopy.

**Age** is a significant predictor of CRC mortality according to some studies [9, 37] while it is not a significant predictor, according to another. In a large study of patients in 2004, it was shown that young rectal cancer patients have equivalent overall and stage-specific survival when stage and grade of the tumor are controlled for [79].

**Race and Socio-economic Status** have been the two commonly visited factors in explaining CRC outcomes. On the one hand, white patients are reported to be more likely to be diagnosed with early-stage CRC and have better overall and stage-specific survival when compared with Black patients, controlling for age, race, sex, and zip code-based income [37, 109]. On the other hand, race has been shown to be only significant in predicting survival of CRC patients after diagnosis in the presence of SES variables and adjusting for age, sex and stage [29, 106].

**Marital Status** is associated with longer survival in some studies. Rectal cancer patients who are married have higher survival rates [47]. However, marriage is reported to not be significant in predicting CRC if diagnostic and treatment variables are controlled for [48].

**Hospital characteristics** have been shown to be important in explaining CRC mortality. Number of operations performed at a hospital, also known as hospital volume, was shown to be important in explaining 2-year survival for rectal cancer patient at stage I-III.
who underwent surgical resection [56]. Greater surgeon and hospital volumes were associated with improved 30-day post-operative survival as well as overall long-term survival for patients undergoing surgery as part of CRC treatment [91]. High Medicaid hospitals have been shown to have higher postoperative colon cancer mortality at 30-day and 1-year but not at 5-year [89].

### 3.1.2 Predictive Modeling

A second approach for describing CRC mortality is predictive modeling. The goal in predictive modeling is to use independent variables and a prediction algorithm to explain the relation between the independent variables and CRC mortality. This is a less explored part of the CRC mortality literature, which can provide insights not accessible through descriptive modeling.

The predictive modeling studies differ in the independent variables considered as well as algorithms used in building prediction models. Partial likelihoods artificial neural network [12, 64] was used to build a prediction model for death within five years of CRC diagnosis. Sixteen independent variables were considered and the difference between predicted survival and Kaplan-Meier curve was used as a base for assessing the quality of prediction [15]. Another study used nine independent variables along with proximal support vector machine (pSVM) [45] and logistic regression to predict rectal cancer survival from the time of diagnosis. The studies reported AUCs of 0.74 and 0.73, respectively [100]. The performance of different prediction models were compared through a study where three algorithms, namely, regression trees, k-nearest neighbors (k-NN) and a regression variant of ANN were used to model patient survival from the time of CRC diagnosis. The comparison of the three models showed that the models’ differences were not significant [2].

### 3.1.3 Challenges of the Literature

The underlying goal of many studies in the literature, whether descriptive or predictive, is to add to medical knowledge and understanding and thereby, help the decision makers, i.e., patients, physicians, and policy makers, make more informed decisions. As it have become clear from the last two sections, the literature, heavy in descriptive modeling, and light on
predictive modeling, have not been successful in achieving this goal in many cases. The following reasons contribute to this lack of success.

1. **Contradictions:** The subset of the literature presented here showcases the kind of contradictions those who seek more understanding face when consulting with the literature. Age, marital status, race, and more independent variables are important and yet, unimportant in explaining CRC mortality depending on the specifics of the studies. In most of the cases, the contradictions can be explained by the use of different reference points for computing mortality across studies, number of patients in the study, presence or absence of other independent variables, and the time horizon considered in the study. Although these contradictions are not unexpected, lack of a common measure in assessing the quality of the studies these contradictions stem from make results confusing, leading to limited insights for the reader.

2. **Generalizability:** In the health-care settings, statistical models are used to explain an otherwise overly complex system. In many practical cases, a small subset of the population in the original complex system is chosen and analyzed. However, there are difficulties in claiming general insights from models based on a limited specific data-set. There are remedies which can be utilized to make the models more generalizable. Many of the results presented in the literature lack such remedies. For example, the results presented are in many cases in-sample results. This makes the models biased and the results are hardly generalizable to other data-sets.

3. **Accuracy:** In many cases, the quality of the results provided is not reported. For example, in the case of predictive modeling, we found very few studies which use AUC or some other measure of predictive accuracy to assess the prediction models.

In the next sections, we build prediction models for CRC mortality and address the above issues through our proposed models.

### 3.2 The Data

The data-set was built by linking two major California data-sets: the California Cancer Registry (CCR) and the California Office of Statewide Health Planning and Development.
(OSHPD). The CCR data-set contains information on tumor registration, required under the state law since 1985, follow up information, as well as the U.S. census block-group information. The OSHPD data-set contains hospital discharge data. The data consists of 117,956 unique patient records. The study period is from January 2001 through December 2011 with the last follow-up date on December 6, 2012. The patients included in the study are those 18 years and older at the time of diagnosis who have been diagnosed with CRC (ICD-10 codes C18-20 and C26) and hospitalized during the study period. The hospitalization can be for the purpose of CRC treatment as well as any other inpatient treatment. As part of the CCR data collection procedure, patient follow-up information is collected on an annual basis. During the follow-up, a physician who has been assigned as a follow-up point of contact was contacted and information about the status of the patient (alive or dead) and the cause of death was collected, if applicable.

**Independent Variables and Outcome:** A total of 153 independent variables were a part of this study. Some of the variables were immediately available in the linked data-set while others were built based on knowledge of the disease, other diagnoses, treatments, and hospital operations. A complete list of variables with their definitions is in Appendix C.

The end-points of the study are 30-day and \( n \)-year survival/mortality where \( n \) is in \( \{0.5, 1, \ldots, 5\} \). The analysis is for death due to CRC, not all-cause mortality, since the effects of variables such as age or sex on CRC mortality cannot be differentiated in all-cause mortality analysis.

**Reference Point:** Analysis of patient mortality requires setting a reference point where survival will be measured from. Given the structure of the data-set and CRC treatment procedures, the time of first hospital discharge after CRC diagnosis has been chosen as the reference point. This reference point is chosen based on the structure of the data-set, i.e., only the records of those who were hospitalized between 2001 and 2010 are in the data-set and predicting patient’s mortality outcome after hospital discharge becomes a relevant question to address.

Figure 3.1 shows the sequence of treatment for patients diagnosed with colorectal cancer. If the tumor is considered an early stage tumor, the patient first receives surgery.
Depending on the positive and examined lymph node counts, chemotherapy might be recommended. On the other hand, if the tumor is considered advanced stage, the patient will first receive radiation or chemotherapy. Surgery is recommended for the patient only if radiation or chemotherapy results in satisfactory results, i.e., the tumor size and spread is controlled. If surgery is recommended, a decision about whether the patient needs to receive chemotherapy after the surgery needs to be made.

![Diagram of CRC treatment sequence](image)

**Figure 3.1.** The sequence of treatment for CRC patients. The solid blue box shows a definite first recommendation of treatment. The dashed boxes are second stage treatments, depending on the result of the prior stage treatment.

### 3.2.1 Censoring

The data is right-censored under two circumstances. First, if a patient is alive at follow-up and the survival horizon considered occurs after the follow-up time. Second, if the patient dies before the end of the time horizon from a non-CRC cause. As a result of censoring, shorter horizons have more data than the longer ones.

### 3.2.2 Randomization

The patient records are randomly assigned to either the training or the test set. The training set includes 80% of the data and it is used for building the prediction models. The test set which contains 20% of the data is used to compute the performance of the prediction models, namely, the ROC curves and the AUCs.
In order to build the training set, we start from the 30-day mortality case and assign 80% of the uncensored patient records to the training set. The rest goes to the test set. For all the other time horizons, we build the training and test sets based on the training and test sets of the previous time horizon, i.e., we remove the patients who are censored under the new time horizon from the training and test sets of the previous time horizon. Using this approach, we assure that if a patient record is used in building the prediction models during the training phase, this record is never used to assess the performance of the prediction models. Keeping the patients in the training set until they are censored has an added advantage that the comparison of the prediction models from one time horizon to the next will be unaffected by the introduction of new patients to the training set had we not kept the same patients in the pool. The number of patient records considered in the training set for each of the eleven time horizons is recorded in Table 3.1.

