Finding Integrative Biomarkers from Biomedical Datasets: An application to Clinical and Genomic Data

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sanjoy Dey

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Adviser: Vipin Kumar
Co-Adviser: Michael Steinbach

August, 2015

Acknowledgements

I would like to express my deepest appreciation to my adviser Prof. Vipin Kumar for his invaluable support and guidance throughout my graduate study. His enthusiasm for high-impact research and drive for excellence conveyed a deep impact on me. He gradually and convincingly instilled into me the spirit of performing great research not only in data mining domain, but also applying the data mining techniques in many real world problems in biomedical domain. Moreover, he inspired and motivated me with great energy and vision every time I faced any difficulties and lost my way. One of the most important things I learnt from him is how to have enthusiasm and patience in together to achieve a greater goal.

I am especially grateful to my co-adviser Dr. Michael Steinbach, who helped me in solving many research problems, no matter how trivial they were. He also helped me a lot in improving my writing skill. I would also like to thank Prof. Rui Kuang, Prof. Jaideep Srivastava, and Prof. Bonnie Westra for being a part of my committee and providing me valuable feedback and suggestions at various levels of my PhD.

My sincere gratitude to my collaborators: Prof. Bonnie Westra, Dr. Gyorgy Simon, Prof. Maneesh Bhargava, Porf. Chris Wendt, Prof. Michael Wilson, Prof. Kelvin O. Lim and Prof. Angus McDonald from various departments of biomedical domain, from whom I learnt many domain challenges and the importance of having biological impact while developing data mining techniques. I am especially indebtful to Prof. Bonnie Westra for her constant support and guidance in many projects in healthcare domain, without which half of my dissertation would not have been possible.

Life as a grad student would not have been the same without my great lab-mates. My special thanks to Gowtham Atluri, Pranjul Yadav, Gaurav Pandey, Gang Fang, Vanja Paunic, Jeremy Weed, Katherine Hauwiller, Jakob Johnson, Andrew Hangsleben, Rohit

Gupta, Wen Wang, Sean Landsman, Varun Mithal, Ashish Garg and Deepthi Cheboli. It was a great experience, being able to interact and share lab space with them.

I would also thank my mentors at Robert Bosch LLC: Dr. Sundararajan Srinivasan, Dr. Jo-Anna Ting and Dr. Juergen Heit for providing me the opportunity of internships twice. These internships not only helped me to gather industry experience, but also gave me an exposure to many software tools and technologies used in data mining domain. I also thank the funding agencies such as NSF and Grad school for supporting me throughout graduate school. I am also grateful to Minnesota Supercomputing Institute (MSI) and Pravakar Roy for providing the technical supports for some of our projects.

I am also indebted to the great support received from my incredible friends, who helped me make Minneapolis my second home. Especially thanks to Dr. Biplob Debnath, Dr. Mohammad Yunus, Dr. Amitava Karmaker, Sauptik Dhar, Shafayat Jamil, Sohini Roy Chowdhury, Ayan Paul, Subhrajit Roychowdhury, Puja Das, Brian Swanson and Mother Taruni Devi Dasi. I couldnt́ have finished my graduate study without their inspiration.

Last, but not the least I would like to thank my parents, my wife and my sister for their constant support in my life. Although words will not suffice their support, I am proud to have them in my life and I will indebtful to them forever.

Dedication

To my Grandparents: Gopendra Kumar Dey, Baidehi R. Purkayastha, Parul Bala Dey and Lila Ghosh.

Abstract

Human diseases, such as cancer, diabetes and schizophrenia, are inherently complex and governed by the interplay of various underlying factors ranging from genetic and genomic influences to environmental effects. Recent advancements in high throughput data collection technologies in bioinformatics have resulted in a dramatic increase in diverse data sets that can provide information about such factors related to diseases. These types of data include DNA microarrays providing cellular information, Single Nucleotide Polymorphisms (SNPs) providing genetic information, metabolomics data in terms of proteins and other metabolites, structural and functional brain data from magnetic resonance imaging (MRI), and electronic health records (EHRs) containing copious information about histo-pathological factors, demographic, and environmental effects. Despite their richness, each of these datasets only provides information about a part of the complex biological mechanism behind human diseases. Thus, effective integration of the partial information of any of these genomic and clinical data can help reveal disease complexities in greater detail by generating new data-driven hypotheses beyond the traditional hypotheses about biomarkers. In particular, integrative biomarkers, i.e., patterns of features that are predictive of disease and that go beyond the simple biomarkers derived from a single dataset, can lead to a customized and more effective approach to improving healthcare.

This thesis focuses on addressing the key issues related to integrative biomarkers by developing new data mining approaches. One very important issue of biomarker discovery is that the models have to easily interpretable, i.e., integrative models have to be not only predictive of the disease, but also interpretable enough so that domain experts can infer useful knowledge from the obtained patterns. In one such effort to make models interpretable, domain information about disease relationships was used as prior knowledge during model development. In addition, a novel metric called I-score was proposed using medical literature to quantify the interpretability of the obtained patterns.

Another key issue of integrative biomarker discovery is that there may be many potential relationships present among diverse datasets. For example, a very important

types of relationship in biomarker discovery is interaction, which are those biomarkers spanning multiple datasets, whose combined features are more indicative of disease than the individual constituent factors. In particular, the individual effects of each type of factor on disease predisposition can be small and thus, remain undetected by most disease association techniques performed on individual datasets. Different types of relationships are explored and an association analysis based framework is proposed to discover them. The proposed framework is especially effective for discovering higher-order relationships, which cannot be found by the existing prominent integrative approaches for the biomarker discovery. When applied on real datasets collected from three different types of data from schizophrenic and normal subjects, this approach yielded significant integrated biomarkers which are biologically relevant.

Disease heterogeneity creates further issues for integrative biomarker discovery, biomarkers obtained from clinicogenomic studies may not be applicable to all patients in the same degree, i.e., a disease consist of multiple subtypes, each occurring in different subpopulations. Some potential reasons responsible for disease heterogeneity are different pathways playing different roles in the same disease and confounding factors such as age, ethnicity and race, or genetic predisposition, which can be available in rich EHR data. Most biomarker discovery techniques use full space model development techniques, i.e., they assess the performance of biomarkers on all patients without finding the distinct subpopulations. In this thesis, more customized models were built depending on patientś characteristics to handle disease heterogeneity.

In summary, several data mining techniques developed in this thesis advance the state-of-the art in integration of diverse biomedical datasets. Moreover, their applications on large-scale EHR yield significant discoveries, which can ultimately lead to generating new data-driven hypotheses for inferring meaningful information about complex disease mechanism.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Integrating Diverse Biomedical Datasets

Human diseases, such as cancer, diabetes and schizophrenia, are inherently complex and governed by the interplay of various underlying factors ranging from genetic and genomic influences to environmental effects [2, 3]. For example, genomic data can elucidate the biological underpinning of a disease, while clinical data can capture mostly the environmental and behavioral aspects of a disease. Recent advancements in high throughput data collection technologies in bioinformatics have resulted in a dramatic increase in diverse data sets that can provide information about such factors related to diseases. Indeed, in last few decades, individual genetic data has become available. A myriad of genomic data has been collected to capture different aspects of genetic and genomic information [4]. For example, microarray data can capture the expression level of genes, SNPs can capture mutational changes and other kinds of polymorphism of DNA sequences, mass spectrometry (MS) data from proteomics can capture the post-translational modifications of the genes, and so on. Furthermore, recently collected electronic health records (EHRs) from hospitals contain copious information about histopathological factors, demographics, treatment history and environmental effects.

Since diseases result from the interaction of a number of these factors, each of the clinical and genomic datasets only provides information about a part of the complex biological mechanism behind human diseases. Thus, effective integration of both genomic and clinical data can help reveal disease complexities in greater detail and is essential to

help realize the goal of personalized medicine [5, 6]. For example, integrative patterns containing features from multiple datasets that are predictive of the disease endpoint can lead to a more customized and more effective approach to improving health care. Although such integrative patterns can be directly useful as potential biomarkers for diagnosis, treatment or prevention of diseases, they can also provide insights into the underlying nature of disease or related biomedical processes. Such usefulness of integrative patterns by fusing multiple diverse clinical and genomic datasets has led to an emerging research area of integrative mining of clinical and genomic data [7] (Figure 1.1).

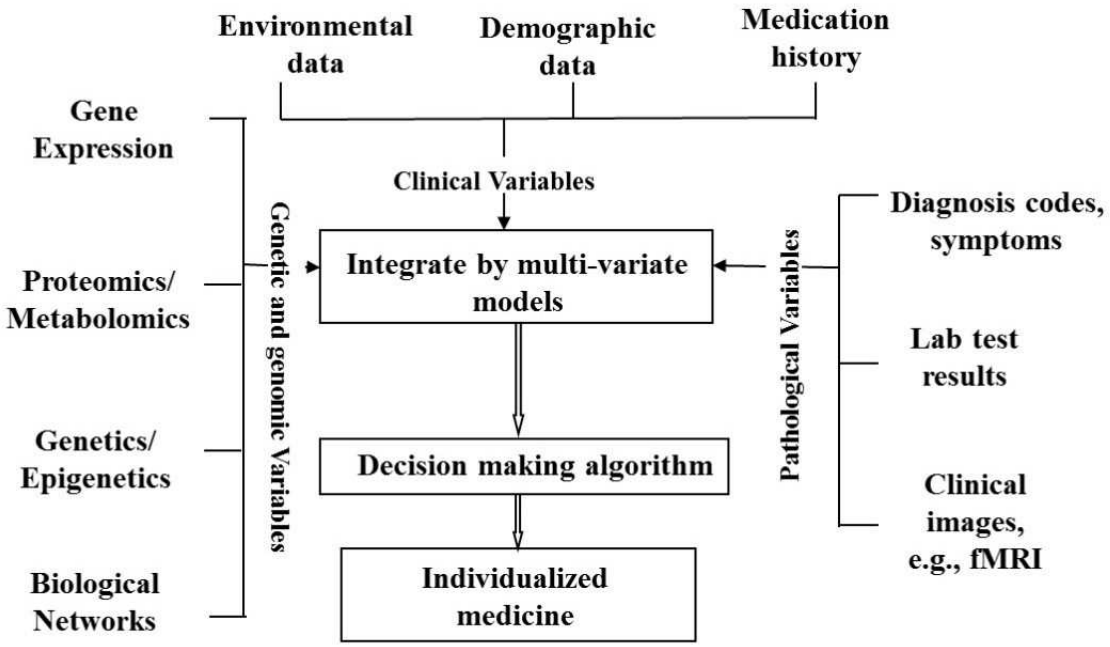


Figure 1.1: The integrative multivariate approach for clinical decision making by combining multiple types of data.

Traditional data mining techniques focus on finding useful information from the data collected from a real-world systems. Most data mining techniques are designed to handle record-based data, where a set of samples are collected from a real-world system for a number of features of interest, which are often represented in records. For

example, patients' data are collected in the form of Electronic Health Records (EHRs) containing lab tests, diagnosis codes and interventions. Several data mining techniques such as predictive modeling, clustering, association analysis and so on can be applied to such record-based data [8]. However, the above mentioned clinical, genetic and genomic datasets are quite diverse, each having their individual properties which data mining techniques have to be cognizant of. Fusing the information available from such datasets has been a challenging problem in the data mining domain and thus provides opportunities for developing different novel data mining techniques.

1.2 Challenges in Data Integration

The integration of multiple datasets poses many challenges for developing data mining techniques. Several key technical challenges are described below.

1. Difference in properties of the datasets: There are significant differences in the properties of clinical and genomic data format, types, dimensionalities, biases and assumptions, and the amount of noise present in the two datasets that need to be dealt with by any integrative methods. First, clinical and genomic variables are mostly categorical or nominal in nature, while genomic data are mostly continuous-valued [9]. Second, genetic and genomic data contain a higher level of missing values because of technological issues [10]. In contrast, clinical data are easy and inexpensive to collect, and so contain fewer missing values. Third, the biases and assumptions of each of the data sets being integrated may be different due to the difference in experimental designs and protocols. In fact, clinical and genomic data have different degrees of reliabilities and usefulness. For example, Clinical variables are gathered more systematically over a longer period of time and they contain less noise. In addition, they are rigorously validated by numerous epidemiology studies [11, 12, 13]. Furthermore, clinical data are cheap and easy to collect. In contrast, genomic data such as gene expression data are less reproducible over independent cohorts because of the high noise, different experimental biases, high-dimensionality and small sample sizes [14, 15, 16, 17]. Therefore, treating both clinical and genomic datasets with equal emphasis often overestimate the performance of genomic variables [11, 12]. Integrating variables

with such different formats, types, structure, dimensionality, and missing values is a challenging problem in the data mining and machine learning domain.

2. Curse of Dimensionality: Another important data related issue in integrative studies is the curse of dimensionality, i.e., the number of samples (n) available in these datasets is much less than the total number of variables (p) coming from multiple datasets. (This is the classical statistical problem of n¡¡p [18]). In that case, the number of samples required to describe the input space defined by high-dimensional datasets increases exponentially, which can lead to overfitting a predictive model, i.e., make a model that conforms closely to limited data and does not generalize [19]. Similarly, biomarker discovery techniques may find potential factors by random chance and need to be cognizant of issues of multiple hypothesis testing [20]. This problem is even more challenging when preforming clinicogenomic integration due to the disparate dimensionalities of clinical and genomic datasets. For example, genomic and genetic datasets have much higher dimensionality in comparison with clinical data which often contains approximately 10-20 variables [15, 21]. Therefore, the clinical variables can be completely lost among the vast number of genomic variables unless the disparity in dimensionality is handled properly [22, 23].

3. Interpretability of the model: Interpretability of the obtained clinicogenomic models is a much desired property, which is critical for their acceptability in clinical practice [24]. However, most integrative models focus on improving the prediction power rather than interpretability. The features selected by the models may provide some interpretation of the models, especially if only a small number of features are selected. In addition, most of the models use components obtained from feature extraction techniques (e.g., PCA [25, 26] or PLS [27]) before developing any integrative models. These components may further reduce the interpretability of the model. If the clinicogenomic models cannot provide interpretable results, the domain expert cannot infer potential knowledge from them so that this knowledge can be used for designing better drug or planning better intervention plans for the disease.

