Novel Statistical Learning Methods for Multi-Modality Heterogeneous Data Fusion in Health Care Applications

by

Xiaonan Liu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2019 by the
Graduate Supervisory Committee:

Jing Li, Chair
Teresa Wu
Rong Pan
Mirek Fatyga

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

With the development of computer and sensing technology, rich datasets have become available in many fields such as health care, manufacturing, transportation, just to name a few. Also, data come from multiple heterogeneous sources or modalities. This is a common phenomenon in health care systems. While multi-modality data fusion is a promising research area, there are several special challenges in health care applications. (1) The integration of biological and statistical model is a big challenge; (2) It is commonplace that data from various modalities is not available for every patient due to cost, accessibility, and other reasons. This results in a special missing data structure in which different modalities may be missed in "blocks". Therefore, how to train a predictive model using such a dataset poses a significant challenge to statistical learning. (3) It is well known that different modality data may contain different aspects of information about the response. The current studies cannot afford to solve this problem. My dissertation includes new statistical learning model development to address each of the aforementioned challenges as well as application case studies using real health care datasets, included in three chapters (Chapter 2, 3, and 4), respectively. Collectively, it is expected that my dissertation could provide a new sets of statistical learning models, algorithms, and theory contributed to multi-modality heterogeneous data fusion driven by the unique challenges in this area. Also, application of these new methods to important medical problems using real-world datasets is expected to provide solutions to these problems, and therefore contributing to the application domains.

This one's dedicated to Lord Jesus and my family,

who've blessed me in more ways than I can tell.

ACKNOWLEDGMENTS

To begin with, I would like to thank my advisor, Jing Li, who taught me how to work independently and who encouraged and supported me during my research and dissertation. Her attitude on life and research profoundly influenced me during the last several years and also will help me through my entire life. I would also like to specially thank Teresa Wu for providing valuable suggestions and kind guidance. It was a great pleasure working with Dr. Wu at the ASU-Mayo Clinic Center for Innovative Imaging (AMCII). I would also like to show my gratitude to Dr. Mirek Fatyga, who supported my research during the last few years. I would also like to express gratitude to members of my committee for their insightful comments, valuable interactions, and advice: Dr. Rong Pan.

Thank you to all members of AMCII, past and present: Na Zou, Shuluo Ning, Kun Wang, Bing Si, Hyunsoo Yoon, Fei Gao, Yinlin Fu, Nathan Gaw, Lujia Wang, Yanzhe Xu, Zhiyang Zheng, Jiajing Huang and Hope Lancaster. I would like to warmly acknowledge my parents for unconditionally supporting me all days and nights.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES viii

CHAPTER

1 INTRODUCTION 1

1.1 Background 1

1.2 Challenges and Expected Original Contributions 2

1.3 Dissertation Outline 5

2 A BIO-STATISTICAL HYBRID MODEL FOR INTEGRATION OF DOSE MAP AND PATIENT CHARACTERISTICS WITH APPLICATION IN RADIATION TOXICITY PREDICTION OF PROSTATE CANCE 6

2.1 Introduction 6

2.2 Integration of Biological and Statistical Models in NTCP Modeling 10

2.3 Case Study 22

2.4 Conclusion 32

3 A TRANSFER LEARNING MODEL FOR MULTI-MODALITY IMAGE DATA FUSION WITH APPLICATION IN ALZHEIMER'S DISEASE EARLY DETECTION 34

3.1 Introduction 34

3.2 Literature Review 37

CHAPTER Page

3.3 Development of the Incomplete-Multimodality Transfer Learning (IMTL) Model 39

3.4 Properties of IMTL 51

3.5 Application Case Study 56

3.6 Conclusion 63

4 A HOTSPOT SEARCH ALGORITHM FOR FUSION OF 3-D RADIATION DOSE MAP AND PATIENT CHARACTERISTICS WITH APPLICATION IN CARDIAC TOXICITY ASSESSMENT OF LUNG CANCER 65

4.1 Introduction 65

4.2 Literature Review 68

4.3 Mathematical Formulation for the Hot Spot Search Problem 70

4.4 Algorithm Development for Hot Spot Search 74

4.5 Case Study 77

4.6 Conclusion 81

5 CONCLUSIONS 82

REFERENCES 83

APPENDIX

A SUPPLEMENTAL MATERIALS FOR CHAPTER 2 91

B SUPPLEMENTAL MATERIALS FOR CHAPTER 3 96

C SUPPLEMENTAL MATERIALS FOR CHAPTER 4 99

LIST OF TABLES

Table Page

1. Comparison between the AUCs of the Integrated Model, Biological Model (LKB) and Statistical Model (l_1-penalized logistic regression) ····· 26
2. Comparison between the Integrated Model, Biological Model, and Statistical Model ($|\alpha_k| = 0.8, \alpha_{p+1} = 0.25, n = 0.2$) . Mean (Standard Deviation) and Confidence Interval are Computed based on 50 Repetitions of the Simulation Experiment ····· 28
3. Comparison between the Integrated Model, Biological Model, and Statistical Model ($|\alpha_k| = 0.6, \alpha_{p+1} = 0.25, n = 0.2$) . Mean (Standard Deviation) and Confidence Interval are Computed Based on 50 Repetitions of the Simulation Experiment ····· 29
4. Comparison between the Integrated Model, Biological Model, and Statistical Model ($|\alpha_k| = 0.8, \alpha_{p+1} = 0.3, n = 0.3$) . Mean (Standard Deviation) and Confidence Interval are Computed based on 50 Repetitions of the Simulation Experiment ····· 29
5. Comparison between the Integrated Model, Biological Model, and Statistical Model ($|\alpha_k| = 0.6, \alpha_{p+1} = 0.3, n = 0.3$) . Mean (Standard Deviation) and Confidence Interval are Computed based on 50 Repetitions of the Simulation Experiment ····· 30

Table Page

6. Out-of-sample Prediction Accuracy on the Test Set with Different Training Sample Sizes (Between-Modality Correlation is Kept as 0.6 in Both Settings) ····· 58

7. Out-of-sample Prediction Accuracy on the Test Set with Different Between-Modality Correlations (Training Sample Size is Kept as 300 in Both Settings) ····· 58

8. Out-of-sample Classification Accuracy on the Test Set with Different Training Sample Sizes (Between-Modality Correlation is Kept as 0.6 in Both Settings) ····· 59

9. Out-of-sample Classification Accuracy on the Test Set with Different Between-Modality Correlations (Training Sample Size is Kept as 300 in Both Settings) ···59

10. Diagnostic and Prognostic Performance: Ave (std) and P Value for Hypothesis Testing that IMTL is Better Than a Competing Method ····· 63

LIST OF FIGURES

Figure Page

1. (a) Cumulative DVH, and (b) Differential DVH on Rectum for a Prostate Cancer Patient Receiving IMRT ····· 11
2. Probability of Complication, $P(Y = 1)$, with respect to $gEUD$ in LKB Model ····· 12
3. A Flow Chart of the Steps for the Proposed Integrated Model ····· 22
4. An Example of the Incomplete Multimodality Dataset (IMD), in Which MRI, FDG-PET, and Amyloid-PET are Considered as Three Modalities. Columns within Each Modality Represent Features Extracted from the Image. Each Sub-cohort Consists of Patients with the Same Availability of Modalities ····· 35
5. Computational Architecture for Between-Institutional/Sub-cohort Collaborative Model Estimation of IMTL without the Need for Data Pooling from Different Sub-cohorts. M-step is Performed by a Global Learner; E-step is Performed Locally at Each Sub-cohort ·····50
6. Percentage of Times Imaging Features are Included in IMTL over 5-fold Cross-Validation and 50 Repeated Experiments ·····63
7. Example of a DVH and Definition of the Fractional Volumetric Variable V_D ···69

Figure Page

8. Hot Spot Sub-Region Found by the Proposed Search Algorithm Overlaid on CT Volume of the Heart (the Hot Spot Sub-region Consists of 19 Units Highlighted in Red of Two Different Shades Used to Represent Different p-value Ranges in the Cox Model Coefficients Corresponding to the Units). ·····79

9. Survival Curves of Two Patients with the Same Covariates but Different $V_{D^*}^{(r^*)}$ ·····84

CHAPTER 1

INTRODUCTION

1.1 Background

With the development of computer and sensing technology, rich datasets have become available in many fields such as health care, manufacturing, transportation, just to name a few. Also, data come from multiple heterogeneous sources or modalities. This is a common phenomenon in health care systems. Here we give two examples:

- In neurological disease diagnosis, such as the Alzheimer's disease (AD), imaging of different modalities provides important and complementary roles. For example, magnetic resonance imaging (MRI) is used to identify structural changes of the brain. Positron emission tomography (PET) imaging is used to identify functional alteration. In addition to imaging, other data modalities such as genetics and cognitive tests also provide important value.
- In radiation oncology, it is an important task to predict radiation toxicity for patients receiving radiation therapy (RT). Various modalities of data can be integrated to accomplish this task, such as radiation dose map, patient demographics, disease characteristics, and concurrent treatment.

Because each data modality contains unique and usually only partial information about the medical problem of interest, statistical learning models based on a single modality can rarely achieve optimal results. Multi-modality data fusion can leverage the joint strength of different modalities and provide a better solution.

1.2 Challenges and Expected Original Contributions

While multi-modality data fusion is a promising research area, there are several special challenges in health care applications.

Challenge I (Integration of biological and statistical models): While statistical models can be used to model any data modality in the general sense, biological models may exist for *some* modality. For example, to predict rectal toxicity for prostate cancer patients receiving radiation therapy, a biological model called the Lyman-Kutcher-Burman (LKB) model exists, which is based on normal the understanding of normal tissue cells' response to injury by ionizing radiation. Biological models contain existing knowledge of the problem of interest, in this case, how radiation dose affects toxicity, but may be based on simplified assumptions. Statistical models, on the other hand, are data-driven, and can leverage the useful information in the data to the maximum extent, but may be limited in obeying biological constraints and providing biological insights or valid interpretation. Therefore, how to combine biological and statistical models is a significant challenge in multi-modality data fusion.

Expected original contribution to address Challenge I: We proposed a bio-statistical hybrid model for integration of 3-D dose map and patient characteristics in prediction of rectal toxicity of prostate cancer patients receiving intensity modulated radiation therapy (IMRT). Here, the 3-D dose map is considered a data modality and patient characteristics including demographics, disease characteristics, and concurrent treatment are considered another data modality because the latter takes a conventional vector format but the former

is 3-D data. Our goal is to preserve the biological knowledge in the LKB model, and meanwhile use statistical modeling strategies to allow for inclusion of patient-specific variables and to better account for patient heterogeneity in the predictive model. The integrated approach is expected to have better generalizability and predictive power than using biological and statistical models in separation.

Challenge II (Modality-wise missing data): It is commonplace that data from various modalities is not available for every patient due to cost, accessibility, and other reasons. For example, in AD diagnosis, MRI is almost available to all the patients because it is part of the routine clinical care. However, FDG-PET may only be available to some patients but not all because the equipment is more expensive and not available in some clinics. Some newly emerged imaging modalities such as amyloid-PET is not covered by some patients' medical insurance so they may also be unavailable to some patients. This results in a special missing data structure in which different modalities may be missed in "blocks". Therefore, how to train a predictive model using such a dataset poses a significant challenge to statistical learning.

Expected original contribution to address Challenge II: A commonly considered approach when there is missing data present in a training dataset is to use missing data imputation algorithms. However, this is not appropriate for datasets with modality-wise missing patterns (i.e., data is missed in blocks as above mentioned), because imputation algorithms typically assume the data is missing at random, which is violated in our problem. Also, there is too much missing data. Imputation will result in a training set with poor quality. We proposed a transfer learning model for integrating multi-modality data with

modality-wise missing patterns. Different from missing data imputation, our transfer learning model does not fill in missing data but utilize whatever data that is available to build a predictive model for each sub-cohort of patients with the same availability of imaging modalities. In this way, knowledge obtained from the modeling of each sub-cohort can be "transferred" to help the modeling of other sub-cohorts.

Challenge III: It is well known that different modality data may contain different aspects of information about the response. For instance, when assessing the heart toxicity among lung cancer patients after RT, the 3-D radiation dose map and patient characteristics are often used. The fusion of the two modalities appears to extract dosimetric/volumetric features from 3-D dose map and combine the features with patient characteristics in the recent studies. However, the extracted dosimetric features lost the spatial information of the radiation dose on heart. Clinicians also want to explore the heart sub-region to the high radiation dose, which is also called hotspot, that is associated with the heart toxicity. This spatial information is contained in the 3-D dose map, which requires researchers to figure out. The current studies cannot afford to solve this problem. Moreover, the dosimetric/volumetric features were determined under exhaustive search, which may cost too much computational time.

Expected original contribution to address Challenge III: We proposed a hotspot search algorithm for the fusion of 3-D radiation dose map and patient characteristics to predict the overall survival of the patients. The proposed algorithm divided the 3-D heart dose map into several regions, determined the dosimetric/volumetric features in each sub region and

finally identified the hotspots of the heart which is associated with overall survival given the patient characteristics.

1.3 Dissertation Outline

My dissertation includes new statistical learning model development to address each of the aforementioned challenges as well as application case studies using real health care datasets, included in three chapters (Chapter 2, 3, and 4), respectively. Collectively, it is expected that my dissertation could provide a new sets of statistical learning models, algorithms, and theory contributed to multi-modality heterogeneous data fusion driven by the unique challenges in this area. Also, application of these new methods to important medical problems using real-world datasets is expected to provide solutions to these problems, and therefore contributing to the application domains.

CHAPTER 2

A BIO-STATISTICAL HYBRID MODEL FOR INTEGRATION OF DOSE MAP AND PATIENT CHARACTERISTICS WITH APPLICATION IN RADIATION TOXICITY PREDICTION OF PROSTATE CANCER

2.1 Introduction

Radiation therapy (RT) is one of the most common treatments for cancer, either by itself or together with other forms of treatments. While the goal of RT is tumor control, it is also important in RT planning to spare the normal (surrounding) tissue from radiation toxicity that could greatly affect patients' quality of life. To understand the impact of radiation toxicity on normal tissue, a model is needed to link radiation dose of the RT with radiation-induced complications. This is known as the Normal Tissue Complication Probability (NTCP) model. There are two types of NTCP models: biological models and statistical models. Biological models are built upon the understanding of normal tissue cells' response to injury by ionizing radiation. Typical works include the Lyman model (Lyman, 1985), the Lyman-Kutcher-Burman (LKB) model (Deasy, 2000), the generalized Lyman model (Tucker et al., 2008), and the relative seriality model (Källman et al., 1992). Statistical models aim to associate features of radiation dose distribution among the receiving tissue with the risk of developing certain complications. There are many statistical models to choose from, since this is a typical classification/prediction problem in statistics so that theoretically speaking, any classification/prediction algorithm can be a potential candidate. Statistical models have been popularly used in recent years for NTCP

modeling (Cella et al., 2013; Liu et al., 2010; Michalski et al., 2010; Gulliford et al., 2012; Boomsma et al., 2012).

Next, we will discuss the advantages and disadvantages of biological and statistical models. One of the most important advantages for biological models is that the results can be generalized beyond a particular study, because biological models are built upon radiobiological principles. This provides convenience for clinical utilization, as clinicians may refer to published results when they do not have the data to build a biological model for their specific practice. Yet, the current knowledge about the biological reaction to radiation is still limited. As a result, biological models usually have a low accuracy in predicting the risk of radiation-induced complications. Another major contributor to the low prediction accuracy is the lack of "personalization", i.e., biological models do not factor in patient-specific information when linking radiation dose with complications. However, mounting evidence has shown that even with the same radiation dose distribution, different patients have different susceptibility to radiation toxicity and thereby different risks of developing complications (Cella et al., 2013; Boomsma et al., 2012).

In contrast, statistical models have the flexibility of incorporating patient-specific variables in addition to radiation dosimetric variables as predictors. In fact, modern development in statistics especially machine learning makes it possible to include a large collection of patient-specific variables such as demographics, health conditions, and even genetic and epigenetic markers. This enables a truly personalized approach in NTCP modeling. In addition to personalization, statistical models can account for the special RT data characteristics better. One distinct characteristic of the RT data is that the sample fraction

of a complication (i.e., the fraction of patients with the complication in the sample dataset used for statistical modeling) is typically much larger than the population fraction of the complication. This is because the latter fraction is usually a very small number (otherwise, the RT would not be permitted for practice). If the same fraction were used in the sample dataset, there would be too few patients with the complication, making the dataset uninformative. Increasing the sample fraction of the complication creates richer information content in the dataset, but the resulting statistical model may be biased and inconsistent. Fortunately, theoretical analysis is possible to quantify the level of inconsistency and bias for most statistical models. The result can further guide the development of effective consistency and bias correction strategies. Because of the aforementioned advantages, statistical models can usually have a better prediction accuracy for complications than biological models. However, a major drawback of statistical models is that the results heavily depend on the particular dataset used in each study, so they may not generalize well.

In this paper, we propose a general framework with specific methods for biological and statistical model integration in NTCP modeling. Our goal is to preserve the biological knowledge in NTCP modeling, and meanwhile use statistical modeling strategies to allow for inclusion of patient-specific variables and to better account for the special RT data characteristic. The integrated approach is expected to have better generalizability and predictive power than using biological and statistical models in separation.

The contributions of this paper are two-folds:

- Novel model development: We propose the first-of-its-kind framework for biological and statistical model integration. Under the framework, we propose the details for developing the integrated model, including a novel model formulation, an efficient algorithm for parameter estimation, and theoretical analysis guided consistency and bias corrections to guarantee that the model has good statistical properties.
- Real-data application: We apply the integrated model to a dataset of prostate cancer patients treated with Intensity Modulated RT (IMRT), an advanced type of RT, at Mayo Clinic Arizona. These patients are at risk of developing a serious complication called the grade 2+ acute rectal complication with symptoms including anal pain, diarrhea, and rectal obstruction. The integrated model achieves higher accuracy in predicting the complication compared with the statistical and biological models used in separation. Also, we perform extensive simulation studies on virtual patients whose data are sampled from the distribution of the real patients. Under various simulation settings, the integrated model outperforms the statistical and biological models in prediction accuracy due to the inclusion of patient-specific variables and in generalizability across different datasets due to the consideration of radiobiological principles.

The rest of this paper is organized as follows: Section 2.2 presents the development of the integrated model; Section 2.3 presents the application and simulation studies. Section 2.4 concludes the paper.

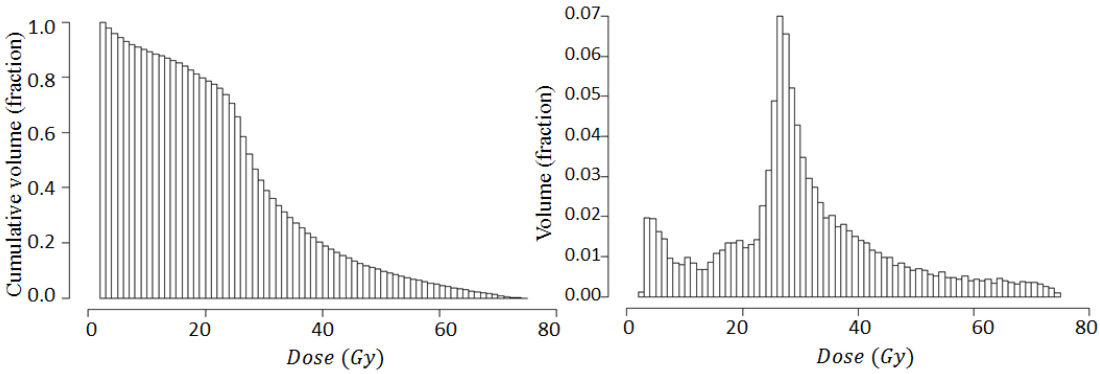
2.2 Integration of Biological and Statistical Models in NTCP Modeling

We propose a model integration framework that includes three major steps: (i) One should start with an in-depth understanding of the biological model, especially the biological meanings of the model parameters. This guides the decision on which parameter(s) are appropriate to be personalized. (ii) Then, the selected biological model parameter(s) is linked with patient-specific variables in an appropriate way, producing an integrated model formulation. This formulation needs to be properly designed to account for the potential high-dimensionality of patient-specific variables and biological constraints on the model parameters. Under the formulation, an optimization algorithm is further developed to estimate the parameters of the integrated model, with considerations on optimality and efficiency. (iii) Finally, statistical properties of the integrated model, such as consistency and bias, are investigated, and corrections are made in order to produce consistent and unbiased estimators for the model parameters. In what follows, we present the details of the proposed approaches for accomplishing steps (i)-(iii) in sub-sections 2.2.1- 2.2.3, respectively.

2.2.1. Understanding the biological model in NTCP modeling

The goal of NTCP modeling is to link the radiation dose delivered to a tissue of interest – a normal tissue to be spared from radiation toxicity – with the risk/probability of developing a complication. For each patient, the radiation dose distribution on the tissue volume can be extracted from the treatment planning software and represented by a Dose-Volume-Histogram (DVH). Figure II.1 shows the DVH on rectum – the normal tissue of interest – for a prostate cancer patient receiving IMRT. There are two types of DVH. A

differential DVH plots the fraction of tissue volume receiving a specific amount of dose in the unit of Gray (Gy). A differential DVH takes the appearance of a typical histogram. A cumulative DVH uses the same x axis as a differential DVH but plots on the y axis the fraction of tissue volume receiving greater than or equal to a specific amount of dose.



(a) (b)

Figure 2.1: (a) Cumulative DVH , and (b) differential DVH on rectum for a prostate cancer patient receiving IMRT

In biological NTCP models, radiobiological principles are typically used to guide (i) the selection or development of a metric from DVH, which captures the essence of normal issue cells' biological response to injury by ionizing radiation; and (ii) the determination of the functional form of the relationship between the metric and the probability of developing the complication. Next, we will present the details of a well-known biological NTCP model, called the LKB model (Gulliford et al., 2012). In the LKB model, a specific metric of DVH is used, called the generalized Equivalent Uniform Dose (gEUD), which takes the form of (II.1):

$$gEUD = \left(\sum_i v_i D_i^{1/n}\right)^n . \quad (2.1)$$

v_i is the fractional tissue volume receiving a dose D_i at the i-th DVH bin. v_i and D_i can be readily obtained from a given DVH. n is a parameter of the LKB model, the meaning of which will be discussed later. It can be seen from (II.1) that $gEUD$ collapses the complex dose distribution represented by a DVH into a single metric. The biological rationale behind this specific form of collapsing is that $gEUD$ represents the uniform dose that, if delivered over the same number of fractions as the real but non-uniform dose distribution, yields the same radiobiological effect (Li et al., 2012). Furthermore, the LKB model links $gEUD$ with the probability for a patient to develop the complication by a sigmoid-shape function, i.e.,

$$P(Y = 1) = \phi\left(\frac{gEUD - TD_{50}}{m \times TD_{50}}\right). \tag{2.2}$$

Here, Y is an indicator variable; $Y = 1$ represents that the patient has the complication and $Y = 0$ otherwise. $\phi(\cdot)$ is the cumulative probability function for the standard normal distribution. TD_{50} and m are two other LKB model parameters. Figure II.2 shows the function of $P(Y = 1)$ with respect to $gEUD$.

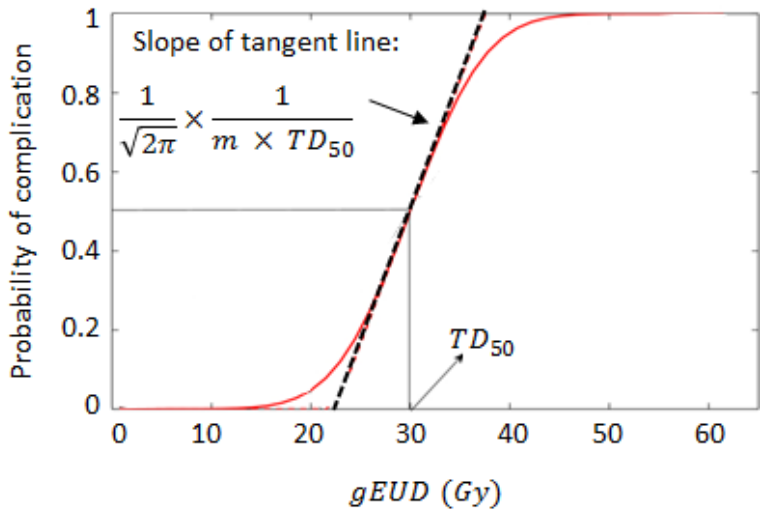


Figure 2.2: Probability of complication, $P(Y = 1)$, with respect to $gEUD$ in LKB model

Next, we will discuss the meanings of the three LKB model parameters, TD_{50}, m, and n. According to (2), TD_{50} is the $gEUD$ given to the normal tissue that results in 50% probability of the complication (Figure 2). Intuitively, TD_{50} is the tolerance dose for developing the complication. That is, a slightly more $gEUD$ than TD_{50} will make the patient having a risk of developing the complication that is greater than a random guess. Furthermore, to understand parameter m, we take the partial derivative of $P(Y = 1)$ with respect to $gEUD$ and evaluate this partial derivative at $gEUD = TD_{50}$, which gives:

$$\frac{\partial\, P(Y=1)}{\partial\, gEUD}\Big|_{gEUD=TD_{50}} = \frac{1}{\sqrt{2\pi}} \times \frac{1}{m \times TD_{50}}. \tag{2.3}$$

(II.3) means that $\frac{1}{\sqrt{2\pi}} \times \frac{1}{m \times TD_{50}}$ is the slope of the "$P(Y = 1)$ vs. $gEUD$" curve at $gEUD = TD_{50}$ (Figure II.2). The slope of this curve reflects sensitivity of the complication probability with respect to a change in $gEUD$. The bigger the slope, i.e., the smaller the $m \times TD_{50}$, the higher the sensitivity. Therefore, the meaning of m is that it is inversely related to the sensitivity of the complication probability with a fixed TD_{50}. Finally, we discuss parameter n. In the LKB model, n is constrained to be between 0 and 1. When $n = 1$, it is obvious from (II.1) that $gEUD$ becomes the average dose received by the normal tissue. When $n = 0$, Proposition 1 shows that $gEUD$ becomes the maximum dose received by the tissue (please see the proof in Appendix 1).