<table>
<thead>
<tr>
<th>horizon</th>
<th>size of training set</th>
<th>size of test set</th>
<th>rate of mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-day</td>
<td>92,649</td>
<td>23,136</td>
<td>0.0165</td>
</tr>
<tr>
<td>0.5-year</td>
<td>89,635</td>
<td>22,385</td>
<td>0.0804</td>
</tr>
<tr>
<td>1-year</td>
<td>87,098</td>
<td>21,732</td>
<td>0.1251</td>
</tr>
<tr>
<td>1.5-year</td>
<td>84,291</td>
<td>20,983</td>
<td>0.1645</td>
</tr>
<tr>
<td>2-year</td>
<td>80,274</td>
<td>20,028</td>
<td>0.2036</td>
</tr>
<tr>
<td>2.5-year</td>
<td>75,290</td>
<td>18,813</td>
<td>0.2384</td>
</tr>
<tr>
<td>3-year</td>
<td>70,360</td>
<td>17,594</td>
<td>0.2679</td>
</tr>
<tr>
<td>3.5-year</td>
<td>65,674</td>
<td>16,356</td>
<td>0.2945</td>
</tr>
<tr>
<td>4-year</td>
<td>61,103</td>
<td>15,246</td>
<td>0.3178</td>
</tr>
<tr>
<td>4.5-year</td>
<td>56,232</td>
<td>14,070</td>
<td>0.3421</td>
</tr>
<tr>
<td>5-year</td>
<td>51,511</td>
<td>12,839</td>
<td>0.3627</td>
</tr>
</tbody>
</table>

Table 3.1. Number of records in the training and test sets and the rate of mortality for each time horizon.

### 3.2.3 Missing Values

The data-set contained missing values. Missing values were addressed differently depending on their nature. For any data point, if the day or month of a follow-up date was missing, the data point was discarded (108 records). For all other independent variables with date
type (such as date of diagnosis, date of staging, etc.), the following procedure was followed: If the only information available about a date was its year, July 1 (the middle of the year) was used. If both year and month were available, the midpoint between the start of that year until the end of the month was used as the date. For example, if the value of date of diagnosis was 2005-06-XX, the date was set to 2005-03-31. This was done to ensure consistency in addressing missing values, as the date cannot be set to 2005-06-15 since this date may go beyond the follow-up date, something which does not appear in the records.

For any other variable with missing values, we handled the case depending on the choice of the prediction algorithm. Classification tree algorithms have mechanisms in place for predicting the missing values where the known values of some of the other variable, called surrogates, can be used to predict the missing value. If the missing value was absent due to the value not being available at the time of data collection, a dummy variable was introduced to represent the missing value. In the regularized logistic regression and relaxed lasso, the missing values for categorical variables were encoded by introducing dummy binary variables. In the case of numerical variables, the median of the variable was used as the value for the missing variables. For example, according to CCR guidelines, stage of the tumor should have been reported. If no stage information is available, specific codes should be used to indicate this. However, for some data points, stage was blank. Therefore, a new variable stage_null was created to represent this category of stage variable.

3.3 Prediction Models for Colorectal Cancer Mortality

This study utilizes three data-mining algorithms to predict CRC mortality: classification trees (GUIDE), regularized logistic regression (lasso), and relaxed lasso. Through the implementation of these prediction models and the use of a comprehensive data-set, the challenges of the current CRC literature, i.e., contradictions in results, generalizability, and lack of quality measures are addressed. Throughout this section, area under the ROC (AUC) is used as a measure of classification performance. Different measures of variable importance are used to compare independent variables and the value they add to predicting CRC mortality.
3.3.1 Mortality from the Time of First Hospital Discharge

Prediction models for CRC mortality from the time of first hospital discharge after CRC diagnosis are presented.

Classification Trees

Classification trees belong to the family of tree-based methods. Tree-based methods predict the outcome by dividing the independent variable space into simple regions and estimate the outcome in each sub-region by a simple function of the training observations in the region. In the case of a classification tree, the sub-regions are high-dimensional rectangles or boxes and the most common function used to predict the outcome in each region is the mean of the training set observed outcomes in that region. To convert the fractional mean value to binary values, the algorithms compare the mean of each terminal node with a threshold computed based misclassification costs.

Classification trees are celebrated in data-mining and statistical analysis. They have a very few tuning parameters and hence are easy to implement. Their output is a binary tree which is easy to understand and interpret, at least as long as the tree is not too large. There are multiple algorithms for classification trees. These algorithms differ in handling missing values and choosing split criterion. In this section, an algorithm called Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) [65] is utilized. GUIDE was chosen over CART [20] since the CRC mortality data-set contains a large number of categorical variables with many levels, a characteristic which triggers CART’s bias in choosing split variables that have large number of levels.

As highlighted in Appendix A, the output of prediction models consists of trees, each corresponding to a specific false-positive and false-negative trade-off. The performance of the family of prediction models is represented by a ROC curve. The ROC curves for 30-day, 1-year, and 5-year prediction models are shown in Figure 3.2. The ROC curves reveal couple of important observations. First, the prediction performance is good, i.e., the AUC varies between 0.96 and 0.88, for the 30-day and 5-year models, respectively. Although there are no results in the literature, which report AUCs for CRC mortality from the time of hospital discharge, one can appreciate the AUCs by understanding the conceptual meaning
of them. An AUC of 0.96 for the 30-day CRC mortality ROC curve means that a tree from the family of trees represented by the ROC can differentiate a patient who died from one who did not with probability 96% [49].

Second, the AUC is higher for the 30-day mortality prediction than 1-year or 5-year mortality. Figure 3.3 shows the AUCs for prediction models as a function of the time horizon, which confirms that except for minor changes in AUCs, shorter time horizon models are generally better in terms of AUC than longer horizons. Finally, the ROC curves for all the three time horizons are steep when the rate of false-positive is low and they become almost flat when the rate of true-positive is high. These two characteristics of the ROCs can be exploited when prediction models are considered as a decision tool for interventions. For example, if it is not possible to offer hospice care to everyone, point (1,1) on the ROC curve, the decision maker can move to a model with slightly lower true-positive rate and much lower false-positive rate, i.e., by accepting that a small number of the true-positive population will not receive hospice care, the decision maker can achieve a much lower false-positive rate. Similarly, if the decision maker decides to apply the intervention

Figure 3.2. ROC curves for 30-day, 1-year, and 5-year mortality using GUIDE.
to a small percentage of the population as opposed to not applying it to anyone, point (0,0) on the ROC curve, she is assured to have a very low false-positive rate and very high true-positive rate, i.e., the first patients who are chosen as part of the intervention are indeed the riskiest patients in the population in terms of the outcome.

A sample of three classification trees from the ROC curves in Figure 3.2 are shown in Figure 3.4(a)-3.4(c). The trees demonstrated correspond to classification models with equal misclassification costs. 3.4(a)-3.4(c).

The two number next to each node of a tree correspond to the number of patients in the training sample who survived (the top number) vs. the number of patients that did not (the bottom number). Green nodes depict a prediction of survival for the patients in the corresponding node while red nodes represent a prediction of mortality. Every node, except for the terminal nodes, corresponds to an independent variable with a splitting criterion. Starting from the root node, if a criterion is met for a patient, the patient moves to the left branch of the tree, otherwise, to the right. Through this process, every patient can be assigned to one of the terminal nodes with specific predicted class value.

Figure 3.3. The AUCs for classification tree over the time horizons, spanning 30-day to 5-year.
Figure 3.4(a) corresponds to the 30-day mortality model where all patients are predicted to survive since the training set does not contain a large number of mortality cases when compared to the number of survivals. Given that the misclassification costs are equal, the model predicts everyone to die in thirty days. This is not the case for the 1-year and 5-year models with a higher rate of mortality in their corresponding training sets.

Figure 3.4(b) shows the classification tree corresponding to the 1-year mortality model. The classification model predicts three groups of patients surviving, a) those with stage 0 or I-III, b) those with stage IV or unstaged who have received chemotherapy (chemotherapy NOS, single agent, multiple agent, or recommended), and c) those with stage IV or unstaged who have not received chemotherapy but received surgery and where discharged to home. Those who die are stage IV or unstaged patients who do not receive chemotherapy or surgery, or those who receive surgery but are discharged to a place other than home.