4. Heterogeneity: Most complex diseases are heterogeneous in nature, i.e., patients with a particular disease may form different subgroups and factors appropriate for one subgroup may not apply to another [28, 29]. Some potential reasons responsible for such disease heterogeneity are different pathways playing different roles in the same disease [3] and confounding factors such as age (Simon, Schrom et al. 2013), ethnicity and race [30, 31, 32], or genetic predisposition [33], which can lead to different degree of association between the other clinicogenomic factors and the disease. Finding such heterogeneous subgroups of population is even more important for clinicogenomic integrative studies, because the clinical and genomic features may measure very different aspects of disease phenomenon such as behavioral, environmental and biological factors. Each sample is not likely to be affected by each of these factors to the same extent. For example, genomic markers can affect a subset of samples and clinical markers may affect a different set of samples. Alternatively, genomic markers can affect different subgroups defined by clinical variables in different way. Clinicogenomic integrative studies thus can provide new opportunities to find insights into disease heterogeneity. Most biomarker discovery techniques use full space model development techniques, i.e., the biosignatures are generated based on how well they can discriminate all patients from the control population and thus cannot find the distinct subpopulations.

5. Different Types of Relationships among markers and disease phenotype: Different types of relationship have been studied in the biological domain, since they can reveal novel insights about the complexity of human disease. For example, for many diseases the 'nature versus nurture' debate has been replaced with into 'nature and nurture' studies that emphasize interactions between diverse genetic, clinical and environmental factors [34, 35, 36]. In fact, one of the most important types of relationship in biomarker discovery is interaction, which is when an integrated biomarker is more indicative of disease than its individual constituent factors. In fact, sometimes the individual (marginal) effects of each of the clinicogenomic factors on disease predisposition can be small and thus can remain undetected by most disease association techniques performed on individual datasets. However, interactions among individual factors may be responsible for increasing the risk of complex disease [37]. For example, neither a gene nor an environmental factor

such as tobacco use may be significantly associated with lung cancer by itself, but together they can increase the risk significantly [30]. Interaction of genes with other types of features such as environments is believed to represent the missing hereditability [38]. This necessitates developing efficient techniques that can find such relationships across multiple types of features spanning multiple datasets.

Note that some challenges arise due to the nature of the data (difference in properties and dimensionalities), which have to be addressed by any integrative studies in biomedical domain and these challenges are common with many other domains (e.g., computer vision, image processing, social networks, etc.) containing multi-modal data sets. Other challenges such as relationships among markers, interpretability of the models and disease heterogeneity are specific to bio-medical data integration.

1.3 Contribution of the Thesis

To address the aforementioned challenges, this thesis aims to advance the state-of-the art in biomedical data integration by developing novel data mining techniques and then applying them in several real world multi-source datasets. Integrating diverse biomedical datasets is a widely studied topic and a few different types of techniques have been proposed in the literature. By and large, the biomedical data integration methodologies can be classified into two broad categories: the predictive models [39, 40] and the biomarker discovery techniques [41, 42]. The primary goal of predictive model based approaches is to increase the prediction power of a disease phenotype by integrating multiple types of biomedical data than the prediction models built on individual datasets. Note that often such techniques do not yield easily interpretable results. The highly significant factors obtained from regression can be interpreted as potential risk factors to some extent. However, these models mostly assess the association of such factors with disease phenotype only one at a time. Such singleton based biomarkers are mostly useful for simple diseases such as Huntingtonś disease [42], but not for complex diseases like cancer, diabetes, etc. In contrast, biomarker discovery techniques assess the individual effect of each factor or a subset of factors using statistical tests or models. We refer to these studies as biomarker discovery techniques, since the factors identified by these studies can be used as potential biomarkers for that disease. Biomarkers that are

constructed using a small number of features can be directly useful in diagnosis, treatment or prevention, but equally as important; they can also provide insights into the underlying nature of the disease or related biomedical processes. Moreover, these techniques are flexible in design, so more non-trivial hypothesis about complex diseases such as the combined effect of multiple factors spanning more than one datasets on a disease can be studied. Hence in this thesis, we primarily focus only on developing integrative biomarker discovery techniques that span multiple datasets. Below, we summarize the contribution of the thesis in terms of addressing some of the challenges of biomedical data integration.

1. Literature Survey: Since the integration of multiple biomedical datasets is quite a diverse topic, several types of integration methodologies have been proposed in the literature stemming from several domains such as Data Mining, Statistics and Bio-medical Informatics. Summarizing all these techniques itself is a challenging research task. In this thesis, I first survey the broad set of issues and challenges that arise in biomedical data integration and discuss how the existing studies address these issues.

2. Finding higher-order relationships among diverse factors: Different types of relationships can be present among different types of markers in addition to the relationships present with the disease phenotype. One of the most important types of relationship in biomarker discovery is interaction, when an integrated pattern is more indicative of disease than its individual constituent factors. I aim to also explore another relationship called coherence, when features of the individual datasets are correlated across multiple samples and also have some prognostic value of the disease. This type of pattern is often useful in explaining possible causal relationships, e.g., the downstream effects of the genetic perturbations that cause functional alteration of brain activity. Most of the traditional biomarker discovery techniques struggle to find such higher-order interaction and coherence type integrated patterns from multiple datasets. This calls for developing a pattern mining based integration framework (PAMIN) to find such relationships. PAMIN tries to find such higher-order relationships using an association analysis based two-step approach. It first finds patterns from individual datasets to capture the

available information and intra-dataset relationship separately, and then combines these patterns to find integrated patterns from multiple datasets. A key advantage of these integrated patterns is that they are supported by the same subjects and PAMIN is generic enough to be applied on any number of datasets. Our results indicate that PAMIN discovers interaction type integrated patterns that cannot be found by other prominent approaches for the integrative biomarker discovery, such as canonical correlation analysis (CCA) and related multi-variate techniques. PAMIN and the CCA based approaches, however, can find coherence type integrative patterns, although CCA based approaches cannot find the subgroups of samples associated with those markers.

3. Enhancing interpretability: In this thesis, I aim to enhance the interpretability of obtained integrative patterns, so that they can be evaluated by the domain experts for clinical decision making. In particular, I utilize the existing medical knowledge in models as prior assumptions to enhance the interpretability of the models. Several types of domain knowledge such as relationships between genes in terms of biological pathways [43] or medical knowledge about the relationship among diagnosis codes [44] are available in biomedical domain. Specifically, I present a multivariate integrative model called Sparse Hierarchical Canonical Correlation Analysis (SHCCA), which incorporates domain information into the model in the context of finding diagnostic groups that are both interpretable and predictive of the urinary incontinence of Home Healthcare patients. In the bio-medical domain, many types of prior information are available from experts opinions. For example, I use two types of domain information as prior knowledge into the model: (i) various clinical information such as demographic, behavioral, physiological, and psycho-social factors available in EHRs for the same patients, and (ii) a standard medical clinical classification system (CCS), which provides a systematic grouping of diagnosis codes into a hierarchical tree structure. In addition, sparsity constraints are imposed on the model to perform the feature selection simultaneously. Finally, I also propose a novel metric called I-score based on the medical literature search to measure the interpretability of obtained diagnostic codes. SHCCA significantly improves the interpretability of the obtained diagnostic codes significantly, without losing the prediction power.

4. Handling patient heterogeneity: I further use information available from multi-source biomedical data to stratify the patient groups more efficiently for handling disease heterogeneity. In particular, the prior risk of the patient during admission is used for determining more homogeneous patient subgroups and then, patterns were found in each subgroup. There are two main challenges for integrative studies on heterogeneous subjects, which is the focus of my current ongoing research. First, the initial health risk of heterogeneous groups can be affected by multiple types of factors, often from multiple datasets, which requires a combinatorial search on the feature space. Second, there can be both generic and specific patterns for each particular homogeneous subgroup. Finding local patterns that are specific to a particular subgroup is computationally hard, since it requires contrasting the local pattern for a particular subgroup with all other subgroups. My research has been to propose an efficient framework to handle the above mentioned two issues. Moreover, I propose a new interestingness measure, which along with domain knowledge, can be leveraged to find both generic and specific local patterns. Application of such framework in a multi-source EHR data resulted in very interesting results, which were evaluated by domain experts.

1.4 Thesis Overview

The organization of this thesis is as follows. A broad overview of existing issues and challenges in biomedical data integration and the existing studies that aim to address such challenges have been discussed in Chapter 2. The problem of finding relationships among clinical and genomic datasets has been discussed in Chapter 3. In Chapter 4, I present a biomarker discovery technique that can take prior knowledge into account for enhancing the model interpretability. In Chapter 5, I discuss several ways to handle disease heterogeneity. I conclude in Chapter 6 with a discussion on future work.

Chapter 2

Biomedical Data Integration

2.1 Introduction

Until the last decade, traditional clinical care and management of complex diseases mainly relied on different clinico-pathological data, such as signs and symptoms, demographic data, pathology results, and medical images. In addition, efforts have been made to capture genetic factors by examining the family history of patients. The effect of such clinical and histo-pathological markers is assessed by cohort based studies conducted on large populations [45] and the knowledge obtained from these studies is summarized in clinical guidelines for the diagnosis, prognosis, monitoring and treatment of human disease, e.g., NPI [46] and Adjuvant! Online [47, 48] for breast cancer and palmOne [49] for prostate cancer. However, this approach still falls short. For example, there are adverse drug reactions for some patients who have risk factors similar to those patients who have been cured by the same therapeutic treatment. This issue stems from the strategy of óne drug fits alláng motivates the need to improve on conclusions drawn from cohort-based studies so that the underlying mechanism of complex diseases can be understood at the individual patient level.

The recent advancement of high-throughput technology has led to an abundance of information for each individual at the micro-molecular level. A myriad of genetic, genomic and metabolomics data have been collected to capture different aspects of cell mechanism that shed light on human physiology. Examples include SNPs, which provide information about the genetic polymorphism of an individual; gene expressions, which

measure transcription; and protein and metabolite abundance, which captures protein abundance and post-translational modifications. These high-throughput datasets have helped answer some complex biological questions for different diseases, such as assessing the prognosis [50, 51, 52, 53], epistasis effects on diseases [54], and discovering new sub-phenotypes of complex diseases [55, 56, 57]. The use of genetic information in epidemiology helped design effective diagnostics, new therapeutics, and novel drugs, which have led to the recent era of personalized medicine (genomic medicine) [58, 59, 60]. However, these genetic factors alone cannot explain all the intricacies of complex diseases. For example, the incidences of cancer vary widely among different countries due to the environmental factors, even for the same ethnic groups, as is illustrated by changes in incidence when people of different ethnicities migrate from one country to another [61, 62].

In recent studies [2, 3], it has been hypothesized that most complex diseases are caused by the combined effects of many diverse factors, including different genetic, genomic, behavioral and environmental factors. For example, cancer, which is the most widely studied disease phenotype in last few decades, is extremely heterogeneous. Different clinical endpoints of cancer, such as the idiosyncrasy of individual tumors, the survival rate of cancer patients after chemotherapy or surgical treatment, development of metastasis, and the effectiveness of drug therapy are governed by different risk factors including multiple mutations of genetic factors (e.g., RAS, RTK, TGF-β, Wnt/signaling pathways), behavioral factors (e.g., tobacco exposure, diet, lifestyle) [62], long-time environmental effects (e.g., stresses, temperature, radiation, oxygen tensions, hydration and tonicity, micro- and macro-nutrients, toxins) [37] and germline variations (e.g. BRCA1/2) [63]. Therefore, clinico-pathological and genomic datasets capture the different effects of these diverse factors on complex diseases in a complementary manner. In a more complicated scenario, a complex genetic network can evolve dynamically under various environmental factors [2]. Using the two diverse perspectives provided by both types of data can potentially reveal disease complexities in greater detail.

It is essential to build integrative models considering both genomic and clinical variables simultaneously so that they can combine the information present in clinical and genomic data [2]. In this chapter, we define clinicogenomic integration as building predictive or descriptive models by integrating clinical and genomic data. Clinical data

refers to a broad category of patients pathological, behavioral, demographic, familial, environmental, and medication history, while genomic data refers to all varieties of a patients genetic information including SNPs, gene expression, and protein and metabolite profiles. Clinicogenomic studies should involve at least one clinical dataset and one genomic dataset for a group of people who are assessed for an outcome of a phenotype of a disease.

The main goal of clinicogenomic integration is to infer more useful information by combining clinical and genomic data about a disease endpoint such as the survival of a disease or the effectiveness of a therapeutic intervention. Some studies aim to achieve this goal by building a predictive model (e.g., a classification [64] or regression model [25]) to assess whether the combined model built on clinical and genomic markers provides better prediction power than the predictive models built either on clinical or genomic data. A different set of integrative studies aim to assess the individual effect of each factor or a subset of factors using statistical tests or models. We refer to these studies as biomarker discovery techniques, since the factors identified by these studies can be used as potential biomarkers for that disease. A unique advantage of these techniques (in contrast to predictive models) is that they can identify factors that are only specific to a particular subgroup of patients. Note that there are many commonalities among these two broad categories in terms of the methods and the issues addressed by them. In particular, many of the common challenges faced by these two studies arise from the differences between the characteristics of clinical and genomic datasets due to the differences between the experimental designs and protocols used to collect such diverse data.

Main Categories		Predictive Modelling	Testing Additional Power
Explicit Dimensionality Reduction	Early Integration	Regression (Li[65], Teschendorff et al.[60], Shedden et al.[66]); Classification (Stephenson et al. [58], Sun et al.[67], Li. et al.[68], Beane et al.[69]); Tree based method (Nevins et al.[70], Pittman et al.[24], Clarke et al.[71], Cao et al.[72])	Tibshirani et al.[73]; Hofling et al.[74]; Boulesteix et al.[11]; Acharya et al.[75], Wang et al.[76], Obulkashim et al.[13]
	Intermediate Integration	Daemen et al.[77], Gevaert et al.[78]	
	Late Integration	Campone et al.[79], Silvaha et al.[80], Futschik et al. 2003 [81]	
Sparse Model	Early Integration	Bovelstad et al.[82], Ma et al.[83]	Binder et al.[22], Boulesteix et al.[84], Kammer et al.[43]

Table 2.1: Taxonomy of different clinicogenomic models. Some branches are missing indicating no studies were observed in that category.

There are a number of issues that are specific to developing integrative predictive models. For instance, the information present in clinical and genomic datasets may not be equally reliable, because clinical and genomic datasets are collected independently and thus the biases and assumptions of each of the datasets may be different due to differences in experimental designs and protocols. In particular, clinical variables have been validated by numerous epidemiology studies [11, 12, 13] with large cohorts and tend to be more reliable than the genomic data. Therefore, treating both datasets equally in predictive models is not desirable.

Building models for biomarker discovery also confronts some unique challenges. For example, biomarkers obtained from clinicogenomic studies may not be applicable to all patients in the same degree due to the issue of disease heterogeneity, i.e., that a disease consist of multiple subtypes, each occurring in different subpopulations. Hence, biomarkers may need to be designed for each subgroup of patients. Finding biomarkers is itself a computationally challenging problem due to the exponential search of all potential combinations of features spanning multiple datasets. Searching for subgroups of patients simultaneously with biomarkers may further complicate the problem of computational complexity.