Proposition 1: $\lim_{n\to 0} gEUD = \lim_{n\to 0}\left(\sum_i v_i D_i^{1/n}\right)^n = \max_i D_i$.

To obtain parameters of the LKB model in a particular clinical study, one needs a collection of patient data, based on which the parameters can be estimated by an Maximum

Likelihood Estimation (MLE). The LKB model has been extensively used to model various types of complications on many normal tissues/organs for major modern RT techniques. To consolidate the results from various studies, the American Association of Physicists in Medicine (AAPM) and the American Society for Radiation Oncology (ASTRO) jointly funded a multidisciplinary study, called Quantitative Analysis of Normal Tissue Effects in the Clinic (QUANTEC), in order to summarize the existing findings and develop clinical guidance. Reference values for the LKB model parameters were provided for clinically significant complications of 16 tissues/organs (Bentzen et al., 2010).

2.2.2 Integration of patient-specific information into the biological model

Formulation

Based on the understanding of the LKB model parameters in the previous section, we now discuss which parameter(s) are more appropriate to be personalized. According to the definition, TD_{50} reflects the radiation dose a patient can tolerate before developing a complication with a chance higher than a random guess. Previous studies have shown that different patients have different tolerances depending on their age, health conditions, and even genetics (Boomsma et al., 2012; Cella et al., 2013). Therefore, it is obvious that TD_{50} should be to patient-specific. Furthermore, $m \times TD_{50}$ reflects the sensitivity of complication probability with respect to a change in radiation dose. There is no clear medical evidence so far to support whether or not this sensitivity should be patient-specific. Therefore, we choose not to personalize $m \times TD_{50}$. Finally, regarding parameter n, there is solid evidence in the literature to suggest that it is more organ-specific than patient-specific (Gulliford et al., 2012; Li et al., 2012). Specifically, in the RT literature, organs

are classified into parallel organs (e.g., lung, kidney, and liver) and serial organs (e.g., spinal cord, intestines, and optic nerves). Sub-units of a parallel organ function relatively independently, so radiation damage to a small region does not make the whole organ dysfunctional. Therefore, the probability of developing a complication for a parallel organ should be more related to the average dose it receives, i.e., with $n \to 1$. On the other hand, a serial organ tends to exhibit the complication if one subunit is incapacitated, so that the probability of developing a complication for a serial organ is more related to the maximum dose, i.e., with $n \to 0$.

Due to these considerations, we propose to incorporate patient-specific variables into TD_{50}. Let $x_{1j}, \dots, x_{pj}$ be p patient-specific variables for patient j. Then,

$$TD_{50_j} = \beta_0 + \sum_{k=1}^{p} \beta_k \, x_{kj}, \tag{2.4}$$

where $\beta_0, \beta_1, \dots, \beta_p$ are parameters to be estimated. Also, because we have decided not to personalize the sensitivity, we can replace the $m \times TD_{50}$ in (II.2) by a new parameter σ. Therefore, (II.2) becomes:

$$P(y_j = 1) = \phi\left(\frac{gEUD_j(n) - \left(\beta_0 + \sum_{k=1}^{p} \beta_k x_{kj}\right)}{\sigma}\right). \tag{2.5}$$

Let $\alpha_{p+1} \triangleq \sigma^{-1}$, $\alpha_0 \triangleq -\beta_0 \sigma^{-1}$, and $\alpha_k \triangleq -\beta_k \sigma^{-1}$. Then, (II.5) becomes:

$$P\big(y_j = 1\big) = \phi\big(\alpha_0 + \sum_{k=1}^{p} \alpha_k \, x_{kj} + \alpha_{p+1} gEUD_j(n)\big). \tag{2.6}$$

Furthermore, using (II.6), the log-likelihood function can be written as

$$l\big(\alpha_0, \alpha_1, \dots, \alpha_p, \alpha_{p+1}, n\big) = \sum_j \big\{ y_j log P(y_j = 1) + \big(1 - y_j\big) log P(y_j = 0) \big\}. \tag{2.7}$$

Note that there could be many patient-specific variables to be included in the model, resulting in too many parameters to be estimated. This runs a risk of overfitting, especially considering that the sample size in this type of studies is usually limited. Also, it is typical that only a few patient-specific variables out of the many are truly relevant to the development of a particular complication. Identification of these relevant variables has important clinical value. Because of these reasons, we propose to add a sparsity-inducing penalty like the l_1-penalty to (II.7). As a result, the parameters can be estimated by minimizing a l_1-penalized negative log-likelihood function, i.e.,

$$\min_{\alpha_0,\alpha_1,\dots,\alpha_p,\alpha_{p+1},n} \left\{-l(\alpha_0,\alpha_1,\dots,\alpha_p,\alpha_{p+1},n) + \lambda \sum_{k=1}^{p}|\alpha_k|\right\}$$

$$s.t. \quad \alpha_{p+1} \geq 0 \tag{2.8}$$

In (8), λ is a tuning parameter. Also note that we add a constraint that $\alpha_{p+1} \geq 0$ to the optimization problem. This is to satisfy the biological validity that radiation dose always has a non-negative effect on the risk of developing a complication.

Estimation

The optimization problem in (2.8) is difficult to solve because of the complicated relationship between the objective function and the parameter n. On the other hand, with a fixed n, $gEUD_j(n)$ can be computed from the DVH of each patient, and consequently (2.8) becomes a convex optimization with a linear constraint that is much easier to solve. This motivates us to treat n as a tuning parameter rather than a parameter to be directly optimized. As a result, (2.8) can be written as (2.9):

$$\min_{\alpha_0,\alpha_1,\dots,\alpha_p,\alpha_{p+1}} \left\{-l_n(\alpha_0,\alpha_1,\dots,\alpha_p,\alpha_{p+1}) + \lambda \sum_{k=1}^{p}|\alpha_k|\right\}$$

$$s.t.\ \ \alpha_{p+1}\ \geq 0 \tag{2.9}$$

where both n and λ are treated as tuning parameters. To solve (II.9), we propose an efficient algorithm based on the result of Proposition 2. The algorithm works by first solving the unconstrained optimization, which can be done by an efficient convex solver, and then using a simple fix to obtain the solution to the constrained optimization. Please see the proof of Proposition 2 in Appendix 2.

Proposition 2: Let $\hat{\alpha}_0, \dots, \hat{\alpha}_{p+1}$ be the solution to the optimization in (2.9). Let $\alpha_0^*, \dots, \alpha_{p+1}^*$ be the solution to unconstrained problem. If $\alpha_{p+1}^* \geq 0$, then $\left(\hat{\alpha}_0, \dots, \hat{\alpha}_{p+1}\right) = \left(\alpha_0^*, \dots, \alpha_{p+1}^*\right)$. If $\alpha_{p+1}^* < 0$, then $\left(\hat{\alpha}_0, \dots, \hat{\alpha}_p\right) = \left(\alpha_0^*, \dots, \alpha_p^*\right)$ and $\hat{\alpha}_{p+1} = 0$.

Finally, we discuss how to choose the two tuning parameters. A grid search can be performed over all combinations of n and λ values within their respective ranges. The optimal n^* and λ^* are those that optimize a criterion. Commonly used criteria include BIC, AIC, and cross-validated deviance. The range of λ, $[\lambda_{min}, \lambda_{max}]$, is chosen such that λ_{max} results in no patient-specific variables being selected (i.e., the sparsest model), and λ_{min} results in all patient-specific variables or the number of patient-specific variables equal to the sample size being selected, whichever is smaller (i.e., the statistically plausible densest model). To set the range for n, we can run the original LKB model and obtain the confidence interval for n. This confidence internal can be used as $[n_{min}, n_{max}]$.

2.2.3 Consistency and bias correction for the integrated model

After the solution to (2.9), i.e., $\hat{\alpha}_0, \dots, \hat{\alpha}_{p+1}$, is obtained, we will use the zero and non-zero patterns in $\hat{\alpha}_1, \dots, \hat{\alpha}_p$ to select patient-specific variables. The selected variables will be used together with $gEUD_j(n^*)$ to fit a non-penalized model. This model can be used to predict

the probability of developing the complication for new patients. Let $\tilde{\alpha}_0, \dots, \tilde{\alpha}_q$ denote the estimated coefficients from this non-penalized model by MLE, i.e.,

$$\begin{aligned}(\tilde{\alpha}_0, \dots, \tilde{\alpha}_q) &= \max_{\alpha_0, \alpha_1, \dots, \alpha_q} l_{n^*}(\alpha_0, \alpha_1, \dots, \alpha_q) \\ &= \max_{\alpha_0, \alpha_1, \dots, \alpha_q} \sum_j \{y_j log P_{n^*}(y_j = 1) + (1 - y_j) log P_{n^*}(y_j = 0)\},\end{aligned} \tag{2.10}$$

where

$$P_{n^*}(y_j = 1) = \phi(\alpha_0 + \sum_{k=1}^{q-1} \alpha_k x_{kj} + \alpha_q gEUD_j(n^*)). \tag{2.11}$$

The form of (2.11) is known as the probit model. The reason of re-fitting a non-penalized model is that the l_1-penalty is known to have a shrinking effect, which makes an l_1-penalized model a good variable selection model but not necessarily a good predictive model (Hastie et al., 2015). Next, we discuss two important statistical properties of the estimators $\tilde{\alpha}_0, \dots, \tilde{\alpha}_q$, i.e., consistency and bias. In statistics, a consistent estimator is one that converges in probability to the true value of the parameter being estimated as the sample size goes to infinity. The bias of an estimator is the difference between the estimator's expected value and the true value of the parameter being estimated. If the bias is zero, the corresponding estimator is called an unbiased estimator. A good estimator should be consistent and unbiased.

Consistency

The population of patients going through a RT is typically heavy imbalanced, i.e., a very small fraction of the population will develop the complication of interest while the large majority will not. Otherwise, the RT would not have been permitted for practice.

When a dataset is sampled from the population for NTCP modeling, a common strategy is to include a larger fraction of patients with the complication in the dataset than what this faction truly is in the population. This is to make sure that the dataset has enough samples with $Y = 1$ (i.e., with the complication) and thus rendering a meaningful statistical analysis (King and Zeng, 2001). However, when this sampling strategy is used, $\tilde{\alpha}_0$ is not a statistically consistent estimator for α_0, although $\tilde{\alpha}_1, \dots, \tilde{\alpha}_q$ still are. This finding is summarized in Proposition 3.

Proposition 3: Let τ and $\bar{y}$ be the fractions of patients with $Y = 1$ in the population and in the sample dataset, respectively. If $\bar{y} \neq \tau$, then $\tilde{\alpha}_0$ is not a consistent estimator for α_0 while $\tilde{\alpha}_1, \dots, \tilde{\alpha}_q$ are consistent estimators for $\alpha_1, \dots, \alpha_q$.

To produce a consistent estimator, we propose a maximum weighted likelihood estimation (MWLE) to estimate the non-penalized model in (II.11). In the MWLE, the log-likelihood function takes a weighted form, i.e.,

$$l_{n^*}^{w}(\alpha_0, \alpha_1, \dots, \alpha_q) = w_1 \sum_{\{y_j=1\}} \log P_{n^*}(y_j = 1) + w_0 \sum_{\{y_j=0\}} \log P_{n^*}(y_j = 0), \tag{2.12}$$

where $w_1 = \tau/\bar{y}$, $w_0 = (1-\tau)/(1-\bar{y})$. By maximizing (2.12), we can obtain estimates for $\alpha_0, \alpha_1, \dots, \alpha_q$, denoted by $\breve{\alpha}_0, \dots, \breve{\alpha}_q$. Proposition 4 shows that $\breve{\alpha}_0, \dots, \breve{\alpha}_q$ are consistent estimators. Proofs of Propositions 3 and 4 share a similar idea to the proof of consistency for the weighted exogenous sampling maximum likelihood (WESML) estimator in Manski and Lerman, 1977, and thus skipped here due to space limit.

Proposition 4: $\breve{\alpha}_0, \dots, \breve{\alpha}_q$ are consistent estimators for $\alpha_0, \alpha_1, \dots, \alpha_q$.

Note that the weighted log-likelihood function proposed in (2.12) assumes a known τ. τ is straightforward to obtain. For example, it can be estimated from a data source that only records patients' complications, such as the Electronic Health Records. Because such a data source does not include the RT dose, it can be easily created and thus including a large patient population to grant an accurate estimate for τ. τ may also be obtained from published epidemiologic studies on the RT, which usually report the fraction of people developing the complication in a large population.

Bias

Although $\breve{\alpha}_0, \dots, \breve{\alpha}_q$ are consistent estimators, they are still biased. Proposition 5 derives the bias of these estimators. Please see the proof in Appendix 3.

Proposition 5: Let $\breve{\boldsymbol{\alpha}} = \left(\breve{\alpha}_0, \dots, \breve{\alpha}_q\right)$. Denote the true values of the parameters being estimated by $\boldsymbol{\alpha} = \left(\alpha_0, \dots, \alpha_q\right)$. The bias of $\breve{\boldsymbol{\alpha}}$ is

$$bias(\breve{\boldsymbol{\alpha}}) = E(\breve{\boldsymbol{\alpha}}) - \boldsymbol{\alpha} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\xi}. \tag{2.13}$$

$\mathbf{X}$ is the $N \times (q+1)$ data matrix of all the predictors including the intercept. $\mathbf{W}$ is an $N \times N$ diagonal matrix with the j-th diagonal element being

$$\mathbf{W}_{jj} = \frac{w_0\phi(\eta_j) + w_1\left(1-\phi(\eta_j)\right)}{\phi(\eta_j)\left(1-\phi(\eta_j)\right)}\varphi^2\left(\eta_j\right),$$

where $\eta_j = \alpha_0 + \sum_{k=1}^{q-1}\alpha_k x_{kj} + \alpha_q gEUD_j(n^*)$ and $\varphi(\cdot)$ is the standard normal probability density function. $\boldsymbol{\xi}$ is a $N \times 1$ vector with the j-th element ξ_j being

$$\xi_i = \frac{\eta_j\phi(\eta_j) - (w_1-1)\varphi(\eta_j)}{2\phi(\eta_j)}\mathbf{Q}_{jj},$$

where $\mathbf{Q}_{jj}$ is the j-th diagonal element of matrix $\mathbf{Q} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T$.

Using the result in Proposition 5, we can obtain an unbiased estimator for $\boldsymbol{\alpha}$, i.e., $\ddot{\boldsymbol{\alpha}} = \breve{\boldsymbol{\alpha}} - \ddot{bias}(\breve{\boldsymbol{\alpha}})$. $\ddot{bias}(\breve{\boldsymbol{\alpha}})$ is an estimate for the theoretical bias in (2.13) by using $\breve{\eta}_j = \breve{\alpha}_0 + \sum_{k=1}^{q-1} \breve{\alpha}_k x_{kj} + \breve{\alpha}_q gEUD_j(n^*)$. Furthermore, it can be shown that $Var(\ddot{\boldsymbol{\alpha}}) = \left(\frac{N-q-1}{N}\right)^2 Var(\breve{\boldsymbol{\alpha}})$. Given that the sample size N is typically larger than the number of parameters, $q+1$, we can get $Var(\ddot{\boldsymbol{\alpha}}) < Var(\breve{\boldsymbol{\alpha}})$. This means that the variance of $\ddot{\boldsymbol{\alpha}}$ is smaller than $\breve{\boldsymbol{\alpha}}$, i.e., the bias correction does not increase the variance of the estimator. Proof for the above variance relationship is similar to that for the case of generalized linear models (Cordeiro and McCullagh, 1991).

Finally, we summarize the steps of our proposed integrated model, as introduced in Sections 2.2.2 and 2.2.3, in Figure 2.3. The R code of our model has been submitted to GitHub website and is publicly available at https://github.com/xliu203/Integration-of-biological-model-and-statistical-model/blob/master/glmnet.probit_v2.R.

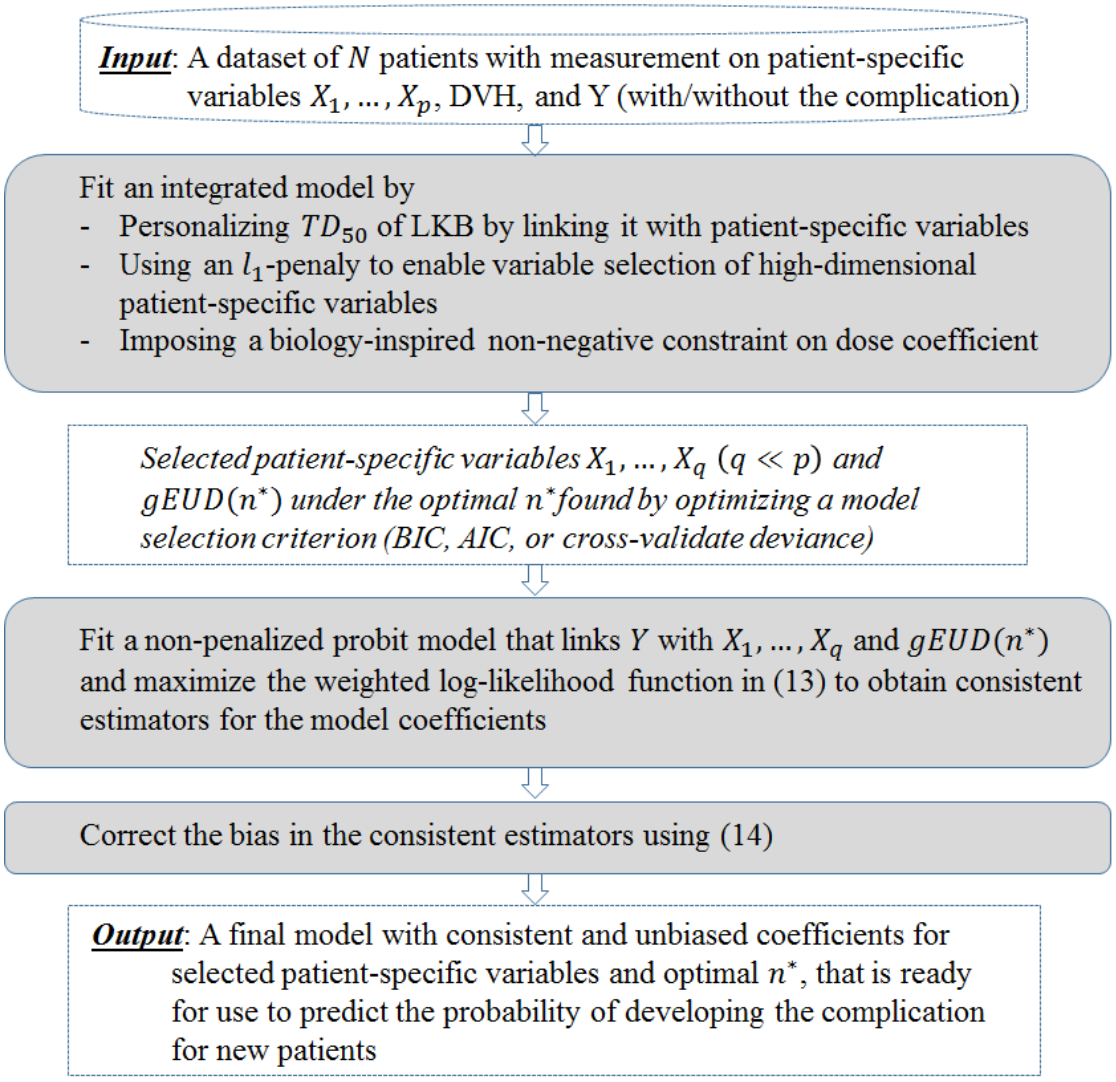


Figure 2.3: A flow chart of the steps for the proposed integrated model

2.3 Case Study

2.3.1 Application to NTCP modeling of acute rectal complication for IMRT treatment of prostate cancer

We present an application in which patients were treated with IMRT for prostate cancer. A serious complication these patients may suffer from after the IMRT is the grade 2+ acute rectal complication with symptoms including anal pain, diarrhea, and rectal obstruction. We obtain a dataset of 86 patients from our collaborative institution,

Department of Radiation Oncology at Mayo Clinic Arizona. The study was approved by the Institutional Review Board (IRB) of Mayo Clinic Arizona and included written informed consent from all subjects. All patients were diagnosed with prostate cancer. The IMRT they received was set up using the following protocol: A static field IMRT technique with 7 coplanar 6MV fields was employed. The whole prostate was designated as a Clinical Target Volume (CTV), and two Planning Target Volumes (PTVs) were created using uniform 3mm and 6mm expansions. A dose of 77.4 Gy in 43 fractions (1.8Gy/fraction) was prescribed to the 3mm expansion, and a dose of 70 Gy to the 6mm expansion. Seminal Vesicles with uniform 7mm expansion were prescribed 54 Gy. A simultaneous integrated boost (SIB) was given to areas suspicious for cancer as demonstrated in a planning multi-parametric magnetic resonance scan which was a combination of T2-weighted imaging, Diffusion Weighted Imaging and Dynamic Contrast-Enhanced imaging. The SIB volume was identified by a diagnostic radiologist specializing in genitourinary imaging, was not expanded, and was prescribed 81-83Gy. All patients were planned using the Eclipse Treatment Planning System (TPS) produced by Varian, Inc.

Because we focus on the rectal complication, the rectum was drawn as a whole organ bounded by ischial tuberosity inferiorly and sigmoid flexure superiorly. Then, DVH on the rectum was extracted for each patient using automated scripts which were written within the Applications Programmer Interface (API) of the Eclipse TPS manufactured by Varian, Inc. Furthermore, we include 11 patient-specific variables that potentially affect the complication, which are age, concurrent treatment status, diabetes status, Gleason score, Androgen Deprivation Therapy (ADT) status, adjuvant ADT status, neoadjuvant ADT

status, prostate-specific antigen (PSA) level prior to treatment, prostate volume, use of statin medications, and stage of the disease (T-stage). After the IMRT, each patient's medical records were reviewed by a physician and his rectal complication was graded according to the Common Terminology Criteria for Adverse Events (CTCAE) v4.0. Among the 86 patients in the dataset, 23 developed the grade 2+ acute rectal complication. We apply the integrated model to the dataset according to the steps in Figure 3. Recall that the integrated model includes a step for consistency correction in case that the population fraction of complication, τ, is smaller than the sample fraction of complication, $\bar{y}$. In our dataset, $\bar{y} = \frac{23}{86} = 26.7\%$, while $\tau < 20\%$ in the published literature on population-based studies (Tucker et al., 2012). Therefore, consistency correction is needed. Furthermore, to choose the optimal tuning parameters, we adopt a model selection criterion called AICc, which includes a correction for the original AIC under small sample sizes (Hurvich and Tsai, 1989). The results from the integrated model are as follows: Among all the patient-specific variables included in the dataset, six are selected using AICc: diabetes status, prostate volume, PSA, statins use, ADT status, T-stage. The optimal n^* is found to be 0.154. These results are consistent with findings in the literature. For example, statins are a class of drugs often prescribed by doctors to help lower cholesterol levels in the blood. Statins use is negatively related to the probability of complication, indicating that the use of statins might be protective against the development of the 2+ acute rectal complication for patients. At least one biological mechanism behind this seemingly protective effect has been suggested (Malek, 2015), and a relatively recent study reported a similar result, namely a negative association between acute rectal complication during pelvic RT and the use of

statins (Wedlake et al., 2012). This corroborates our finding. PSA is a blood test that is commonly used to detect prostate cancer; the higher level of PSA, the higher chance the patient has prostate cancer. Finally, knowing that the range of n is between 0 and 1, the optimal $n^* = 0.154$ found by the integrated model is small. This is consistent with prior findings (Bentzen et al., 2010) and agrees with the clinical expectation that because rectum is a serial organ, it is well-known that serial organs tend to have small n (Gulliford et al., 2012; Li et al., 2012).

Furthermore, we would like to assess the prediction/classification accuracy of the integrated model in comparison with the biological and statistical models used alone. The best strategy in accuracy assessment of a statistical model without running the risk of overfitting is to divide the entire dataset into a training set and a test set. Samples in the test set are not used in training but only used to compute the classification accuracy of the training model. Realizing that we have a small dataset, we put all but one samples in the training set and the remaining sample in the test set. This will allow us to compute the classification accuracy on one test sample. We repeat this training-test split over all the samples, which will allow us to compute the test accuracy on all samples. Note that this scheme is different from leave-one-out cross validation, because the latter would report the best accuracy optimized on the test set while our scheme assumes the test set is completely unseen at the training stage. Using this scheme, Table 1 show the test accuracy of the integrated model in comparison with those of the biological and statistical models. The accuracy metric is Area Under the Curve (AUC). Use of AUC avoids having to choose a cutoff for the predicted probabilities and therefore provides a more objective measure for

the prediction accuracy. Recall that LKB model fitting only uses the DVH and cannot take patient-specific variables into consideration. As a result, the AUC is low. The commonly used statistical model in the NTCP literature is logistic regression (Cella et al., 2013; Boomsma et al., 2012; Deville et al., 2012). We follow the convention of NTCP modeling and build a logistic regression that includes $D_{10\%}$, $D_{15\%}$, $D_{20\%}$, $\cdots$, $D_{90\%}$ and patient-specific variables as predictors. Here, $D_{\alpha\%}$ is the radiation dose such that $\alpha\%$ of the rectal volume receives this dose level or higher. For a fair comparison with the integrated model, we also use an l_1-penalty for variable selection and select the penalty parameter using AICc. The AUC of the statistical model is around 0.76, which is also lower than the integrated model.