Figure 3.4(c) shows the classification tree for the 5-year CRC mortality model. Similar to the 1-year model, the stage of the tumor is the first variable the tree splits on. Those at stage IV or unstaged, at stage III who have been dispositioned to the same hospital units, another hospital, SN/IC, or against medical advice are predicted not to survive beyond 5 years. Furthermore, patients with stage III tumors or tumors that are unstaged, those whose disposition is home or home health service, had a lymph node ratio greater than 30% and had more than 7 positive lymph nodes are not predicted to survive.

An interesting question to ask is whether the most important variables for the three time horizons when the costs of misclassification are equal are those who appear as splitting variables in Figure 3.4. The answer is indeed no. Intuitively, this is clear since if a prediction model such as 30-day mortality does not have any split variables, this should not translate to no variable being important in predicting survival. GUIDE’s measure of relative importance [65] (RI) was used to measure the importance of each variable. For each independent variable $X_i$, the importance measure is computed as

$$\sum_t \sqrt{N(t)} W_M(t, i)$$

where $N(t)$ is the training sample size in node $t$ and $W_M(t, i)$ is the Wilson-Hilfery marginal chi-squared value of $X_i$ at $t$. The corresponding value for each variable is then normalized.
such that the most important variable has RI value of 100. The results are shown in Tables 3.2 and 3.3.

As shown in the table, for the 30-day mortality model, whether the patient has received surgery has the highest relative importance while for the 1-year and 5-year models, the highest relative importance is the stage of the tumor. It is also interesting to note that not all the variables shown in the trees in Figure 3.4 are present as the top twenty variables in terms of relative importance. For example, the tree for 1-year mortality branches on chemotherapy while chemotherapy is not in the list of top twenty variables.

To the best of my knowledge, this is the first time classification trees were used in the context of predicting CRC mortality. The results presented in this section are important for three reasons.

First, there were no pre-assumptions about the independent variables used in the models. All the independent variables in Appendix C were considered. Yet, the variables chosen by the model in the final tree are those familiar to us from the medical literature, namely, stage of the tumor, chemotherapy, surgery, lymph node ratio, surgery, and the number of positive lymph nodes. Second, specific variable thresholds are introduced as a result of the splitting mechanism of classification trees. For example, a threshold of 30% is used in the 5-year mortality model to differentiate those who survive from the rest based on the lymph node ratio. Lymph node ratio is a popular measure in CRC diagnosis and treatment. It is known to be a predictive of CRC recurrence and overall survival in those where more than 10 lymph node have been examined [11]. Using classification trees, the ratio cut-off is automatically generated for specific subsets of the population. The tree predicts stage III tumors or those for whom the stage information is not available, and whose disposition is home or home health service will survive if the lymph node ratio is more than 30% and the number of positive lymph nodes is less than or equal to 7. This translated to a total number of lymph nodes examined less than 23. Therefore, although the literature suggests that lymph node ratio is not predictive of survival when less than 10 lymph nodes are examined, the classification tree in Figure 3.4(b) suggests otherwise. Third, the performance of classification trees measured by AUC for predicting CRC mortality is between 0.87-0.96. Given the simple structure of decision trees and their performance, they can be easily implemented and used in practice.
Figure 3.4. 30-day, 1-year, and 5-year CRC mortality prediction models using GUIDE for the case of equal mis-classification costs.
<table>
<thead>
<tr>
<th>rank</th>
<th>30-day RI</th>
<th>variable name</th>
<th>1-year RI</th>
<th>variable name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>surgery</td>
<td>100</td>
<td>stage</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>number of days between first date of treatment and diagnosis</td>
<td>96.4</td>
<td>surgery</td>
</tr>
<tr>
<td>3</td>
<td>65.6</td>
<td>tumor histology</td>
<td>64.8</td>
<td>extensive surgery</td>
</tr>
<tr>
<td>4</td>
<td>63.6</td>
<td>stage</td>
<td>52.9</td>
<td>lymph node ratio</td>
</tr>
<tr>
<td>5</td>
<td>60.9</td>
<td>disposition</td>
<td>49.1</td>
<td>tumor histology</td>
</tr>
<tr>
<td>6</td>
<td>58.6</td>
<td>extensive surgery</td>
<td>45.6</td>
<td>number of days between first date of treatment and diagnosis</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
<td>source of admission</td>
<td>42.8</td>
<td>disposition</td>
</tr>
<tr>
<td>8</td>
<td>43.3</td>
<td>number of days between surgery and diagnosis</td>
<td>38.4</td>
<td>number of days between surgery and diagnosis</td>
</tr>
<tr>
<td>9</td>
<td>43</td>
<td>surgery removing regional lymph nodes</td>
<td>37.6</td>
<td>source of admission</td>
</tr>
<tr>
<td>10</td>
<td>39.7</td>
<td>positive to examined lymph node ratio</td>
<td>33.4</td>
<td>surgery removing regional lymph nodes</td>
</tr>
<tr>
<td>11</td>
<td>27.7</td>
<td>number of positive lymph nodes more than 90</td>
<td>22.9</td>
<td>type of admission to hospital</td>
</tr>
<tr>
<td>12</td>
<td>22.8</td>
<td>secondary malignant neoplasm of liver</td>
<td>22.3</td>
<td>secondary malignant neoplasm of liver</td>
</tr>
<tr>
<td>13</td>
<td>22.2</td>
<td>age at diagnosis</td>
<td>21.4</td>
<td>number of positive lymph nodes more than 90</td>
</tr>
<tr>
<td>14</td>
<td>21.6</td>
<td>route of admission</td>
<td>21</td>
<td>degree of differentiation of the tumor</td>
</tr>
<tr>
<td>15</td>
<td>20.5</td>
<td>type of admission</td>
<td>20.2</td>
<td>age at diagnosis</td>
</tr>
<tr>
<td>16</td>
<td>19.8</td>
<td>chemotherapy</td>
<td>17.7</td>
<td>age at discharge</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>location where the tumor originated</td>
<td>17.5</td>
<td>route of admission</td>
</tr>
<tr>
<td>18</td>
<td>17.7</td>
<td>age at discharge</td>
<td>17</td>
<td>metastasis</td>
</tr>
<tr>
<td>19</td>
<td>15.4</td>
<td>degree of differentiation of the tumor</td>
<td>15.9</td>
<td>location where the tumor originated</td>
</tr>
<tr>
<td>20</td>
<td>10.4</td>
<td>radiation</td>
<td>14.6</td>
<td>length of stay in the hospital</td>
</tr>
</tbody>
</table>

Table 3.2. Top twenty variables from GUIDE; 30-day and 1-year mortality models.
### Table 3.3

Top twenty variables in GUIDE; 5-year mortality model

<table>
<thead>
<tr>
<th>rank</th>
<th>RI</th>
<th>variable name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>stage</td>
</tr>
<tr>
<td>2</td>
<td>42.8</td>
<td>extensive surgery</td>
</tr>
<tr>
<td>3</td>
<td>41.6</td>
<td>lymph node ratio</td>
</tr>
<tr>
<td>4</td>
<td>38.9</td>
<td>surgery</td>
</tr>
<tr>
<td>5</td>
<td>32.6</td>
<td>disposition</td>
</tr>
<tr>
<td>6</td>
<td>30.1</td>
<td>tumor histology</td>
</tr>
<tr>
<td>7</td>
<td>27.3</td>
<td>source of admission</td>
</tr>
<tr>
<td>8</td>
<td>21.9</td>
<td>number of days between surgery and diagnosis</td>
</tr>
<tr>
<td>9</td>
<td>20.5</td>
<td>number of positive lymph nodes</td>
</tr>
<tr>
<td>10</td>
<td>19.9</td>
<td>type of admission</td>
</tr>
<tr>
<td>11</td>
<td>19.6</td>
<td>metastasis</td>
</tr>
<tr>
<td>12</td>
<td>17.8</td>
<td>secondary malignant neoplasm of liver</td>
</tr>
<tr>
<td>13</td>
<td>16.8</td>
<td>number of days between first date of treatment and diagnosis</td>
</tr>
<tr>
<td>14</td>
<td>16</td>
<td>length of stay in the hospital</td>
</tr>
<tr>
<td>15</td>
<td>15.2</td>
<td>chemotherapy</td>
</tr>
<tr>
<td>16</td>
<td>13.5</td>
<td>age at discharge</td>
</tr>
<tr>
<td>17</td>
<td>12.7</td>
<td>age at diagnosis</td>
</tr>
<tr>
<td>18</td>
<td>11.4</td>
<td>degree of differentiation</td>
</tr>
<tr>
<td>19</td>
<td>10.8</td>
<td>route of admission</td>
</tr>
<tr>
<td>20</td>
<td>9.8</td>
<td>surgery removing regional lymph nodes</td>
</tr>
</tbody>
</table>
Regularized Logistic Regression with Lasso and Relaxed Lasso