Given the importance of the topic, a large number of researchers have developed approaches for integrating clinical and genomic data. In this review chapter, we discuss the broad set of issues and challenges that arise in clinicogenomic integration and discuss how the existing studies address these issues, as well as directions for future research. Table 2.1 summarizes all these papers in different categories of predictive models and biomarker discovery techniques. Some previous review articles have focused on specific challenges of clinicogenomic integration. For example, Boulesteix et al. [21] performed a survey of techniques that try to validate the additional predictive power of genomic markers over traditional clinical variables. Correa et al. [41] and Sui et al. [85] reviewed some of the integration approaches that find relationships between datasets using correlation measure. A few studies [31, 48, 38] surveyed primarily interaction-type relationships between genes and environmental factors. Table 2.2 summarizes the different aspects of clinicogenomic studies covered by these articles. It also includes some articles that survey research on integrating the omics datasets only [86, 87, 88],

since they address some of the challenges that are common with clinicogenomic integration. Note that, our review does not cover other related topics such as architectures and platforms for integrating clinical and genomic datasets to enhance interoperability [89, 90, 91, 92] and mutual incorporation of genomic data into electronic health records (e.g., [93, 94]).

In brief, the key contributions of this chapter are (i) it identifies the key issues and challenges that arise in integrating clinical and genomic data and organizes them into three broad categories: data specific issues (Section 2), predictive modeling issues (Section 3) and biomarker discovery related issues (Section 4), (ii) it surveys the existing computational techniques from multiple bio-medical domains (e.g. bioinformatics, computational biology and neuroscience) that aim to address each of the above mentioned three issues, and (iii) it discusses additional issues in clinicogenomic integration and outlines future work (Section 5).

Studies	Challenges	Data Issues		Predictive Model Issues		Biomarker Issues		Validation	Dataset Used	
		Stages	Dimensionalities	Generic Predictive model	Added Predictive Value	Disease heterogeneity	Finding relationships		'Omic'	Clinical
Boulesteix et al [21]								x	x	x
Thomas et al. [107]; Manuck et al.[38]; Hunter et. al.[31]							Interactions		x	Environmental
Correa et al.[41], Sui et al.[85]							Correlations		x	Lab Images
Hamid et al.[86]	x	x		x					x	
Tsiliki et al.[87]	x			x			x		x	
Bebek et al.[88]							x		x	
Our review	x	x	x	x	x	x	x	x	x	x

Table 2.2: Comparative analysis on different review articles.

2.2 Data Related Issues

There are significant differences in the nature of clinical and genomic data format, types, properties, dimensionalities, biases and assumptions, and the amount of noise

present in the two datasets that need to be dealt with by any integrative methods. For example, clinical and genomic variables are mostly categorical or nominal in nature, while genomic data are mostly continuous-valued [9]. Clinical variable has generally fewer missing values and noise in comparison to genomic data [95]. Another big data related issue in integrative studies is the curse of dimensionality, i.e., the number of samples (n) available in these datasets is much less than the total number of variables (p) coming from multiple datasets. (This is the classical statistical problem of n¡¡p [18]). In that case, the number of samples required to describe the input space defined by high-dimensional datasets increases exponentially, which can lead to overfitting a predictive model, i.e., make a model that conforms closely to limited data and does not generalize [19]. Similarly, biomarker discovery techniques may find potential factors by random chance and need to be cognizant of issues of multiple hypothesis testing [20]. This problem is even more challenging when performing clinicogenomic integration due to the disparate dimensionalities of clinical and genomic datasets. For example, genomic and genetic datasets have much higher dimensionality in comparison with clinical data which often contains approximately 10-20 variables [96, 21]. Therefore, the clinical variables can be completely lost among the vast number of genomic variables unless the disparity in dimensionality is handled properly [22, 23]. In this section, we will discuss several clinicogenomic models that try to address the above mentioned two data related issues using various techniques from statistics, machine learning, and data mining. First, we will discuss several stages of integration that are used to address the disparate natures of clinical and genomic data. Second, we will describe how the disparate dimensionalities of clinical and genomic datasets are handled.

2.2.1 The Stages of Integration

The differences in the nature of the clinical and genomic data create several challenges for developing integrative models. In general, integration of such diverse datasets can be performed in different stages depending on how disparate the natures of the individual data sources are. More specifically, integration can be performed on the data level by merging all datasets, on the decision level after merging the models built individually on each dataset, or on some intermediate representation of data. These three types of integrative studies were first proposed by Pavlidis et al. [1] in a seminal study of biomedical

data integration and were called early, late and intermediate integration, respectively. Figure 2.1 shows the detailed steps of the three stages of data integration. Although these three types of models were developed in predictive modeling context, the concept of integrating the studies in different stages is generic, and early and intermediate integration can be applied to biomarker discovery studies as well. We will describe how these three generic categories are applied in the context of clinicogenomic integration, their advantages and disadvantages, and the future research required in this context.

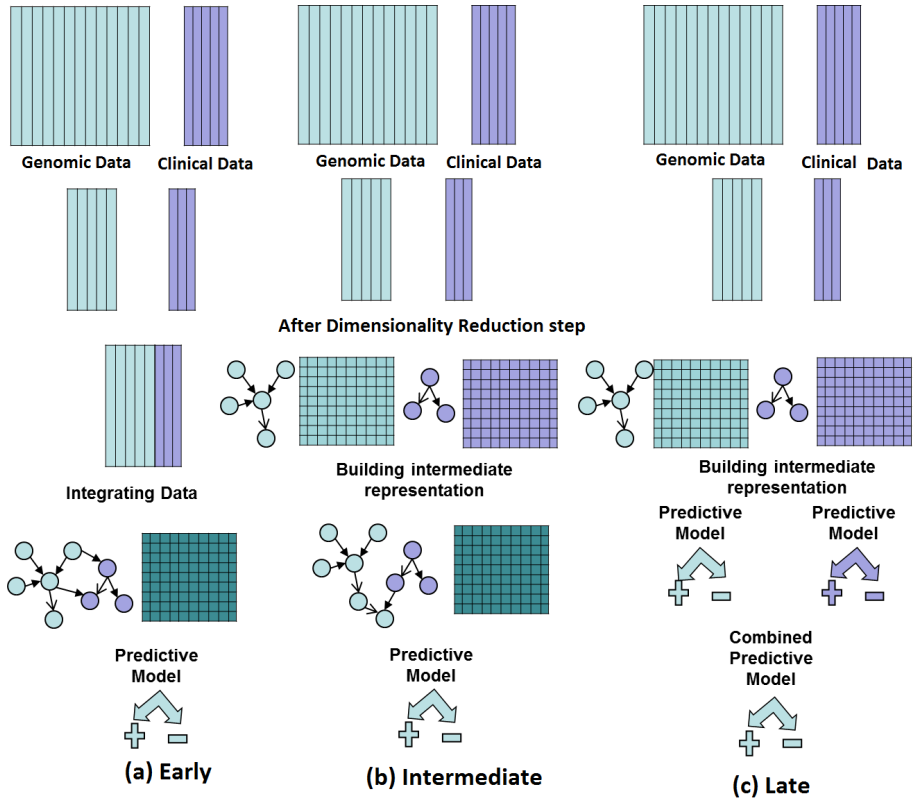


Figure 2.1: Pictorial representation of three stages of integrations inspired by [1].

2.2.1.1 Early Integration

In general, early integrative approaches merge the independent data sources together before performing any kind of data analysis. In a simplistic case, the individual data matrices are simply combined into a larger matrix if both of the datasets have same

set (or subset) of samples. Thus, the integration of the individual datasets, which are clinical and genomic data in our case, is performed at an early stage of the overall analysis. Once the combined data matrix is prepared, any types of models can be developed based on the two goals of the clinicogenomic studies. The unique assumption of this type of integration is that both of the datasets are similar in nature, i.e., most of the properties of the datasets such as data type, formats, structure, and dimensionality, are either similar or preprocessed to be as similar as possible. In reality, clinical and genomic data have different natures, and thus, early integration loses the individual properties of each dataset. Some clinicogenomic predictive models perform a significant amount of preprocessing, such as dimensionality reduction, missing value imputation and data discretization, before integrating the individual datasets to make them more homogeneous, but this typically leads to a loss of information. Similarly, in the context of biomarker discovery, especially in the neuroscience domain, multivariate statistical models such as SVD or ICA [97, 98, 99] have been developed, where the model is applied after combining individual datasets with significant pre-processing.

Advantage: Early integration is the simplest approach, since any standard model can be applied on the integrated dataset to achieve any of the objectives. Therefore, most of the clinicogenomic studies fall in this category [Table 2.1]. Moreover, they can preserve any kind of inter-data relationships. For example, if some clinical and genomic variables are correlated, the model developed after data integration can take the correlation structure into account.

Disadvantage: Early integration loses the individual properties of each dataset such as the structure and the different degree of information when merged together into an augmented dataset. The dimensionality of the augmented dataset also increases. Thus, the model may also suffer from high dimensionality and low statistical significance of the obtained results.

2.2.1.2 Late Integration

Late integration first develops predictive models separately for each of the individual data sources and then merges the individual decisions of all predictive models into a final score as the prediction for the outcome variable. As opposed to early integration, this type of integration actually merges the classifier decision rather than the original

dataset. The main assumption of late integration is that the individual datasets are independent and there is no inter-dataset relationship.

The biggest challenge of late integration is how to merge the decision of classifiers obtained from individual datasets. Several strategies like majority voting, linear aggregation and weighted average have been applied for this purpose. For example, two breast cancer studies conducted by Campone et al. 2008 [79] and Silhava et al. 2009 [80] simply summed up the individual decision coming from genomic and clinical data. Campone et al. applied the Cox regression model to summarize the topmost 15 discriminating genes into a single genomic score and then added it to the traditional clinical score of breast cancer, NPI, to get the final score for assessing the effect of adjuvant chemotherapy. On the other hand, Silhava et al.[80] applied two different predictive models: logistic regression and BionomialBoosting(BB) [100] to get the genomic and clinical score, respectively, before summing them.

However, simple summation is not always appropriate because the contribution of the individual data sources to the overall clinicogenomic model may be different. Alternatively, the contribution from each individual dataset towards the disease phenotype can be assessed and the scores obtained from the individual models can be weighted accordingly. For example, Futschik et al, 2003 [81] used parameterized learning for merging the individual decisions of the clinical (Bayesian classifier) and genomic data (evolutionary fuzzy artificial neural network (EFuNN [101]) into a final decision. Furthermore, they also tested statistical independence of the outputs of two independent models using the mutual information [102], which is a key assumption for late integration. In a more complicated scenario, with many datasets being integrated, the more general problem arises when some of the models built on individual datasets produce binary class decisions and some of the predictive models generate continuous-valued scores. Several approaches, including majority vote and its more generic version called consensus learning [103] or Bayes rule based aggregation [104, 105] have been studied in many other domains such as image processing and social networks.

Advantage: The individual structure and the nature of each dataset are preserved in late integration, since model is developed on each dataset separately. Moreover, different models can be used for different datasets depending on the individual nature of

each dataset. Late integration is particularly useful when each of the datasets is completely heterogeneous, i.e., the datasets cannot be transformed into a common format for integration.

Disadvantage: Late integration misses any kind of possible relationship, such as correlation or interactions, which may be present among the datasets. Moreover, late integration generates a different hypothesis for each of the datasets as opposed to a single hypothesis for the integrated dataset. Interpretation and validation of these different types of hypotheses is not trivial.

2.2.1.3 Intermediate Integration

This approach tries to address the limitations of the above two approaches. It transforms each dataset into a common suitable intermediate structure, such as a kernel (pairwise similarities) or graph, depending on the nature of each dataset, and then merges these structures into a final model. Therefore, it can handle data sets of very different kinds (unlike early integration) and take into account the possible relationships between the datasets to some extent (unlike late integration). The main assumption of this approach is that there is an appropriate intermediate representation for each dataset preserving the individual properties of that dataset and the intermediate representations can be combined easily. Even though most existing studies used either early or late integration, intermediate integration is most suitable for different types of clinical and genomic datasets. The most popular intermediate integration techniques are based on kernel based machine learning approaches [1, 39, 106, 107, 108], because it is more generic and better preserves the individual properties of each data type and source. A kernel approach first transforms the original features of a dataset into pairwise similarities between any two samples, and then the individual kernels built on each dataset are merged before building any model. Note that, merging kernels is much easier than merging decisions in late integration. (Refer to the review paper [40] for a more theoretical description of kernel fusion methodologies.) Daemen et al. [9] further designed kernels separately for each of the clinical data categories nominal, categorical, ordinal and numeric to account for the wide-variety of ranges present among such categories before merging them with the kernel developed on genomic data.

One advantage of such kernel based integration is that the weights corresponding to

an individual dataset can denote the relative contribution towards the final prediction. Similarly, the kernel versions of the multivariate statistical models, kernel CCA [109] and kernel ICA [110], have been applied for biomarker discovery. However, choosing an appropriate kernel for a particular dataset is not trivial. Moreover, kernels are not easily interpretable, although a recent study [111] attempted to enhance the interpretation of kernel-based fusion by visualizing the contribution of the variables to principal components (PCA [26]).

In contrast, graph based techniques can provide more interpretable models for intermediate integration. For example, Gevaert et al. [78] used a partial integration approach conceptually similar to intermediate integration. They first inferred a directed acyclic graph (DAG) for each dataset and then merged the two DAGs using a Bayesian network. Although such graph-based intermediate integration provides more interpretable models, merging the structures (DAGs) obtained from each dataset is not as straightforward as fusing the kernels. Alternatively, a distance based intermediate approach [112] has been proposed to account for the different scales and data types of clinicogenomic variables, where individual distance metrics are defined for each data type and data source, and then they are combined using a related metric scaling method. However, this intermediate representation can only be used for clustering and distance based classification models. Regardless, in all of the previous studies, intermediate integration showed better performance than early and late integration.

Advantage: Intermediate integration can preserve the individual properties of a dataset. Moreover, inter-dataset relationships such as correlation and redundancy can also be taken into account during final model creation, although this depends on many issues such as the choice of kernel and how such relationships are preserved during kernel fusion.