Table 2.1: Comparison between the AUCs of the integrated model, biological model (LKB), and statistical model (l_1-penalized logistic regression)

	Integrated model	Biological model	Statistical model
AUC on test data	0.82	0.66	0.76

Finally, we report the AUCs of the integrated model under other model selection criteria. The AUCs under AIC and BIC are 0.82 and 0.78, respectively, which are still higher than the biological and statistical models used alone. The AUC under AIC is the same as that under AICc (Table II.1), implying that the integrated model is not sensitive to small sample correction that is accounted for by AICc. The AUC under BIC is relatively lower than other criteria. BIC has been known to most suitable for cases when there is a very large number of predictors, while our study involves only 11 predictors.

2.3.2 Simulation experiments

We would like to compare the performance of the integrated model with biological and statistical models under different parameter settings using simulation data. A significant challenge of this study is how to simulate DVH for each virtual patient. To make sure that the simulated DVH has similar data characteristics to the DVH of real patients, we simulate the DVH of each virtual patient from the DVH measurements of the 86 real patients. Specifically, we fit a multivariate normal distribution for a random vector of $(D_{1\%}, D_{2\%}, \dots, D_{99\%})^T$ using the DVH of the 86 patients. Next, we sample from the fitted distribution to create the DVH for a virtual patient. Then, the DVH is used to compute $gEUD(n)$ using (1). We set $n = 0.2$ and 0.3 in our simulation experiments. Furthermore, we sample from a multivariate normal distribution, $\left(X_1, \dots, X_p\right)^T \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, to create data for p patient-specific variables of the virtual patient. $\mathbf{\Sigma}$ has all the diagonal elements being one and off-diagonal elements being $\mathbf{\Sigma}_{ij} = 0.2^{|i-j|}$ to account for possible correlation between the variables. We set $p = 14$ in our experiment, which is close to the number of patient-specific variables included in the real-data application in Section II.3.1.

Furthermore, to set the coefficients for the patient-specific variables and $gEUD(n)$, we refer to the model in (II.14) for proper ranges of the coefficients. 10 out of the 14 patient-specific variables are set to have zero coefficients. In this way, we can test the accuracy of the model in selecting patient-specific variables. In the remaining four patient-specific variables with non-zero coefficients, two are set to have positive coefficients and the other two are set to have negative coefficients. The magnitude of non-zero coefficients is set to be $|\alpha_k| = 0.8, k = 1,2,3,4$. We also run experiments on a smaller magnitude of

0.6. The coefficient for $gEUD(n)$ is set to be $\alpha_{p+1} = 0.25$ and 0.3. The intercept α_0 is set to achieve a desired population fraction of complication, τ. $\tau = 20\%$ is used in our experiments.

With simulated data for a virtual patient and under a particular setting of the model coefficients, we use the right-hand side of (II.6) to generate the probability that the patient develops the complication. Then, this probability is used as the parameter of a Bernoulli distribution from which a binary variable y can be sampled. Furthermore, to mimic the reality that the sample fraction of complication in a dataset, $\bar{y}$, is usually higher than the population fraction of complication, τ, we set $\bar{y} = 40\%$. The sample size of the dataset is 200.

We apply the integrated model, LKB, and l_1-penalized logistic regression to the simulation datasets. The results are shown in Tables II.2-5. The following observations can be drawn:

Table 2.2: Comparison between the integrated model, biological model, and statistical model ($|\alpha_k| = 0.8, \alpha_{p+1} = 0.25, n = 0.2$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment.

	LOOCV-AUC	Sensitivity in selecting patient-specific variables	Specificity in selecting patient-specific variables	95% Confidence interval for n
Integrated model	0.91 (0.009) p <0.001*	1 (0.000)	0.85 (0.146)	0.2 ∈[0.078, 0.368]
Biological model (LKB)	0.68 (0.005)	—	—	0.2 ∉[0.060, 0.184]
Statistical model (logistic regression)	0.86 (0.016)	1 (0.000)	0.84 (0.142)	—

p<0.001* is the p value for a one-sided hypothesis testing that the integrated model has a higher LOOCV-AUC than the biological or statistical model whichever has a higher LOOCV-AUC.

Table 2.3: Comparison between the integrated model, biological model, and statistical model ($|\alpha_k| = 0.6, \alpha_{p+1} = 0.25, n = 0.2$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment.

	LOOCV-AUC	Sensitivity in selecting patient-specific variables	Specificity in selecting patient-specific variables	95% Confidence interval for n
Integrated model	0.89 (0.015) p <0.001*	0.98 (0.076)	0.86 (0.134)	0.2 $\in$[0.105, 0.342]
Biological model (LKB)	0.71 (0.003)	—	—	0.2 $\notin$[0.114, 0.176]
Statistical model (logistic regression)	0.81 (0.031)	0.96 (0.105)	0.86 (0.152)	—

p<0.001* is the p value for a one-sided hypothesis testing that the integrated model has a higher LOOCV-AUC than the biological or statistical model whichever has a higher LOOCV-AUC.

Table 2.4: Comparison between the integrated model, biological model, and statistical model ($|\alpha_k| = 0.8, \alpha_{p+1} = 0.3, n = 0.3$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment.

	LOOCV-AUC	Sensitivity in selecting patient-specific variables	Specificity in selecting patient-specific variables	95% Confidence interval for n
Integrated model	0.93 (0.009) p <0.001*	1 (0.000)	0.83 (0.141)	0.3 $\in$[0.131, 0.474]
Biological model (LKB)	0.75 (0.001)	—	—	0.3 $\notin$[0.166, 0.184]
Statistical model (logistic regression)	0.84 (0.030)	0.98 (0.106)	0.85 (0.099)	—

p<0.001* is the p value for a one-sided hypothesis testing that the integrated model has a higher LOOCV-AUC than the biological or statistical model whichever has a higher LOOCV-AUC.

Table 2.5: Comparison between the integrated model, biological model, and statistical model ($|\alpha_k| = 0.6, \alpha_{p+1} = 0.3, n = 0.3$). Mean (standard deviation) and confidence interval are computed based on 50 repetitions of the simulation experiment.

	LOOCV-AUC	Sensitivity in selecting patient-specific variables	Specificity in selecting patient-specific variables	95% Confidence interval for n
Integrated model	0.88 (0.029) p <0.001*	0.94 (0.210)	0.87 (0.121)	$0.3 \in [0.211, 0.389]$
Biological model (LKB)	0.74 (0.001)	—	—	$0.3 \notin [0.166, 0.178]$
Statistical model (logistic regression)	0.77 (0.052)	0.94 (0.190)	0.86 (0.126)	—

p<0.001* is the p value for a one-sided hypothesis testing that the integrated model has a higher LOOCV-AUC than the biological or statistical model whichever has a higher LOOCV-AUC.

First, the LOOCV-AUC of the integrated model is significantly higher than LKB and logistic regression (p value <0.001) across all the simulation settings. The LOOCV-AUC of the integrated model becomes lower when the magnitude of coefficients for patient-specific variables, $|\alpha_k|$, gets smaller, but it is little affected by n and the coefficient for $gEUD(n)$, α_{p+1}. Moreover, although LKB has the lowest mean LOOCV-AUC among the three models, the standard deviation of the LOOCV-AUC over 50 repetitions of the simulation experiment for LKB is the smallest. This is expected because LKB, as a biological model, is built upon radiobiological principles such that its performance is less

affected by sampling variability. On the other hand, logistic regression has the largest standard deviation of the LOOCV-AUC. This is because statistical models are pure data-driven and therefore the performance is more variable across different datasets. By integrating the biological and statistical models, the proposed integrated model can achieve a high accuracy due to the inclusion of patient-specific variables and a more robust performance against sampling variability due to the consideration of radiobiological principles.

Furthermore, the integrated model achieves high sensitivity and specificity in selecting the patient-specific variables across all the experiments. Here, sensitivity is the proportion of non-zero coefficients for patient-specific variables that are correctly identified as non-zero. Specificity is the proportion of zero coefficients for patient-specific variables that are correctly identified as zero. The sensitivity and specificity becomes lower when the magnitude of coefficients for patient-specific variables, $|\alpha_k|$, gets smaller. The same phenomenon is observed when we keep $|\alpha_k|$ unchanged but decrease the sample size. These observations are consistent with the findings of existing variable selection approaches (Huang et al., 2012). Logistic regression achieves a similar level of sensitivity and specificity to the integrated model. Considering this result together with that on AUC, we can conclude that logistic regression may perform as well as the integrated model on a single dataset, which is reflected by the sensitivity and specificity in selecting patient-specific variables. However, it performs worse than the integrated model when one wants to apply the model trained on one dataset to another dataset (i.e., weaker generalizability), which is reflected by the AUC.

Finally, we compare the estimated parameter n between the integrated model and LKB. A universally true observation across all the experiments is that the confidence interval of n includes the true value in the integrated model but not in LKB. Because LKB fails to account for the effect of patient-specific variables on the probability of complication, its parameter estimation is compromised. This result corroborates the low LOOCV-AUC of LKB.

2.4 Conclusion

In this paper, we proposed an integrated model for NTCP modeling. We developed the model by starting with an in-depth understanding of the biological model (i.e., LKB) parameters. Among all the parameters, TD_{50} reflects the radiation dose a patient can tolerate before developing complication with a chance higher than a random guess, and therefore should be patient-specific. We proposed to link patient-specific variables with TD_{50} by a linear model and used this personalized TD_{50} to replace the original TD_{50} in LKB. This resulted in an integrated model formulation. We further added to the formulation a sparsity-inducing penalty to enable variable selection from high-dimensional patient-specific variables, and a biological constraint on the model coefficients to account for the fact that radiation dose always poses at least "some" risk of complication to the normal tissue. Next, we developed an efficient algorithm to estimate the parameters of the integrated model. Furthermore, we performed theoretical analysis and proposed modified approaches to ensure that the integrated model have statistical consistent and unbiased coefficient estimators. Finally, we applied the integrated model to a real dataset of prostate cancer patients treated with IMRT who are at risk of developing the grade 2+ acute rectal

complication. The integrated model had higher prediction accuracy measured by LOOCV-AUC than the biological (i.e., LKB) and statistical models (i.e., l_1-penalized logistic regression) used in separation. Also, three patient-specific variables were selected by the integrated model as important predictors to the complication, including age, statins use, and PSA. Age and statins use have also been found to be related to the risk of complication in the existing RT literature. The finding about PSA was novel and further investigation on the biological mechanism is yet to be performed. Various simulation studies were also conducted, showing that the integrated model significantly outperformed both biological and statistical models in LOOCV-AUC. The variable selection sensitivity and specificity of the integrated and statistical models were comparable. These results indicated that the statistical model may perform as well as the integrated model on a single dataset, but it has worse generalizability to other studies. In addition, the integrated model accurately estimated the organ parameter n while LKB was not able to.

There are some limitations of the present study, which drive future investigation: The proposed model was demonstrated on a small dataset consisting of 86 patients with 23 individuals having complications. More data are needed to further validate the model and findings. Also, our study found PSA to be a significant predictor for rectal complication of prostate cancer patients. We are not aware of any prior study that reported this relationship, although there are plenty of studies that correlate PSA with the existence of prostate cancer (Partin et al., 1996; Catalona et al., 2000; Catalona et al., 1997). Further investigation is needed to validate this finding with more data and discover the biological mechanism behind this relationship if still held true.

CHAPTER 3

A TRANSFER LEARNING MODEL FOR MULTI-MODALITY IMAGE DATA FUSION WITH APPLICATION IN ALZHEIMER'S DISEASE EARLY DETECTION

3.1 Introduction

Multimodality datasets are becoming increasingly common in various domains to provide complementary information for predictive analytics. For example, in health care, images of different types such as structural magnetic resonance imaging (MRI) and fludeoxyglucose positron emission tomography (FDG-PET) provide complementary information about the organ of interest, which allows for building a predictive model to accurately detect a certain disease (Jack et al. 2009; Lowe et al. 2009; Clark et al. 2011). In manufacturing, data collected from multiple different types of sensors provide complementary information about the process and product, allowing for more accurate assessment of process and product quality (Basir and Yuan 2007).

One important challenge for integration of multimodality datasets in building a predictive model is that the multiple different modalities are not universally available for all the samples. Take the diagnosis of the Alzheimer's disease (AD) – a fatal neurological disorder – using multimodality images as an example. Figure 1 shows the special "incomplete multimodality dataset (IMD)" we are focusing on in this paper, which includes three complementary diagnostic image modalities, i.e., MRI, FDG-PET, and amyloid-PET for detection of AD at an early stage of the disease called Mild Cognitive Impairment (MCI) (Jack et al. 2012). In the recently published expert consensus criteria by the National Institute of Aging and Alzheimer's Association, the use of multimodality images for early

detection of AD has been highly recommended (Albert et al. 2011). In Figure III.1, each sub-cohort consists of patients who have the same availability of modalities. Different sub-cohorts have different missing modality patterns. The reasons for the existence of IMD are multifold: some imaging equipment such as PET is costly and only available in limited clinics; some modalities are not accessible to patients due to insurance coverage; it is not safe to put patients with some pre-existing conditions through a certain imaging examination.

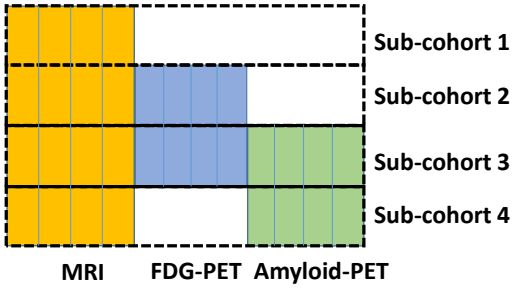


Figure 3.1. An example of the incomplete multimodality dataset (IMD), in which MRI, FDG-PET, and amyloid-PET are considered as three modalities. Columns within each modality represent features extracted from the image. Each sub-cohort consists of patients with the same availability of modalities.

If we applied existing methods to model IMD, there would be three options.

1) Missing data imputation followed by a statistical model that uses imputed features to predict a response variable of interest. While there are many imputation algorithms (He et al. 2017; Beaulieu-Jones, Moore, and CONSORTIUM 2017; Ordóñez Galán et al. 2017), they are not suitable for use to fill in the missing data in IMD because of several reasons. First, as seen in the example in Fig. 1, there is a substantial amount of missing data in IMD. Imputation will result in a dataset with poor quality. Second,

data in IMD are missed as the entire modality/modalities, which violates the fundamental missing-at-random assumption that most imputation algorithms have to assume.

2) Separate modeling (SM). This approach would build a separate predictive model for each sub-cohort, e.g., four separate models for the dataset in Figure III.1. The limitation of SM is obvious: because each model can only use the data specific to the corresponding sub-cohort, sample size shortage may prevent building a robust model.

3) All available data modeling (AADM). This approach would build four models for the dataset in Figure III.1: an MRI-alone model using the data in all sub-cohorts; an MRI & FDG-PET model using the data in sub-cohorts 2 and 3; an MRI & amyloid-PET model using sub-cohorts 3 and 4; an MRI & FDG-PET & amyloid-PET model using sub-cohort 3. AADM requires data pooling from different sub-cohorts which typically contain patients from different institutions. This is practically difficult, if not impossible, due to the need for multi-institutional collaboration agreement and patient privacy concerns.

In this paper, we propose a novel Incomplete-Multimodality Transfer Learning (IMTL) model to tackle the aforementioned limitations of existing methods. Different from missing data imputation, IMTL does not fill in missing data but utilize whatever data that is available to build a predictive model for each sub-cohort. Different from SM and AADM, IMTL estimates the sub-cohort-specific models jointly rather than in separation. In this way, knowledge obtained from the modeling of each sub-cohort can be "transferred" to help the modeling of other sub-cohorts. This makes IMTL a transfer learning model. Also,

the model estimation of IMTL does not require data pooling from different sub-cohorts like AADM. We propose a computational architecture that that includes iterative communication between global and local learners to allow for between-institutional collaborative model estimation without the need for data pooling. This is particularly important for patient privacy preservation in health care applications of IMTL. Finally, we would like to stress that although IMTL is developed in the context of multimodality data in health care, it can be effortlessly extended to other non-medical domains that fusion of multimodality datasets is common and much needed, including but not limited to manufacturing (Basir and Yuan 2007) and transportation (Xia, Li, and Shan 2013).

The remainder of the paper is organized as follows: Section 3.2 provides a literature review. Section 3.3 presents the development of IMTL. Section 3.4 investigates unique properties of IMTL. Section 3.5 presents case studies. Section 3.6 is the conclusion.

3.2 Literature Review

This paper primarily intersects with the research area of statistical and machine learning models using data missed in chunks of modalities, termed as IMD in this paper. To our best knowledge, this area only has limited work. In what follows, we review each related paper in detail.

Yuan et al. (2012) proposed an incomplete multisource feature learning method (iMSF), which used an l_{21} penalty to enforce same features within each modality to be selected across different sub-cohorts. One limitation of iMSF is that it cannot do "out-of-sample prediction". That is, if a modality-wise missing pattern is not included in training data, iMSF cannot make a prediction on new samples with that missing pattern. Also, the

l_{21} enabled feature selection scheme is most effective if different modalities have little correlation. To overcome the limitations of iMSF, Xiang et al. (2014) proposed an incomplete source-feature selection (ISFS) model. The main idea was to estimate a set of common coefficients across different sub-cohorts and specific coefficients to account for the uniqueness of each sub-cohort. To gain this flexibility, ISFS needs to estimate many parameters.

Thung et al. (2014) developed a matrix completion method, which selected samples and features in the original dataset to produce a smaller dataset. This was done by using the group-lasso based multitask learning algorithm twice on features and samples, respectively. Then, standard missing data imputation algorithms were applied to the reduced dataset and classifiers were built on the imputed data. While the proposed idea of data reduction is novel, imputation would still have to be used.

Liu et al. (2017) proposed a view-aligned hypergraph learning (VAHL) method. VAHL divided the dataset into several views according to the availability of different modalities. A hypergraph of subjects was constructed on each view. Then, the hypergraphs were fused by a view-aligned regularizer under a classification framework. VAHL had a novel perspective of exploiting subject relationship using hypergraphs to naturally get around the issue of missing modalities. Also because of this "subject" perspective, the model has to be re-trained from scratch every time new data are becoming available. Also, VAHL has many parameters to estimate.

Li et al. (2014) proposed a deep learning (DL) framework specifically for imaging data. The basic idea was to train a 3-D convolutional neutral network (CNN) between two

imaging modalities, MRI and FDG-PET, based on a dataset in which both modalities were available. The CNN was then used to predict a "pseudo" PET from an MRI image for any patient whose PET is missing. This work represents one of the pioneers that introduced DL into imaging-based AD diagnosis and prognosis. On the other hand, DL models are black-box models with difficulty for interpretation. Also, while crafting pseudo PET from MRI is possible from a pure data-driven perspective, the idea needs further validation by imaging physics.

In summary, limited work has been done to develop statistical models for IMD data. All the above-reviewed models, despite their specific weakness, share some common limitations: 1) Most models cannot do out-of-sample prediction, which limits broader utilization; 2) Model estimation needs data pooling from different sub-cohorts. If the sub-cohorts correspond to different health institutions, which is typically the case, protection of patient privacy is a concern. Also, the institutions have to establish data sharing agreement before the modeling can take place, which is a lengthy process if not impossible. 3) While showing empirically good performance on specific datasets, there is a lack of theoretical study on why the performance is guaranteed.

3.3 Development of the Incomplete-Multimodality Transfer Learning (IMTL) Model

For notation simplicity, we present our model development in the context of three modalities, while the model is easily generalizable to any number of modalities. For example, the three modalities can be MRI, FDG-PET and amyloid-PET as shown in Figure 3.1. Under this structure, there are four patient sub-cohorts corresponding to different

availabilities of the modalities: 1) MRI alone; 2) MRI & FDG-PET; 3) MRI & amyloid-PET; 4) all three modalities.

Let k be the index for modalities, $k = 1,2,3$; l be the index for sub-cohorts, $l = 1,2,3,4$; and i be the index for samples/patients, $i = 1, \dots, n$. Denote the sample size of each sub-cohort by n_l. Let $\mathbf{x}_i^{(kl)}$ contain features in modality k for patient i in sub-cohort l. Let $y_i^{(l)}$ be the response variable (e.g., a diagnostic or prognostic result) for patient i in sub-cohort l. We propose two IMTL models, one for a continuous response variable (i.e., a predictive model) and the other for a binary response variable (i.e., a classification model).

3.3.1. IMTL predictive modeling

1) Formulation and estimation

Consider the joint distribution of $y_i^{(l)}$, $\mathbf{x}_i^{(2l)}$, and $\mathbf{x}_i^{(3l)}$ given $\mathbf{x}_i^{(1l)}$ to be multivariate normal, i.e.,

$$(y_i^{(l)}, \mathbf{x}_i^{(2l)}, \mathbf{x}_i^{(3l)}) \mid \mathbf{x}_i^{(1l)} \sim MVN\left(\boldsymbol{\mu}\left(\mathbf{x}_i^{(1l)}\right), \quad \boldsymbol{\Sigma}\right). \tag{3.1}$$

Here, we consider features in modality 1, $\mathbf{x}_i^{(1l)}$, to be covariates instead of random variables because $\mathbf{x}_i^{(1l)}$ contains no missing features. Making it covariates does not affect the performance of IMTL and meanwhile has the benefit of reducing the number of parameters to be estimated.

In (3.1), $\boldsymbol{\mu}(\cdot)$ is a vector function of covariates. Although $\boldsymbol{\mu}(\cdot)$ can take any form in theory, we focus on a linear function in this paper, i.e.,

$$\boldsymbol{\mu}\left(\mathbf{x}_i^{(1l)}\right) = \left(\mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 + \beta_0,\ \mathbf{x}_i^{(1l)}\mathbf{A}_2 + \mathbf{b}_2,\ \mathbf{x}_i^{(1l)}\mathbf{A}_3 + \mathbf{b}_3\right), \tag{3.2}$$

where $\boldsymbol{\beta}_1, \beta_0, \mathbf{A}_2, \mathbf{b}_2, \mathbf{A}_3, \mathbf{b}_3$ are coefficients. The conditional covariance matrix $\boldsymbol{\Sigma}$ in (III.1) can be written in a more explicit format to include sub-matrices of covariance between the response and each modality and between the modalities, i.e.,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \boldsymbol{\Sigma}_{y2} & \boldsymbol{\Sigma}_{y3} \\ \boldsymbol{\Sigma}_{2y} & \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{3y} & \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix}, \tag{3.3}$$

Let $\boldsymbol{\Theta} = (\boldsymbol{\Sigma}, \boldsymbol{\beta}_1, \beta_0, \mathbf{A}_2, \mathbf{b}_2, \mathbf{A}_3, \mathbf{b}_3)$ contain all the unknown parameters for the model in (III.1). We can write down the negative log-likelihood function (NLLF), i.e.,

$$l(\boldsymbol{\Theta}) = n\, log|\boldsymbol{\Sigma}| + \sum_{l=1}^{4}\sum_{i=1}^{n_l}\Big(y_i^{(1)} - \mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 - \beta_0\,, \mathbf{x}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2 - \mathbf{b}_2, \mathbf{x}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3 - \mathbf{b}_3\Big)\boldsymbol{\Sigma}^{-1}\Big(y_i^{(1)} - \mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 - \beta_0\,, \mathbf{x}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2 - \mathbf{b}_2, \mathbf{x}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3 - \mathbf{b}_3\Big)^T. \tag{3.4}$$

Since $l(\boldsymbol{\Theta})$ includes missing features from modalities 2 and 3, we cannot directly optimize $l(\boldsymbol{\Theta})$ to estimate the parameters, but resort to the Expectation-Maximization (EM) algorithm. The general EM framework includes an E-step and an M-step. The E-step is to find the expectation of NLLF with respect to the missing data given the observed data and the current parameter estimates. The M-step is to update the parameter estimates by minimizing the expectation in the E-step. The two steps are iterated until convergence. The challenges in using the general EM framework in a specific model estimation are to derive the expectation specific to that model formulation in the E-step and to solve the specific optimization problem in the M-step. In what follows, we will develop the details for the E-step and M-step for our specific model.