A common regression analysis technique used in the medical literature is logistic regression. The choice of logistic regression is not accidental. It is a powerful technique when the outcome of interest is a binary variable, a common variable type in medical applications. It is also easy to implement and analyze when the number of independent variables is much smaller than the number of observations. However, when this is not the case, logistic regression models tend to be hard to interpret partly because of large number of independent variables in the model. Furthermore, in the applications where binary independent variables with few of one class versus the other are present, perfect separation of an independent variable and the outcome results in very large coefficient values for these variables, which mask the effect of other independent variables on the outcome. The above two shortfalls are results of logistic regression’s loss function, i.e., least squares. In practice, the least squares is unlikely to yield coefficients with value equal to 0 [59]. Also, the loss function lacks a term for controlling the size of coefficients. Therefore, logistic regression models become complicated if the number of independent variables considered by the algorithm is very large.

There are three main approaches for handling the large number of non-zero coefficients: subset selection, shrinkage, and dimension reduction. In this section, regularized logistic regression with lasso penalty or for brevity, lasso, is used [102]. Lasso’s loss function encourages the shrinkage in size of the coefficients through penalizing their size. All independent variables are considered in building the prediction model, however, only a subset of these variables have coefficients greater than zero in the final model. The lasso’s loss function is the least squares function plus an $l_1$ norm penalty as follows

$$
\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) + \lambda \sum_{j=1}^{p} |\beta_j|,
$$

where $n$ is the size of the training set, $y_i$ is the outcome of the $i$-th observation, $x_{ij}$ is the value of the $j$-th independent variable for the $i$-th observation, $\beta_j$ is the $j$-th coefficient corresponding to the $j$-th independent variable, $p$ is the total number of independent variables, and $\lambda$ is a complexity parameter to be determined by minimizing an estimate prediction
error plus one standard error based on 10-fold cross-validation [102]. As \( \lambda \) approaches zero, the loss function approaches the least squares and all coefficients are non-zero. As the value of \( \lambda \) increases, the loss function will be penalized more by the size of coefficients. As a result, some coefficients will be driven to zero.

There are two challenges in using lasso in practice. First, the number of selected variables based on the \( \lambda \) found through cross-validation is usually large. Second, estimates of non-zero coefficients are biased toward zero and inconsistent, i.e., as the sample size increases, the estimates do not necessarily converge to their true value. An alternative is to use relaxed lasso [70], which is a two stage algorithm based on lasso. In the first stage, lasso is run. The second stage uses the variables with non-zero coefficients in from the first stage and runs lasso on these subset of variables again. As before, the best model is chosen by cross-validation. Relaxed lasso results in fewer non-zero coefficients and furthermore, it increases the accuracy of the predictions [70].

Lasso and relaxed lasso were used to build CRC mortality prediction models for the eleven specified time horizons. Figure 3.5 shows the ROC curves for 30-day, 1-year, and 5-year models by relaxed lasso. As it is clear from the ROC curves when compared to Figure 3.2, a similar decrease in predictability occurs with relaxed lasso as with classification trees, i.e., the AUC for ROCs decrease as the time horizon increases.

Figure 3.6 shows the AUC as a function of the time horizons. Both AUC functions decrease as the time horizon increases, i.e., as time horizon increases, the predictive performance decreases. The decrease is expected since the sample size decreases as a function of horizon because of censoring and long term survival is more complicated to predict from the disease standpoint. Through Figure 3.6 we can also observe that relaxed lasso does better than regular lasso when the measure of performance is the AUC although the difference in terms of performance is in the order of 1% of the measure in the worst case, for the 5-year prediction. For the rest of this chapter, the models refer to the output of the relaxed lasso, unless otherwise specified.

The top twenty most important variables for 1-year and 5-year CRC mortality using relaxed lasso, based on their coefficient size are shown in Table 3.4. All coefficients for the 30-day mortality model were zero. Therefore, GUIDE and relaxed lasso for 30-day mortality reduce to the same model. The sign of a coefficient indicates an increase or decrease
Figure 3.5. ROC curves for 30-day, 1-year, and 5-year mortality with relaxed lasso.

Figure 3.6. The AUC for lasso and relaxed lasso over the time horizons.
in the dependent variable, CRC mortality, as the value of the corresponding independent variable increases. In the context of CRC mortality, a positive coefficient indicates a variable which increases the probability of mortality. The reverse relation is true for a negative coefficient. The most important variable for 1-year CRC mortality is the number of days between admission to the hospital and CRC diagnosis. This variable was not present in the top twenty most important variables in GUIDE, as shown in Table 3.2. On the other hand, stage of the disease which was also ranked the most important variable by GUIDE is chosen as the most important variable in 5-year CRC mortality by relaxed lasso. Both in the case of the 1-year and the 5-year models, relaxed lasso ranks second the variable, which captures whether the principal language of the patient is English or not as second. Although this variable was used in the tree, it was not present in the top twenty most important variables.

**Classification Tree vs. Relaxed Lasso**

The classification trees and relaxed lasso models are compared below in terms of performance, number of variables, and interpretability.

- **Performance**
  
  In terms of the overall performance of the prediction models measured based on the AUCs, relaxed lasso does better compared to classification trees. The AUC across all time horizons ranges from 0.96 to 0.91 for relaxed lasso (Figure 3.6) while this range is from 0.96 to 0.87 for classification trees (Figure 3.3). In other words, the two models do equally well for the 30-day time horizon, but relaxed lasso does better for all other time horizons. However, it is important to note that the performance improvement in the most pronounced case, the 4.5-year time horizon, is close to 5% in terms of the area under the ROC. Therefore, classification trees may be the method of choice for the reasons explained below.

- **Number of variables in the final model**
  
  The number of independent variables considered when building classification trees and lasso were 153 and 594, respectively. The difference is a result of the way lasso handles categorical variables, i.e., all categorical variables are represented by dummy variables, each corresponding to one
level of a categorical variable. In general, the number of variables used in the classification trees were smaller than relaxed lasso, even when different categories of one variable used in relaxed lasso were counted as a single variable. For example, as we saw in Figure 3.4, the classification tree for the 1-year CRC mortality uses only four independent variables while relaxed lasso uses 59 (146 if the dummy variables from a single independent variable are not combined) for same time horizon and misclassification cost.

- **Interpretability** There are two reasons for trees to be a more favorable option when it comes to interpretability. First, they use fewer independent variables. Second, the final model is easily understandable when plotted. Although the interpretation of odds ratio for relaxed lasso coefficients may seem to compete with trees, the easy to understand structure of trees make them appealing for practical purposes and implementations.

In short, if the goal of classification is to have prediction models which are easy to interpret and understand, classification trees should be chosen given that the loss in performance in the worst case is only 5%.
### Table 3.4. Top twenty important variables from relaxed lasso.