Disadvantage: Finding interaction type of relationship and causality is difficult by this approach. Even a Bayesian approach learns the dependency structures (the putative causal variables) independently for each dataset and thus, cannot find dependencies across two data sources. Although intermediate integration can be applied theoretically for very disparate data sources, they have mostly been applied for clinicogenomic datasets with similar record based format. However, clinicogenomic datasets

may contain other formats besides the traditional record based datasets. Examples include protein-protein interaction networks, which contains pairwise relationship between two genomic features. In these cases, new intermediate integration techniques need to be developed. For example, networks can be inferred from the record based datasets and then they can be merged with existing genomic networks. Alternatively, genomic networks can be incorporated as prior knowledge in clinicogenomic models. Furthermore, some data sets may contain structures, e.g., measurements across time or across a genetic sequence, that are not present in others. Building clinicogenomic models at multiple time-points and finding trends of the combined clinicogenomic biomarkers require developing new computational techniques.

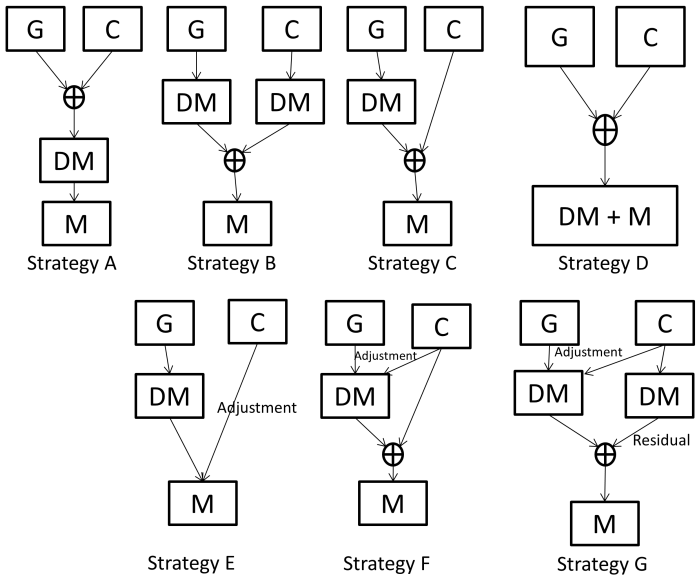


Figure 2.2: Different Strategies of dimensionality reduction (DM) and predictive modeling (M) combining Genomic (G) and Clinical data (C). Strategies A-D are generic to both predictive models and biomarker discovery, while the later strategies (D-F) are specific to predictive models assessing additional power of genomic variables.

2.2.2 Handling Curse of Dimensionality

In this section, we will describe two techniques to address the curse of dimensionality. First, we describe techniques that rely on reducing the dimensionality of the high-dimensional datasets and that are also cognizant of the disparate dimension of clinical and genomic datasets. Second, we describe multi-site methods that aim at increasing sample size by integrating multiple similar types of datasets from multiple cohorts with or without dimensionality reduction.

2.2.2.1 Dimensionality Reduction

The easiest way to perform any dimensionality reduction technique-either feature selection or feature extraction is to merge the two datasets before performing any integrative study (strategy A). However, most of the time clinical variables are completely lost among the vast number of genomic variables in this nave framework as shown in Figure 2.2(a) (following the terminology of Boulesteix et al. [22, 23]). Therefore, most of the clinicogenomic studies were designed carefully to handle the unique challenge of disparate dimensionalities of the clinical and genomic data. We discuss three different ways to perform dimensionality reduction for clinicogenomic integration: explicit dimensionality reduction, implicit feature selection via regularized models, and dimensionality reduction based on prior knowledge.

Explicit dimensionality reduction: The easiest way to handle the disparate dimensionalities of individual datasets is to develop a two-step approach, namely, perform explicit dimensionality reduction separately for each dataset in a first step and then build the integrative model in a second step (strategy B) as shown in Figure 2.2(b). (The specific design issues for dimensionality reduction and predictive models are discussed in detail in Section 3). For example, neuroscience studies combine multiple high-dimensional genetic and pathological datasets using various feature extraction techniques before developing any integrative models [113, 41]. Another popular strategy is to apply dimensionality reduction techniques solely on the genomic dataset since clinical demographic variables are already low dimensional (Strategy C in Figure 2.2(c)). Note that both feature selection (e.g., univariate regression model [114, 82]) and feature extraction based dimensionality reduction (e.g., PCA [26] or PLS[27]) techniques can be

used in these strategies. (Reader is referred to [7] for an extended discussion of different dimensionality reduction techniques). Moreover, the selected top few discriminative genomic and clinical features are either used directly or combined into a global score [23, 21] before integrating them in the model development step.

However, there are some disadvantages to this approach. First, determining the appropriate number of features is challenging [11]. Second, the top predictive genomic variables selected in the first step may be highly correlated with the clinical variables and thus, may not provide additional predictive power for model development in the second step. In that case, the subtle contributions of many genes to prediction will be missed by the dominant genomic features that are correlated with the clinical variables [84]. The later issue is especially important when the goal is to assess the additional power of genomic data over clinical variables as described in Section 3.2. A recent study [23] aimed to handle the above mentioned two issues. A cross-validation (CV) [115, 25] framework was used to select the appropriate number of genomic features. In addition, regression models were used to select only those genes that have additional predictive power after adjusting for clinical variables. These clinical variables were included as the baseline in the regression model so that such redundancy of the genomic variables with clinical variables was removed.[S7]

Implicit feature selection via Regularized Statistical Models: This approach merges the steps of dimensionality reduction and model development into a single step by using regularization based statistical models (Strategy D in Figure 2.2(c)) [25]. In general, regularized techniques introduce an extra penalty term for model complexity in addition to the original loss function of the predictive model. By preferring less complex models, regularized models can increase the generalization power of a predictive model and as a result, are more effective for reducing the overfitting of high-dimensional data. In addition, sparse regularized models can perform variable selection by imposing a special type regularization (L1) to force most of the coefficients to be zero and consequently, remove the need for an explicit variable selection step.

Incorporating this regularized regression techniques for clinicogenomic study is not straightforward because of the disparate dimensionalities of clinical and genomic datasets. This necessitates the modification of the original regularized regression framework so that regularization is performed to different degrees for the two datasets. The first

type of approach incorporates the clinical variables as mandatory variables without imposing any penalty on them. An example of such an approach is CoxBoost [116, 22], which is a sparse component-wise boosting based Cox regression model. The second type of approach penalizes the inclusion of variables in the two datasets differently. For example, Ma et al. [83] proposed a more generalized regularization method called Cov-TGDR (Covariate-Adjusted Threshold Gradient Descent Regularization [117]) with two different penalty structures for clinical and genomic variables.

The main advantage of these models is that they can implicitly take into account the redundancy present between genomic and clinical datasets, since both datasets are considered together during model development. This property makes these models more suitable than explicit dimensionality reduction approaches for assessing the additional predictive performance of genomic features over the clinical variables [84]. However, each of the regularized model approaches has their own assumptions and requires learning several parameters, which results in higher computational complexity. Bovelstad et al. [82] provided a methodological comparison of different dimensionality reduction techniques designed for Cox regression in survival studies. Among various explicit and implicit techniques, they observed that modified ridge regression performed the best when applied to three different clinicogenomic datasets. However, they did not compare it to the Cov-TGDR methods.

Incorporating prior biological knowledge in dimensionality reduction: There are some issues with discriminative genomic features selected by the dimensionality reduction techniques described in last two paragraphs. First, the genomic features are often not reproducible among different studies [14, 15, 16, 17]. Second, often they are hard to interpret and thus, do not provide any meaningful biological knowledge. Third and more importantly, complex diseases are often caused by the activation of a large group of functionally related genes, such as those in biological pathways [118, 119], , with individually small risk factors. For example, cancer is often activated by a group of oncogenes or by the deregulation of a group of tumor-suppressor genes. Bringing such prior knowledge into biomarker discovery has been very popular in the biological domain [120, 121, 122] since it helps resolve the previous three issues. Similarly, Kammers et al. [43] recently used one such form of biological knowledge for grouping genes, namely, gene ontologies (GO) [123]. In particular, they clustered the genes belonging

to each GO group into a few clusters and then used the combined effect of each cluster (as captured by the first principal component) as a feature for penalized Cox regression in a manner similar to CoxBoost.

Although pathway based analyses can provide interpretations for the obtained biomarkers to some degree, they only succeed in testing the association with the already known pathways or gene ontologies rather than discovering new biological knowledge. Therefore, more sophisticated computational methods are needed which do not solely rely on the prior knowledge, but rather combine known biomarkers with the data driven discriminative features for inferring new knowledge.

Study	Dimensionality reduction technique	Predictive method	Integration type	Testing additive performance of genomic variables	Clinical Endpoint	Disease
Acharya et al. 2009	Gene Clustering and preselection based on prior knowledge	Hierarchical Clustering	Early, Semi-supervised	Yes	Relapse free survival (may be distant)	Breast Cancer
Shedden et al. 2008	Comparative study among 8 dimensionality reduction techniques	Cox hazard model	Early	No	Survival data	Lung Cancer
Teschendorff et al. 2006	Common genes across 6 datasets	Univariate Cox model	Early	No	Survival vs. death; Development of metastases	ER+ breast cancer
Teschendorff et al. 2006	Common genes across 5 datasets	Unsupervised and semi-supervised clustering	Early	Yes	Survival, Time to distant metastasis ER	breast cancer

Table 2.3: Summary of multi-site clinicogenomic studies.

2.2.2.2 Multi-cohort Clinicogenomic Integration

Integrating multiple cohorts of patients with same phenotype can increase the sample size significantly and thus, is very popular for developing reproducible genomic biomarkers [124]. Indeed, genomic studies are often criticized for the lack of reproducible results produced from independent cohorts due to small sample cohorts size, selection bias during sample inclusion and annotation, different protocols for sample preparation and data preprocessing, and heterogeneous clinical endpoints [66]. For example, very few overlapping genes were observed between the biomarker genes of two well-known breast cancer studies of Vant́ Veer et al. 2002 [180, 162] and Wang et al. 2005 [125] by other independent studies [14, 15, 16, 17]. Inspired by the advantages demonstrated by multi-cohort genomic studies, a few studies built universal multi-cohort clinicogenomic models by integrating datasets not only from multiple modalities (integrating heterogeneous datasets), but also multiple similar types of datasets (integrating homogeneous datasets). Despite their advantages, multi-cohort clinicogenomic studies face numerous challenges. Two are discussed here: 1) minimizing experimental biases present in samples from multiple cohorts and 2) merging features collected from multiple cohorts. Table 2.3 summarizes the details of individual studies.

Minimizing the experimental biases: Several strategies have been undertaken to minimize the experimental biases present in samples collected by multiple cohorts. In one such study, Teschendorff et al. 2006 [60] used a recently developed statistical evaluation measure called the D-index [126] instead of traditional classification accuracy to build a universal marker from six breast cancer datasets. The D-index depends only on the relative risk ordering of the test samples rather than relying on the absolute value of the outcome variable, which makes it suitable for assessing performance across test samples coming from different cohorts with diverse characteristics. In another approach, which was used for predicting the survival of lung adenocarcinoma patients, Shedden et al. [66] tried to directly minimize the experimental bias among different cohorts by generating their own datasets from six different institutions using a robust and reproducible protocol [127].

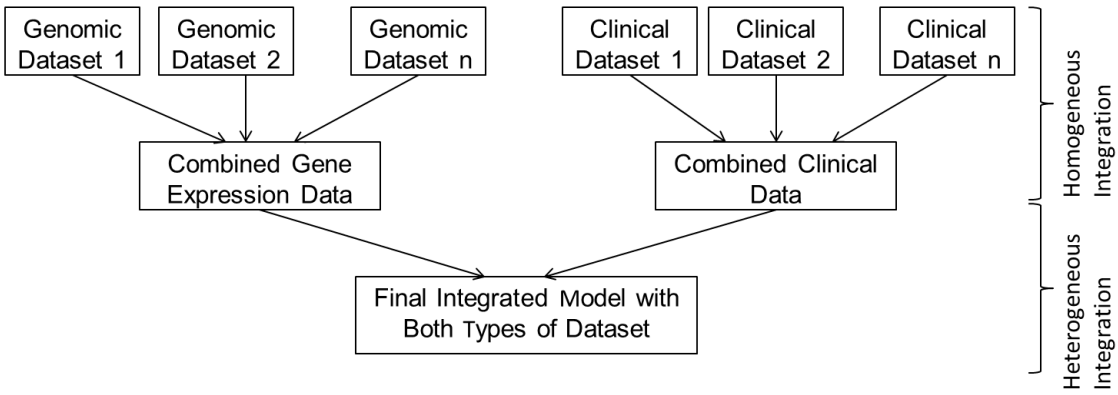


Figure 2.3: The two stages integration of the multi-site clinicogenomic models

Merging features: Existing multi-cohort clinicogenomic models proceed in two steps to merge features (Figure 2.3). For homogeneous integration, most of studies take the approach of retaining only those features that are common in all of the datasets. Then, the heterogeneous integration of clinical and genomic data is performed using any of the techniques described earlier. Most existing multi-cohort studies [60, 66] select the genes from multiple datasets by keeping the common features among all datasets (data level early integration as mentioned in Section 2.1). In contrast, Acharya et al. [75] used prior biological knowledge represented by pathways (similar to the pathway based analysis described in Section 2.2) to select the genes that are common across multiple platforms. Techendorff et al. 2007 [128] used a semi-supervised approach along with imputation of missing values to select common features across studies for finding prognostic molecular subtype for the ER breast cancer. On the other hand, spatio-temporal neuroscience data collected from magnetic resonance imaging (MRI) pose more challenges for such feature selection. Since there is auto-correlation across nearby regions or consecutive times, they are often summarized into group level activities before merging multiple samples. Group ICA techniques [99] have been proposed to concatenate the features in multiple spatio-temporal levels.

All the approaches mentioned above assume that all samples contain the same set of features. However, they cannot use the full potential of each dataset, especially

if each dataset has different pools of features, which is often true for genomic datasets containing genes and proteins. In fact, many of the features will be lost due to differences among the datasets. Functional relevance among the biological features (e.g., Tag SNPs [129]) among different studies can help retain more features after feature pooling [31]. A more sophisticated intermediate or late integration approach can be used to fuse information available in individual datasets rather than selecting the common genes. For example, the information available in each dataset can be learnt first and then the prediction of each dataset can be merged together using a Bayesian framework similar to that used for multi-cohort gene expression datasets [130, 131, 132]. Additionally, the genomic information can be integrated at a higher level, such as the pathway level, rather than at the individual gene level in order to use the full potential of each dataset.