When the likelihood function is based on a distribution in the exponential family, Little and Rubin (2002) showed that the E-step becomes finding expectations of sufficient statistics. Our likelihood function is based on a multivariate normal distribution in (3.1). Therefore, the goal of the E-step is to find the sufficient statistics associate with (3.1) and derive their expectations. Let S be a collection of the sufficient statistics. It can be shown that

$$S = \left\{ \mathbf{x}_i^{(21)}, \mathbf{x}_i^{(31)}, \mathbf{x}_i^{(32)}, \mathbf{x}_i^{(24)}, \begin{pmatrix} (\mathbf{x}_i^{(21)})^T \mathbf{x}_i^{(21)} & \left(\mathbf{x}_i^{(31)}\right)^T \mathbf{x}_i^{(21)} \\ (\mathbf{x}_i^{(21)})^T \mathbf{x}_i^{(31)} & \left(\mathbf{x}_i^{(31)}\right)^T \mathbf{x}_i^{(31)} \end{pmatrix}, \right\}.$$
$$(\mathbf{x}_i^{(32)})^T \mathbf{x}_i^{(32)}, (\mathbf{x}_i^{(24)})^T \mathbf{x}_i^{(24)}$$

In the E-step at the $(t+1)$-th iteration, we need to drive the expectation of S given the previous estimate for $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}^{(t)}$, and observed data, $\mathbf{x}_i^{obs}$ and y_i^{obs}, i.e.,

$$E\left(S \middle| \mathbf{x}_i^{obs}, y_i^{obs}, \boldsymbol{\Theta}^{(t)}\right). \tag{3.5}$$

In (III.6)-(III.8) below, we first derive the expectations of the 1st -order elements in S. The detailed derivation process is skipped due to space limit.

$$\left(\tilde{\mathbf{x}}_i^{(21)}, \tilde{\mathbf{x}}_i^{(31)}\right) = E\left[\left(\mathbf{x}_i^{(21)}, \mathbf{x}_i^{(31)}\right) \middle| \mathbf{x}_i^{(11)}, y_i^{(1)}, \boldsymbol{\Theta}^{(t)}\right]$$
$$= \left(\mathbf{x}_i^{(11)}\mathbf{A}_2^{(t)} + \mathbf{b}_2^{(t)}, \mathbf{x}_i^{(11)}\mathbf{A}_3^{(t)} + \mathbf{b}_3^{(t)}\right) + \left(\boldsymbol{\Sigma}_{2y}^{(t)}, \boldsymbol{\Sigma}_{3y}^{(t)}\right)\left(\sigma_y^{(t)2}\right)^{-1}\left(y_i^{(1)} - \mathbf{x}_i^{(11)}\boldsymbol{\beta}_1^{(t)} - \beta_0^{(t)}\right), \tag{3.6}$$

$$\tilde{\mathbf{x}}_i^{(32)} = E\left[\mathbf{x}_i^{(32)} \middle| \mathbf{x}_i^{(12)}, \mathbf{x}_i^{(22)}, y_i^{(2)}, \boldsymbol{\Theta}^{(t)}\right]$$
$$= \mathbf{x}_i^{(12)}\boldsymbol{\beta}_1^{(t)} + \beta_0^{(t)} + \left(\boldsymbol{\Sigma}_{3y}^{(t)}, \boldsymbol{\Sigma}_{32}^{(t)}\right)\begin{pmatrix} \sigma_y^{(t)2} & \boldsymbol{\Sigma}_{y2}^{(t)} \\ \boldsymbol{\Sigma}_{2y}^{(t)} & \boldsymbol{\Sigma}_{22}^{(t)} \end{pmatrix}^{-1}\begin{pmatrix} y_i^{(2)} - \mathbf{x}_i^{(12)}\beta_1^{(t)} - \beta_0^{(t)} \\ \mathbf{x}_i^{(22)} - \mathbf{x}_i^{(12)}\mathbf{A}_2^{(t)} - \mathbf{b}_2^{(t)} \end{pmatrix} \tag{3.7}$$

$$\tilde{\mathbf{x}}_i^{(24)} = E\left[\mathbf{x}_i^{(24)} \middle| \mathbf{x}_i^{(14)}, \mathbf{x}_i^{(34)}, y_i^{(4)}, \boldsymbol{\Theta}^{(t)}\right]$$

$$= \mathbf{x}_i^{(14)}\boldsymbol{\beta}_1^{(t)} + \beta_0^{(t)} + \left(\boldsymbol{\Sigma}_{2y}^{(t)}, \boldsymbol{\Sigma}_{23}^{(t)}\right)\begin{pmatrix}\sigma_y^{(t)2} & \boldsymbol{\Sigma}_{y3}^{(t)} \\ \boldsymbol{\Sigma}_{3y}^{(t)} & \boldsymbol{\Sigma}_{33}^{(t)}\end{pmatrix}^{-1}\begin{pmatrix} y_i^{(4)} - \mathbf{x}_i^{(14)}\boldsymbol{\beta}_1^{(t)} - \beta_0^{(t)} \\ \mathbf{x}_i^{(34)} - \mathbf{x}_i^{(14)}\mathbf{A}_3^{(t)} - \mathbf{b}_3^{(t)}\end{pmatrix} \quad (3.8)$$

Using (3.6) -(3.8), we can further derive the expectations of the 2nd-order elements in S as:

$$E\left(\begin{pmatrix}(\mathbf{x}_i^{(21)})^T\mathbf{x}_i^{(21)} & \left(\mathbf{x}_i^{(31)}\right)^T\mathbf{x}_i^{(21)} \\ (\mathbf{x}_i^{(21)})^T\mathbf{x}_i^{(31)} & \left(\mathbf{x}_i^{(31)}\right)^T\mathbf{x}_i^{(31)}\end{pmatrix} \middle| \mathbf{x}_i^{(11)}, y_i^{(1)}, \boldsymbol{\Theta}^{(t)}\right) =$$

$$\begin{pmatrix}(\tilde{\mathbf{x}}_i^{(21)})^T\tilde{\mathbf{x}}_i^{(21)} & (\tilde{\mathbf{x}}_i^{(31)})^T\tilde{\mathbf{x}}_i^{(21)} \\ (\tilde{\mathbf{x}}_i^{(21)})^T\tilde{\mathbf{x}}_i^{(31)} & (\tilde{\mathbf{x}}_i^{(31)})^T\tilde{\mathbf{x}}_i^{(31)}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\Sigma}_{22|y}^{(t)} & \boldsymbol{\Sigma}_{23|y}^{(t)} \\ \boldsymbol{\Sigma}_{32|y}^{(t)} & \boldsymbol{\Sigma}_{33|y}^{(t)}\end{pmatrix}, \quad (3.9)$$

$$E\left[\left(\mathbf{x}_i^{(24)}\right)^T\mathbf{x}_i^{(24)}\middle|\mathbf{x}_i^{(14)}, \mathbf{x}_i^{(34)}, y_i^{(4)}, \boldsymbol{\Theta}^{(t)}\right] = (\tilde{\mathbf{x}}_i^{(24)})^T\tilde{\mathbf{x}}_i^{(24)} + \boldsymbol{\Sigma}_{22|3y}^{(t)}, \quad (3.10)$$

$$E\left[\left(\mathbf{x}_i^{(32)}\right)^T\mathbf{x}_i^{(32)}\middle|\mathbf{x}_i^{(12)}, \mathbf{x}_i^{(22)}, y_i^{(2)}, \boldsymbol{\Theta}^{(t)}\right] = (\tilde{\mathbf{x}}_i^{(32)})^T\tilde{\mathbf{x}}_i^{(32)} + \boldsymbol{\Sigma}_{33|2y}^{(t)} \quad (3.11)$$

where

$$\begin{pmatrix}\boldsymbol{\Sigma}_{22|y}^{(t)} & \boldsymbol{\Sigma}_{23|y}^{(t)} \\ \boldsymbol{\Sigma}_{32|y}^{(t)} & \boldsymbol{\Sigma}_{33|y}^{(t)}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{22}^{(t)} & \boldsymbol{\Sigma}_{23}^{(t)} \\ \boldsymbol{\Sigma}_{32}^{(t)} & \boldsymbol{\Sigma}_{33}^{(t)}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\Sigma}_{2y}^{(t)} \\ \boldsymbol{\Sigma}_{3y}^{(t)}\end{pmatrix}\left(\sigma_y^{(t)2}\right)^{-1}(\boldsymbol{\Sigma}_{y2}^{(t)}\ \boldsymbol{\Sigma}_{y3}^{(t)}),$$

$$\boldsymbol{\Sigma}_{22|3y}^{(t)} = \boldsymbol{\Sigma}_{22}^{(t)} - \left(\boldsymbol{\Sigma}_{2y}^{(t)}, \boldsymbol{\Sigma}_{23}^{(t)}\right)\begin{pmatrix}\sigma_y^{(t)2} & \boldsymbol{\Sigma}_{y3}^{(t)} \\ \boldsymbol{\Sigma}_{3y}^{(t)} & \boldsymbol{\Sigma}_{33}^{(t)}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{\Sigma}_{y2}^{(t)} \\ \boldsymbol{\Sigma}_{32}^{(t)}\end{pmatrix},$$

$$\boldsymbol{\Sigma}_{33|2y}^{(t)} = \boldsymbol{\Sigma}_{33}^{(t)} - \left(\boldsymbol{\Sigma}_{3y}^{(t)}, \boldsymbol{\Sigma}_{32}^{(t)}\right)\begin{pmatrix}\sigma_y^{(t)2} & \boldsymbol{\Sigma}_{y2}^{(t)} \\ \boldsymbol{\Sigma}_{2y}^{(t)} & \boldsymbol{\Sigma}_{22}^{(t)}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{\Sigma}_{y3}^{(t)} \\ \boldsymbol{\Sigma}_{23}^{(t)}\end{pmatrix}.$$

In the M-step, we find the new estimate for $\boldsymbol{\Theta}$, $\boldsymbol{\Theta}^{(t+1)}$, as the solution to the following equation:

$$E(S|\mathbf{x}_i^{obs}, \mathrm{y}_i^{obs}, \boldsymbol{\Theta}^{(t)}) = E\left(S\middle|\ \boldsymbol{\Theta}^{(t)}\right). \quad (3.12)$$

Equation (3.12) includes a set of simultaneous equations corresponding to the elements in S. Directly solving (3.12) is not trivial. By using a notational trick, we convert the solving of (3.12) into obtaining least square estimators for regression coefficients. Specifically, denote $\mathbf{x}_i^{(kl)}$ by $\tilde{\mathbf{x}}_i^{(kl)}$ if $\mathbf{x}_i^{(kl)}$ does not contain missing features. Then, $(\boldsymbol{\beta}_1, \beta_0, \mathbf{A}_2, \mathbf{b}_2, \mathbf{A}_3, \mathbf{b}_3)$ are coefficients of the following regressions:

$$\begin{cases} y_i^{(l)} \sim \mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 + \beta_0 \\ \tilde{\mathbf{x}}_i^{(2l)} \sim \mathbf{x}_i^{(1l)}\mathbf{A}_2 + \mathbf{b}_2 \\ \tilde{\mathbf{x}}_i^{(3l)} \sim \mathbf{x}_i^{(1l)}\mathbf{A}_3 + \mathbf{b}_3 \end{cases},$$

for which the coefficients can be obtained by least square estimation, i.e.,

$$\begin{cases} \begin{pmatrix} \beta_0^{(t+1)} \\ \boldsymbol{\beta}_1^{(t+1)} \end{pmatrix} = \left(\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1, \mathbf{x}_i^{(1l)}\right)^T\left(1, \mathbf{x}_i^{(1l)}\right)\right)^{-1}\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1, \mathbf{x}_i^{(1l)}\right)^T y_i^{(l)} \\ \begin{pmatrix} \mathbf{b}_2^{(t+1)} \\ \mathbf{A}_2^{(t+1)} \end{pmatrix} = \left(\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1, \mathbf{x}_i^{(1l)}\right)^T\left(1, \mathbf{x}_i^{(1l)}\right)\right)^{-1}\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1, \mathbf{x}_i^{(1l)}\right)^T \tilde{\mathbf{x}}_i^{(2l)} \\ \begin{pmatrix} \mathbf{b}_3^{(t+1)} \\ \mathbf{A}_3^{(t+1)} \end{pmatrix} = \left(\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1, \mathbf{x}_i^{(1l)}\right)^T\left(1, \mathbf{x}_i^{(1l)}\right)\right)^{-1}\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1, \mathbf{x}_i^{(1l)}\right)^T \tilde{\mathbf{x}}_i^{(3l)} \end{cases}. \tag{3.13}$$

Let $\mathbf{z}_i^{(l)} = \left(y_i^{(1)} - \mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1^{(t+1)} - \beta_0^{(t+1)}, \tilde{\mathbf{x}}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2^{(t+1)} - \mathbf{b}_2^{(t+1)}, \tilde{\mathbf{x}}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3^{(t+1)} - \mathbf{b}_3^{(t+1)}\right)$.

Using (3.13), we can further estimate $\boldsymbol{\Sigma}$ as follows:

$$\boldsymbol{\Sigma}^{(t+1)} = \frac{1}{n}\left\{\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(\mathbf{z}_i^{(l)}\right)^T\mathbf{z}_i^{(l)} + n_4\begin{pmatrix}0 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22|3y}^{(t)} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\end{pmatrix} + n_2\begin{pmatrix}0 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{33|2y}^{(t)}\end{pmatrix} + n_1\begin{pmatrix}0 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{22|y}^{(t)} & \boldsymbol{\Sigma}_{23|y}^{(t)}\\ \mathbf{0} & \boldsymbol{\Sigma}_{32|y}^{(t)} & \boldsymbol{\Sigma}_{33|y}^{(t)}\end{pmatrix}\right\}. \quad (3.14)$$

2) Prediction

At the convergence of the above EM iterations, we can obtain the estimated parameters $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\beta}}_1, \hat{\beta}_0, \widehat{\mathbf{A}}_2, \hat{\mathbf{b}}_2, \widehat{\mathbf{A}}_3, \hat{\mathbf{b}}_3)$. Then, these parameters can be used to predict on new samples. Consider a new sample i^*. Depending on what available modality/modalities this sample has, we can use the following model to predict the response variable of the sample:

$\hat{y}_{i^*} = \mathbf{x}_{i^*}^{(11)}\widehat{\boldsymbol{\beta}}_1 + \hat{\beta}_0,$ if $i^* \in sub-cohort\ 1$;

$\hat{y}_{i^*} = \mathbf{x}_i^{(12)}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\mathbf{A}}_2\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{2y}\right) + \mathbf{x}_i^{(22)}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{2y} + (\hat{\beta}_0 - \hat{\mathbf{b}}_2\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\widehat{\boldsymbol{\Sigma}}_{2y}),$ if $i^* \in sub-cohort\ 2$;

$$\hat{y}_{i^*} = \mathbf{x}_{i^*}^{(13)}\left(\widehat{\boldsymbol{\beta}}_1 - (\widehat{\mathbf{A}}_2, \widehat{\mathbf{A}}_3)\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{22} & \widehat{\boldsymbol{\Sigma}}_{23}\\ \widehat{\boldsymbol{\Sigma}}_{32} & \widehat{\boldsymbol{\Sigma}}_{33}\end{pmatrix}^{-1}\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{2y}\\ \widehat{\boldsymbol{\Sigma}}_{3y}\end{pmatrix}\right) + \left(\mathbf{x}_{i^*}^{(23)}, \mathbf{x}_{i^*}^{(33)}\right)\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{22} & \widehat{\boldsymbol{\Sigma}}_{23}\\ \widehat{\boldsymbol{\Sigma}}_{32} & \widehat{\boldsymbol{\Sigma}}_{33}\end{pmatrix}^{-1}\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{2y}\\ \widehat{\boldsymbol{\Sigma}}_{3y}\end{pmatrix} + \left(\hat{\beta}_0 - (\hat{\mathbf{b}}_2, \hat{\mathbf{b}}_3)\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{22} & \widehat{\boldsymbol{\Sigma}}_{23}\\ \widehat{\boldsymbol{\Sigma}}_{32} & \widehat{\boldsymbol{\Sigma}}_{33}\end{pmatrix}^{-1}\begin{pmatrix}\widehat{\boldsymbol{\Sigma}}_{2y}\\ \widehat{\boldsymbol{\Sigma}}_{3y}\end{pmatrix}\right),$$ if $i^* \in sub-cohort\ 3$;

$\hat{y}_{i^*} = \mathbf{x}_{i^*}^{(14)}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\mathbf{A}}_3\widehat{\boldsymbol{\Sigma}}_{33}^{-1}\widehat{\boldsymbol{\Sigma}}_{3y}\right) + \mathbf{x}_{i^*}^{(34)}\widehat{\boldsymbol{\Sigma}}_{33}^{-1}\widehat{\boldsymbol{\Sigma}}_{3y} + (\hat{\beta}_0 - \hat{\mathbf{b}}_3\widehat{\boldsymbol{\Sigma}}_{33}^{-1}\widehat{\boldsymbol{\Sigma}}_{3y}),$

if $i^* \in sub-cohort\ 4$.

3.3.2. IMTL classification model

1) Formulation and estimation

In a classification model, $y_i^{(l)}$ can take the values of 0 or 1 that represent two classes. Within each class, consider the joint distribution $\mathbf{x}_i^{(2l)}$ and $\mathbf{x}_i^{(3l)}$ given $\mathbf{x}_i^{(1l)}$ to be multivariate normal, i.e.,

$$(\mathbf{x}_i^{(2l)}, \mathbf{x}_i^{(3l)}) | \mathbf{x}_i^{(1l)}, y_i^{(l)} = 1 \sim MVN(\boldsymbol{\mu}_1\left(\mathbf{x}_i^{(1l)}\right), \boldsymbol{\Sigma}), \tag{3.15}$$

$$(\mathbf{x}_i^{(2l)}, \mathbf{x}_i^{(3l)}) | \mathbf{x}_i^{(1l)}, y_i^{(l)} = 0 \sim MVN(\boldsymbol{\mu}_0\left(\mathbf{x}_i^{(1l)}\right), \boldsymbol{\Sigma}), \tag{3.16}$$

We focus on linear functions of $\boldsymbol{\mu}_1(\cdot)$ and $\boldsymbol{\mu}_0(\cdot)$ in this paper:

$$\boldsymbol{\mu}_1\left(\mathbf{x}_i^{(1l)}\right) = \left(\mathbf{x}_i^{(1l)}\mathbf{A}_2 + \mathbf{b}_{21}, \mathbf{x}_i^{(1l)}\mathbf{A}_3 + \mathbf{b}_{31}\right),$$

$$\boldsymbol{\mu}_0\left(\mathbf{x}_i^{(1l)}\right) = \left(\mathbf{x}_i^{(1l)}\mathbf{A}_2 + \mathbf{b}_{20}, \mathbf{x}_i^{(1l)}\mathbf{A}_3 + \mathbf{b}_{30}\right).$$

Also, we can write $\boldsymbol{\Sigma}$ in a more explicit form to facilitate subsequent discussion, i.e.,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{22} & \boldsymbol{\Sigma}_{23} \\ \boldsymbol{\Sigma}_{32} & \boldsymbol{\Sigma}_{33} \end{pmatrix}.$$

Furthermore, we consider the distribution of $y_i^{(l)}$ given $\mathbf{x}_i^{(1l)}$ to be Bernoulli, i.e.,

$$y_i^{(l)} = 1 | \mathbf{x}_i^{(1l)} \sim Bernoulli\ \left(\frac{1}{1+\exp\{-\mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 - \beta_0\}}\right). \tag{3.17}$$

Let $\widetilde{\boldsymbol{\Theta}} = (\boldsymbol{\Sigma}, \boldsymbol{\beta}_1, \beta_0, \mathbf{A}_2, \mathbf{b}_2, \mathbf{A}_3, \mathbf{b}_3)$ be the unknown parameters for the model in (3.15)-(3.17). The NLLF is:

$$l(\widetilde{\boldsymbol{\Theta}}) = n \log|\boldsymbol{\Sigma}| + \sum_{l=1}^{4}\sum_{i=1}^{l}\Big(y_i^{(l)}\Big(\mathbf{x}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2 - \mathbf{b}_{21}, \mathbf{x}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3 - \mathbf{b}_{31}\Big)\boldsymbol{\Sigma}^{-1}\Big(\mathbf{x}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2 - \mathbf{b}_{21}, \mathbf{x}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3 - \mathbf{b}_{31}\Big)^T + (1 - y_i^{(l)})\Big(\mathbf{x}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2 -$$

$$\mathbf{b}_{20}, \mathbf{x}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3 - \mathbf{b}_{30}\Big)\boldsymbol{\Sigma}^{-1}\Big(\mathbf{x}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2 - \mathbf{b}_{20}, \mathbf{x}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3 - \mathbf{b}_{30}\Big)^T\Big) +$$

$$y_i^{(l)}\Big(\mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 + \beta_0\Big) - \log\Big(1 + \exp\Big(\mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 + \beta_0\Big)\Big).$$

(3.18)

Equation (3.18) can be decomposed into a logistic regression and a conditional multivariate normal distribution. As a result, we can estimate $(\boldsymbol{\beta}_1, \beta_0)$ and the remaining parameters in $\widetilde{\boldsymbol{\Theta}}$ separately. Specifically, $(\boldsymbol{\beta}_1, \beta_0)$ are coefficients of the logistic regression model:

$$logit\Big(P\Big(y_i^{(l)} = 1\Big)\Big) = \mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 + \beta_0.$$

This model does not involve missing data, which means that $(\boldsymbol{\beta}_1, \beta_0)$ can be estimated by iteratively reweighted least squares (IRLS) estimation.

Furthermore, let $\boldsymbol{\Theta}$ be the parameters in $\widetilde{\boldsymbol{\Theta}}$ excluding $(\boldsymbol{\beta}_1, \beta_0)$. $\boldsymbol{\Theta}$ can be estimated by a similar EM algorithm to the predictive model in Section III.3.1. Here, we skip the details and present the result of derivations in the E-step and M-step. Specifically, the E-step derives the following expectations:

$$\Big(\tilde{\mathbf{x}}_i^{(21)}, \tilde{\mathbf{x}}_i^{(31)}\Big) = E\left[\Big(\mathbf{x}_i^{(21)}, \mathbf{x}_i^{(31)}\Big)\Big|\mathbf{x}_i^{(11)}, y_i^{(1)}, \boldsymbol{\Theta}^{(t)}\right]$$

$$= y_i^{(1)}\Big(\mathbf{x}_i^{(11)}\mathbf{A}_2^{(t)} + \mathbf{b}_{21}^{(t)}, \mathbf{x}_i^{(11)}\mathbf{A}_3^{(t)} + \mathbf{b}_{31}^{(t)}\Big) + (1 - y_i^{(1)})\Big(\mathbf{x}_i^{(11)}\mathbf{A}_2^{(t)} +$$

$$\mathbf{b}_{20}^{(t)}, \mathbf{x}_i^{(11)}\mathbf{A}_3^{(t)} + \mathbf{b}_{30}^{(t)}\Big),$$

$$\tilde{\mathbf{x}}_i^{(32)} = E\Big[\mathbf{x}_i^{(32)}\Big|\mathbf{x}_i^{(12)}, \mathbf{x}_i^{(22)}, y_i^{(2)}, \boldsymbol{\Theta}^{(t)}\Big]$$

$$= y_i^{(2)}\Big(\mathbf{x}_i^{(12)}\mathbf{A}_3^{(t)} + \mathbf{b}_{31}^{(t)}\Big) + (1 - y_i^{(2)})\Big(\mathbf{x}_i^{(12)}\mathbf{A}_3^{(t)} + \mathbf{b}_{30}^{(t)}\Big) + \boldsymbol{\Sigma}_{32}^{(t)}\Big(\boldsymbol{\Sigma}_{22}^{(t)}\Big)^{-1}\Big(\mathbf{x}_i^{(22)} -$$

$$y_i^{(2)}\Big(\mathbf{x}_i^{(12)}\mathbf{A}_2^{(t)} + \mathbf{b}_{21}^{(t)}\Big) - (1 - y_i^{(2)})\Big(\mathbf{x}_i^{(12)}\mathbf{A}_2^{(t)} + \mathbf{b}_{20}^{(t)}\Big)\Big),$$

$$\tilde{\mathbf{x}}_i^{(24)} = E\left[\mathbf{x}_i^{(24)} \middle| \mathbf{x}_i^{(14)}, \mathbf{x}_i^{(34)}, y_i^{(4)}, \boldsymbol{\Theta}^{(t)}\right]$$

$$= y_i^{(4)}\left(\mathbf{x}_i^{(14)}\mathbf{A}_2^{(t)} + \mathbf{b}_{21}^{(t)}\right) + (1 - y_i^{(4)})\left(\mathbf{x}_i^{(14)}\mathbf{A}_2^{(t)} + \mathbf{b}_{20}^{(t)}\right) + \boldsymbol{\Sigma}_{23}^{(t)}\left(\boldsymbol{\Sigma}_{33}^{(t)}\right)^{-1}\left(\mathbf{x}_i^{(34)} - y_i^{(4)}\left(\mathbf{x}_i^{(14)}\mathbf{A}_3^{(t)} + \mathbf{b}_{31}^{(t)}\right) - (1 - y_i^{(4)})\left(\mathbf{x}_i^{(14)}\mathbf{A}_3^{(t)} + \mathbf{b}_{30}^{(t)}\right)\right).$$