<table>
<thead>
<tr>
<th>rank</th>
<th>1-year variable name</th>
<th>coefficient value</th>
<th>5-year variable name</th>
<th>coefficient value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>days between admission to hospital and CRC diagnosis</td>
<td>-1.1881</td>
<td>stage iv</td>
<td>0.8087</td>
</tr>
<tr>
<td>2</td>
<td>principal language English</td>
<td>-0.7664</td>
<td>principal language English</td>
<td>-0.5979</td>
</tr>
<tr>
<td>3</td>
<td>stage iv</td>
<td>0.5925</td>
<td>age</td>
<td>0.3503</td>
</tr>
<tr>
<td>4</td>
<td>age at diagnosis</td>
<td>0.3916</td>
<td>secondary malignant neoplasm of liver</td>
<td>0.3177</td>
</tr>
<tr>
<td>5</td>
<td>ninety or more lymph nodes positive</td>
<td>0.3243</td>
<td>days between admission to hospital and CRC diagnosis</td>
<td>-0.3092</td>
</tr>
<tr>
<td>6</td>
<td>lymph node ratio</td>
<td>0.2596</td>
<td>lymph node ratio</td>
<td>0.3054</td>
</tr>
<tr>
<td>7</td>
<td>stage 0</td>
<td>-0.2376</td>
<td>stage i</td>
<td>-0.2879</td>
</tr>
<tr>
<td>8</td>
<td>unscheduled</td>
<td>0.2105</td>
<td>stage 0</td>
<td>-0.2659</td>
</tr>
<tr>
<td>9</td>
<td>stage i</td>
<td>-0.2053</td>
<td>metastasis</td>
<td>0.2139</td>
</tr>
<tr>
<td>10</td>
<td>metastasis</td>
<td>0.1891</td>
<td>number of positive lymph nodes</td>
<td>0.1962</td>
</tr>
<tr>
<td>11</td>
<td>no chemotherapy</td>
<td>0.1710</td>
<td>ninety or more nodes positive</td>
<td>0.1921</td>
</tr>
<tr>
<td>12</td>
<td>home disposition</td>
<td>-0.1704</td>
<td>unscheduled</td>
<td>0.1738</td>
</tr>
<tr>
<td>13</td>
<td>grade iii or poorly differentiated tumor</td>
<td>0.1701</td>
<td>home disposition</td>
<td>-0.1730</td>
</tr>
<tr>
<td>14</td>
<td>single agent chemotherapy</td>
<td>-0.1567</td>
<td>no surgery</td>
<td>0.1562</td>
</tr>
<tr>
<td>15</td>
<td>secondary malignant neoplasm of liver</td>
<td>0.1496</td>
<td>surgery was performed</td>
<td>-0.1437</td>
</tr>
<tr>
<td>16</td>
<td>no surgery</td>
<td>0.1493</td>
<td>carcinoid tumor</td>
<td>-0.1267</td>
</tr>
<tr>
<td>17</td>
<td>surgery was performed</td>
<td>-0.1356</td>
<td>source of admission is home with no ER involvement</td>
<td>-0.1186</td>
</tr>
<tr>
<td>18</td>
<td>number of positive lymph nodes</td>
<td>0.1241</td>
<td>days between the start of the first course of treatment and diagnosis</td>
<td>-0.1155</td>
</tr>
<tr>
<td>19</td>
<td>number of lymph nodes examined</td>
<td>-0.1193</td>
<td>grade iii or poorly differentiated tumor</td>
<td>0.1119</td>
</tr>
<tr>
<td>20</td>
<td>SN/IC disposition</td>
<td>0.1186</td>
<td>B-cell</td>
<td>-0.1096</td>
</tr>
</tbody>
</table>
3.4 Variable Importance

In medicine and bioinformatics, assessing the importance of independent variables in predicting the outcome is of special interest. This is specially true because of the advancements in genomics and the need for an ongoing assessment of variable importance with the hope of finding better and more accurate variables predicting outcomes such as acute cardiac disease [58], cancer [31, 63], and symptomatic peripheral arterial disease [90].

Different measures of variable importance are used in the literature, each with its own advantages and disadvantages. One of the most basic measures is p-value. Although this measure has been criticized for its lack of clinical significance when a variable is identified as statistically significant [32], it is unlikely for a variable to be considered in building a prediction model when it is not related to the outcome in a statistically significant manner [84]. A series of measures have been introduced to capture how well a model discriminates between two events corresponding to the outcome, such as mortality or survival, in the presence compared to the absence of an independent variable [28]. Until recently, the most popular of these measures has been the AUC [4, 28] with a downside that for models which already have independent variables explaining the outcome reasonably well, a new independent variable needs to have a very large impact on the outcome for it to influence the AUC measure [86, 108]. As a response to the problem with AUC measure, reclassification measure and predictiveness curves [26, 85], percentage of correct reclassifications [26], net reclassification improvement and integrated discrimination improvement [84] were developed.

In this section, three measures of variable importance are utilized to assess the importance of a sample of independent variables in predicting CRC mortality. These measures are coefficient size of relaxed lasso, sequence of entrance to coefficient profile plot in relaxed lasso, and a measure developed for the purpose of this work, adjusted improvement in correct reclassification. The first two measures are not applicable to classification trees while the latter measure can be used across models for comparing the variable importance as observed in different models. For the brevity of results, we show the results for this measure only for relaxed lasso models. All measures are computed based on models with equal misclassification costs.
Coefficient Size

The coefficient size for each independent variable in relaxed lasso was used to rank variables in terms of their importance in predicting CRC mortality for 1-year and 5-year models. The 30-day model was not considered since relaxed lasso predicts survival for all the patients and as a result, all coefficients are zero in this model. All models were built on standardized data in order to allow for direct coefficient comparison. The results were presented in Table 3.4. The list of variables reveals the following insights.

- Although the majority of the variables important in predicting CRC at the 1-year model are important in predicting CRC in the 5-year model, this is not the case for all the variables; There are variables such as “no chemotherapy” (1 if the patient did not receive chemotherapy as the first course of treatment; 0 otherwise,) with positive coefficient, which are important in the 1-year model but not in the 5-year model.

- The magnitude of the coefficients of independent variables changes from 1-year to 5-year CRC mortality model. For example, number of days between admission to hospital and CRC diagnosis is much more important for the 1-year model than the 5-year model (-1.1881 vs. -0.3092).

- New variables not previously seen in the literature are identified as important. These include number of days between CRC diagnosis and hospitalization as well as the principal language of the patient.

Figure 3.7 shows coefficient magnitude of four independent variable in lasso as a function of horizon. The variables are chosen based on the their importance in Table 3.4. These variables indicate whether the tumor was at stage IV at the time of diagnosis, surgery performed as the first course of treatment, the patient was diagnoses with secondary malignant neoplasm of liver, and patient’s age at the time diagnosis. Based on the figures, as the mortality horizon increases, whether the tumor was diagnosed at stage IV or not plays a more important role. Whether the surgery was performed on the patient and age at diagnosis have almost a constant coefficient size throughout. However, their effect on CRC mortality is reverse, i.e., the higher the age, the more the chance of mortality. However, if surgery
is performed on a patient, the chances of CRC mortality decreases on the patient. Secondary malignant neoplasm of liver has a direct effect on CRC mortality, i.e., if the patient is diagnosed with metastasis to liver, the patient has higher chance of CRC mortality.

**Figure 3.7.** Coefficient size for relaxed lasso over time, for four variables stage IV tumor, surgery, metastasis to the liver, and age.
3.4. Variable Importance

Coefficient Profiles

As mentioned in the introduction to section 3.3.1, when the value of lasso penalty, $\lambda$, decreases, more coefficients become non-zero in the lasso and relaxed lasso models. The sequence at which the independent variables enter the model as the value of $\lambda$ decreases was used as a measure of variable importance; the variables which enter the model earlier may be interpreted to be more important in predicting the outcome.

Figure 3.8 shows coefficient profile plots under relaxed lasso and for the case of 30-day, 1-year, and 5-year CRC prediction models. Only the top 20 variables are plotted. Table 3.5 shows the list of variables as their coefficients become non-zero with the decrease in $\lambda$. It is interesting to note that many of the variables common across the three time horizons have the same coefficients sign in all the coefficient profiles. However, this is not the case for all the variables. For example, stage IV has negative coefficient in the 30-day model while it has positive coefficients in the other two models. From this, we learn that if a patient is diagnosed with stage IV CRC tumor and is hospitalized after diagnosis, the patient has a higher probability of surviving the first 30 days, while the patient will not likely survive the 1-year and 5-year horizons. The rankings from coefficient profiles do not match the rankings from coefficient magnitudes in Table 3.4, except for stage IV variable in the 5-year model. However, there are common variables, among the two measures. For example, whether the tumor is diagnosed as stage IV, metastasis to liver, lymph node ratio, metastasis, age, ninety or more positive lymph nodes, and more. Other variables which are important based on coefficient size but not coefficient profile are days between admission to hospital and CRC diagnosis, principal language of the patient, stage 0 and I, and more. From Figure 3.8 we also learn that whether the tumor is stage IV or not may be a much more important variable in predicting mortality in the 5-year model since the penalty $\lambda$ should decrease significantly before coefficient of a new variable (lymph node ratio) becomes non-zero. It is also important to note that although coefficient size did not provide a ranking for the 30-day model, coefficient profile provides such a ranking. In this sense, coefficient profiles can be more useful than coefficient sizes.
Adjusted Percentage of Improvement in Reclassification

Percentage of improvement in reclassification has been used in the medical literature as a measure that captures clinical importance as opposed to mere statistical significance [110].