2.3 Predictive Model Related Issues

A key question that any integrative study wants to answer is whether integrating multiple data sources provides additional information beyond that provided by the individual data sources, i.e., whether the datasets contain complementary information. The information present in a dataset is assessed based on how well it can predict the disease endpoint. Therefore, a combined clinicogenomic model is assessed based on its improvement of prognosis power of the corresponding disease endpoint over the models built on either clinical or genomic data independently. However, treating both clinical and genomic datasets with equal emphasis often overestimate the performance of genomic variables [11, 12]. In fact, clinical and genomic data have different degrees of reliabilities and usefulness. Clinical variables are gathered more systematically over a longer period of time and they contain less noise. In addition, they are rigorously validated by numerous epidemiology studies [11, 12, 13]. Furthermore, clinical data are cheap and easy to collect. In contrast, genomic data such as gene expression data are less reproducible over independent cohorts because of the high noise, different experimental biases, and high-dimensionality and small sample sizes [14, 15, 16, 17]. Such domain information has led to a different set of predictive models, which aim to include genomic variables only if they provide additional information over traditional clinical variables. We first discuss the generic clinicogenomic predictive models that treat clinical and

genomic variables in the same manner and then the models that assess the additional power added by genomic variables over the clinical variables in an unbiased manner.

2.3.1 Improving the Prognostic Power Only

The primary goal of the predictive clinicogenomic models is to improve the prognosis of diseases by integrating clinical and genomic datasets, i.e., hopefully the combined clinicogenomic model provides better prediction power for disease endpoints than is provided by individual clinical or genomic data sources. The phenotype can be either binary class labels, such as cancer vs. no cancer, or continuous variables, e.g., the survival time after chemotherapy or other therapeutic treatments. A number of predictive models have been applied in this context, each having their own advantages and limitations that clinicians should consider [64, 133]. Moreover, the predictive models have to be designed properly with a proper evaluation technique so that the gain of the combined model over the individual models can be measured accurately. In this section, we will first describe the overall study designs for predictive models and then we will briefly describe different predictive models and a comparative study (The details of these techniques can be found in Appendix Table A.1).

2.3.1.1 Design Strategy

Most clinicogenomic techniques used a dimensionality reduction technique described in Section 2.2 to address the different dimensionalities of clinical and genomic datasets (Strategy B-D in Figure 2.2(b-d) before developing integrative predictive models. This creates a challenge for model evaluation. Most of the evaluation techniques such as random sampling [71, 82], bootstrapping [25] and systematic cross-validation (CV) [115, 25] framework [24, 68, 58, 83, 67, 22, 82] must be applied to the predictive models with an explicit supervised dimensionality reduction with special care, i.e., the explicit supervised dimensionality reduction should be performed on the training data rather than whole data [134, 135, 136]. The most frequent metrics used during these evaluation steps to assess the gain in prediction power are mostly area under the ROC curve (AUC), C-index, and Brier score. Furthermore, a recent study [137] proposed an additional metric called explained variation to specifically quantify the gain of prediction power of genomic (or clinical or combined) model in comparison to the baseline prediction

accuracy of a null model without containing any variable to allow comparison among multiple predictive models and also among multiple datasets.

2.3.1.2 Linear Models with Explicit Dimensionality Reduction

The choice of the particular predictive model differs based on the clinical endpoints of the disease, i.e., whether the target variable is discrete or continuous. If the response variable is continuous, such as survival of patients after a particular therapeutic treatment or the development of metastasis after surgery, then the regression based methods are deployed for model development. For example, the Cox proportional hazard model estimates the lifetime (survival or failure) of an event associated with the covariates using two parameters: a hazard function describing the changes of hazard (risk) over time at the baseline level of covariates and the co-efficients describing the effect of each variable on survival [138, 114] . In one such clinicogenomic study, Lexin Li [65] used the Cox model for predicting the survival of the patients with diffuse large-B-Cell lymphoma (DLBCL) after chemotherapy. In addition to the genomic features (selected by a supervised dimensionality reduction [139]), they included a well-established clinical factor called international prognosis index (IPI) [140], which combines different clinical factors of DLBCL.

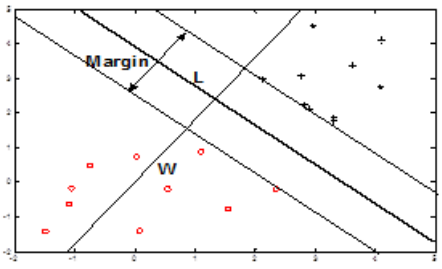


Figure 2.4: Discriminant models with linear decision boundary. SVM tries to maximize the separation between the two classes (red and black).

Classification techniques are used to build clinicogenomic models when the output variable has discrete categories. This includes mostly binary two-class variables, e.g., diseased vs. healthy group, successful vs. unsuccessful treatment, recurrent vs. non-recurrent, survival vs. death after certain time point, and metastasis vs. relapse free

outcome. Among the wide-variety of classification schemes, discriminant models, which aim at learning a discriminative function to separate the two classes, are widely used. Discriminant models learn a discriminant function $L = g(x) = w^T x + w_0$, where w is the coefficients for each variable of x (as shown in Figure 2.4) for two-dimensional dataset x (x denotes the genomic variables here, but can also denote the clinical variables c and clinicogenomic variables z). Linear discriminant analysis (LDA) chooses the parameters $\mathbf{w}$ and w_0, such that the samples from the two classes are well-separated, maximizing the between class variances[141]. Sun et al. [67] used LDA [141] for combining the current clinical guidelines for breast cancer prognosis such as St. Gallen [47], [142] and NIH [143] with genomic information to predict the survival of breast cancer patients.

In addition to learning linear decision boundary, logistic regression [144],[141] learns the posterior probability of the outcome variable by a logistic function. Logistic regression is a generalized linear model that summarizes the contributions of all predictors into a single variable, which is fed into a sigmoid transfer function to produce the final predicted probability of outcome event y. Most of the clinicogenomic studies [58, 69] use a stepwise logistic regression model where each predictive variable is added successively to the model until the optimal model is achieved using criterion such as Akaikeś information criteria (AIC) [145]. In one such model, Beane et al. [69] combined the gene expression profiles of lung epithelial cells of potential lung cancer patients using bronchoscopy [146] with the clinical and demographic data to make better diagnostic decisions. Similarly, Stephenson et. al. [58] used step-wise logistic regression to predict the recurrence of prostate cancer after a radical prostatectomy (RP) using a well-established clinical marker called nomogram [147, 148, 149, 150, 151] that includes diagnostic variables such as PSA level, Gleason grade, margin status, and pathological stage along with gene expression data. For avoiding model over-fitting, a goodness of fit measure like AIC is used to select the optimal model.

On the other hand, SVM tries to learn the decision boundary in such a way that it maximizes the separation between the two classes (measured by the soft margin). Li et al [68] applied SVM with linear kernel to predict the survival of advanced-stage ovarian cancer after platinum-based Chemotherapy.

2.3.1.3 Nonlinear Models with Explicit Dimensionality Reduction

Although logistic regression and LDA provide simpler discriminant models, they are typically confined to finding linear decision boundaries only. Support vector machines (SVM) [152] can circumvent this problem to learn more generalized non-linear decision boundary, by utilizing the power of kernel machines. Kernel machines first map the original feature space into higher dimensions by a non-linear mapping function and then, linear SVM is applied in that higher dimensional space. Thus, learning a linear decision boundary in the higher dimensional space yields a non-linear decision boundary in the original space, which was used for developing an intermediate integration described earlier (Section 2.2).

Other types of non-linear models have also been applied for the integrative purpose. For example, tree based methods [25] are very popular because of two properties. First, they can be easily represented as classification rules which are more interpretable to clinicians and can be tested for inferring new domain knowledge. Second, these methods are based on recursive partitioning of all available samples into more homogeneous subgroups with respect to the binary class variable, therefore they can capture the non-linear interactions between the variables of a tree. Because of these two properties, [153] used a multi-step decision tree to find the interactions between 81 clinical covariates and genomic variables to predict the treatment of asthma patients. Other clinicogenomic studies include Pittman et al. 2004 [70, 24] for enhancing the prognostic power for breast cancer patients relative to long-term recurrence and Clarke et al, 2008 [71] for the survival prediction of ovarian cancer.

One problem with tree based methods is that there is no single optimal tree because they are built using heuristic search criteria. To circumvent this problem, all these clinicogenomic studies used ensemble learning [154],[155] and model averaging [156, 157, 158] techniques to generate a forest of trees and then, estimate the final prediction by taking the weighted average of the individual predictions of each tree. Such techniques not only boost the predictive performances by combining many weak learners (trees), but also provide a confidence interval for the prediction estimated from the individual models. This property is extremely useful in the context of an integrative clinicogenomic study for capturing the clinical uncertainties [30, 159] arising from different clinical processes such as variability of tissue processing, hybridization measures, small sample

size, and sample selection [70, 24].

Also, such model uncertainty may capture potential conflicting predictions either within or between the clinical and genomic factors, which can be very important for complex heterogeneous diseases. Similarly, mixture of expert (ME) [160] is another non-linear method that combines several expert trees using a convex weighted sum of all the outputs produced by them. However, each expert can be trained on different partitions of the input data with possible overlaps among them (soft split) as opposed to hard split of the data used by CART. Cao et al. [72] applied the ME method for integrating categorical clinical variables directly with continuous-valued gene expression data without any discretization. Furthermore, ME provided better results than random forest based approached used by [11].

2.3.1.4 Sparse Models with Implicit Dimensionality Reduction

Some clinicogenomic studies leverage the strength of sparse modeling technique to perform model development and feature selection in a single step by considering clinical and genomic data simultaneously. For example, Ma et al, 2007 [83] extended one such iterative boosting approach called Threshold Gradient Directed Regularization (TGDR [117]) into a more generalized framework (Cov-TGDR) for two generalized linear models: logistic regression and the Cox survival model. Cov-TGDR iteratively optimized the gradient of negative log-likelihood considering as the loss function. Moreover, in each iteration the component-wise gradient was updated only for only a few variables controlled by a regularization parameter. Thus, the components with lower gradient values are not updated in each iteration and these results in a sparse representation of the solution. Moreover, variable selection was performed separately for the two datasets to respect their individual properties of the data using two parameters L_1 and L_2 for the two datasets. Finally, this study applied the Cox proportional model for the survival of follicular lymphoma [161] and logistic regression for the binary prediction of the development metastasis of breast cancer ([162]).

2.3.1.5 Comparative Studies

van Vilet et al. [163] performed a recent comparative study of the two-step predictive models to systematically assess whether combining clinical and genomic data help

improve the prediction power of breast cancer. They consider three simple classifiers such as nearest mean classifier (NMC), Nave Bayes, Nearest neighbor, and two more complex classifiers such as SVM (similar to [77]) and tree based classifier. All of these models were developed in three different stages (early, intermediate and late) along with no integration (built on clinical and genomic variables). The original tree based classifiers proposed by [24] were modified for intermediate integration by restricting one dataset at the top node. For all these classifiers, integration improved the prediction power for breast cancer significantly, and simple classifiers performed better than complex classifiers (with NMC with OR-type late integration performing the best) which may be an effect of small sample size. Moreover, either late or intermediate strategies performs the best, which confirms the previous studies [78, 77]). Unlike the previous study by [162], this study found that clinical data has slightly better information than genomic data, which they believe that is mainly because of more comprehensive clinical features such as matrix information, central fibrosis, etc. Moreover, the genomic and clinical features obtained from this study perform better than the markers found by previous four studies in different cell lines [162, 164, 165, 166]. However, they did not assess the effect of different feature selection techniques in the model development stage. Bovelstad et al. [82] provided a methodological comparison of different dimensionality reduction techniques designed for Cox regression in survival studies. They covered both explicit and implicit dimensionality reduction approaches (Section 2.4) in their model development and they observed that modified ridge regression performed the best when applied to three different clinicogenomic datasets. However, they did not compare it to the Cov-TGDR methods.

2.3.1.6 Advantages and Disadvantages of the Predictive Models

The main advantage of predictive models is that they are easy to develop and simple from a methodological perspective. Any model that is applicable on either clinical or genomic data can be applied directly (for two-step approaches) or with minor modifications (for regularized methods) to the combined dataset. These models build unbiased models on clinical and genomic data sets without any prior information and bias towards any of the datasets being integrated. Therefore, the predictive model can test whether the datasets being integrated are complementary in nature based on the improvement of

the predictive power of the combined model over the individual models. However, the final clinicogenomic models may select a completely different set of clinical and genomic variables than those selected by independent models. Hence, comparing the predictive power of clinical and genomic features grossly in dataset level cannot assess directly how much additional power genomic features possess given the traditional clinical variables.

2.3.2 Assessing the Additional Prediction Power of Genomic Variables

Clinical variables are considered to be more reliable than genomic features in the clinicogenomic domain [21, 167]. The generic predictive models described in Section 3.1 do not consider the different amount of information and reliability present in clinical and genomic data and therefore, they often over-estimate the prognosis power of genomic data. Using synthetic datasets[SD11], Trutzner et al. [12] systematically showed that the genes selected by the generic predictive models are less reproducible in the independent test datasets and concluded that such over-estimation of the predictive power of genomic data happens because of small sample size combined with too many free parameters associated with large numbers of genes. To remove such over-estimation of genomic variables, care must be taken in both dimensionality reduction and model development phase of Strategy B-D. In this section, we will first discuss the proper design strategy and then the three different specific predictive modeling techniques for assessing additional power of genomic variables.

2.3.2.1 Design Strategy

Most generic clinicogenomic studies with explicit supervised dimensionality reduction techniques (Strategies B-C in Figure 2.2) perform both dimensionality reduction and predictive model development on the same training dataset, and finally assess the performance on a test or external validation dataset using CV. However, using the same training data for both dimensionality reduction and predictive modeling will bias the model towards high-dimensional genomic data, because the predictive models will favor genomic features that have already seen the class label during dimensionality reduction [167]. This problem can be solved by designing a two-fold CV framework for these two steps, which uses three separate datasets for dimensionality reduction, integrative model development and testing the additional performance of the integrative model

respectively. This framework has been evaluated as being very effective for assessing additional power of genomic variables by De Bin et al. [167]. Similarly, strategy D can also be fit into this two-fold CV framework: the inner one for estimating the parameters of model and the outer CV for assessing the additional performance.