In the M-step, the parameters in $\boldsymbol{\Theta}$ can be updated as

$$\mathbf{A}_2^{(t+1)} = \left(\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(\mathbf{x}_i^{(1l)}\right)^T\mathbf{x}_i^{(1l)}\right)^{-1}\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(\mathbf{x}_i^{(1l)}\right)^T\tilde{\mathbf{x}}_i^{(2l)},$$

$$\mathbf{b}_{21}^{(t+1)} = \frac{\sum_{l=1}^{4}\sum_{i=1}^{n_l} y_i^{(l)}\left(\tilde{\mathbf{x}}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2^{(t+1)}\right)}{\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(y_i^{(l)}\right)},$$

$$\mathbf{b}_{20}^{(t+1)} = \frac{\sum_{l=1}^{4}\sum_{i=1}^{n_l}(1-y_i^{(l)})\left(\tilde{\mathbf{x}}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2^{(t+1)}\right)}{\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1-y_i^{(l)}\right)},$$

$$\mathbf{A}_3^{(t+1)} = \left(\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(\mathbf{x}_i^{(1l)}\right)^T\mathbf{x}_i^{(1l)}\right)^{-1}\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(\mathbf{x}_i^{(1l)}\right)^T\tilde{\mathbf{x}}_i^{(3l)},$$

$$\mathbf{b}_{31}^{(t+1)} = \frac{\sum_{l=1}^{4}\sum_{i=1}^{n_l} y_i^{(l)}\left(\tilde{\mathbf{x}}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3^{(t+1)}\right)}{\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(y_i^{(l)}\right)},$$

$$\mathbf{b}_{30}^{(t+1)} = \frac{\sum_{l=1}^{4}\sum_{i=1}^{n_l}(1-y_i^{(l)})\left(\tilde{\mathbf{x}}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3^{(t+1)}\right)}{\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(1-y_i^{(l)}\right)},$$

$$\boldsymbol{\Sigma}^{(t+1)} = \frac{1}{n}\left\{\sum_{l=1}^{4}\sum_{i=1}^{n_l}\left(y_i^{(l)}\left(\tilde{\mathbf{x}}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2^{(t+1)} - \mathbf{b}_{21}^{(t+1)}, \tilde{\mathbf{x}}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3^{(t+1)} - \mathbf{b}_{31}^{(t+1)}\right)^T\left(\tilde{\mathbf{x}}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2^{(t+1)} - \mathbf{b}_{21}^{(t+1)}, \tilde{\mathbf{x}}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3^{(t+1)} - \mathbf{b}_{31}^{(t+1)}\right) + \left(1 - y_i^{(l)}\right)\left(\tilde{\mathbf{x}}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2^{(t+1)} - \mathbf{b}_{20}^{(t+1)}, \tilde{\mathbf{x}}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3^{(t+1)} - \mathbf{b}_{30}^{(t+1)}\right)^T\left(\tilde{\mathbf{x}}_i^{(2l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_2^{(t+1)} -\right.\right.$$

$$\mathbf{b}_{20}^{(t+1)}, \tilde{\mathbf{x}}_i^{(3l)} - \mathbf{x}_i^{(1l)}\mathbf{A}_3^{(t+1)} - \mathbf{b}_{30}^{(t+1)}\Big)\Big) + n_4\begin{pmatrix}\boldsymbol{\Sigma}_{22}^{(t)} - \boldsymbol{\Sigma}_{23}^{(t)}\left(\boldsymbol{\Sigma}_{33}^{(t)}\right)^{-1}\boldsymbol{\Sigma}_{32}^{(t)} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} +$$

$$n_2\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{33}^{(t)} - \boldsymbol{\Sigma}_{32}^{(t)}\left(\boldsymbol{\Sigma}_{22}^{(t)}\right)^{-1}\boldsymbol{\Sigma}_{23}^{(t)}\end{pmatrix} + n_1\boldsymbol{\Sigma}^{(t)}\Big\}.$$

2) Prediction

At the convergence of the above EM iterations, we can obtain the estimated parameters $\widehat{\boldsymbol{\Theta}} = (\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\beta}}_1, \hat{\beta}_0, \widehat{\mathbf{A}}_2, \hat{\mathbf{b}}_2, \widehat{\mathbf{A}}_3, \hat{\mathbf{b}}_3)$. These parameters can be used in a logistic regression model to predict on new samples. Consider a new sample i^*. Depending on what available modality/modalities this sample has, i.e., which sub-cohort the sample belongs to, we can use the following model to predict the response variable of the sample:

$$P\left(y_{i^*}^{(1)} = 1\middle|\mathbf{x}_{i^*}^{(11)}\right) = f\left(\mathbf{x}_{i^*}^{(11)}\widehat{\boldsymbol{\beta}}_1 + \hat{\beta}_0\right), \qquad \text{if } i^* \in sub-cohort1;$$

$$P\left(y_{i^*}^{(2)} = 1\middle|\mathbf{x}_{i^*}^{(12)}, \mathbf{x}_{i^*}^{(22)}\right) = f\left(\mathbf{x}_{i^*}^{(12)}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\mathbf{A}}_2\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\left(\hat{\mathbf{b}}_{21} - \hat{\mathbf{b}}_{20}\right)\right) + \mathbf{x}_{i^*}^{(22)}\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\left(\hat{\mathbf{b}}_{21} - \hat{\mathbf{b}}_{20}\right) + \hat{\beta}_0 - \frac{1}{2}\hat{\mathbf{b}}_{21}^T\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\mathbf{b}}_{21} + \frac{1}{2}\hat{\mathbf{b}}_{20}^T\widehat{\boldsymbol{\Sigma}}_{22}^{-1}\hat{\mathbf{b}}_{20}\right), \qquad \text{if } i^* \in sub-cohort2;$$

$$P\left(y_i^{(3)} = 1\middle|\mathbf{x}_{i^*}^{(13)}, \mathbf{x}_{i^*}^{(23)}, \mathbf{x}_{i^*}^{(33)}\right) = f\left(\mathbf{x}_{i^*}^{(13)}\left(\widehat{\boldsymbol{\beta}}_1 - (\widehat{\mathbf{A}}_2, \widehat{\mathbf{A}}_3)\widehat{\boldsymbol{\Sigma}}^{-1}\begin{pmatrix}\hat{\mathbf{b}}_{21} - \hat{\mathbf{b}}_{20}\\ \hat{\mathbf{b}}_{31} - \hat{\mathbf{b}}_{30}\end{pmatrix}\right) + \left(\mathbf{x}_{i^*}^{(23)}, \mathbf{x}_{i^*}^{(33)}\right)\widehat{\boldsymbol{\Sigma}}^{-1}\begin{pmatrix}\hat{\mathbf{b}}_{21} - \hat{\mathbf{b}}_{20}\\ \hat{\mathbf{b}}_{31} - \hat{\mathbf{b}}_{30}\end{pmatrix} + \hat{\beta}_0 - \frac{1}{2}(\hat{\mathbf{b}}_{21}, \hat{\mathbf{b}}_{31})\widehat{\boldsymbol{\Sigma}}^{-1}\begin{pmatrix}\hat{\mathbf{b}}_{21}\\ \hat{\mathbf{b}}_{31}\end{pmatrix} + \frac{1}{2}(\hat{\mathbf{b}}_{20}, \hat{\mathbf{b}}_{30})\widehat{\boldsymbol{\Sigma}}^{-1}\begin{pmatrix}\hat{\mathbf{b}}_{20}\\ \hat{\mathbf{b}}_{30}\end{pmatrix}\right),$$

$$\text{if } i^* \in sub-cohort3;$$

$$P\left(y_i^{(4)} = 1\middle|\mathbf{x}_{i^*}^{(14)}, \mathbf{x}_{i^*}^{(34)}\right) = f\left(\mathbf{x}_{i^*}^{(14)}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\mathbf{A}}_3\widehat{\boldsymbol{\Sigma}}_{33}^{-1}\left(\hat{\mathbf{b}}_{31} - \hat{\mathbf{b}}_{30}\right)\right) + \mathbf{x}_{i^*}^{(34)}\widehat{\boldsymbol{\Sigma}}_{33}^{-1}\left(\hat{\mathbf{b}}_{31} - \hat{\mathbf{b}}_{30}\right) + \hat{\beta}_0 - \frac{1}{2}\hat{\mathbf{b}}_{31}^T\widehat{\boldsymbol{\Sigma}}_{33}^{-1}\hat{\mathbf{b}}_{31} + \frac{1}{2}\hat{\mathbf{b}}_{30}^T\widehat{\boldsymbol{\Sigma}}_{33}^{-1}\hat{\mathbf{b}}_{30}\right), \qquad \text{if } i^* \in sub-cohort4.$$

$f(\cdot)$ is the sigmoid function.

3.3.3. Collaborative model estimation without data pooling

As mentioned previously, one reason leading to generation of IMD data in health care applications is that each sub-cohort corresponds to a different institution. The availability of modalities varies across the different institutions due to accessibility and cost. In the IMTL models proposed in Section 3.3.1-3.2, model estimation is assumed to happen at a centralized place into which the training data from different institutions (i.e., sub-cohorts) have been deposited. This requires a multi-institutional data sharing agreement that has been previously established – a process known to be time- and effort-intensive. A more commonly encountered scenario is that the different institutions would like to collaborate on model estimation without having to share their respective patients' data. In this section, we address the latter scenario by proposing a modification on the EM algorithms in Section 3.3.1-3.3.2. Figure 3.2 shows the computational architecture to support between-institution

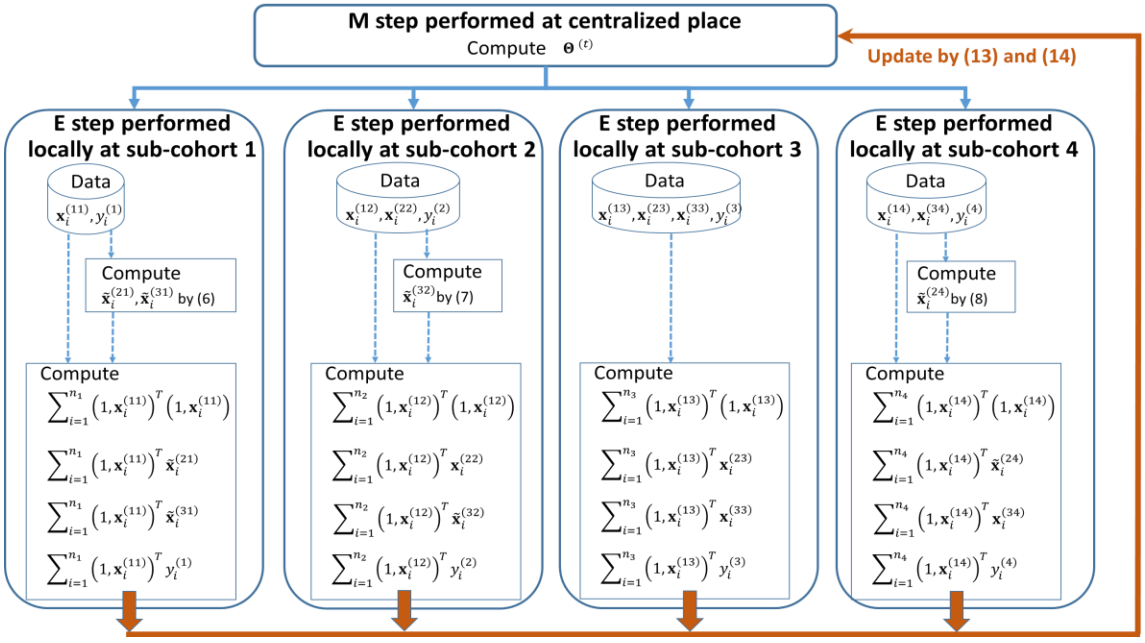


Figure 3.2. Computational architecture for between-institutional/sub-cohort collaborative model estimation of IMTL without the need for data pooling from different sub-cohorts. M-step is performed by a global learner; E-step is performed locally at each sub-cohort.

collaborative model estimation for IMTL without the need for data pooling. The key to this architecture is to consider the M- and E-steps as a global and a local learner, respectively. The global learner resides in a centralized place while the local learners reside in each sub-cohort. A local learner can only "see" and perform computation on the data in its respective sub-cohort. Results from local computation, not the data, are sent to the global learner for integration. Iterative communication between the global and local learners complete the model estimation. Specifically, the estimated parameters at the t-th iteration of the M-step, $\boldsymbol{\Theta}^{(t)}$, is sent to each sub-cohort/institution. Although $\boldsymbol{\Theta}^{(t)}$ is estimated using data from all the sub-cohorts, the data are no longer identifiable after $\boldsymbol{\Theta}^{(t)}$ is obtained. After receiving the $\boldsymbol{\Theta}^{(t)}$, each sub-cohort performs its own E-step *locally* using $\boldsymbol{\Theta}^{(t)}$ and the data specific to the sub-cohort. No communication is needed between the sub-cohort-wise E steps. Once the E-steps computations are completed, the results are sent back to M-step to compute an updated $\boldsymbol{\Theta}^{(t+1)}$. The E- and M-steps iterate until convergence. Note that the proposed computational architecture not only enables privacy preservation but also speeds up the EM iterations because sub-cohort-wise E-steps can be performed on parallel computing resources.

3.4 Properties of IMTL

In this section, we discuss two unique properties of IMTL: 1) the ability for out-of-sample prediction; 2) a theoretical guarantee for a larger Fisher information compared with models without transfer learning, which explains the outperformance of IMTL from a theoretical point of view.

3.4.1. Ability for out-of-sample prediction

Definition: Consider a training set that includes L sub-cohorts, $D_{tr} = \{D_{tr,1}, \dots, D_{tr,L}\}$. Each sub-cohort corresponds to a specific missing modality pattern. Further consider a test set D_{te} that includes a sub-cohort whose missing modality pattern is different from all the sub-cohorts in D_{tr}. If a model trained on D_{tr} can be used predict D_{te}, the model is called capable of *out-of-sample prediction*.

For example, a training set can include only sub-cohorts 1, 2, and 4 in Fig. 1 while the test set includes sub-cohort 3. It is obvious that the two competing methods to IMTL, i.e., SM and AADM, cannot do out-of-sample prediction, because they need inclusion of *some* data from sub-cohort 3 in the model training. In contrast, IMTL is capable of out-of-sample prediction. Next, we provide an illustrative proof for this capability of IMTL. We focus on the predictive model in Section III.3.1. Also, for notation simplicity, each modality is assumed to contain one feature. The case of multivariate features is a straightforward extension.

Consider a sample i^* in the test set who belongs to sub-cohort 3. To predict the response variable of this sample, (III.19) will be used, i.e.,

$$\hat{y}_{i^*}^{(3)} = x_{i^*}^{(13)} \left(\hat{\beta}_1 - (\hat{A}_2, \hat{A}_3) \begin{pmatrix} \hat{\Sigma}_{22} & \hat{\Sigma}_{23} \\ \hat{\Sigma}_{32} & \hat{\Sigma}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\Sigma}_{2y} \\ \hat{\Sigma}_{3y} \end{pmatrix} \right) +$$

$$\left(x_{i^*}^{(23)}, x_{i^*}^{(33)} \right) \begin{pmatrix} \hat{\Sigma}_{22} & \hat{\Sigma}_{23} \\ \hat{\Sigma}_{32} & \hat{\Sigma}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\Sigma}_{2y} \\ \hat{\Sigma}_{3y} \end{pmatrix} + \left(\hat{\beta}_0 - (\hat{b}_2, \hat{b}_3) \begin{pmatrix} \hat{\Sigma}_{22} & \hat{\Sigma}_{23} \\ \hat{\Sigma}_{32} & \hat{\Sigma}_{33} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\Sigma}_{2y} \\ \hat{\Sigma}_{3y} \end{pmatrix} \right).$$

(3.19)

The parameters of the model in (3.19) are estimated from a training set that includes only sub-cohorts 1, 2, and 4 but not 3. It is easy to understand why estimation of other parameters

is possible except Σ_{23}. Intuitively, since Σ_{23} is the covariance between features in modalities 2 and 3, one would expect to have at least *some* training data from sub-cohort 3, which have both modalities 2 and 3 available, in order to estimate Σ_{23}. However, this is not our case. Therefore, the key to demonstrating that IMTL can do out-of-sample prediction is to demonstrate that the estimation for Σ_{23} is possible by IMTL even without any data from sub-cohort 3 in the training set. To show this, consider the estimation for Σ_{23} by the EM algorithm. At convergence, it can be derived that Σ_{23} can be estimated by (3.20). The detailed derivation is skipped due to space limit.

$$\hat{\Sigma}_{23} = \frac{1}{n-n_1-\tilde{n}_2-\tilde{n}_4}(\tilde{n}_1 - n_1)\hat{\Sigma}_{2y}\left(\hat{\sigma}_y^2\right)^{-1}\hat{\Sigma}_{y3} + \frac{1}{k_2}\sum_{i=1}^{n_2}\left(x_i^{(22)} - x_i^{(12)}\hat{A}_2 - \hat{b}_2\right)\left(y_i^{(2)} - x_i^{(12)}\left(\hat{\beta}_1 - \hat{A}_2\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{2y}\right) - x_1^{(22)}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{2y} - \left(\hat{\beta}_0 - \hat{b}_2\,\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{2y}\right)\right)\hat{\Sigma}_{y3} + \frac{1}{k_3}\sum_{i=1}^{n_4}\left(x_i^{(34)} - x_i^{(14)}\hat{A}_3 - \hat{b}_3\right)\left(y_i^{(4)} - x_i^{(14)}(\hat{\beta}_1 - \hat{A}_3\hat{\Sigma}_{33}^{-1}\hat{\Sigma}_{3y}) - x_1^{(34)}\hat{\Sigma}_{33}^{-1}\hat{\Sigma}_{3y} - (\hat{\beta}_0 - \hat{b}_3\,\hat{\Sigma}_{33}^{-1}\hat{\Sigma}_{3y})\right)\hat{\Sigma}_{y2},$$

(3.20)

where

$$k_2 = \hat{\sigma}_y^2 - \hat{\Sigma}_{y2}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{2y},$$

$$k_3 = \hat{\sigma}_y^2 - \hat{\Sigma}_{y3}\hat{\Sigma}_{33}^{-1}\hat{\Sigma}_{3y},$$

$$\tilde{n}_1 = \sum_{i=1}^{n1}\frac{\left(y_i^{(1)} - x_i^{(11)}\hat{\beta}_1 - \hat{\beta}_0\right)^2}{\hat{\sigma}_y^2},$$

$$\tilde{n}_2 = \sum_{i=1}^{n_2}\left(x_i^{(22)} - x_i^{(12)}\hat{A}_2 - \hat{b}_2\right)^2\left(\hat{\Sigma}_{22}^{-1} + \frac{1}{k_2}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{2y}\hat{\Sigma}_{y2}\hat{\Sigma}_{22}^{-1}\right) - \frac{1}{k_2}\sum_{i=1}^{n_2}\left(x_i^{(22)} - x_i^{(12)}\hat{A}_2 - \hat{b}_2\right)^T\left(y_i^{(2)} - x_i^{(12)}\hat{\beta}_1 - \hat{\beta}_0\right)\hat{\Sigma}_{y2}\hat{\Sigma}_{22}^{-1},$$

$\tilde{n}_4 = \sum_{i=1}^{n_4}\left(x_i^{(34)} - x_i^{(14)}\hat{A}_3 - \hat{b}_3\right)^2\left(\hat{\Sigma}_{33}^{-1} + \frac{1}{k_3}\hat{\Sigma}_{33}^{-1}\hat{\Sigma}_{3y}\hat{\Sigma}_{y3}\hat{\Sigma}_{33}^{-1}\right) - \frac{1}{k_3}\sum_{i=1}^{n_2}\hat{\Sigma}_{33}^{-1}\hat{\Sigma}_{3y}\left(y_i^{(4)} - x_i^{(14)}\hat{\beta}_1 - \hat{\beta}_0\right)\left(x_i^{(34)} - x_i^{(14)}\hat{A}_3 - \hat{b}_3\right)$.

Equation (3.20) indicates that although training data from sub-cohort 3 are not available, Σ_{23} can be estimated *indirectly* through a summation of three parts: The first part, $(\tilde{n}_1 - n_1)\hat{\Sigma}_{2y}\left(\hat{\sigma}_y^2\right)^{-1}\hat{\Sigma}_{y3}$, contributes to estimating the covariance between modalities 2 and 3 through exploiting their respective covariances with y. The second part leverages the training data in sub-cohort 2, and contributes to estimating Σ_{23} by exploring the covariance between residual modality 2 and residual modality 3. Here, residual modality 2 is modality 2 after factoring out modality 1; residual modality 3 is the residual of the response variable regressing on modalities 1 and 2. Both residual modalities are computed using the training data in sub-cohort 2. Similarly, the third part leverages the training data in sub-cohort 4, and contributes to estimating Σ_{23} by exploring the covariance between residual modality 2 and residual modality 3 that are computed on the training data in sub-cohort 4.

3.4.2. Fisher information performance

The next section shows better performance of IMTL than SM and AADM (i.e., models without transfer learning) on real data. In this section, we would like to explain the observed performance from a theoretical standpoint. Theorem 1 shows that the better performance of IMTL can be attributed to a larger Fisher information. Because Fisher information characterizes the variance of maximum likelihood estimators for the parameters of a model, a larger Fisher information means that the parameter estimation of IMTL has a smaller variance. For clarity of presentation, Theorem 1 is presented within

the context of a two-modality IMD structure with modality 2 having missing data. This reduces the IMTL model to:

$$(y_i^{(l)}, \mathbf{x}_i^{(2l)}) \mid \mathbf{x}_i^{(1l)} \sim MVN\left(\boldsymbol{\mu}\left(\mathbf{x}_i^{(1l)}\right),\ \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{yy} & \sigma_{y2} \\ \sigma_{2y} & \sigma_{22} \end{pmatrix}$ and $\boldsymbol{\Omega} \triangleq \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \theta_{yy} & \theta_{y2} \\ \theta_{2y} & \theta_{22} \end{pmatrix}$.

Theorem 1: Let $I_{IMRL}(\theta_{ij})$ be the Fisher information for each element in $\boldsymbol{\Omega}$ under IMTL. Let $I_{SM}(\theta_{ij})$ and $I_{AADM}(\theta_{ij})$ be the Fisher information under SM and AADM, respectively. Then,

$$I_{IMRL}(\theta_{ij}) > I_{SM}(\theta_{ij}) = I_{AADM}(\theta_{ij}),$$

if the following condition holds:

$$\frac{-n_1+2p_1+\sqrt{(n_1-2p_1)^2+n_1p_1}}{4p_1} < \frac{\sigma_{2y}^2}{\sigma_{22}\sigma_{yy}}\ , \tag{III.21}$$

where n_1 is the sample size of sub-cohort 1 (i.e., the sub-cohort with only modality 1 available) and p_1 is number of features of modality 1.

Please see the proof in Appendix A. If considering n_1 and p_1 to be fixed (i.e., the left side of (III.21) is a constant), Theorem 1 indicates that the correlation between modality 2 and the response variable must be sufficiently large (i.e., larger than the constant) in order for IMTL to have a larger Fisher information than SM and AADM. The practical insight this Theorem provides is that IMTL will be most effective when the modality with missing data is a significant predictor for the response. If the modality contains largely noise with little predictive value, IMTL may not perform as well as models without transfer learning because it runs the risk of transferring noise and thus hurting the model performance. This

problem is known as "negative transfer" in the literature (Pan and Yang 2010). In essence, Theorem 1 identifies a condition to prevent negative transfer for IMTL.

3.5 Application Case Study

In this section, we apply IMTL to simulated and real-world datasets. Simulation experiments are presented in Section 3.5.1, with purposes of demonstrating the out-of-sample prediction ability of IMTL, which the competing methods (i.e., SM and AADM) do not possess. Section 3.5.2 presents an application of AD diagnosis and prognosis of MCI patients using the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Here, "diagnosis" means detection of the existence of AD pathology in the brain of an MCI patient. "Prognosis" means prediction if an MCI patient will progress to AD by a certain year of interest, e.g., 6 years. Both tasks are important for treatment and management of the patients.

3.5.1. Simulation experiments

1) Out-of-sample prediction by IMTL predictive model

We conduct simulation experiments for the IMTL predictive model and classification model. For the predictive model, we first generate data for three modalities, i.e., $\mathbf{x}_i^{(1l)}, \mathbf{x}_i^{(2l)}, \mathbf{x}_i^{(3l)}$, from a zero-mean multivariate normal distribution $MVN(\mathbf{0}, \mathbf{\Sigma})$. The number of features in each modality is set to be $p_1 = 10$, $p_2 = p_3 = 5$, which are close to the size of features in the real-world data presented in Sec. 5.1. All diagonal elements of $\mathbf{\Sigma}$ are set to be one. $\mathbf{\Sigma}$ includes two parts: within-modality correlation and between-modality correlation. The former has been found to have little impact on the model performance and therefore is set to be 0.6. We investigate two settings for between-modality correlation: 0.6

and 0, which represent moderately strong correlation and no correlation. Furthermore, we investigate two training sample sizes: 300 and 150.

Once the data for features are generated, we generate the response variable $y^{(l)}$ by a linear model, $y_i^{(l)} = \mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 + \mathbf{x}_i^{(2l)}\boldsymbol{\beta}_2 + \mathbf{x}_i^{(3l)}\boldsymbol{\beta}_3 + \beta_0 + \epsilon$. Here, $\beta_0 = 2$; elements in $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$ are set to be 0.2; $\epsilon \sim N(0,1)$. Then, the simulated training data are equally separated into three sub-cohorts, $l = 1, 2, 4$, corresponding to sub-cohorts 1, 2, and 4 in Fig. 1. To obtain the incomplete modality pattern in each sub-cohort, we remove the training data of modalities 2 and 3 for sub-cohort 1, remove modality 3 for sub-cohort 2, and remove modality 2 for sub-cohort 4. Because our intention of this experiment is to assess the out-of-sample prediction capability of IMTL, we generate data in a test data that includes only sub-cohort 3, i.e., all modalities are available. The sample size of the test set is 100.

IMTL is trained on the training set that includes only data from sub-cohorts 1, 2, and 4. Then, the trained model is used to predict on the test set that only includes samples from sub-cohort 3. The predicted response variables of the test set is compared with the true responses to compute a prediction mean square error (PMSE) and a Pearson correlation (PC). We repeat the entire experiment for 100 times. Table III.1 summarizes the results. As expected, increasing the training sample size significantly improves PMSE and PC ($p<0.001$). The correlation between modalities also helps improves PMSE and PC ($p<0.001$). This is consistent with the theoretical discovery in Section 3.4.1, in which we found that the key to out-of-sample prediction was to be able to estimate Σ_{23} from the training data. From (3.20), it is known that the estimation of Σ_{23} is affected by the

correlation between modality 2 and 3. Even though the training data does not include samples with both modality 2 and 3 available, Σ_{23} can still be estimated indirectly by IMTL through exploiting the between-modality correlation and the relationship between modalities and the response variable.