**Figure 3.8.** Coefficient profiles for relaxed lasso. The best λ, and therefore the number of non-zero coefficients, was chosen based on 10-fold cross validation. log(λ) is -7.6, -5.9, and -3.4, corresponding to 30-day, 1-year, and 5-year horizons.
Table 3.5. Variable coefficients as they become nonzero when $\lambda$ decreases for 30-day, 1-year, and 5-year prediction models with relaxed lasso.

It measures how much improvement an independent variable can make in terms of the number of correct classifications when compared to a base model. Figure 3.9 shows how the
The green and red rectangles in the base model correspond to the correct and incorrect classifications. Once the independent variable is added to the model, some of the observations which were correctly classified under the base model may be incorrectly classified (hatched red rectangle) while some of the previously incorrectly classified observations may now be classified correctly (hatched green rectangle).

As it is clear from Figure 3.9-(a), the measure used in the literature overestimates the improvement a variable makes in the model by ignoring the hatched red rectangle. I propose using adjusted percentage improvement in reclassification, AIR. Two cases are considered. First, the improvement that the inclusion of a variable can bring to the model is compared to the constant guess, i.e., the case where no independent variable is used. Second, the improvement that the inclusion of the variable can bring into the model is measured by comparing the adjusted correct reclassification when all the variables are included vs. when all but the specific variable is included in the model. The results of these two extreme cases for a sample of variables are shown in Figure 3.10. The orange plot corresponds to the first case and the blue to the second. Figure 3.10 shows AIR for a selected set of independent variables used in relaxed lasso. AIR is plotted over the time horizon to show how the importance of the variable changes as the CRC time horizon changes. The two cases can be considered as two extremes in capturing the most and the least improvements a variable can offer to the classification problem. Stage of the disease has the most dominant contribution to classification improvement when compared to lymph node ratio, secondary malignant neoplasm of liver, and receiving surgery as the first course of treatment. In the absence of any other variable in the model, the stage of the disease captures 79% of the accuracy while the accuracy raises to 87% when all the other variables are included in the model. As shown by the blue curve in 3.10-(a), the effect of stage is not completely substituted by other variables in the model in its absence for time horizons more than two years. When stage is removed from the model, lymph node ratio which has a correlation of 0.3 with stage partially covers the absence of stage. The other three independent variables, namely, lymph node ratio, secondary malignant neoplasm of liver, and are almost completely substituted in by other variables, as indicated by the flat and very close to 0 AIR for these variables in Figure 3.10.

Not all the variables are as important in predicting CRC mortality as stage, lymph node
3.4. Variable Importance

\[ \text{percentage of improvement in reclassification} = \frac{\text{adjusted percentage of improvement in reclassification}}{\text{base model}} \times 100 \]

(a) Percentage of improvement in reclassification

(b) Adjusted percentage of improvement in reclassification

Figure 3.9. Two measures of improvement in reclassification.

ratio, and surgery. For example, it is interesting to note that the type of admission to the hospital (whether the patient had a scheduled or unscheduled hospitalization) and race do not predict CRC mortality in the majority of the horizons considered. AIR for these two variables is shown in Figure 3.11. From the two variables, type of hospital admission is in the list of top twenty most important variables using coefficient size and coefficient profile. None of the measures used found race to be an important variable in predicting CRC mortality.
Chapter 3. Colorectal Cancer Mortality

Figure 3.10. Adjusted percentage of improvement in reclassification. The blue plot corresponds to the case where the improvement is relative to the model where all the independent variables are present. The orange plot corresponds to the case where the base case corresponds to constant guess.
Figure 3.11. Adjusted percentage of improvement in reclassification for type of admission and race
In this dissertation, prediction models were utilized to build models that can predict patient non-attendance at medical appointments as well as colorectal cancer mortality. Through the first application, a new methodology was introduced to optimize operationalizing prediction models in practice. Namely, I showed how the costs and benefits of interventions can be taken into account to find the optimal intervention policy given a family of prediction models.

In the second application, prediction models for colorectal cancer mortality were built, variable importance measures were studied, and a new variable importance measure was introduced. There were two main outcomes from the chapter. First, inclusion of all the independent variables available in the model can help us find new variables which may predict CRC mortality but would not be found through traditional channels of including only a limited number of variables. Second, prediction models with very high ROC performance measured by AUC were built.

An important question to ask is how the prediction models and ROC curves developed in Chapter 3 can be utilized to help the different decision makers in the health-care system, i.e., patients, physicians, providers, and policy makers. Two scenarios need to be considered to address this question. In the first scenario, the cost and benefit corresponding to one or more actions which need to be taken based on a chosen model are known. For example, the cost of screening colorectal cancer in people over age 40 along with its benefit. Under this scenario, the methodology developed in Chapter 2 can be utilized to
find optimal intervention policies, i.e., the optimal number of people in the population who should be screened. The second scenario is one where the cost and benefit information is not available. This is not an uncommon case in health applications. For example, the value of extra year of lower quality life is not necessarily quantifiable for many who need to make decisions about the type of treatment to receive. Under this second scenario, different mechanisms for finding an optimal solution may need to be developed and different circumstances might require different mechanisms for decision making.

With the advancements in prognostic, diagnostic, and treatment procedures, specially in the case of cancer, both scenarios mentioned above require solutions that account for specific patient and disease characteristics, which may influence mortality, as well as other outcome measures such as recurrence of the disease. Building reliable and accurate prediction models is a key step in achieving this goal.
A ROC curve depicts the relation between false-positive and true-positive joint probabilities. The area under the ROC curve (AUC), also known as c-statistic, is commonly used as a performance measure for prediction algorithms. AUC is a measure of how well the model can be used to discriminate between two distinct outcomes, one called “positive” and the other “negative”. Examples of positive and negative outcomes in the medical context are: survival vs death, high blood pressure vs low blood pressure, etc. Any prediction algorithm that has not been severely overfit has an AUC in $[0.5, 1.0]$ where 0.5 corresponds to a performance as good as a random guess and 1.0 to perfect prediction performance. The quality of any model is affected by numerous factors. One such factor is whether the threshold for labeling an observation as positive or negative is used in building the prediction models. We address this issue in more details.

A prediction algorithm assigns a probability to each observation. Whether the user interprets a certain probability assigned by the algorithm to an observation as “positive” or “negative” depends on a user selected threshold, which is reflective of the false-positive and false-negative trade-off. There are two approaches to using the threshold in building a family of prediction models, resulting in two ROC curves.

In the first case, a single model is built and the threshold is only used at the final stage of the algorithm for converting the probabilities of the predicted outcome being positive to 0
or 1, corresponding to negative or positive outcome, respectively. In this case, the threshold is varied between 0 and 1. For a threshold value $t$, all observations with a predicted value above $t$ are labelled as 1; all others as 0. For example, when $t$ is 0.5, observations with predicted probability greater than 0.5 will be labeled as class 1 while other observations will be labeled as class 0.

In the second case, multiple models are built, each minimizing a cost function that reflects a false-positive and false-negative trade-off. In this case, the threshold $t$ is defined for each model by the costs of false-positive and false-negative used in the objective function of the model. When these mis-classification costs are equal, $t$ is 0.5. When they are not, following Bayes rule, one can find $t$ by dividing the cost of false-positive error by the some of the misclassification costs.