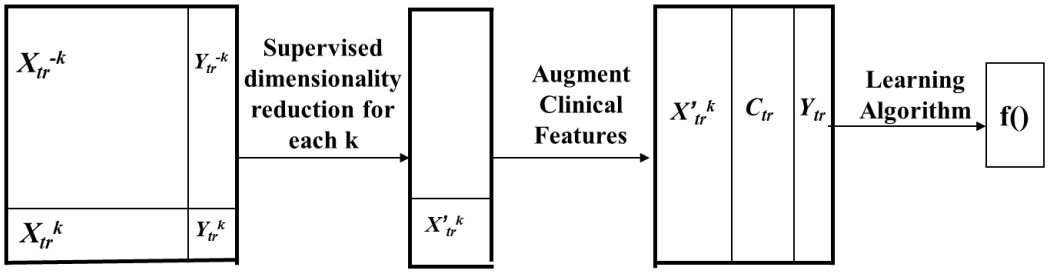


Figure 2.5: Schematic diagram of Pre-validation as suggested by Tibshirani et al. The dimensionality reduction step is repeated k times to get the full X'^{k}_{tr} matrix. Here X^{k}_{tr} represents the k-th part of training set. In particular, the available training data (X_{tr}) from the outer CV was further divided into two sets in each of the iteration of k-fold inner CV. One set (K-1 parts denoted by X^{-k}_{tr}) was used for the supervised dimensionality. Later, the same set of selected genes or feature creation rules was applied to the left-out k-th samples (X^{k}_{tr}) to create a new genomic feature (X'^{k}_{tr}). This process will be repeated for each k-th part of the data in rotation to get a new genomic feature for each of the sample (X'^{k}_{tr}).

2.3.2.2 Integrative Models based on Statistical Testing

An effective way to reduce the over-optimism of genomic variables during integrative modeling is to assess the additive performance of genes in a hypothesis testing framework by answering the question of "Do genomic variables boost the performance of models given the clinical variables?" in comparison to the null-hypothesis of 'no additional value'. Tibshirani et al. [73, 74] performed some seminal works for removing the bias towards genomic variables using a variant of above-mentioned two-fold CV framework (called pre-validation). One problem with such two-step CV frameworks is that the available data in each of the nested folds becomes too small. To fully utilize the potential of all available data, the first CV uses the training data for selecting the

most discriminative genomic features and then uses the test data to summarize the contribution of selected genes into a new fairer genomic score using approaches like PCA (See detailed description in Figure 2.5). In a second independent CV, the statistical significance of the newly obtained pre-validated genomic score was tested in a logistic regression model using either a standard statistical test [73] or empirically by using permutation testing [74]. In both cases, the pre-validated genomic score was less significant than the original overestimated genomic score without pre-validation in a landmark breast cancer study [162]. However, such statistical tests based on p-value do not directly assess the additional predictive power of genomic variables over clinical variables [167].

2.3.2.3 Integrative Predictive Model for Assessing Predictive Power of Genomic Variables

Several clinicogenomic studies develop predictive models by favoring the clinical variables [21, 23] in some way. The strategies described earlier (Strategies B-D in Figure 2.2) can be modified to develop such predictive models for assessing the additional power of genomic variables.

The studies with explicit dimensionality reductions on both clinical and genomic data (Strategy B) can further be modified to favor clinical variables. For example, De Bin et al. [23] favored the clinical variables over the genomic variables by forcing the number of genomic features selected during dimensionality reduction step to a small number (e.g., 25). Another popular technique is to modify strategy C as it becomes strategy E, -to include all clinical features as mandatory variables and then add the gene signature components successively as long as the prediction power is improved. In one such study, Boulesteix et al. [11] first used a pre-validation framework based on partial least squares [168] to reduce the genomic features, which were then successively added to the model as long as the prediction power was improved, as assessed by out-of-bag error [169].

Note that the favoring techniques that perform separate explicit dimensionality reduction on individual datasets (Strategy E, modified Strategy B) are not be able to remove the redundancy between clinical and genomic variables completely and thus they are only partially successful in assessing the additional prognosis power of genomic

features. For example, the selected genomic features, even after the pre-validation strategy, may be correlated with clinical variables, since the clinical variables are not taken into account while performing dimensionality reduction (Section 2.2.1). Moreover, some subtle contributions of genes can be missed by the dominant genomic features that are already correlated with the clinical variables. An alternative study design can check the additional power of genomic variables in the early stage of the dimensionality reduction step by adjusting them for clinical variables as described in Section 2.2.1 (Strategy F). The regularized models with implicit feature selection (Strategy D) (e.g., CoxBoost [116, 22] and (De Bin, Sauerbrei et al. 2014[23])) that favor clinical variables by mandatory inclusion in the objective function, are able to remove any types of redundancy between the two types of variables.

2.3.2.4 Additional power based on the residual of clinical variables

A more rigorous one-step strategy can be designed (Strategy G) where the model fully exploits the clinical data and fits the residual to the omics data, thus using omics data only if needed as proposed by Boulesteix et al. [23]. The main idea of the method is to include not only the clinical variables, but also the contribution of those clinical variables as mandatory (offset) variable in the model, so that genomic variables cannot influence the clinical contribution. More specifically, this method first fits a generalized linear model on the clinical variables only, and then the impact (measured by coefficients) of those clinical variables is fed into the final combined clinicogenomic model as a fixed offset [170] that is not changed during the iterative learning of the final integrative model. Note that the genes of the final integrative model can also be preselected based on their additive powers in an independent dimensionality reduction step [23] (Section 2.2.1). Finally, the integrated model be tested for statistical significance similar to [74] as performed by Globalboosttest [84] and/or be evaluated by the metrics for measuring prediction power [23]. In a later study [171], pre-validation testing and Globalboosttest were tested more rigorously by generating several synthetic datasets with different amounts of correlation between clinical and genomic markers. As expected, if the informative genes are highly correlated with the clinical variables, Globalboosttest is more conservative in selecting genomic features than pre-validation.

2.3.2.5 Future considerations for predictive models

Most predictive models focus only on improving predictive power rather than enhancing the interpretability of the model. For example, often the top most factors obtained from the clinicogenomic models are different than the biomarkers obtained from the individual models. As a result, it is difficult to conclude that the integrative models provide better markers than the individual models, although there may be a gain in prediction power. More systematic studies are required to compare combined clinicogenomic models with individual models, not only in term of prediction power, but also in terms the stability of the obtained biomarkers. Moreover, predictive models typically focus on assessing the additional predictive value of genomic variables besides the clinical variables assuming they are better-validated than genomic variables. However, this assumption may not be justified in the future as genomic data becomes more easily available and is validated more rigorously in multiple independent studies.

2.4 Biomarker Related Issues

Some integrative studies involving clinical and genomic variables are performed from the biomarker discovery perspective, i.e., the main purpose is to find the clinical and genomic factors that are associated with the disease phenotype. Note that, here we use the term biomarker as defined by World Health Organization (WHO) as any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease [172, 173]. Unlike predictive models, biomarker studies traditionally aim to assess the association of each factor with the disease phenotype directly on a particular population or subpopulation mostly using statistical tests in a hypothesis testing framework. However, the advent of large scale multi-modal datasets provides an opportunity to discover new data-driven hypotheses beyond the traditional hypotheses. For example, new hypothesis on the effectiveness of a marker on a particular subpopulation due to disease heterogeneity or the relationships present among diverse factors can be studied to understand the disease phenomenon in greater details. These hypothesis discovery techniques mostly use either semi-supervised techniques or modified unsupervised techniques such as clustering, association analysis and statistical methods to take the class label into account (Table 2.4). In the rest

of this section, we discuss two important topics of clinicogenomic biomarker discovery: population heterogeneity and finding relationships among biomarkers.

2.4.1 Heterogeneity

Most complex diseases are heterogeneous in nature, i.e., patients with a particular disease may form different subgroups and factors appropriate for one subgroup may not apply to another [28, 29]. Some potential reasons responsible for such disease heterogeneity are different pathways playing different roles in the same disease [3] and confounding factors such as age [174], ethnicity and race [30, 31], or genetic predisposition [33], which can lead to different degree of association between the other clinicogenomic factors and the disease. Finding such heterogeneous subgroups of population is even more important for clinicogenomic integrative studies, because the clinical and genomic features may measure very different aspects of disease phenomenon such as behavioral, environmental and biological factors. Each sample is not likely to be affected by each of these factors to the same extent. For example, genomic markers can affect a subset of samples and clinical markers may affect a different set of samples. Alternatively, genomic markers can affect different subgroups defined by clinical variables in different way. Clinicogenomic integrative studies thus can provide new opportunities to find insights into disease heterogeneity.

Most biomarker discovery techniques use full space model development techniques, i.e., the bio-signatures are generated based on how well they can discriminate all patients from the control population and thus cannot find the distinct subpopulations. Recently, several subspace based studies based on both univariate and multivariate methods aim to find such patient subgroups so that they can classify complex diseases such as cancer into more homogeneous sub-types (Table 2.4). Subspace clinicogenomic studies can be divided into two main groups. The first group of studies is more generic in the sense that they try to find all possible subgroups of samples corresponding to any clinical and genomic markers and build predictive model for each subgroup separately. The second group of studies is stricter and aims to leverage the complementary strengths of the two different datasets. In particular, clinical variables can be good at classifying a particular group of patients who cannot be classified well by genomic features and vice versa. These studies try to find only two distinct patient subgroups corresponding to

the two types of markers.

2.4.1.1 Generic subspace models

The generic subspace model aims to stratify the samples into a certain number of subgroups and to find clinical and genomic factors that are specific to a particular subgroup of samples. The easiest way to find all subgroups of samples associated with a set of biomarkers is to design a two-step study, where the subgroups of samples are identified after biomarker discovery occurs in the first phase [175]. For example, Schwarz et al. [176] first found all the clinicogenomic factors using univariate analysis for Schizophrenia patients and then built a two layer bi-partite graph representing all clinicogenomic biomarkers in one partition and all patients in the other partition, with an edge across the two layers representing the association between them. Finally, a network clustering technique called Markov Chain clustering [177] was used to find network modules containing homogeneous subgroups of Schizophrenia patients having a common abnormality in serum primary fatty acid.

An alternative approach based on association rule mining technique [178] can find the combinations of markers along with the corresponding patient subgroups in a single step. In particular, such patient subgroups are called patterns by this algorithm. Berlingerio et al., [179] used this technique along with a post-processing step to remove the non-discriminative patterns for discovering the demographic, pathological (e.g., hepatic cirrhosis) and genomic factors responsible for the allograft rejection of liver transplant. Another advantage of the association pattern mining or the network based approach described here is that they are non-parametric models that can capture non-linear interactions easily. This may be extremely useful for integrating heterogeneous types of data where the same kinds of model assumptions may not hold for all data types.

2.4.1.2 Biased subspace models

These studies find at least two distinct subgroups related to clinical and genomic variables respectively and then build a global predictive model utilizing the complementary strength of the subgroups. However, most clinicogenomic studies gave priority to clinical variables. These studies first find the subgroups based on the predictive power of the clinical variables. Subsequently, genomic variables (and possibly other left out clinical

variables from the first step) were used to find biomarkers for the rest of the patients that cannot be predicted well by clinical variables. Clinical variables that are used to stratify populations are included based on prior knowledge. Examples include estrogen receptor status [125, 60, 128], tumor grade and tumor status [180] for breast cancer; cirrhosis and vascular invasion of hepato-cellular carcinoma (HCC) [76] and a prognostic score [75] generated by a clinical software Adjuvant! Online [47, 48]. In the second stage, genomic data is incorporated for the subgroups of patients with poor clinical outcome. A few studies attempted to further subgroup those patients based on genomic variables using clustering techniques [75] or other unsupervised techniques [128]. Finally, the clinicogenomic variables were used for a particular subgroup of patients depending on the corresponding subgroups.

A few studies aim to build a final predictive model based on the subgroups of patients. Conceptually, the studies that stratify the initial patient groups based on clinical variables are similar to a decision tree where the topmost nodes are restricted to clinical variables. In fact, van Vilet et al [163] explicitly developed such a hybrid tree based on an intermediate approach where the prediction obtained from a clinical classifier was used for the topmost node. On the other hand, Obulkasim et al. [13] used a step-wise classification model to automatically determine the most effective subgroups for the molecular data given the clinical data and successively were included in the final model.

Alternatively, a few studies used a population stratification strategy to remove the well-known confounding factors directly [181]. Recently, a few other techniques [182, 174] based on survival analysis and association based techniques have been proposed to first remove the effects of the confounding clinical factors such as age and ethnicity, before performing any association study.

2.4.1.3 Generic models vs. biased subspace models

Generic models can find any number of homogeneous subgroups of patients that are available in the data. Therefore, generic subspace based approaches can be used for new hypothesis discovery, especially minor causal factors that are represented in very few samples and thus, overlooked by full-space models [183]. However, at the same time, a lot of spurious patterns and modules can be discovered due to the noise present in the data. Thus, the obtained patterns may not be statistically significant since they

are associated with only a few samples. Building classification models based on such small subgroups may be extremely difficult and may lead to overfitting. Therefore, the observed patterns require more robust validation both statistically and clinically before considering the patterns and modules as potential factors. On the other hand, biased subspace models are well-validated, since the goal of these studies is to improve classification. Therefore, they cannot find all subgroups of patients in the data and thus, cannot be used for hypothesis discovery.

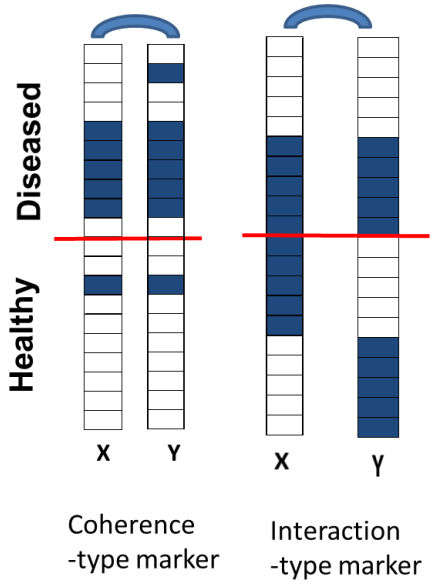


Figure 2.6: Coherence and Interaction type pattern between X and Y as binary variables. Here columns represent features and rows represent samples of two groups: disease (case) and healthy (control). A shaded cell in these data matrices indicates the presence of a feature for a subject and an integrated pattern contains features from both datasets. In (a), a coherence pattern is represented where both of the features (X and Y) are discriminative both individually and in together. Moreover there is a high correlation between the two features because they are represented in five samples together. In (b), an interactive pattern is trol where both features (X and Y) are present in four samples in cases but not in control.