Table 3.1A Out-of-sample prediction accuracy on the test set with different training sample sizes (between-modality correlation is kept as 0.6 in both settings)

Training size	PMSE: $ave(std)$	PC: $ave(std)$
300	1.174 (0.175)	0.945 (0.010)
150	1.469 (0.289)	0.931 (0.017)

Table 3.1B Out-of-sample prediction accuracy on the test set with different between-modality correlations (training sample size is kept as 300 in both settings)

Between-modality correlation	PMSE: $ave(std)$	PC: $ave(std)$
0.6	1.174 (0.175)	0.945 (0.010)
0	1.300 (0.187)	0.866 (0.028)

2) Out-of-sample prediction by IMTL classification model

The data generation process of this experiment is the same as the previous section except that we use a logistic regression model to link the response variable with predictors/features. Specifically, we first simulate a linear predictor $z_i^{(l)} = \mathbf{x}_i^{(1l)}\boldsymbol{\beta}_1 + \mathbf{x}_i^{(2l)}\boldsymbol{\beta}_2 + \mathbf{x}_i^{(3l)}\boldsymbol{\beta}_3 + \beta_0 + \epsilon$. Then, $y_i^{(l)}$ is generated from a Bernoulli distribution with success probability equal to $1/\left(1 + e^{-z_i^{(l)}}\right)$. Test accuracy is reported as the Area Under the Curve (AUC). Table 3.2 summarizes the results. Doubling the training sample size does not seem to dramatically improve the AUC although this improvement is still statistically significant ($p<0.001$). The correlation between modalities also helps improve the AUC significantly ($p<0.001$).

Table 3.2A Out-of-sample classification accuracy on the test set with different training sample sizes (between-modality correlation is kept as 0.6 in both settings)

Training size	AUC: $ave(std)$
300	0.882 (0.037)
150	0.832 (0.05)

Table 3.2B Out-of-sample classification accuracy on the test set with different between-modality correlations (training sample size is kept as 300 in both settings)

Between-modality correlation	AUC: $ave(std)$
0.6	0.882 (0.037)
0	0.781 (0.046)

3.5.2. Early diagnosis and prognosis of AD

1) Introduction to ADNI

ADNI (http://adni.loni.ucla.edu) was launched in 2003 by the NIH, FDA, private pharmaceutical companies, and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W.Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. For up-to-date information, see http://www.adni-info.org/.

2) Patient inclusion and diagnostic/prognostic end points

Our study includes 214 MCI patients from ADNI through our collaborative intuition, Banner Alzheimer's Institute (BAI), with which two co-authors are affiliated. BAI is a member of ADNI PET core (PI, William Jagust UC Berkeley). Multimodality image data include MRI, FDG-PET, amyloid-PET, which follow the IMD structure in Fig. 1. Each sub-cohort has the same sample size. For *diagnosis*, we use Aβ positivity is an indicator for high-risk AD. We follow the recommendation by Fleisher et al. (2011) and use a threshold of mean SUVR greater than or equal to 1.18 to define Aβ positivity. According to this criterion, there are 87 and 127 patients in class 1 (high-risk) and 0 (otherwise). For *prognosis*, the purpose is to predict when an MCI patient will convert to AD. We searched the ADNI database for the 214 patients from the time when their imaging data were collected up to six years' follow up, and found that 46 converted to AD, i.e., there are 46 and 168 converters (class 1) and non-converters (class 0).

3) Image processing and feature computation

For MRI images, we use the FreeSurfer (http://surfer.nmr.mgh.harvard.edu/) software to extract volumetric measurements for pre-defined regions of interest (ROIs). We focus on three ROIs including hippocampal, ventricle, and entorhinal volumes relative to intracranial volume. All three have been widely reported to be related to AD (Devanand et al. 2007; Thompson et al. 2004) . Both FDG-PET and amyloid-PET are PET images, so they share the same image processing step in which we use SPM8 (http://www.fil.ion.ucl.ac.uk/spm/) to spatially normalize each patient's PET images into the common Montreal Neurological Institute (MNI) altas space. Then, we extract features from each type of PET image separately. From FDG-PET, the features include

hypometabolic convergence index (HCI) (Chen et al. 2011), statistical region of interest (sROI) (Chen et al. 2010), and regional precuneus metabolism and posterior cingulate metabolism. All these features have been previously reported to be related to AD (Bailly et al. 2015; Del Sole et al. 2008). From amyloid-PET, the features include SUVRs from six brain regions including orbital frontal cingulate, temporal cortex, anterior cingulate, posterior cingulate, parietal lobe, and precuneus. These regions are known to be associated with amyloid depositions and AD (Fleisher et al. 2011). Because the six SUVRs are highly correlated, we apply principal component analysis (PCA) and include the first PC as a feature for amyloid-PET.

4) Inclusion of clinical variables and feature screening

We also include the following clinical variables which could potentially help the early diagnosis and prognosis of AD: age, gender, years of education, APOE e4 status, and cognitive test scores from several commonly used instruments such as mini-mental state examination (MMSE), AD assessment scale-cognitive (ADAS-Cog), clinical dementia rating (CDR), and auditory verbal learning test (AVLT). No patient has missing data for these clinical variables so they are used in the same way as MRI features in our model. Furthermore, we put all the features through a feature screening module using the approach by (Fan and Lv 2008). Note that feature screening is only applied to the training set not the entire dataset to avoid overfitting.

5) Application of IMTL

Within each sub-cohort, the samples are divided into five folds. We combine four folds from each sub-cohort into a training set and use the remaining data as the test set. We apply

IMTL to the training set and then use the trained model to predict on the test set. We exhaust all four-fold combinations in training, which produces a 5-fold cross validation procedure for evaluating the accuracy of IMTL. This process is repeated for 50 times. For comparison, two competing methods are applied on the same data: SA and AADM. Table 3.3 summarizes the results. IMTL has significantly higher AUC and sensitivity than both competing methods in both diagnosis and prognosis. Notably, competing methods have low AUC and sensitivity in prognosis. This is greatly improved by IMTL. Prognosis is more challenging than diagnosis because the former has a heavily imbalanced dataset (46 converters vs. 168 non-converters). Clearly, IMTL is more robust to sample imbalance. All models achieve similar levels of specificity. Finally, we show the contribution of each imaging feature to diagnosis and prognosis by plotting the percentage of times a feature is included in the IMTL model. The result is shown in Figure 3.3. Hippocampal volume from MRI and the first PC of six SUVRs from amyloid-PET are almost always included in both diagnostic and prognostic models. This is consistent with findings in the literatures that hippocampal atrophy and amyloid-PET SUVRs provide most important biomarkers for AD (Fleisher et al. 2011; Devanand et al. 2007). Other features that are selected for over 50% of the time include HCI, sROI and precuneus metabolism from FDG-PET for diagnosis; and ventricle volume from MRI and HCI and sROI from FDG-PET for prognosis.

Table 3.3 Diagnostic and prognostic performance: ave (std) and p value for hypothesis testing that IMTL is better than a competing method

	DIAGNOSIS			PROGNOSIS		
	IMTL	SM	AADM	IMTL	SM	AADM
AUC	**0.93**(0.03)	0.86(0.06)	0.90(0.04)	**0.85**(0.05)	0.72(0.09)	0.78(0.07)
		p<0.001	p<0.001		p<0.001	p<0.001
SENSITIVITY	**0.91**(0.06)	0.84(0.09)	0.88(0.06)	**0.96**(0.09)	0.58(0.18)	0.76(0.17)
		p<0.001	p=0.03		p<0.001	p<0.001
SPECIFICITY	0.87(0.06)	0.82(0.06)	0.86(0.05)	0.85(0.05)	0.86(0.05)	0.86(0.05)
		p<0.001	p=0.27		p=0.78	p=0.34

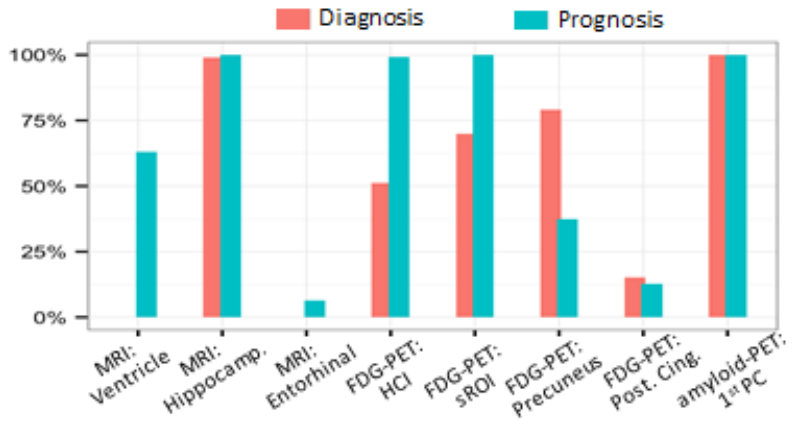


Figure 3.3. Percentage of times imaging features are included in IMTL over 5-fold cross-validation and 50 repeated experiments.

3.6 Conclusion

In this paper, we proposed IMTL to build predictive and classification models for IMD data. We developed an EM algorithm for parameter estimation of IMTL and further

extended it to achieve between-institutional collaborative model estimation without the need for data pooling. We demonstrated that IMTL was capable of out-of-sample prediction and proved that it has a larger Fisher information than models without transfer learning under mild conditions. This explained the outperformance of IMTL from a theoretical standpoint. Simulation experiments demonstrated high accuracy in using IMTL for out-of-sample prediction and classification. IMTL was applied for AD early diagnosis and prognosis, i.e., at the MCI stage, using incomplete multimodality imaging data. Significantly higher AUC and sensitivity were achieved in both diagnosis and prognosis compared with competing methods. Image features selected to include in the models were widely-reported in the literature to be related to AD. Future research may include extension to non-linear models and categorical features as well as application of IMTL to other application domains.

CHAPTER 4

A HOTSPOT SEARCH ALGORITHM FOR FUSION OF 3-D RADIATION DOSE MAP AND PATIENT CHARACTERISTICS WITH APPLICATION IN CARDIAC TOXICITY ASSESSMENT OF LUNG CANCER

4.1 Introduction

Lung cancer is a fatal disease which caused 150,000 deaths in year 2018 alone. Low survival rate is a main characteristic of lung cancer. In the U.S., around 15% of patients with lung cancer can survive for five years after the diagnosis, and this number is much lower in the developing world (Majumdar, 2009). One of the most common treatment for lung cancer is the radiation therapy (RT). However, the side effect of the RT is that the radiation dose may also cause injury on normal tissues while killing cancer cells. As location of lungs is close to the heart, radiation damage on the heart has been known as "cardiac toxicity" (Belliere et al., 2013; Hardy et al., 2010; Cella et al., 2014). Cardiac toxicity can cause heart damage that may overrule the benefit of RT in treating the lung cancer, leading to shortening of the overall survival (OS). In fact, recent clinical trials on using radiation dose escalation protocols to treat lung cancer found unexpected results of decrease in OS, despite the fact that dose escalation has been found to be an effective approach in treating other cancers (Zelefsky et al., 2008). This suggests that lung cancer patients receiving dose escalation may have suffered from early onset cardiac toxicity because of the tumor location, which may reduce the OS (Lancellotti et al., 2013; Bradley et al., 2015; Chun et al., 2017).

There is an urgent need to understand how radiation dose on the heart affects the OS of lung cancer patients, which will help optimize RT treatment planning to minimize damage to the heart. Not much work has been done so far, because this is an emerging area that has just recently caught the attention of medicine. Among the few existing studies, the commonly adopted approach is to collapse the 3-D dose map of the heart into a 1-D dose-volume histogram (DVH), and then build univariate or multivariate models to correlate features extracted from the histogram with OS. Here, a 3-D dose map contains the dose level (in unit of Gy) on each voxel of a 3-D image (e.g., CT or MRI) of the heart. A 1-D histogram is constructed by counting the frequency of voxels of the heart (y axis) that have a certain dose level (x axis). The conversion from a 3-D dose map to a 1-D histogram has an obvious limitation that spatial information of the dose distribution is lost.

This paper has a different perspective from the existing histogram-based studies. We would like to avoid the conversion and the associated loss of spatial information, and use the 3-D dose map directly. In particular, our objective is to find if there is any sub-region of the heart, referred to as a hot spot, which is a significant biomarker for cardiac toxicity.

The contribution of this study is multifold:

- Methodological development: We propose a novel mathematical formulation of the 3-D spatial hot spot search problem, by integration generalized linear model (GLM) and a proposed definition of the hot spot, which includes both a location/sub-region of the heart and a dose threshold. This definition facilitates inclusion of the found hot spot into the existing

treatment planning software to minimize cardiac toxicity. Furthermore, under the proposed mathematical formulation, we develop a novel search algorithm for the hot spot, in which search of the dose threshold is theoretically proven to be equivalent to the search of this threshold on the whole heart. Under the found optimal dose threshold, search of the hot spot location can be further accomplished by scanning over the units of the heart using a unit-wise GLM.

- Application impacts: We apply the proposed algorithm to a dataset containing patients with Non-Small Cell Lung Cancer (NSCLC) who received Intensity Modulated Radiation Therapy (IMRT) at Mayo Clinic Arizona. The hot spot sub-region found by our algorithm harbors the sinoatrial node of the electronic conduction system of the heart. Damage of the sinoatrial node by radiation toxicity disrupts the crucial function of the heart, leading to shortening of OS. Our finding suggests that protective strategies may be developed to spare the susceptible hot spot, and thus helping RT planning achieve the optimal result in treating lung cancer patients.

The remainder of this paper is structured as follows: Section 2 provides the literature review and points out limitations of the existing research, which motivates the research in this paper. Section 3 presents the mathematical formulation of the hot spot search problem. Section 4 presents the algorithm development for hot spot search. Section 5 presents a case study based on real data. Section 6 is conclusion.

4.2 Literature Review

Cardiac toxicity has been studied for lung cancer. As locations of the lungs are close to the heart, radiation therapy intended to treat the cancer inevitably spills radiation dose to the heart. Because radiation dose maps of the heart are 3-D, a challenging issue is what features to extract from a 3-D map to correlate with the response variable (e.g., OS, development of heart disease, etc). To tackle this challenge, most existing research converts the 3-D map into a DVH, which plots the fractional volume of the heart that receives a certain dose level (Fig. 4.1). Then, features are extracted from the DVH. Specifically, denote the DVH by $f(x)$, where x is the dose variable ranging from 0 to the maximum dose, max_x, and f is the fractional volume of the heart that receives dose level x. A typical feature extracted from $f(x)$ is the fractional volume of the heart that receives dose greater than a threshold D, denoted by V_D (Fig. 4.1). By definition, $V_D = P(x > D)$. The rationale behind such a feature is that there is a tipping- point dose level D^*, which if exceeded, may lead to heart tissue damage that is hard to repair (Hendry, 2015). However, because D^* is unknown, different values of D may be used, resulting in a set of V_D's, which are linked with the response variable by a statistical model. Depending on the type of the response variable, classification models (for categorical response), predictive models (for continuous response), or cox models (for censored data) may be used. Next, we provide more detailed review for the existing research on cardiac toxicity. Note that because the area of cardiac toxicity of lung cancer is fairly new, there are not many papers yet and most papers were published in the past two years.

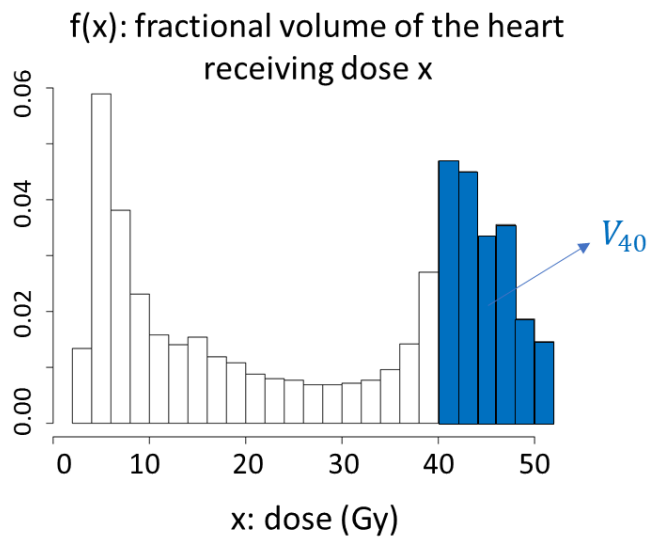


Figure. 4.1: Example of a DVH and definition of the fractional volumetric variable V_D

Speirs et al. (2017) investigated cardiac toxicity of 251 patients with locally advanced NSCLC. Fractional volumetric variables V_5 to V_{75} by 5Gy apart were extracted from the DVH of the heart and then binarized using cutoffs. Using a univariate Cox model, the paper found that OS would be worsened if the binarized factional volumetric variables from V_5 to V_{60} were increased. Moreover, a multivariate Cox model was also fit, which included V_{50}, heart volume, mean lung dose, bilateral information, and the left lower lobe tumor location as predictors. V_{50} was found to be associated with OS. Wang et al. (2017) studied cardiac toxicity of 112 patients with stage III NSCLC. Symptomatic cardiac events and time to death or last follow-up were used as clinical endpoints. Fractional volumetric variables V_5 , V_{30} and mean dose were extracted from the DVH of the heart, all of which were found to be significant in univariate survival analysis. The paper concluded that cardiac events were associated with radiation doses to the heart. Stam et al. (2017)

investigated cardiac toxicity of 803 patients with stage I-II NSCLC, who were treated with Stereotactic Body Radiation Therapy (SBRT). A multivariate Cox model was used to identify doses to the heart structure that may be associated with non-cancer death. The paper found that maximum dose to the left atrium and the dose to 90% of the superior vena cava were significant associated with non-cancer death. Johnson et al. (2017) proposed a match-pair analysis between NSCLC IIIB patients receiving standard dose RT (43 patients) and those receiving high dose RT (43 patients). The paper performed univariate and multivariate analysis using V_5 and V_{30} of the heart, and found that V_{30} was associated with OS.

However, the existing studies share the same limitation that they built models using features extracted from DVH of the heart. As DVH collapses the 3-D dose map into a 1-D histogram, spatial distribution of radiation dose on the heart is lost. Because different parts of the heart have different functions, there is regional heterogeneity within the heart in terms of susceptibility to radiation damage. Therefore, a new model is needed that can search a sub-region of the heart – referred to as the "hot spot" in this paper – whose fractional volumetric variables are significant predictors to inform cardiac toxicity. This has not been studies in the existing literature and is the goal of this paper.

4.3 Mathematical Formulation for the Hot Spot Search Problem

We first propose the definition of a hot spot as follows.

Definition 1 (hot spot): Let r be a sub-region of the heart. Let $V_D^{(r)}$ be the fractional volumetric variable extracted from the DVH of region r. That is, $V_D^{(r)}$ is the fractional volume of region r that receives dose greater than D. A hot spot is a minimum region r^*

for which there exists a D^* such that $V_{D^*}^{(r^*)}$ is positively associated with cardiac toxicity, i.e., the larger that the $V_{D^*}^{(r^*)}$, the higher the toxicity.

According to the definition, the concept of a hot spot includes both a sub-region of the heart, r^*, and a dose threshold D^*. Therefore, we use to (r^*, D^*) to represent a hot spot in this paper. Definition 1 also reveals the challenge for a hot spot search algorithm, i.e., the algorithm will need to search not only a sub-region r^* of the heart but also the dose threshold D^*. Furthermore, we would like to clarify a few important points regarding Definition 1:

(1) A hot spot is not a sub-region receiving high radiation dose. Instead, it is a sub-region that is *sensitive* to radiation dose, i.e., the larger the fractional volume of this sub-region that receives dose greater than D^*, the more severe of the cardiac toxicity. A sub-region receiving high radiation dose for a particular patient may not be a hot spot if that sub-region is insensitive to dose, i.e., how much dose is received and how the dose is distributed spatially within that sub-region will have little impact on toxicity.

(2) A hot spot is not a sub-region whose *average* dose is positively associated with cardiac toxicity. Instead, it is a sub-region whose volumetric feature $V_{D^*}^{(r^*)}$ is associated with cardiac toxicity. The latter is biologically justified as being the factional volume of a sub-region receiving radiation dose greater that a tipping point D^* that may lead to tissue damage beyond repair. Also, $V_{D^*}^{(r^*)}$, once found, can be incorporated into the current treatment planning optimization software which uses volumetric features as a constraint when developing the optimal RT plan for a patient.

(3) A hot spot is a *minimum* sub-region for which the stated condition in Definition 1 holds. It is possible for a superset sub-region of the hot spot sub-region to also satisfy the condition, but the superset is not the hot spot we are looking for because it includes subset sub-regions other than the hot spot for which the condition does not hold. Finding the minimum sub-region has a clear benefit for treatment planning, i.e., the smaller a sub-region needs to be protected from radiation damage, the easier for the treatment planning software to accommodate this constraint without compromising dose delivery to the tumor.

(4) In order to operationalize Definition 1 to guide the search of the hot spot, one will need to define the clinical endpoint the measures cardiac toxicity. There are multiple ways to measure cardiac toxicity. Direct measures include physiological changes of the heart after being exposed to radiation, such as cardiac ischemia, pleural, pericardial effusions, and heart failure (Beukema et al., 2015). However, unless through specially designed clinical trials to assess cardiac toxicity, these direct measures are not collected in routine clinical practice for treating cancer patients. Indirect measures are more commonly used, with the most common one being the OS. Most existing studies on cardiac toxicity as reviewed in Section 2 used OS. In this paper, our proposed method can be applied to any direct or indirect measure of cardiac toxicity, which is generally called *the response variable*.

Next, we introduce the mathematical formulation for the hot spot search problem. Let y_i, M_i, and $\mathbf{Z}_i$ be the response variable, 3-D dose map of the heart, and covariates for patient i, respectively. y_i can be categorical/binary (e.g., development of heart diseases, death by a pre-specified number of years), continuous (e.g., severity of the heart diseases),

or censored (e.g., OS). Given r and D, $V_{D,i}^{(r)}$ that can be computed from M_i. $\mathbf{Z}_i$ contains demographic variables and other clinical variables that may potentially impact the response variable beyond M_i, such as cancer stage, location of tumor, and concurrent therapy. We adopt a generalized linear model (GLM) to link y_i with predictors $\mathbf{Z}_i$ and $V_{D,i}^{(r)}$, in order to accommodate all the aforementioned types of the response variable y_i. In a GLM, y_i has a distribution from exponential family that takes the following form:

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{4.1}$$

where ϕ and θ_i are called the dispersion parameter and the canonical parameter, respectively; $a(\cdot), b(\cdot)$ and $c(\cdot)$ are known functions. To link y_i with predictors $\mathbf{Z}_i$ and $V_{D,i}^{(r)}$, a link function $g(\cdot)$ is used, i.e.,

$$g(\mu_i) = \boldsymbol{\beta}^T\mathbf{Z}_i + \gamma V_{D,i}^{(r)}, \quad \text{or equivalently } \mu_i = g^{-1}(\boldsymbol{\beta}^T\mathbf{Z}_i + \gamma V_{D,i}^{(r)}), \tag{4.2}$$

where μ_i is the mean of y_i. It is known that θ_i in (4.1) can be written as a function of μ_i, i.e., $\theta_i = \theta(\mu_i)$. Inserting the μ_i in (2) into $\theta(\mu_i)$, we can get

$$\theta_i = \theta\left(g^{-1}(\boldsymbol{\beta}^T\mathbf{Z}_i + \gamma V_{D,i}^{(r)})\right). \tag{4.3}$$

Inserting (4.3) into (4.1), we can write the log-likelihood function for the unknown parameters of the model as

$$l(\boldsymbol{\beta}, \gamma, r, D) = \sum_{i=1}^{n} \frac{y_i\theta\left(g^{-1}(\boldsymbol{\beta}^T\mathbf{Z}_i + \gamma V_{D,i}^{(r)})\right) - b\left(\theta\left(g^{-1}(\boldsymbol{\beta}^T\mathbf{Z}_i + \gamma V_{D,i}^{(r)})\right)\right)}{a(\phi)} + c(y_i, \phi). \tag{4.4}$$

We can use maximum likelihood estimation (MLE) to estimate the unknown parameters in (4.4) by solving the following optimization:

$$(\boldsymbol{\beta}^*, \gamma^*, r^*, D^*) = argmax\ l(\boldsymbol{\beta}, \gamma, r, D). \tag{4.5}$$

(4.5) is not straightforward to solve because it includes two additional unknown parameters, r and D, beyond the regular MLE for a GLM.

4.4 Algorithm Development for Hot Spot Search

4.4.1 Finding D^* in the hot spot search

The optimization problem in (4.5) would be easier to solve if either r^* or D^* were known ahead of time. Theorem 1 reveals a nice property of our problem formulation, which implies that D^* can be found from the DVH of the *whole* heart. Then, the found D^* can be "plugged" into (4.5) to solve for $\boldsymbol{\beta}^*, \gamma^*$ and r^*.

Proposition 1: Let (r^*, D^*) be a hot spot. Let $Cov\left(g(\mu), V_{D^*}^{(r^*)} | \mathbf{Z}\right)$ be the partial covariance between the linear predictor of the GLM, $g(\mu)$, and $V_{D^*}^{(r^*)}$, given the patient covariates $\mathbf{Z}$. $Cov\left(g(\mu), V_{D^*}^{(r^*)} | \mathbf{Z}\right) > 0$ according to Definition 1. Then, we have

$$Cov\left(g(\mu), V_{D^*}^{(H)} | \mathbf{Z}\right) > 0,$$

where $V_{D^*}^{(H)}$ denotes the fractional volumetric variable extracted from the DVH of the whole heart.