Figure A.1 shows the ROC curves corresponding to the two approaches. Here, the dashed line is the ROC curve generated using the first approach, i.e., equal cost of false-positive and false-negative were considered when building the prediction model and the cut-off threshold was varied after the model was built. The solid line is the ROC curve for the second case where different costs of misclassification are considered in building the prediction models. In this case, for each point on the ROC curve the threshold is computed following the Bayes rule. We observe that in terms of prediction model performance, AUC, the second approach does better. The results used in this thesis are based on the second approach.
Figure A.1. Two ROC curves generated using the two approaches for varying the cut-off threshold. The ROC curve corresponding to the second approach (Bayes Threshold) does better than the one corresponding to the first approach (Simple threshold).
Appendix B

List of Independent Variables for Non-Attendance at Medical Appointments

The variables used for predicting non-attendance in Section 2.3 are listed below. Some of the variables listed are directly provided through the dataset. However, most of the variables are a result of combining a number of data elements. For example, any variable with information about patients’ previous appointment day was created based on data on the history of the patient in terms of treatment, diagnosis, general state of health, and commitment to previous appointments. There are a total of 104 variables used in the analysis.

Treatment

- Any or all of the following treatments received in the previous visit: AC/Taxol, Avastin/Folfiri, Avastin-Support, Carboplatin/Taxol, Carboplatin-VP16, Folfox, Gemcitabine, Herceptin, Taxotere, Velcade, Rituxan.

- During the last visit the patient was treated with: cancer treatment, cancer treatment side-effect control, non-cancer treatment, vascular disease treatment, vitamin treatment, pain treatment, nausea/vomiting or fatigue medication.

- Maximum and minimum cycle numbers of treatments to be received during the hospital visit.
• Infection and allergy indicators present during the previous visit. Maximum and minimum phase sequence number for the current visit.

• Previous visit’s treatment type: agent, toxicity, consult, or report.

Severity of the Disease

• Stage of the most recent cancer diagnosed.

• Number of previous ICD codes and cancer ICD codes.

• Number of treatments received during the previous visit.

• ICD codes with more than 1000 times occurrence (162.3, 162.9, 174.1, 174.8, 174.9, 185, 197.7, 198.3, 198.5, 285.21, 285.9).

• Number of days between the day of last visit and the current visit, between the onset date of the most recent cancer and today, the onset date of the most recent diagnosis code and today, the minimum number of days between the onset day and today, and the maximum number of days between onset day and today.

Commitment

• First time patient.

• If the previous visit was non-attended, the reason for it: patient availability, patient decision, patient condition, patient deceased, provider availability, provider decision, location availability, machine availability, transportation issue, inclement weather, rescheduled, other.

• Number of days between the day of the last treatment initiation and the visit date.

• Fraction and number of previous cancellations, no-shows, rescheduled visits, rescheduled visits attended, and visits.
Day of the Visit

- Visit start time.
- Day of the week, Day of the month, Day of the year, Week of the year, and month.

General Information

- Age at the most recent diagnosis, at the most recent cancer diagnosis, and at the time of appointment.
- Sex.
- Distance from home to hospital.
Appendix C

List of Colorectal Cancer Study Independent Variables

All Census variables are based on the block group the patient lives in at the time of diagnosis. Variables marked with “***” are defined as part of this research while the others were directly used from the corresponding data-base.

Table C.1. Patient demographics and General Information

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age at the time of tumor diagnosis</td>
</tr>
<tr>
<td>o_ageyears</td>
<td>Age at the time of the last hospital discharge</td>
</tr>
<tr>
<td>marstat</td>
<td>Marital status when the patient was first diagnosed with the tumor</td>
</tr>
<tr>
<td>race08</td>
<td>Race/ethnicity of the patient including Hispanic ethnicity at the time of diagnosis</td>
</tr>
<tr>
<td>sex</td>
<td>Sex of the patient at the time of tumor diagnosis</td>
</tr>
<tr>
<td>city</td>
<td>City the patient lived in at the time of diagnosis</td>
</tr>
<tr>
<td>county</td>
<td>County the patient lived in at the time of diagnosis</td>
</tr>
</tbody>
</table>
### Table C.2. Socio-economic status variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bellow200</td>
<td>Percent of population who are at or below the 200% of poverty line (Census)</td>
</tr>
<tr>
<td>college</td>
<td>Proportion of those 25 years of age and over who have a college degree (Census)</td>
</tr>
<tr>
<td>nojob</td>
<td>Proportion of those 16 years of age or older in the labor force that are unemployed (Census)</td>
</tr>
<tr>
<td>mdinc</td>
<td>Median household income (Census)</td>
</tr>
<tr>
<td>yostscl</td>
<td>Kathy Yosts index of socioeconomic status</td>
</tr>
<tr>
<td>quinyost</td>
<td>Quintile of YOSTSCL score</td>
</tr>
<tr>
<td>payer</td>
<td>Primary source of payment to the hospital</td>
</tr>
<tr>
<td>payer2</td>
<td>Secondary source of payment to the hospital</td>
</tr>
<tr>
<td>distancereshosp*</td>
<td>Distance from patient’s residence to the hospital</td>
</tr>
</tbody>
</table>
### Table C.3. Comorbidities

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>Charlson comorbidity score</td>
</tr>
<tr>
<td>renal</td>
<td>Renal disease</td>
</tr>
<tr>
<td>mi</td>
<td>Myocardial infarction</td>
</tr>
<tr>
<td>chf</td>
<td>Congestive heart failure</td>
</tr>
<tr>
<td>pvd</td>
<td>Peripheral vascular disease</td>
</tr>
<tr>
<td>cvd</td>
<td>Cerebrovascular disease</td>
</tr>
<tr>
<td>dementia</td>
<td>Dementia</td>
</tr>
<tr>
<td>copd</td>
<td>Chronic pulmonary disease</td>
</tr>
<tr>
<td>rheum</td>
<td>Connective tissue disease</td>
</tr>
<tr>
<td>pud</td>
<td>Peptic ulcer disease</td>
</tr>
<tr>
<td>mildliver</td>
<td>Mild liver disease</td>
</tr>
<tr>
<td>dmcomp</td>
<td>Diabetes Mellitus</td>
</tr>
<tr>
<td>plegia</td>
<td>Hemiplegia</td>
</tr>
<tr>
<td>malignancy</td>
<td>Malignant lymphoma</td>
</tr>
<tr>
<td>severeliver</td>
<td>Sever liver disease</td>
</tr>
<tr>
<td>mets</td>
<td>Metastasis</td>
</tr>
<tr>
<td>hiv</td>
<td>AIDS</td>
</tr>
<tr>
<td>o_diags_1 to 24</td>
<td>ICD-9 diagnosis code for disease diagnosed prior to hospitalization or while in hospital</td>
</tr>
<tr>
<td>o_diags_pa_1 to 24</td>
<td>Whether the diagnosis code was present at admission</td>
</tr>
<tr>
<td>o_prin.diag</td>
<td>Diagnosis code which is known to be the reason for hospitalization</td>
</tr>
<tr>
<td>xnode.tu</td>
<td>Number of regional lymph nodes examined</td>
</tr>
<tr>
<td>pnode.tu</td>
<td>Number of positive lymph nodes</td>
</tr>
<tr>
<td>pnode_to_xnode*</td>
<td>Lymph node ratio</td>
</tr>
</tbody>
</table>
Table C.4. Hospital discharge data

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>o_lenstay</td>
<td>Length of stay in the hospital during the last visit</td>
</tr>
<tr>
<td>o_srceadm</td>
<td>Source of admission to the hospital</td>
</tr>
<tr>
<td>o_typeadm</td>
<td>Type of admission (scheduled, unscheduled, etc.)</td>
</tr>
<tr>
<td>o_disppat</td>
<td>Type of disposition at the time of discharge (routine, within this hospital, etc.)</td>
</tr>
<tr>
<td>o_exptpay</td>
<td>Expected source of payment (Medicare, Medi-cal, etc.)</td>
</tr>
<tr>
<td>o_paypln</td>
<td>Name of plan if licensed under Knox-Keene Healthcare Service Plan Act of 1975 or MCOHS</td>
</tr>
<tr>
<td>o_totchrg</td>
<td>Total charges of hospitalization</td>
</tr>
<tr>
<td>o_mdc</td>
<td>Major diagnostic category</td>
</tr>
<tr>
<td>o_visitno</td>
<td>Appointment number in the year of hospitalization for the patient</td>
</tr>
<tr>
<td>o_typecovg</td>
<td>Type of coverage (managed care, traditional coverage)</td>
</tr>
<tr>
<td>o_plsid1</td>
<td>The preferred language of the patient</td>
</tr>
<tr>
<td>o_typcare</td>
<td>Type of care (acute, skilled nursing/immediate care, psychiatric care, chemical dependency recovery care, physical rehabilitation care)</td>
</tr>
<tr>
<td>o_source</td>
<td>Source of admission</td>
</tr>
<tr>
<td>o_srcsite</td>
<td>The site of source of admission (home, ambulatory, and so on)</td>
</tr>
<tr>
<td>o_srclicens</td>
<td>The licensure of site of admission (this hospital, another hospital, not a hospital)</td>
</tr>
<tr>
<td>o_srcroute</td>
<td>Route of admission (hospital’s emergency room, not emergency room or another hospital’s emergency room)</td>
</tr>
</tbody>
</table>
### Table C.5. Hospital-specific information