2.4.2 Relationships among Markers

Integration of multiple biomedical datasets can provide a great opportunity to unravel relationships present among the diverse factors which cannot be obtained by looking at the datasets independently. Note that the predictive models described earlier focused on the overall relationship between clinical and genomic variables at the model level, but cannot shed light on the detailed relationships present among the individual markers. Different types of relationship have been studied in the biological domain, since they can reveal novel insights about the complexity of human disease. For example, for many diseases the nature versus nurture debate has been replaced with into nature and nurture studies that emphasize interactions between diverse genetic, clinical and environmental factors [34, 35, 36]. In fact, one of the most important types of relationship in biomarker discovery is interaction, which is when an integrated biomarker is more indicative of disease than its individual constituent factors. In fact, sometimes the individual (marginal) effects of each of the clinicogenomic factors on disease predisposition can be small and thus can remain undetected by most disease association techniques performed on individual datasets (Figure 2.6). However, interactions among individual factors may be responsible for increasing the risk of complex disease [37]. For example, neither a gene nor an environmental factor such as tobacco use may be significantly associated with lung cancer by itself, but together they can increase the risk significantly [30]. Interaction of genes with other types of features such as environments is believed to represent the missing hereditability [38].

Another potential relationship among diverse factors that can decipher novel biological relationship is coherence (correlation). Formally, coherence is when features of the individual datasets (i.e., two sets of clinical and genomic variables) are correlated across multiple samples (both in healthy and diseased population) and also have some prognostic value for the disease (Figure 2.6). This type of relationship is often useful in explaining possible causal relationships between two-types of data, and this can be observed as the correlation present among the mediator variables of the causal pathway [184]. Examples include clinical variables, such as pathological and behavioral effects, as the downstream effect of causal genomic features in cancer [185, 186, 165] or the downstream effects of the genetic perturbations that cause functional alteration of brain activity studied in neuroscience [41].

In this section, we will focus on computational methods for the two most studied types of relationships among markers: coherence (correlation) and interaction, which can explain the disease in more details and thus, can be used as potential biomarkers. The most prominent application of approaches to finding relationships among markers is in the neuroscience domain, where abundant clinico-pathological data are collected using MRI technology that measures functional [187] and structural brain activities [188]. Another prominent domain is the interaction between genetic and environmental factors [31].

2.4.2.1 Coherence relationships among factors

The most prevalent technique for finding coherence type of patterns from neuroscience dataset is based on multivariate blind-source separation (BSS) techniques [189, 187, 190, 188] (Refer to [85] for detailed review with application on brain images). In brief, blind source separation techniques such as independent component analysis (ICA) [191] and principal component analysis (PCA) [25] generate two matrices from a dataset: the modulation profile and the component maps, where component maps represent the sources by a linear combination of factors and the modulation profile denotes the association of each individual with those components.

These original BSS methods have been extended for integrating multiple datasets with the goal of finding relationships among multiple datasets. For example, joint ICA (jICA) [97, 99] and linked ICA [192] perform ICA after combining the two datasets into an augmented matrix, as in the early integration technique. However, it cannot differentiate the effect of each type of dataset on each sample. Alternatively, several statistical methods, such as canonical correlation analysis (CCA) [193] and parallel ICA (pICA) [98, 194] provide a more natural framework for data integration where the relationship between different components found from multiple datasets is defined as the inter-subject variability measured by correlation. Parallel ICA is a two-step approach where components are found by ICA from each dataset separately, such that they are maximally correlated across the datasets. Alternatively, CCA merges these two steps into a single optimization approach to find a linear transformation of each dataset (modulation profile) such that they have maximum correlation after transformation. CCA has been used extensively for various purposes [41]. An extension of CCA for

integrating more than two datasets called multi-set CCA (MCCA) [195, 196] has been applied for integrating fMRI, EEG and sMRI datasets [197]. Sparse CCA [198, 199, 200, 201] has also been proposed based on a regularization framework for simultaneously performing variable selection and reducing model overfitting issues that arise for high-dimensional data with small sample size. Another recent extension is a supervised technique called discriminative CCA (DCCA) [202], which can take the class labels into account while finding canonical components. Each of these multivariate models has their own assumptions [41]. Recently, an effort has been made to combine these two techniques to minimize assumptions [203, 204].

2.4.2.2 Interaction relationships among factors

As mentioned, interaction is defined as the situation where the joint effect of two or more markers is more indicative of disease than the individual constituent factors. Many types of models have been proposed to find gene-gene interaction, although multiplicative and additive models are the most popular [31]. Multiplicative models measure interaction as the departure from multiplication of the individual main effect of the two genes under consideration (often referred as epistasis) using statistical models such as regression. In contrast, additive interaction (e.g. synergy [54]) is assessed by the difference between the joint effect and the sum of the main effects. Please refer to [205, 206, 207] for detailed description of the popular methods and to [208] for different types of interactions in genomic studies. However, there are some unique issues that the above mentioned methods have to address when applied in clinicogenomic integration. First, the statistical power is a great challenge for any interaction study. For example, the number of samples required for detecting an interaction between two simple binary factors is four times higher than that of main effect [31]. For integrative study, it is even higher because of the increased dimensionalities leading to the increased number of hypotheses. Second, often a mixture of clinical, behavioral and environmental factors interacts with multiple genetic factors as many as up to ten SNPs [209]. The computational complexity and required samples for such higher-order interactions increase exponentially as more factors are added. Third, there may be both within and across dataset interactions, which creates unique challenge for integrative studies.

Most of the integrative interaction studies first preselect a few candidate genes using univariate methods depending on their relevance (e.g., using linkage disequilibrium for SNPs [210, 207]) or prior knowledge (e.g. pathways [48]) or based on the marginal effect of the individual factors [211, 48], and then test for interactions of those with other factors to avoid the problem of multiple hypothesis testing and poor statistical significance, especially for large-scale genome-wide studies. However, these studies will miss factors that have poor individual marginal effect but strong interactions with other factors [206, 48, 38]. Moreover, it is computationally hard to search for the higher-order interaction both within and across datasets. Although multivariate regularized regression models, tree methods, and pattern mining have been applied to find higher-order gene-gene interactions [206, 212, 38, 213], they have not been applied for integrative purposes. Furthermore, often interactions are confounded by the coherence type relations, since they are not considered simultaneously by most of the studies [38]. Recently, Dey et al. [214] proposed a pattern mining framework called PAMIN to find higher-order relationships from both within and across datasets. PAMIN first finds patterns from individual datasets to capture the available information separately and then combines these patterns to find integrated patterns (IPs) consisting of variables from multiple datasets. In addition, this study further characterized the IPs into two groups: coherent and interaction based on two different association measures. Using both synthetic data and a neuroscience dataset containing fMRI, sMRI and SNPs, the authors showed that PAMIN can discover interaction type patterns that competing approaches like CCA and discriminative CCA [202, 41] cannot find.

2.4.2.3 Other types of relationships

The genetic influences in complex diseases are often intricate exhibiting penetrance levels in between zero and one [206] (the probability of exhibiting phenotype, e.g., sensitivity of genetic influences in present of a moderator [38], protective versus recessive [31]). This creates a problem for defining relationships between genetic and other factors and their biological interpretation. Therefore, wide varieties of relationships can be defined between interaction and coherence, e.g., synergy, moderator, marginally interactive and coherence [211]. Moreover, causal relationships may be present in several datasets, for example, sometimes clinical variables such as pathological and behavioral effects can

be the downstream effect of causal genomic features. Causal makers can be extremely useful, i.e., drugs can be designed in a better way to target the original causal factor or preventive health care can be designed in a more intelligent way. Causal markers are generally correlated, but correlation does not imply causality [11, 215]. Better data mining techniques need to be developed for finding such diverse types of relationships from multiple datasets. Moreover, mixed types of relationships may be present, e.g., some of the variables may be correlated, while others may have causal relationships. Finding such mixed relationships creates further challenges for building computational biomarker discovery techniques.

2.5 Issues Cutting Across Predictive Models and Biomarker Discoveries

There are three issues that cut across both of predictive and biomarker discovery techniques: interpretability, validation and use of prior domain knowledge (Table 2.4). Here we provide a discussion of these issues along with direction to future research.

2.5.1 Interpretability

Interpretability of the obtained clinicogenomic models is a much desired property, which is critical for their acceptability in clinical practice [24]. However, most predictive models focus on improving the prediction power rather than interpretability. The features selected by the models may provide some interpretation of the models, especially if only a small number of features are selected. In addition, components obtained from feature extraction based techniques (e.g., PCA or PLS) may further reduce the interpretability of the model. Models based on biomarker discovery techniques can lead to more interpretable knowledge, since they tend to derive a more direct relationship between the target phenotype and a small number of features. Although coherence and interactive relationships provide some understanding of the dynamics of complex diseases, further research is required to understand the causal relationships present among the diverse clinicogenomic factors to understand the disease phenomenon in greater details.

2.5.2 Validation

Results obtained from individual studies spanning both categories require rigorous validation to assess their consistency across external datasets [65, 69, 21, 163, 167]. However, in most cases, data is scarce and therefore, the available dataset is used in a CV framework, whose test data mimics the independent validation data. However, the observed results from many clinicogenomic studies are not consistent. For example, van Vilet [163] found that clinical data has slightly better information than genomic data in contrast to previous breast cancer studies [162, 164, 165, 166]. One of the main hurdles for such validation is the unavailability of large-scale benchmark datasets covering wide ranges of clinical and genomic data [6], often due to the privacy related constraints [216]. Although some initial efforts for such data collection have been made [89, 217], they mostly contain genomic datasets. More comprehensive benchmark datasets containing both multiple genomic and clinical data collected from EHRs are needed and comparative studies need to be performed to validate clinicogenomic markers. Moreover, more rigorous prospective studies need to be designed to re[S13]move potential population selection bias and the recall bias of previous clinical and behavioral data [218, 31]. Moreover, any significant findings need to be finally tested for biological significance before its use as biomarkers.

2.5.3 Use of prior domain knowledge as model assumption

Utilizing existing medical knowledge in models as prior assumptions has the potential to enhance both the interpretability and validity of clinicogenomic models. Several types of domain knowledge have been incorporated into both predictive models and biomarker discovery techniques as summarized in Table 2.4. Other types of knowledge such as relationships between genes in terms of biological pathways [43] or medical knowledge about the relationship among diagnosis codes [44] or intervention plans [219] can also be incorporated into multiple stages of clinicogenomic model development. However, biasing the models too much based on prior knowledge may hinder the discovery of new knowledge. More systematic studies are required for deciding the optimum level of incorporation of domain knowledge into the models.

2.6 Conclusion

In spite of the great potential of clinicogenomic integration, it remains a hard problem that is still in a rudimentary phase. Most existing clinicogenomic studies consider only gene expression, clinical and pathological data. Other genomic datasets such as post-transcriptional modification, protein synthesis and epigenetic data need to be incorporated to get the complete understanding of cellular mechanism (see [220] for a recent such study). It is important to note that these genomic datasets are inherently related and thus, by considering such relationships one could drastically reduce the discovery of biologically spurious, albeit statistically significant associations. Similarly, electronic health records (EHR) contain rich medical information about patients treatment history, notes and diagnosis codes, where existing studies have mostly made use of demographic and pathological data. All these datasets will give rise to new challenges and research goals, for which novel computational techniques are yet to be designed. Furthermore, the focus of existing studies has been on generating a prognostic risk score using clinicogenomic features at a single point of time, often during the initial assessment of a disease. Advanced techniques are needed to make use of relationships present among clinical and genomic factors and their evolution over time.

The ultimate goal of integrating clinical and genomic datasets is to use it during every step of clinical decision making to enable personalized medicine. However, techniques surveyed in this paper only help discover potential hypotheses which need to be evaluated through independent randomized control trials before deploying them into clinical decision making as a complement to the existing medical knowledge gathered through evidence based guidelines. The effectiveness of such guidelines can now be re-analyzed or new guidelines can be designed as more complete knowledge regarding patients health emerges through the advancement of clinicogenomic studies, so that they can be used finally for personalized medicine.

Categories	Goals	Model Assumptions	Role of domain knowledge	Models mostly used
Generic Predictive Models	Assess whether integration of diverse datasets increase the prognostic power of the disease	None	None	Predictive Models
Predictive Models with Additional Values of Genomic Variables	Assessing the additional power of the genomic variables over clinical variables	The clinical variables are more reliable than the genomic variable	Clinical variables are treated especially into the global model development	Predictive Models with modifications
Generic Subspace Models	Finding subspace of population with different set of biomarkers	Clinicogenomic markers are specific to a subgroup of patients	None	Unsupervised Clustering, pattern mining
Biased subspace Models	Finding subpopulation and building models with them	Clinical variables are useful to stratify the patient population	Clinical variables are used to stratify initial samples and finally build the local models	Unsupervised clustering or semi-supervised
Relationship Finding	Finding relationship among markers	There are relationships among different types of markers	Biological relationships are used for generating new hypothesis	Unsupervised multivariate statistical models

Table 2.4: Models and assumptions of the clinicogenomic studies.

Chapter 3

Finding Relationships across Datasets

3.1 Introduction

Integration of multiple biomedical datasets can provide a great opportunity to unravel relationships present among the diverse factors which cannot be obtained by looking at the datasets independently. Different types of relationship have been studied in the biological domain, since they can reveal novel insights about the complexity of human disease. One of the most important types of relationship in biomarker discovery is interaction, when an integrated biomarker is more indicative of the disease than its individual constituent factors. In fact, sometimes the individual (marginal) effects of each of the clinicogenomic factors on disease predisposition can be small and thus can remain undetected by most disease association techniques performed on individual datasets. However, interactions among those individual factors may be responsible for increasing the risk of complex disease [221]. For example, neither a gene nor an environmental factor such as tobacco use may be significantly associated with lung cancer by itself, but together they can increase the risk significantly [222]. Interaction of genes with other types of features such as environments is believed to represent the missing hereditability [223].

Another potential relationship among diverse factors that can decipher novel biological relationship is the coherence (correlation) [224]. Formally, coherence is defined

when features of the individual datasets (i.e., two sets of clinical and genomic variables) are correlated across multiple samples (both in healthy and diseased population) and also have prognostic value for the disease (Figure 3.1). This type of relationship is often useful in explaining possible causal relationships, and can be observed as the correlation present among two types of data. Examples include clinical variables, such as pathological effects as the downstream effect of causal genomic features in cancer [225, 226, 227] or the downstream effects of the genetic perturbations that cause functional alteration of brain activity studied in neuroscience [228].

In this section, I will focus on the computational methods for the two most studied types of relationships among markers: coherence (correlation) and interaction, which can explain the disease in more details and thus, can be used as potential biomarkers [224]. The most prominent application of approaches to finding relationships among markers is in the neuroscience domain, where abundant clinico-pathological data are collected using EEG [229] and magnetic resonance image (MRI) technology that measures functional [230], structural brain activities [231]. Another prominent domain is the interaction between genetic and environmental factors [232]. I first define this two different types of relationships formally and then discuss the computational methods to find them.