Please see the proof of Theorem 1 in the Appendix. Driven by Theorem 1, we can find the optimal D^* of the optimization problem in (4.5) by solving (4.5) on the whole heart (i.e., consider r to be the whole heart), i.e.,

$$\left(\boldsymbol{\beta}^{(H)}, \gamma^{(H)}, D^*\right) = argmax \quad l(\boldsymbol{\beta}, \gamma, D | r = H). \tag{4.6}$$

To solve (4.6), we can alternate between two sub-optimizations: Given D, (6) becomes the MLE for a regular GLM; given $(\boldsymbol{\beta}, \gamma)$, the optimization with respect to D can be solved by a Quasi-Newton method (Byrd et al., 1995; Zhu et al., 1997). The two sub-optimizations

will be solved iteratively until convergence. The convergence property is guaranteed by Proposition 2, which shows that the likelihood function in (4.6) keeps increasing through the iterative steps between the two sub-optimizations. The proof of Proposition 2 is skipped. We summarize the steps in solving (4.6) in Algorithm 1.

Proposition 2: Let $l(\boldsymbol{\beta}_j, \gamma_j, D_j | r = H)$ be the likelihood function in (6) at the j-th iteration of the optimization algorithm. Let $l(\boldsymbol{\beta}_{j+1}, \gamma_{j+1}, D_j | r = H)$ be the likelihood function under fixed D_j while $\boldsymbol{\beta}_{j+1}$ and γ_{j+1} are obtained by MLE. Let $l(\boldsymbol{\beta}_{j+1}, \gamma_{j+1}, D_{j+1} | r = H)$ be the likelihood function under fixed while $\boldsymbol{\beta}_{j+1}$ and γ_{j+1} while D_{j+1} is obtained by the Quasi-Newton method. Then we have

$$l(\boldsymbol{\beta}_j, \gamma_j, D_j | r = H) \leq l(\boldsymbol{\beta}_{j+1}, \gamma_{j+1}, D_j | r = H) \leq l(\boldsymbol{\beta}_{j+1}, \gamma_{j+1}, D_{j+1} | r = H).$$

Algorithm 1 (Find D^* in the hot spot search)

Input: $y_i(response)$, $M_i(3 - D\ dose\ map\ of\ the\ heart)$, and $\mathbf{Z}_i(covariates)$ for a set of patients $i = 1, \dots, n$.

Initialization: $(\boldsymbol{\beta}, \gamma, D) = (\boldsymbol{\beta}_0, \gamma_0, D_0)$

Iterate between 1) and 2) till convergence; at the j -th iteration, do the following:

1) Given $D = D_j$, use MLE to solve the optimization in (6) and get estimation $(\boldsymbol{\beta}_{j+1}, \gamma_{j+1})$;
2) Given $(\boldsymbol{\beta}, \gamma) = (\boldsymbol{\beta}_{j+1}, \gamma_{j+1})$, use the Quasi-Newton method to solve the optimization in (4.6) and get D_{j+1}.

Output: D^* at convergence ($\boldsymbol{\beta}^{(H)}$, $\gamma^{(H)}$ can also be obtained but they are side products that will not be used in the subsequent estimation.)

4.4.2 Finding r^* in the hot spot search

Under the D^* found by Algorithm 1, we will need to further find the hot spot sub-region r^* the heart. Suppose the 3-D heart H is composed by N indexed units, i.e., $H = \{s_k\}_{k=1,\dots,N}$. Theoretically speaking, a unit of the heart can be as small as a voxel on the 3-D dose map. In practice, a unit should include a sufficient number of voxels to allow for construction of a DVH for the unit. On the other hand, a too large unit is also not appropriate because if the size of the hot spot is even smaller than the unit, we will not be able to find the shape of the hot spot to a good precision. Determining an appropriate unit is study-dependent.

A hot spot r^* is composed by a subset of units of the heart. In order to identify the units included in the hot spot, we can compute $Cov\left(g(\mu), V_{D^*}^{(s_k)} | \mathbf{Z}\right)$ for each unit of the heart, $k = 1, \dots, N$, and identify the units with $Cov\left(g(\mu), V_{D^*}^{(s_k)} | \mathbf{Z}\right) > 0$. Denote the collection of these units by Ω, i.e., $\Omega = \left\{s_k | Cov\left(g(\mu), V_{D^*}^{(s_k)} | \mathbf{Z}\right) > 0, k = 1, \dots, N\right\}$. By definition, we know that Ω is the hot spot r^*.

An empirical approach for finding units with $Cov\left(g(\mu), V_{D^*}^{(s_k)} | \mathbf{Z}\right) > 0$ is to fit a GLM with $V_{D^*}^{(s_k)}$ and $\mathbf{Z}$ as predictors. If the coefficient for $V_{D^*}^{(s_k)}$ is significantly greater than zero, it is evidence that $Cov\left(g(\mu), V_{D^*}^{(s_k)} | \mathbf{Z}\right) > 0$. We summarize steps in finding the hot spot sub-region in Algorithm 2.

Algorithm 2 (Find r^* in the hot spot search)

Input: $y_i(response)$, $M_i(3-D\ dose\ map\ of\ the\ heart)$, and $\mathbf{Z}_i(covariates)$ for a set of patients $i = 1, \dots, n$; D^* found by Algorithm 1.

Initialization: divide the 3-D heart into N indexed units $s_k, k = 1, \dots, N$; let Ω to be the empty set.

For each unit s_k, do the following:

1) Fit a GLM with $V_{D^*}^{(s_k)}$ and $\mathbf{Z}$ as predictors
2) If the coefficient for $V_{D^*}^{(s_k)}$ is significantly greater than zero, add s_k to Ω.

Output: Ω as the hot spot r^*.

4.5 Case Study

Our dataset contains 134 patients with stage III NSCLC who were treated at Mayo Clinic Arizona. All the patients were treated with Intensity Modulated Radiation Therapy (IMRT). The 3-D dose map of the heart for each patient was retrieved from treatment planning software. We consider the 3-D heart to be composed by 64 units, i.e., 4 equally spaced intervals in each of the x, y, and z directions. We focus on finding the hot spot significantly associated with OS, i.e., the GLM included in our hot spot search algorithm is a Cox model. Three covariates were included in the Cox model, including age, stage of the lung cancer, and use of chemotherapy, according to the AIC criterion. Average follow-up time of these patients is 1.5 years with 53/81 patients being alive/deceased at the time of the follow-up.

To find the hot spot, we first use Algorithm 1 on the whole heart and find the optimal dose threshold D^*. D^* is found to be $54.88\ Gy$. Then, under this D^*, we use

Algorithm 2 and find the hot spot sub-region, r^*, which consists of 19 units that are connected and located at the right-superior of the heart. Fig. IV.2 shows locations of the 19 units on the heart that are highlighted in red. Different shades of red are used to show the significant level of each unit in association with OS in the Cox model.

Furthermore, we fit a Cox model with $V_{D^*}^{(r^*)}$ and the three aforementioned covariates, with $D^* = 54.88\,Gy$ and r^* being the hot spot that includes the 19 units highlighted in Fig. 4.2. The coefficient of $V_{D^*}^{(r^*)}$ in this Cox model is 4.33 with p value = 0.0005. The positiveness of the coefficient means that the more fractional volume of the hot spot sub-region with dose greater than $54.88\,Gy$, the higher the risk of death. The small p value means that this relationship holds with high confidence. Specifically, we show the survival curves of two patients with the same covariates (both are 68 years old, underwent chemotherapy, and had stage III-b lung cancer) but different values for $V_{D^*}^{(r^*)}$, i.e., $V_{D^*}^{(r^*)} = 0$ and 0.39 for patient A and B, respectively. It is clear from Fig. 3 that patient B has lower survival probability (higher risk of death) than A.

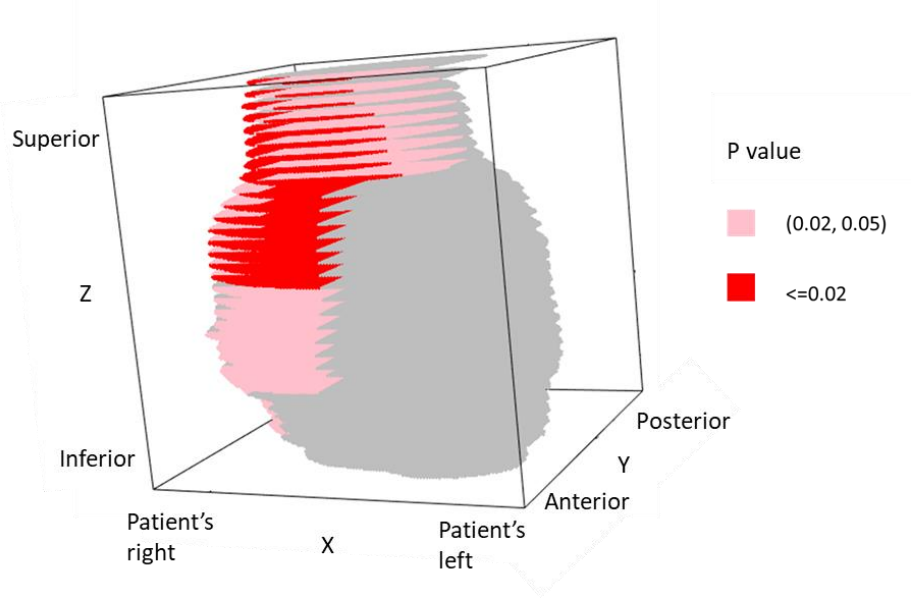


Figure 4.2: Hot spot sub-region found by the proposed search algorithm overlaid on CT volume of the heart (the hot spot sub-region consists of 19 units highlighted in red of two different shades used to represent different p-value ranges in the Cox model coefficients corresponding to the units).

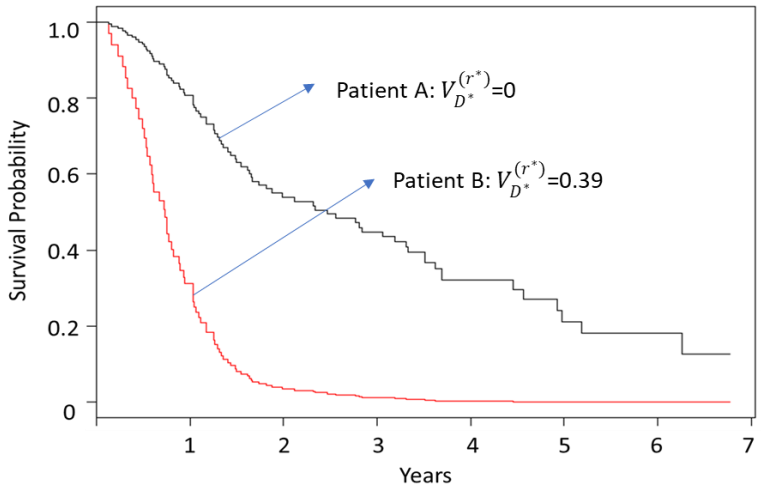


Figure 4.3: Survival curves of two patients with the same covariates but different $V_{D^*}^{(r^*)}$

Interpretation of the Hot Spot: As shown in Fig. 4.2, location of the hot spot is at the right-superior of the heart. This sub-region harbors a crucial structure, known as the sinoatrial (SA) node of the electronic conduction system of the heart (James, 1961). The SA node generates electronic signals and stimulates the heart to contract so that the heart can pump normally. From our findings, we suspect that radiation dose may cause injury on the SA node and therefore disrupt the normal function of the electronic conduction system of the heart, which further shorten OS. The potential risk of radiation dose on the electronic conduction system has been reported in the literature (Darby et al., 2013; Jaworski et al., 2013).

Comparison Experiments: For comparison purposes, we apply two commonly adopted approaches for studying cardiac toxicity to the same dataset as the proposed algorithm. One approach is to fit a Cox model with covariates and V_D extracted from the whole heart (Johnson, 2017). We try different values for D and find the most significant one happens at $D = 55Gy$ with p value =0.04. This p value is far less significant than that corresponds to the hot spot found by our algorithm (p=0.0005). Another commonly used approach is to segment the heart into four anatomically-defined compartments, i.e., left ventricle, left atrium, right ventricle and right atrium (Kang et al., 2012). Then, a Cox model is fit for each compartment using V_D extracted from that compartment and covariates. We try different values for D and find the most significant one happens at $D = 55Gy$ with p value =0.024 for compartment right ventricle. This p value is more significant than the one obtained from the whole heart analysis, but still far less significant than that corresponds

to the hot spot found by our algorithm. These results using the existing approaches demonstrate the utility of the proposed hot spot search algorithm.

4.6 Conclusion

In this paper, we developed the mathematical formulation and associated algorithms for searching the hot spot of the heart susceptible to cardiac toxicity and subsequently shortening OS for patients with lung cancer under RT. Our finding of applying the algorithms to real data revealed that the hot is located at the right-superior of the heart, which harbors the SA node that is crucial to the heart function. This explained the mechanism as to how radiation damages the heart. The found location and extent of the hot spot together with the dose threshold may be included in the treatment planning software as constraints to help optimize the treatment plan to maximally kill cancer cells while minimizing cardiac toxicity. Future work may include applying the hot spot search algorithms to study direct measures of cardiac toxicity and integration of the hot spot search with treatment optimization.

CHAPTER 5

CONCLUSIONS

In healthcare, rich datasets have been available from multiple heterogeneous sources or modalities. There is great need for heterogeneous data fusion to leverage the information from different modalities to make better diagnosis or prognosis in healthcare. Although it is a promising method, there are several challenges in healthcare application. In Chapter 2, we proposed a bio-statistical hybrid model for integration of 3-D dose map and patient characteristics in prediction of rectal toxicity of prostate cancer patients who receives IMRT. In Chapter 3, we proposed an IMTL model to deal with missing modality problem. And in Chapter 4, we proposed a hotspot search algorithm for the fusion of 3-D radiation dosemap and patient characteristics to predict the overall survival of the patients. The proposed algorithm divided the 3-D heart dose map into several regions, determined the dosimetric/volumetric features in each sub region and finally identified the hotspots of the heart which is associated with overall survival given the patient characteristics. My dissertation provided the first-of-its-kind rigorous statistical models to multi-modality heterogeneous data fusion, which worked well in the specific application domain. Future work may include the feature selection method in the novel heterogeneous data fusion methods.

REFERENCES

Akimoto, T., et al. (2004). Rectal bleeding after hypofractionated radiotherapy for prostate cancer: correlation between clinical and dosimetric parameters and the incidence of grade 2 or worse rectal bleeding. *International Journal of Radiation Oncology• Biology• Physics*, **60**(4), 1033-1039.

American Cancer Society (2018), Cancer Facts & Figures 2018.

Albert, Marilyn S., Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, et al. (2011) The Diagnosis of Mild Cognitive Impairment due to Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's and Dementia* **7**(3).

Bailly, Matthieu, Christophe Destrieux, Caroline Hommet, Karl Mondon, Jean-Philippe Cottier, Emilie Beaufils, Emilie Vierron, et al. (2015) Precuneus and Cingulate Cortex Atrophy and Hypometabolism in Patients with Alzheimer's Disease and Mild Cognitive Impairment: MRI and 18F-FDG PET Quantitative Analysis Using FreeSurfer, Precuneus and Cingulate Cortex Atrophy and Hypometabolism in P." *BioMed Research International, BioMed Research International* 2015.

Basir, Otman, and Xiaohong Yuan. (2007) Engine Fault Diagnosis Based on Multi-Sensor Information Fusion Using Dempster-Shafer Evidence Theory. *Information Fusion* 8 (4): 379–86. doi:10.1016/j.inffus.2005.07.003.

Beaulieu-Jones, Brett K, Jason H Moore, and THE POOLED RESOURCE OPEN-ACCESS ALS CLINICAL TRIALS CONSORTIUM. (2017) Missing Data Imputation in the Electronic Health Record Using Deeply Learned Autoencoders. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 22: 207–18. doi:10.1142/9789813207813_0021.

Belliere, A., Girard, N., Chapet, O., Khodri, M., Kubas, A., Souquet, P. J., & Mornex, F. (2009). Feasibility of high-dose three-dimensional radiation therapy in the treatment of localised non-small-cell lung cancer. *Cancer/Radiothérapie*, **13**(4), 298-304.

Bentzen, S. M., et al. (2010). Quantitative analyses of normal tissue effects in the clinic (QUANTEC): an introduction to the scientific issues. *International Journal of Radiation Oncology• Biology• Physics*, **76**(3), S3-S9.

Beukema, J. C., van Luijk, P., Widder, J., Langendijk, J. A., & Muijs, C. T. (2015). Is cardiac toxicity a relevant issue in the radiation treatment of esophageal cancer?. *Radiotherapy and Oncology*, **114**(1), 85-90.

Boomsma, M. J., et al. (2012). A prospective cohort study on radiation-induced hypothyroidism: development of an NTCP model. *International Journal of Radiation Oncology• Biology• Physics*, **84**(3), e351-e356.

Bradley, J.D., Paulus, R., Komaki, R., Masters, G., Blumenschein, G., Schild, S., Bogart, J., Hu, C., Forster, K., Magliocco, A. and Kavadi, V., (2015). Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial phase 3 study. *The Lancet Oncology*, **16**(2), pp.187-199.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**(5), 1190-1208.

Catalona, William J., et al. (2000). Comparison of percent free PSA, PSA density, and age-specific PSA cutoffs for prostate cancer detection and staging. *Urology*, **56**(2), 255-260.

Catalona, William J., Deborah S. Smith, and David K. Ornstein. (1997). Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng/mL and benign prostate examination: enhancement of specificity with free PSA measurements. *Jama*, **277**(18), 1452-1455.

Cella, Laura, et al. (2013). Multivariate normal tissue complication probability modeling of gastrointestinal toxicity after external beam radiotherapy for localized prostate cancer. *Radiation Oncology*, **8**(1), 221.

Cella, L., Palma, G., Deasy, J.O., Oh, J.H., Liuzzi, R., D'Avino, V., Conson, M., Pugliese, N., Picardi, M., Salvatore, M. and Pacelli, R. (2014). Complication probability models for radiation-induced heart valvular dysfunction: do heart-lung interactions play a role?. *PloS one*, **9**(10), e111753.

Chen, Kewei, Napatkamon Ayutyanont, Jessica B.S. Langbaum, Adam S. Fleisher, Cole Reschke, Wendy Lee, Xiaofen Liu, et al. (2011). Characterizing Alzheimer's Disease Using a Hypometabolic Convergence Index. *NeuroImage* **56** (1).

Chen, Kewei, Jessica B.S. Langbaum, Adam S. Fleisher, Napatkamon Ayutyanont, Cole Reschke, Wendy Lee, Xiaofen Liu, et al. (2010) Twelve-Month Metabolic Declines in Probable Alzheimer's Disease and Amnestic Mild Cognitive Impairment Assessed Using an Empirically Pre-Defined Statistical Region-of-Interest: Findings from the Alzheimer's Disease Neuroimaging Initiative. *NeuroImage* **51** (2).

Chun, S.G., Hu, C., Choy, H., Komaki, R.U., Timmerman, R.D., Schild, S.E., Bogart, J.A., Dobelbower, M.C., Bosch, W., Galvin, J.M. and Kavadi, V.S., 2017. Impact of intensity-modulated radiation therapy technique for locally advanced non-small-cell lung cancer: a

secondary analysis of the NRG Oncology RTOG 0617 randomized clinical trial. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, **35**(1), pp.56-62.

Clark, Christopher M., Julie A. Schneider, Barry J. Bedell, Thomas G. Beach, Warren B. Bilker, Mark A. Mintun, Michael J. Pontecorvo, et al. (2011) Use of Florbetapir-PET for Imaging β-Amyloid Pathology. *JAMA* 305 (3). American Medical Association: 275. doi:10.1001/jama.2010.2008.

Cordeiro, Gauss M., and Peter McCullagh. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(3), 629-643.

Darby, S.C., Ewertz, M., McGale, P., Bennet, A.M., Blom-Goldman, U., Brønnum, D., Correa, C., Cutter, D., Gagliardi, G., Gigante, B. and Jensen, M.B. (2013). Risk of ischemic heart disease in women after radiotherapy for breast cancer. *New England Journal of Medicine*, **368**(11), pp.987-998.

Deasy, J. O. (2000). Comments on the use of the Lyman-Kutcher-Burman model to describe tissue response to nonuniform irradiation. *Int J Radiat Oncol Biol Phys*, **47**(5):1458–1459.

Devanand, D. P., G. Pradhaban, X. Liu, A. Khandji, S. De Santi, S. Segal, H. Rusinek, et al. (2007) Hippocampal and Entorhinal Atrophy in Mild Cognitive Impairment: Prediction of Alzheimer Disease. *Neurology* **68**(11): 828–36. doi:10.1212/01.wnl.0000256697.20968.d7.

Deville, Curtiland, et al. (2012). Acute gastrointestinal and genitourinary toxicity of image-guided intensity modulated radiation therapy for prostate cancer using a daily water-filled endorectal balloon. *Radiation Oncology*, **7**(1), 76.

Fan, Jianqing, and Jinchi Lv. (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **70**(5): 849–911. doi:10.1111/j.1467-9868.2008.00674.x.

Fleisher, Adam S., Kewei Chen, Xiaofen Liu, Auttawut Roontiva, Pradeep Thiyyagura, Napatkamon Ayutyanont, Abhinay D. Joshi, et al. (2011) Using Positron Emission Tomography and Florbetapir F 18 to Image Cortical Amyloid in Patients with Mild Cognitive Impairment or Dementia due to Alzheimer Disease. *Archives of Neurology* **68** (11): 1404–11. doi:10.1001/archneurol.2011.150.

Garcia-Salicetti, Sonia, Charles Beumier, Gérard Chollet, Bernadette DorizziJean, Leroux les Jardins, Lunter Jan, Yang Ni, and Dijana Petrovska-Delacrétaz. (2003) *Audio- and Video-Based Biometric Person Authentication*.

Gulliford, Sarah L., et al. (2012). Parameters for the Lyman Kutcher Burman (LKB) model of normal tissue complication probability (NTCP) for specific rectal complications observed in clinical practise. *Radiotherapy and Oncology*, **102**(3), 347-351.

Hardy, D., Liu, C. C., Cormier, J. N., Xia, R., & Du, X. L. (2010). Cardiac toxicity in association with chemotherapy and radiation therapy in a large cohort of older patients with non-small-cell lung cancer. *Annals of oncology*, **21**(9), 1825-1833.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, Boca Raton, FL.
Huang, Jian, Patrick Breheny, and Shuangge Ma. (2012). A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, **27**(4) ， 481-499.

He, Dakuo, Zhengsong Wang, Le Yang, and Wanwan Dai. (2017) Study on Missing Data Imputation and Modeling for the Leaching Process." *Chemical Engineering Research and Design* 124. Institution of Chemical Engineers: 1–19.

Hendry, J. H. (2015). Threshold doses and circulatory disease risks. *Annals of the ICRP*, **44**(1_suppl), 69-75.

Hurvich, Clifford M., and Chih-Ling Tsai. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**(2), 297-307.

Källman, P., Ågren, A., & Brahme, A. (1992). Tumour and normal tissue responses to fractionated non-uniform dose delivery. *International journal of radiation biology*, **62**(2), 249-262.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, **9**(2), 137-163.

Jack, Clifford R., Val J. Lowe, Stephen D. Weigand, Heather J. Wiste, Matthew L. Senjem, David S. Knopman, Maria M. Shiung, et al. 2009. "Serial PIB and MRI in Normal, Mild Cognitive Impairment and Alzheimers Disease: Implications for Sequence of Pathological Events in Alzheimers Disease." *Brain* 132 (5): 1355–65. doi:10.1093/brain/awp062.

Jack, Clifford R., David S. Knopman, Stephen D. Weigand, Heather J. Wiste, Prashanthi Vemuri, Val Lowe, Kejal Kantarci, et al. (2012) An Operational Approach to National Institute on Aging-Alzheimer's Association Criteria for Preclinical Alzheimer Disease. *Annals of Neurology* **71** (6). Wiley Subscription Services, Inc., A Wiley Company: 765–775. doi:10.1002/ana.22628.

James, T. N. (1961). Anatomy of the human sinus node. *The Anatomical Record*, **141**(2), 109-139.

Jaworski, C., Mariani, J. A., Wheeler, G., & Kaye, D. M. (2013). Cardiac complications of thoracic irradiation. *Journal of the American College of Cardiology*, **61**(23), 2319-2328.

Johnson, M. D., Sura, K., Mangona, V. S., Glick, A., Wallace, M., Ye, H., & Grills, I. S. (2017). Matched-Pair Analysis of High Dose Versus Standard Dose Definitive Chemoradiation for Locally Advanced Non–Small-Cell Lung Cancer. *Clinical lung cancer*, **18**(2), 149-155.

Kang, D., Woo, J., Kuo, C. J., Slomka, P. J., Dey, D., & Germano, G. (2012). Heart chambers and whole heart segmentation techniques. *Journal of Electronic Imaging*, **21**(1), 010901.

Lancellotti, P., Nkomo, V.T., Badano, L.P., Bergler-Klein, J., Bogaert, J., Davin, L., Cosyns, B., Coucke, P., Dulgheru, R., Edvardsen, T. and Gaemperli, O. (2013). Expert consensus for multi-modality imaging evaluation of cardiovascular complications of radiotherapy in adults: a report from the European Association of Cardiovascular Imaging and the American Society of Echocardiography. *European Heart Journal–Cardiovascular Imaging*, **14**(8), pp.721-740.