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>teachrurl</td>
<td>teaching or rural hospital</td>
</tr>
<tr>
<td>bedlic</td>
<td>Number of licensed beds</td>
</tr>
<tr>
<td>bedstf</td>
<td>Number of staffed beds</td>
</tr>
<tr>
<td>daymcaltr</td>
<td>Patient days Medi-Cal - traditional</td>
</tr>
<tr>
<td>daymcalmc</td>
<td>Patient days Medi-Cal - managed</td>
</tr>
<tr>
<td>daymcartr</td>
<td>Patient days Medicare - traditional</td>
</tr>
<tr>
<td>daymcarmc</td>
<td>Patient days Medicare - managed</td>
</tr>
<tr>
<td>daytot</td>
<td>Total number of patient days</td>
</tr>
<tr>
<td>dismcartr</td>
<td>Number of discharged Medicare - traditional</td>
</tr>
<tr>
<td>dismcarmc</td>
<td>Number of discharged Medicare - managed</td>
</tr>
<tr>
<td>dismcaltr</td>
<td>Number of discharged Medi-Cal - traditional</td>
</tr>
<tr>
<td>dismcalmc</td>
<td>Number of discharged Medi-Cal - managed</td>
</tr>
<tr>
<td>distot</td>
<td>Total number of discharges</td>
</tr>
<tr>
<td>disacute</td>
<td>Total number of acute discharges</td>
</tr>
<tr>
<td>netincome</td>
<td>Net income</td>
</tr>
<tr>
<td>oproom</td>
<td>Number of operating rooms</td>
</tr>
<tr>
<td>surgip</td>
<td>Number of inpatient surgeries</td>
</tr>
<tr>
<td>surgop</td>
<td>Number of outpatient surgeries</td>
</tr>
<tr>
<td>prdhrmgnt</td>
<td>Number of productive hours - management and supervision</td>
</tr>
<tr>
<td>prdhrtch</td>
<td>Number of productive hours - technical and specialist</td>
</tr>
<tr>
<td>percdaymcar*</td>
<td>Percentage of days served Medicare patients</td>
</tr>
<tr>
<td>percdaymcal*</td>
<td>Percentage of days served Medi-Cal patients</td>
</tr>
<tr>
<td>percdismcar*</td>
<td>Percentage of Medicare discharges</td>
</tr>
<tr>
<td>percdismcal*</td>
<td>Percentage of Medi-Cal discharges</td>
</tr>
<tr>
<td>perstaffedbed*</td>
<td>Percentage of staffed beds</td>
</tr>
<tr>
<td>typecntrl</td>
<td>Type of control of the hospital</td>
</tr>
</tbody>
</table>
Table C.6. Tumor Diagnosis

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>site_02</td>
<td>Location where the tumor originated (ICD-O-3 codes)</td>
</tr>
<tr>
<td>differn2</td>
<td>The grade or differentiation of the tumor (Sixth digit of ICD-O-3)</td>
</tr>
<tr>
<td>histo_m3</td>
<td>Malignancy or behavior of this tumor (The fifth digit of the ICD-O-3)</td>
</tr>
<tr>
<td>histo_t3</td>
<td>Histology of the tumor (The first four digits of the ICD-O-3 morphology code)</td>
</tr>
<tr>
<td>marker3</td>
<td>Records LDH for testicular cancer</td>
</tr>
<tr>
<td>markerca</td>
<td>Breast cancer tumor marker for California: Her2/neu</td>
</tr>
<tr>
<td>o_prindiag_mod*</td>
<td>The patients principal diagnosis is the condition established to be the chief cause of the admission to the hospital</td>
</tr>
<tr>
<td>stage_seer</td>
<td>Stage of the tumor at diagnosis as reported by SEER</td>
</tr>
</tbody>
</table>
Table C.7.  Treatment

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>chemosum</td>
<td>Chemotherapy as the first course of treatment</td>
</tr>
<tr>
<td>hormsum</td>
<td>Hormonotherapy as the first course of treatment</td>
</tr>
<tr>
<td>radbstmod</td>
<td>Volume or anatomic target of the most clinically significant regional radiation therapy</td>
</tr>
<tr>
<td>radseq</td>
<td>Sequence of radiation therapy with surgery</td>
</tr>
<tr>
<td>radsum</td>
<td>Radiation therapy as the first course of treatment</td>
</tr>
<tr>
<td>norad</td>
<td>Reason why the first course of treatment did not include radiotherapy</td>
</tr>
<tr>
<td>nosurg</td>
<td>Reason why the first course of treatment did not include surgery</td>
</tr>
<tr>
<td>surgprim</td>
<td>Most extensive type of surgery as the first course of treatment</td>
</tr>
<tr>
<td>scope</td>
<td>Surgery removing regional lymph nodes during the first course of treatment</td>
</tr>
<tr>
<td>o_oprocs</td>
<td>Procedures operated on the patient while hospitalized</td>
</tr>
<tr>
<td>o_dnr</td>
<td>Do not resuscitate</td>
</tr>
<tr>
<td>o_admtype</td>
<td>Admission type to hospital (scheduled or unscheduled)</td>
</tr>
<tr>
<td>no_doctors*</td>
<td>Number of physicians the patient was in contact with for treatment</td>
</tr>
<tr>
<td>nouniquedoctors*</td>
<td>Number of unique physicians the patient was in contact with for treatment</td>
</tr>
<tr>
<td>perc_doctors*</td>
<td>Percentage of unique physicians the patient was in contact with for treatment</td>
</tr>
</tbody>
</table>
### Table C.8. Days Between Major Events

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>datedx_diff_datedxcolrec1*</td>
<td>Number of days between the most recent CRC diagnosis and the first CRC tumor diagnosis</td>
</tr>
<tr>
<td>odateadm_diff_datedxcolrec1*</td>
<td>Number of days between the most recent date of admission* to the hospital and the first CRC diagnosis</td>
</tr>
<tr>
<td>o_datedis_diff_datedxcolrec1*</td>
<td>Number of days between the most recent discharge from the hospital and the first CRC diagnosis</td>
</tr>
<tr>
<td>o_datepp_diff_datedxcolrec1*</td>
<td>Number of days between the date of principal procedure during the last hospital visit and the first CRC diagnosis</td>
</tr>
<tr>
<td>rxdater_diff_datedxcolrec1*</td>
<td>Number of days between the date radiation therapy started as the first course of treatment and the first CRC diagnosis</td>
</tr>
<tr>
<td>rxdateh_diff_datedxcolrec1*</td>
<td>Number of days between the date hormonotherapy started as the first course of treatment and the first CRC diagnosis</td>
</tr>
<tr>
<td>rxdatesn_diff_datedxcolrec1*</td>
<td>Number of days between the date of diagnosis or staging procedure and the first CRC diagnosis</td>
</tr>
<tr>
<td>surgdate_diff_datedxcolrec1*</td>
<td>Number of days between the date the earliest definitive surgery was performed and the first CRC diagnosis</td>
</tr>
<tr>
<td>rxdate_diff_datedxcolrec1*</td>
<td>Number of days between the first course of definitive treatment and the first CRC diagnosis</td>
</tr>
<tr>
<td>rxdatec_diff_datedxcolrec1*</td>
<td>Number of days between the start date of chemotherapy as the first course of treatment and the first CRC diagnosis</td>
</tr>
</tbody>
</table>
Bibliography