Li, X. Allen, et al. (2012). The use and QA of biologically related models for treatment planning: Short report of the TG-166 of the therapy physics committee of the AAPM. *Medical physics*, **39**(3), 1386-1409.

Li, Rongjian, Wenlu Zhang, Heung-il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. 2014. “Deep Learning Based Imaging Data Completion for Improved Brain Disease Diagnosis,” 305–12.

Little, Roderick J a, and Donald B Rubin. 2002. *Statistical Analysis with Missing Data, Second Edition*. *Statistical Analysis with Missing Data, Second Edition*. John Wiley & Sons. doi:10.2307/1533221.

Liu, Mitchell, et al. (2010). Normal tissue complication probability (NTCP) modeling of late rectal bleeding following external beam radiotherapy for prostate cancer: a test of the QUANTEC-recommended NTCP model. *Acta Oncologica*, **49**(7), 1040-1044.

Liu, Mingxia, Jun Zhang, Pew Thian Yap, and Dinggang Shen. 2017. “View-Aligned Hypergraph Learning for Alzheimer’s Disease Diagnosis with Incomplete Multi-Modality Data.” *Medical Image Analysis* 36. Elsevier B.V.: 123–34. doi:10.1016/j.media.2016.11.002.

Lowe, V. J., B. J. Kemp, C. R. Jack, M. Senjem, S. Weigand, M. Shiung, G. Smith, et al. 2009. “Comparison of 18F-FDG and PiB PET in Cognitive Impairment.” *Journal of Nuclear Medicine* 50 (6): 878–86. doi:10.2967/jnumed.108.058529.

Lyman, J. T. (1985). Complication probability as assessed from dose-volume histograms. *Radiation Research*, **104**(2), 13-19.

Majumdar, S. (2009) *Stem Cells and Cancer*, Springer, New York, NY.

Malek, Karim. (2015). Contribution to the clinical and laboratory study of tissue reactions observed radiosensitivity after radiotherapy of prostate cancer: potential radioprotective effect of statins. Doctoral Thesis, University Joseph Fourier – Grenoble.

Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica: Journal of the Econometric Society*, **45**(8), 1977-1988.

Michalski, J. M., Gay, H., Jackson, A., Tucker, S. L. and Deasy, J. O. (2010). Radiation dose–volume effects in radiation-induced rectal injury. *International Journal of Radiation Oncology• Biology• Physics*, **76**(3), S123-S129.

Ordóñez Galán, Celestino, Fernando Sánchez Lasheras, Francisco Javier de Cos Juez, and Antonio Bernardo Sánchez. 2017. “Missing Data Imputation of Questionnaires by Means of Genetic Algorithms with Different Fitness Functions.” *Journal of Computational and Applied Mathematics* 311. Elsevier B.V.: 704–17. doi:10.1016/j.cam.2016.08.012.

Pan, Sinno Jialin, and Qiang Yang. 2010. “A Survey on Transfer Learning.” *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–59. doi:10.1109/TKDE.2009.191.

Partin, A. W., Catalona, W. J., Southwick, P. C., Subong, E. N., Gasior, G. H. and Chan, D. W. (1996). Analysis of percent free prostate-specific antigen (PSA) for prostate cancer detection: influence of total PSA, prostate volume, and age. *Urology*, **48**(6), 55-61.

Skwarchuk, M. W., et al. (2000). Late rectal toxicity after conformal radiotherapy of prostate cancer (I): multivariate analysis and dose–response. *International Journal of Radiation Oncology• Biology• Physics*, **47**(1), 103-113.

Speirs, C.K., DeWees, T.A., Rehman, S., Molotievschi, A., Velez, M.A., Mullen, D., Fergus, S., Trovo, M., Bradley, J.D. and Robinson, C.G. (2017). Heart dose is an independent dosimetric predictor of overall survival in locally advanced non–small cell lung cancer. *Journal of Thoracic Oncology*, **12**(2), pp.293-301.

Sole, Angelo Del, Francesca Clerici, Arturo Chiti, Michela Lecchi, Claudio Mariani, Laura Maggiore, Lisa Mosconi, and Giovanni Lucignani. 2008. “Individual Cerebral Metabolic Deficits in Alzheimer’s Disease and Amnestic Mild Cognitive Impairment: An FDG PET Study.” *European Journal of Nuclear Medicine and Molecular Imaging* 35 (7): 1357–66. doi:10.1007/s00259-008-0773-6.

Stam, B., Peulen, H., Guckenberger, M., Mantel, F., Hope, A., Werner-Wasik, M., Belderbos, J., Grills, I., O'Connell, N. and Sonke, J.J. (2017). Dose to heart substructures is associated with non-cancer death after SBRT in stage I–II NSCLC patients. *Radiotherapy and Oncology*, **123**(3), pp.370-375.

Thompson, Paul M., Kiralee M. Hayashi, Greig I. De Zubicaray, Andrew L. Janke, Stephen E. Rose, James Semple, Michael S. Hong, et al. 2004. “Mapping Hippocampal and Ventricular Change in Alzheimer Disease.” *NeuroImage* 22 (4): 1754–66. doi:10.1016/j.neuroimage.2004.03.040.

Thung, Kim Han, Chong Yaw Wee, Pew Thian Yap, and Dinggang Shen. 2014. “Neurodegenerative Disease Diagnosis Using Incomplete Multi-Modality Data via Matrix Shrinkage and Completion.” *NeuroImage* 91. Elsevier Inc.: 386–400. doi:10.1016/j.neuroimage.2014.01.033.

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

Tucker, S. L., et al. (2008). Analysis of radiation pneumonitis risk using a generalized Lyman model. *International Journal of Radiation Oncology• Biology• Physics*, **72**(2), 568-574.

Tucker, Susan L., et al. (2012). Use of fractional dose–volume histograms to model risk of acute rectal toxicity among patients treated on RTOG 94-06. *Radiotherapy and Oncology*, **104**(1), 109-113.

Vavassori V, Fiorino C, Rancati T, et al. (2007). Predictors for rectal and intestinal acute toxicities during prostate cancer high-dose 3D-CRT: results of a prospective multicenter study. *International Journal of Radiation Oncology• Biology• Physics*, **67**(5), 1401-1410.

Wang, K., Eblan, M.J., Deal, A.M., Lipner, M., Zagar, T.M., Wang, Y., Mavroidis, P., Lee, C.B., Jensen, B.C., Rosenman, J.G. and Socinski, M.A. (2017). Cardiac Toxicity After Radiotherapy for Stage III Non-Small-Cell Lung Cancer: Pooled Analysis of Dose-Escalation Trials Delivering 70 to 90 Gy. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, **35**(13), pp.1387-1394.

Wedlake L, Silia F, Benton B, et al. (2012). Evaluating the efficacy of statins and ACE-inhibitors in reducing gastrointestinal toxicity in patients receiving radiotherapy for pelvic malignancies. *European Journal of Cancer*, **48**(14), 2117-2124.

Xia, Yingjie, Xiumei Li, and Zhenyu Shan. 2013. "Parallelized Fusion on Multisensor Transportation Data: A Case Study in CyberITS." *International Journal of Intelligent Systems* 28 (6): 540–64.

Xiang, Shuo, Lei Yuan, Wei Fan, Yalin Wang, Paul M. Thompson, and Jieping Ye. 2014. "Bi-Level Multi-Source Learning for Heterogeneous Block-Wise Missing Data." *NeuroImage* 102 (P1). Elsevier Inc.: 192–206. doi:10.1016/j.neuroimage.2013.08.015.

Yuan, Lei, Yalin Wang, Paul M. Thompson, Vaibhav A. Narayan, and Jieping Ye. 2012. "Multi-Source Feature Learning for Joint Analysis of Incomplete Multiple Heterogeneous Neuroimaging Data." *NeuroImage* 61 (3). Elsevier Inc.: 622–32. doi:10.1016/j.neuroimage.2012.03.059.

Zelefsky, M.J., Yamada, Y., Fuks, Z., Zhang, Z., Hunt, M., Cahlon, O., Park, J. and Shippy, A. (2008). Long-term results of conformal radiotherapy for prostate cancer: impact of dose escalation on biochemical tumor control and distant metastases-free survival outcomes. *International Journal of Radiation Oncology* Biology* Physics*, **71**(4), pp.1028-1033.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, **23**(4), 550-560.

APPENDIX A

SUPPLEMENTAL MATERIALS FOR CHAPTER 2

1. Proof of Proposition 1:

Let $a = 1/n$ and $D = \max_i D_i$. By definition, $D_i \leq$ D for $\forall i$. Therefore, $(\sum_i v_i D_i^a)^{1/a} \leq (\sum_i v_i D^a)^{1/a} = D$, i.e.,

$$\limsup_{a\to\infty} (\sum_i v_i D_i^a)^{1/a} \leq D. \tag{A-1}$$

Furthermore, let $S_\delta = \{i | D \leq D_i + \delta\}$. Then, $(\sum_i v_i D_i^a)^{1/a} \geq (\sum_{i\in S_\delta} v_i)^{1/a}(D - \delta)$. By letting $\delta \to 0$ and $a \to \infty$, we can get

$$\liminf_{a\to\infty} (\sum_i v_i D_i^a)^{1/a} \geq D. \tag{A-2}$$

Combining (A-1) and (A-2), we get $\lim_{a\to\infty} (\sum_i v_i D_i^a)^{1/a} = D$.

Δ

2. Proof of Proposition 2:

If $\alpha_{p+1}^* \geq 0$, the constraint in the optimization problem in (9) is automatically satisfied by the optimal solution to the unconstrained optimization problem. This means that the optimal solution to the unconstrained problem is just that to the constrained problem, i.e., $(\hat{\alpha}_0, \dots, \hat{\alpha}_{p+1}) = (\alpha_0^*, \dots, \alpha_{p+1}^*)$.

If $\alpha_{p+1}^* < 0$, let $\boldsymbol{\alpha}^* = (\alpha_0^*, \cdots, \alpha_{p+1}^*)$ and $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \cdots, \hat{\alpha}_{p+1})$ be the optimal solutions to the unconstrained and constrained optimization problems, respectively. We need to prove that $(\hat{\alpha}_0, \dots, \hat{\alpha}_p) = (\alpha_0^*, \dots, \alpha_p^*)$ and $\hat{\alpha}_{p+1} = 0$. To prove this, we start by constructing a feasible solution to the constrained optimization in (9), i.e.,

$$\tilde{\boldsymbol{\alpha}} = \beta\hat{\boldsymbol{\alpha}} + (1 - \beta)\boldsymbol{\alpha}^*, \tag{A-3}$$

where $\beta = -\frac{\alpha_{p+1}^*}{\hat{\alpha}_{p+1} - \alpha_{p+1}^*}$. Here, β is valid because we have known that $\alpha_{p+1}^* < 0$ and $\hat{\alpha}_{p+1} \geq 0$ such that the denominator of β must not be zero. Also, $\tilde{\boldsymbol{\alpha}}$ is a feasible solution to (9) because

$$\tilde{\alpha}_{p+1} = \beta\hat{\alpha}_{p+1} + (1-\beta)\alpha_{p+1}^* = 0, \tag{A-4}$$

which satisfies the constraint in (9). Then, the following inequality holds:

$$g(\tilde{\boldsymbol{\alpha}}) \geq g(\hat{\boldsymbol{\alpha}}) \geq g(\boldsymbol{\alpha}^*). \tag{A-5}$$

The first "≥" in (A-5) holds because $\tilde{\boldsymbol{\alpha}}$ is a feasible solution and $\hat{\boldsymbol{\alpha}}$ is the optimal solution. The second "≥" holds because $\hat{\boldsymbol{\alpha}}$ is the optimal solution to the constrained optimization while $\boldsymbol{\alpha}^*$ is that to the unconstrained optimization.

Furthermore, let $g(\alpha_0, \cdots, \alpha_{p+1}) = -l(\alpha_0, \cdots, \alpha_{p+1}) + \lambda \sum_{k=1}^{p} |\alpha_k|$, which is the objective function of (9). $g(\cdot)$ a convex function. Therefore, we can get

$$\begin{aligned} g(\tilde{\boldsymbol{\alpha}}) &= g(\beta\hat{\boldsymbol{\alpha}} + (1-\beta)\boldsymbol{\alpha}^*) \\ &\leq \beta g(\hat{\boldsymbol{\alpha}}) + (1-\beta) g(\boldsymbol{\alpha}^*) \\ &\leq \beta g(\hat{\boldsymbol{\alpha}}) + (1-\beta) g(\hat{\boldsymbol{\alpha}}) = g(\hat{\boldsymbol{\alpha}}). \end{aligned} \tag{A-6}$$

The fact that $g(\tilde{\boldsymbol{\alpha}}) \geq g(\hat{\boldsymbol{\alpha}})$ in (A-5) and $g(\tilde{\boldsymbol{\alpha}}) \leq g(\hat{\boldsymbol{\alpha}})$ in (A-6) leads to $g(\tilde{\boldsymbol{\alpha}}) = g(\hat{\boldsymbol{\alpha}})$. This means that $\tilde{\boldsymbol{\alpha}}$ is the optimal solution to (9), i.e., $\hat{\boldsymbol{\alpha}} = \tilde{\boldsymbol{\alpha}}$. Then, $\hat{\alpha}_{p+1} = \tilde{\alpha}_{p+1} = 0$ according to (A-4). Furthermore, using (A-3), we can get $(\hat{\alpha}_0, \ldots, \hat{\alpha}_p) = (\alpha_0^*, \ldots, \alpha_p^*)$.

Δ

3. Proof of Proposition 5:

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator for the linear coefficients $\boldsymbol{\beta}$ in a generalized linear model (GLM). It can be derived that under a finite sample size, the bias of $\hat{\boldsymbol{\beta}}$ is

$$bias(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\xi}. \quad \text{(A-7)}$$

Here, $\mathbf{X^T W X} = -E[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}]$, where $l(\boldsymbol{\beta})$ is the log-likelihood function. The j-th element of $\boldsymbol{\xi}$ is $\xi_j = -\frac{1}{2}\frac{\mu_j''}{\mu_j'}\mathbf{Q}_{jj}$, where $\mu_j' = \partial\mu_j/\partial\eta_j$, $\mu_j'' = \partial^2\mu_j/\partial\eta_j^2$, μ_j is the mean of the distribution from the exponential family the GLM corresponds to, η_j is the linear predictor of the GLM, and $\mathbf{Q}_{jj}$ is the j-th diagonal element of matrix $\mathbf{Q} = \mathbf{X}(\mathbf{X^T W X})^{-1}\mathbf{X^T}$. Next, we use the result in (A-7) to derive the bias of our model, i.e., $bias(\tilde{\boldsymbol{\alpha}})$.

Specifically, the log-likelihood function of our model, $l_{n^*}^w(\boldsymbol{\alpha})$, is given in (12). The first derivative of $l_{n^*}^w(\boldsymbol{\alpha})$ is

$$\frac{\partial l_{n^*}^w(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{j=1}^N \left\{\frac{w_1 y_j}{\phi(\eta_j)} - \frac{w_0(1-y_j)}{1-\phi(\eta_j)}\right\}\varphi(\eta_j)\mathbf{x}_j =$$

$$\sum_{j=1}^N \left\{\frac{[w_1 y_j - w_0\phi(\eta_j)] - (w_1 - w_0)y_j\phi(\eta_j)}{\phi(\eta_j)(1-\phi(\eta_j))}\right\}\varphi(\eta_j)\mathbf{x}_j,$$

where $\mathbf{x}_j$ is the j-th column of $\mathbf{X}$. Furthermore, we can get

$$\mathbf{X^T W X} = -E\left[\frac{\partial^2 l_{n^*}^w(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\right] = \sum_{j=1}^N \left\{\frac{w_0\phi(\eta_j) + w_1\left(1-\phi(\eta_j)\right)}{\phi(\eta_j)\left(1-\phi(\eta_j)\right)}\right\}\varphi^2(\eta_j)\mathbf{x}_j\mathbf{x}_j^T.$$

(A-8)

It is clear from (A-8) that $\mathbf{W}$ is a diagonal matrix with the j-th element being

$$\mathbf{W}_{jj} = \frac{w_0\phi(\eta_j) + w_1\left(1-\phi(\eta_j)\right)}{\phi(\eta_j)\left(1-\phi(\eta_j)\right)}\varphi^2(\eta_j). \quad \text{(A-9)}$$

Next, to derive ξ_i, we need to derive μ_i' and μ_i'' for our model. Specifically, $\mu_i = \phi(\eta_i)^{w_1}$. Therefore, the first derivative of μ_i is

$$\mu_i' = w_1\phi(\eta_i)^{w_1 - 1}\varphi(\eta_i),$$

and the second derivative of μ_i is

$$\mu_i'' = -w_1\phi(\eta_i)^{w_1-1}\varphi(\eta_i)\eta_i + w_1(w_1-1)\phi(\eta_i)^{w_1-2}\varphi^2(\eta_i).$$

Therefore,

$$\xi_j = -\frac{1}{2}\frac{\mu_j''}{\mu_j'}\mathbf{Q}_{jj} = \frac{\eta_j\phi(\eta_j)-(w_1-1)\varphi(\eta_j)}{2\phi(\eta_j)}\mathbf{Q}_{jj}. \quad \text{(A-10)}$$

Finally, by inserting (A-9) and (A-10) into (A-7), we can obtain the $bias(\breve{\boldsymbol{\alpha}})$ in (13).

Δ

APPENDIX B

SUPPLEMENTAL MATERIALS FOR CHAPTER 3

Proof of Theorem 1

Under IMTL, the complete-data log-likelihood function is:

$$llh_{IMTL} = L_1 + L_2, \tag{B-1}$$

where

$$L_1 = -\frac{n_1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n_1}\Big(x_i^{(21)} - \mathbf{x}_i^{(11)}\mathbf{A}_2 - b_2, y_i^{(1)} - \mathbf{x}_i^{(11)}\boldsymbol{\beta}_1 - \beta_0\Big)\boldsymbol{\Sigma}^{-1}\begin{pmatrix} x_i^{(21)} - \mathbf{x}_i^{(11)}\mathbf{A}_2 - b_2 \\ y_i^{(1)} - \mathbf{x}_i^{(11)}\boldsymbol{\beta}_1 - \beta_0 \end{pmatrix},$$

$$L_2 = -\frac{n_2}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n_2}\Big(x_i^{(22)} - \mathbf{x}_i^{(12)}\mathbf{A}_2 - b_2, y_i^{(2)} - \mathbf{x}_i^{(12)}\boldsymbol{\beta}_1 - \beta_0\Big)\boldsymbol{\Sigma}^{-1}\begin{pmatrix} x_i^{(22)} - \mathbf{x}_i^{(12)}\mathbf{A}_2 - b_2 \\ y_i^{(2)} - \mathbf{x}_i^{(12)}\boldsymbol{\beta}_1 - \beta_0 \end{pmatrix}.$$

Let $D_{obs} = \left\{\mathbf{x}_i^{(11)}, \mathbf{x}_i^{(12)}, x_i^{(22)}, y_i^{(1)}, y_i^{(2)} \mid i = 1, \dots, n\right\}$ and $D_{mis} = \left\{x_i^{(21)} | i = 1, \dots, n\right\}$ denote the observed and missing data in (B-1). The observed Fisher information for each element θ_{ij} in $\boldsymbol{\Omega}$ is:

$$\zeta_{IMTL}\left(\theta_{ij}\middle|D_{obs}\right) = -\frac{\partial^2 llh_{IMTL}}{\partial\theta_{ij}^2} - E_{D_{mis}}\left\{\left(\frac{\partial\, llh_{IMTL}}{\partial\theta_{ij}}\right)^2\middle| D_{obs}\right\} + E_{D_{mis}}\left(\frac{\partial\, llh_{IMTL}}{\partial\theta_{ij}}\middle| D_{obs}\right)^2.$$

Through some algebra, we can derive $\zeta_{IMTL}\left(\theta_{ij}\middle|D_{obs}\right)$ as

$$\zeta(\theta_{22}|D_{obs})_{MMTL} = \frac{1}{2}n_2\sigma_{22}^2 + (2\tilde{n}_1 - n_1)\left(\sigma_{2y}^2/\sigma_{yy}\right)\sigma_{22} + 2(n_1 - \tilde{n}_1)\left(\sigma_{2y}^2/\sigma_{yy}\right)^2 - \frac{1}{2}n_1\sigma_{22}^2, \tag{B-2}$$

$$\zeta\left(\theta_{2y}\middle|D_{obs}\right)_{MMTL} = \frac{1}{2}n_2\left(\sigma_{22}\sigma_{yy} + \sigma_{2y}^2\right) + \frac{1}{2}(n_1 - \tilde{n}_1)\sigma_{22}\sigma_{yy} + \frac{1}{2}(n_1 + \tilde{n}_1)\sigma_{2y}^2, \tag{B-3}$$

$$\zeta\left(\theta_{yy}\middle|D_{obs}\right)_{MMTL} = \frac{1}{2}(n_1 + n_2)\sigma_{yy}^2, \tag{B-4}$$

where

$$\tilde{n}_1 = \frac{1}{\sigma_{yy}} \sum_{i=1}^{n_1} \left(y_i^{(1)} - \mathbf{x}_i^{(11)} \boldsymbol{\beta}_1 - \beta_0 \right)^2 .$$

Under the two-modality IMD, SM and AADM has the same log-likelihood function: $llh_{SM} = llh_{AADM} = L_2$. We can derive their observed Fisher information (only SM is shown here):

$$\zeta(\theta_{22}|D_{obs})_{SM} = \frac{1}{2} n_2 \sigma_{22}^2, \tag{B-5}$$

$$\zeta\left(\theta_{2y}\middle|D_{obs}\right)_{SM} = \frac{1}{2} n_2 \left(\sigma_{22}\sigma_{yy} + \sigma_{2y}^2\right), \tag{B-6}$$

$$\zeta\left(\theta_{yy}\middle|D_{obs}\right)_{SM} = \frac{1}{2} n_2 \sigma_{yy}^2 . \tag{B-7}$$

Furthermore, taking the expectation of (B-2) through (B-7) with respect to D_{obs}, we can get the Fisher information, $I_{IMTL}(\theta_{ij})$ and $I_{SM}(\theta_{ij})$. Subtracting the respective Fisher information of the two models, we have

$$I_{IMTL}(\theta_{22}) - I_{SM}(\theta_{22}) = (n_1 - 2p_1)\left(\sigma_{2y}^2/\sigma_{yy}\right)\sigma_{22} + 2p_1\left(\sigma_{2y}^2/\sigma_{yy}\right)^2 - \frac{1}{2} n_1 \sigma_{22}^2, \tag{B-8}$$

$$I_{IMTL}(\theta_{2y}) - I_{SM}(\theta_{2y}) = n_1 \sigma_{2y}^2 + \frac{1}{2} p_1 |\boldsymbol{\Sigma}|, \tag{B-9}$$

$$I_{IMTL}(\theta_{2y}) - I_{SM}(\theta_{2y}) = \frac{1}{2} n_1 \sigma_{yy}^2 . \tag{B-10}$$

(B-8) and (B-9) are positive. Therefore, $I_{IMTL}(\theta_{ij}) > I_{SM}(\theta_{ij})$ if (B-8) is positive, i.e.,

$$\frac{-n_1 + 2p_1 + \sqrt{(n_1 - 2p_1)^2 + n_1 p_1}}{4p_1} < \frac{\sigma_{2y}^2}{\sigma_{22}\sigma_{yy}} . \qquad \blacksquare$$

APPENDIX C

SUPPLEMENTAL MATERIALS FOR CHAPTER 4

Proof of Proposition 1

Let H denote the whole heart. With r^* denoting the sub-region of H corresponding to the hot spot, let r_c^* denote the rest of H excluding r^*. Furthermore, let $x^{(H)}$, $x^{(r^*)}$, and $x^{(r_c^*)}$ denote the dose variable defined for H, r^*, and r_c^*. Let $f\left(x^{(H)}\right), f\left(x^{(r^*)}\right) and\ f(x^{(r_c^*)})$ denote the probability distributions of these variables. Empirical distributions of these variables can be obtained by the DVHs computed from the 3-D dose map of the whole heart, localized to r^*, and localized to r_c^*, respectively. According to the definitions of these variables, we know that $f\left(x^{(H)}\right)$ is a mixed distribution of , $f\left(x^{(r^*)}\right)$ and $f(x^{(r_c^*)})$ in the form of

$$f\left(x^{(H)}\right) = \pi f\left(x^{(r^*)}\right) + (1-\pi) f(x^{(r_c^*)}),$$

which leads to

$$P\left(x^{(H)} > D^*\right) = \pi P\left(x^{(r^*)} > D^*\right) + (1-\pi) P(x^{(r_c^*)} > D^*)$$

i.e.,

$$V_{D^*}^{(H)} = \pi V_{D^*}^{(r^*)} + (1-\pi) V_{D^*}^{(r_c^*)}.$$

Taking the partial covariance between $g(\mu)$ and $V_{D^*}^{(H)}$, we can get

$$Cov\left(g(\mu), V_{D^*}^{(H)} | \mathbf{Z}\ \right) = \pi Cov\left(g(\mu), V_{D^*}^{(r^*)} | \mathbf{Z}\ \right) + (1-\pi) Cov\left(g(\mu), V_{D^*}^{(r_c^*)} | \mathbf{Z}\ \right)$$

$Cov\left(g(\mu), V_{D^*}^{(r^*)} | \mathbf{Z}\ \right) > 0$ by the definition of a hot spot (i.e., Definition 1). $Cov\left(g(\mu), V_{D^*}^{(r_c^*)} | \mathbf{Z}\ \right) \geq 0$. Therefore, we have $Cov\left(g(\mu), V_{D^*}^{(H)} | \mathbf{Z}\ \right) > 0$. Δ