3.2 Related Work

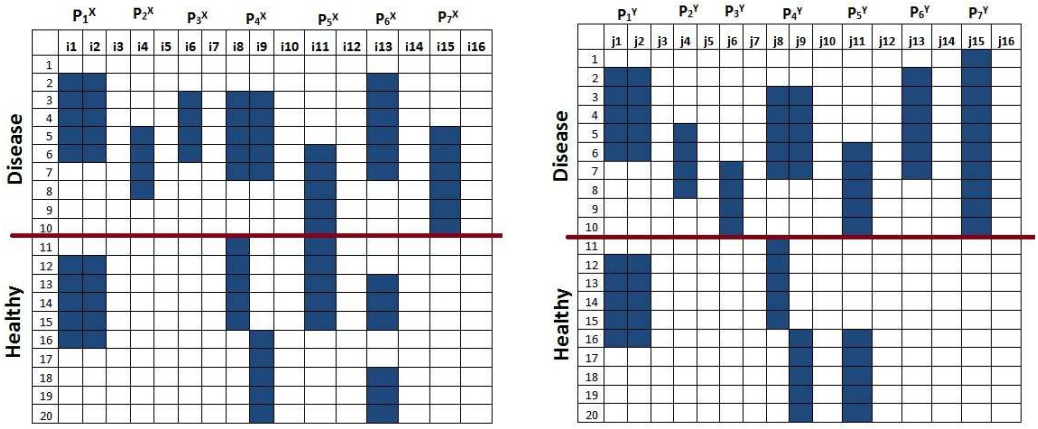
Multivariate statistical models including blind source separation techniques such as variants of independent component analysis (ICA) [233] and canonical correlation analysis (CCA) [228] have been developed to find the relationship between variables across the datasets directly. In general, these models look for components in each of the available datasets such that those components have some relationships across multiple datasets. The original ICA framework, which cannot combine multiple datasets directly, has been extended in several ways for integration purpose. Examples include joint ICA [234], parallel ICA [233], and group ICA [235]. On the other hand, canonical correlation analysis (CCA) and its extensions [236] provide a natural framework for integration where the relationship between different components found from multiple datasets is defined in terms of the inter-subject variabilities. It has been also shown that CCA has fewer

model assumptions than ICA based techniques [228]. Therefore, CCA has become very popular for integrating datasets, including neuroscience datasets [228] and various biological datasets [237]. The original CCA is unsupervised in nature and thus, the discrimination power of the obtained components are assessed in a post-processing step. Discriminative CCA (DCCA) [236] has recently been proposed to combine these two steps to find discriminative components directly. Multi-set CCA, a generalized CCA that can integrate more than two datasets, has recently been applied for integrating functional MRI(fMRI), EEG and structural data [238].

Note that these approaches can only find biomarkers whose individual features are discriminative and correlated. However, as we will show in this paper, these techniques are unable to find integrative biomarkers that consist of features that are not correlated or individually discriminative, but can distinguish between the healthy and disease groups when taken together. Such biomarkers, referred to in the rest of the paper as *interaction-type* integrative biomarkers, are important due to their ability to combine complementary information from different data sets. Several other statistical models [239, 240] that look for such interactions are unable to find higher-order interactions. Alternative techniques, such as multi-factorial dimension reduction (MDR) [241], aim to achieve this goal in a brute-force manner and so are only suitable for small datasets or for selected markers that already have sufficient individual effects. Furthermore, most of these techniques are unable to find different subgroups of population associated with those interactions due to disease heterogeneity.

3.3 Problem Formulation

In this section, we define the types of integrated patterns that are relevant to the biomarker discovery problem. Consider two binary matrices X and Y, representing two case-control datasets as shown in Figure 3.1(a) and Figure 3.1(b), respectively. Each of the two datasets has 16 features (represented by columns) and 20 subjects (represented by rows) with equal representations from healthy and diseased groups (separated by a horizontal line). A shaded cell in these data matrices indicates the presence of a feature i.e., has a value 1 for the corresponding subject, while a white cell indicates that a feature has a value 0 for a corresponding subject. We define a *pattern* as a combination of a



(a) Synthetic dataset X with inserted patterns (b) Synthetic dataset Y with inserted patterns

Figure 3.1: Two synthetic datasets.

subset of features and a subset of subjects (also referred to as supporting subjects) in which all features in the subset are present. X and Y have a set of 7 patterns represented as $\{P_i^X\}_{i=1}^7$ and $\{P_i^Y\}_{i=1}^7$, respectively. P_2^X pattern is supported by 4 subjects in the disease group and none in the healthy group. Similarly, $P_4^X = \{i8, i9\}$ is supported by 5 samples from the disease group but none in the healthy group. Note that a pattern can consist of single feature such as P_2^X or multiple features such as P_4^X.

A number of measures exist to quantify the interestingness of a discriminative pattern [242]. We use *diffsup* to measure the discrimination power of a pattern [243, 244]. *Diffsup* is defined as the absolute difference between the fraction of samples supported by a pattern in two classes and is more formally defined in Section 3.5.

We will use IP_{ij} to denote an *integrated pattern (IP)* as the union of the features present in pattern P_i^X (found in X) and the features in pattern P_j^Y (found in Y). The support of the *integrated pattern* is the fraction of subjects that are common in both of the constituent patterns. For example, the support of IP_{22} is 0.4 and 0 in diseased and healthy group, respectively. On the other hand, the support of IP_{33} in diseased group is 0 since there is no common subject between constituent patterns P_3^X and P_3^Y.

An IP is useful as a potential biomarker only if it is has more supporting subjects in one group than the other (i.e., higher discrimination power). For example, IP_{11} is not interesting, since both the IP and the constituent patterns (P_1^X and P_1^Y) in individual datasets have same number of supporting subjects (0.5) in both classes ($diffsup = 0$).

Similarly, IP_{33} is not discriminative since its support in both classes is zero, although its constituent patterns (P_3^X and P_3^Y) are discriminative with $diffsup = 0.4$ before integration. On the other hand, IP_{22} is interesting since both IP_{22} and its constituent patterns have more supporting subjects in the disease group with $diffsup = 0.4$. In contrast, integrated pattern IP_{55} demonstrates a situation that is totally opposite to that of IP_{33}. Here, the constituent patterns P_5^X and P_5^Y have same number of supporting subjects in both the groups, leading to a $diffsup = 0$. However, together they cover the same disease subjects but different healthy subjects. Thus, IP_{55} has more supporting subjects in the disease group. All these IPs are summarized in Table 3.1 based on their discrimination power before and after integration. Among the four IPs described above, only IP_{22} and IP_{55} are discriminative after integration and can serve as potential biomarkers which will be referred to as discriminative IPs in this paper.

Discriminative Before	DisriminativeAfter	IP
No	No	IP_{11}
Yes	No	IP_{33}
Yes	Yes	IP_{22}
No	Yes	IP_{55}

Table 3.1: Different types of IPs based on discrimination power before and after integration.

We now describe two subtypes of integrated patterns that are of interest in biomarker discovery: *coherence-type* IP and *interaction-type* IP. More specifically, a *coherence-type IP* has same degree of discrimination as that of the constituent individual patterns (e.g., IP_{22}). These patterns are interesting for biomarker discovery in scenarios when the upstream effects, such as genetic perturbations, can be validated by downstream effects, such as changes in protein abundance in metabolomics data [245]. Thus, it can potentially elucidate an underlying causal/cascade relationship among different biomarkers coming from individual datasets. In contrast, we also define an *interaction-type IP* as one whose constituent individual patterns have a degree of discrimination that is lower than that of the integrative pattern (e.g., IP_{55}).

Note that the coherence and interaction type relations can be present both within and across the datasets. Lastly, we will demonstrate one such pattern. $IP_{44} = \{P_4^X, P_4^Y\}$ represents a special type of discriminative IP where the individual patterns

$P_4^X = \{i8, i9\}$ and $P_4^Y = \{j8, j9\}$ capture the within dataset interaction for datasets X and Y, respectively. Therefore, it is a potential biomarker and also an *coherence-type integrated pattern* similar to IP_{22}.

3.4 Preliminaries and definition of measures

Let D be a dataset with a set of m items (binary variables), $I = \{i_1, i_2, ..., i_m\}$, and n samples from two classes S^+ and S^-. Each sample can be represented as a vector $(\vec{x_i}, y_i)$ for $i = [1, ..., n]$, where $\vec{x_i} \subseteq I$ is a set of items and $y_i \in \{S^+, S^-\}$. The two sets of samples that respectively belong to the two classes S^+ and S^- are denoted by D^+ and D^-. We define a pattern as $P^D = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ of dataset D, where l is the length of the pattern and $\alpha_i \in I$, $\forall i \in \{1...l\}$. The set of samples from the two classes that contain P are denoted by $D_P^+ \subseteq D^+$ and $D_P^- \subseteq D^-$. The ratio of the samples covered by P in a particular class to the total samples of that class is defined as *RelSup*. For example, $RelSup^+(P^D) = \frac{|D_P^+|}{|D^+|}$ for the positive class S^+. The absolute difference of the relative support of P between two classes is defined to be diffsup as in [243].

$$diffsup(P^D) = |RelSup^+(P^D) - RelSup^-(P^D)| \tag{3.1}$$

The above mentioned $diffsup$ measure can be used to assess the discriminative support of both patterns from a single dataset and IPs from multiple datasets. Furthermore, we want to differentiate between the coherence-type and interaction-type IPs. To quantify interaction, we want to measure the minimum increase in the discrimination power of an integrated pattern from those of any subsets of features. More formally, a measure called *improvement* is defined as below.

Definition. For a pattern $P^D = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ in a dataset D, the improvement is defined as

$$improvement(P^D) = diffsup(P^D) - \max_{q^D \subset P^D}(diffsup(q^D)) \tag{3.2}$$

A pattern P^D is called an *interaction*-type pattern if its $diffsup(P^D) > \beta$ and $improvement(P^D) > \gamma$, for parameters $\beta > 0$ and $\gamma > 0$. We will use this measure to find both within dataset interactions (interaction patterns) and across dataset interactions (interaction-type IPs).

Similarly, we aim to measure the association of the constituent patterns of an IP to find the *coherence-type* IP. However, this is more challenging for two reasons. First,

each dataset has its own properties and a different amount of information, and thus, having same amount of support may not indicate the same association for each dataset. Consider two IPs: $IP_{22} =< P_2^X, P_2^Y >$ and $IP_{77} =< P_7^X, P_7^Y >$ with same support in disease group as shown in Figure 3.1. The first IP has a true association, but the second one is more likely to occur by random chance since the individual patterns themselves have high support. Second, the two patterns may contribute unequally to the joint association. For example, for IP_{77}, P_7^X contributes more than P_7^Y, since the former has lower support than the latter ($conf(P_7^X \rightarrow P_7^Y) = 0.66$, but $conf(P_7^Y \rightarrow P_7^X) = 0.5$). From the wide variety of interestingness measures [246], we select IS for assessing associations because it has two important properties. First, it can measure the association relative to the baseline supports of constituent markers in each dataset (expected association). Second, the IS measure combines the contribution of each individual marker towards the joint associations from the constituent markers using the geometric mean [8]. Thus, if the contribution from any of the two directions(confidence measure) is low, then the IS measure is also low.

$$IS(A,B) = \frac{relsup^+(A,B)}{\sqrt{relsup^+(A) \times relsup^+(B)}} \tag{3.3}$$

A pattern P^D is called an *coherence*-type pattern if $diffsup(P^D) > \beta$ and $IS(P^D) > \alpha$, for parameters $\beta > 0$ and $\alpha > 0$.

In the following section, we propose an approach to discover both interaction type IPs and coherence type IPs that pass a user provided discriminative power threshold from a given a set of different datasets.

3.5 An integration framework

In this section, we describe a generic pattern mining based integration (PAMIN) framework to find the two types of IPs from multiple diverse datasets. A straightforward way to analyze multiple datasets is to augment them into a common matrix format, and then apply discriminative pattern mining on the combined dataset. Unfortunately, differences in the nature of the data, such as differences in format, semantics, type of variables, dimensionalities and the amount of information present in each dataset, can create numerous challenges to taking such an approach. In addition, we will find many patterns

that consist of variables from only one dataset which are not interesting for integration and thus, have to be filtered in a post-processing step. Furthermore, such an approach will generate too many potential hypotheses and thus the statistical significance of the IPs will be poor due to the increased Type-I error. Generally, augmenting such disparate datasets limits our ability to apply the most relevant pattern finding techniques, muddles the underlying semantics of the data, increases computational complexity, and reduces the statistical significance of the discovered patterns.

To address the above challenges, we propose a two-step framework (Refer to [214] for detailed algorithm). The idea is to first find the discriminative patterns from individual datasets, taking into account the individual characteristics of each dataset, and then combine them into integrated patterns that can distinguish the disease group from the healthy population (Figure 3.2).

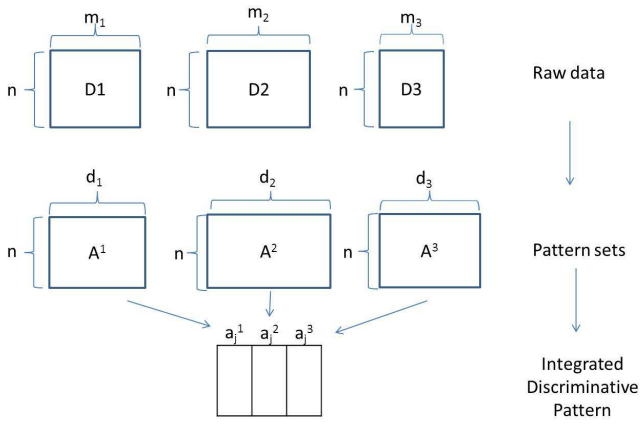


Figure 3.2: The generic two-step framework for finding integrated discriminative patterns.

3.5.1 Finding patterns from individual datasets.

In this step, we will generate all discriminative patterns from each dataset being integrated based on the *diffsup* score. Among different types of discriminative patterns described in [247], we will only retain interaction type discriminative patterns that show within dataset interactions, e.g., $IP_{44} =< \{P_4^X, P_4^Y\} >$ in Figure 3.1. We use an approach similar to that used in [248] to find all interaction patterns. More specifically, we

first mine for the discriminative patterns with $diffsup > \delta$ using (SupMaxPair) SMP [244] and then look at the improvement score of the discriminative patterns with non-negative scores. However, the improvement score is not anti-monotonic in nature [248]. This can potentially lead to many missed interactions present across the datasets. For example, a singleton variable from a particular dataset may not be discriminative, but may have interactions with the individual variables or patterns found in other datasets. Therefore, we also retain all the singletons along with the interaction patterns obtained from each datasets. We denote the set of all patterns found from Dataset D as a pattern set, $PS(D) = \{P_j^D\}_{j=1}^{d} \bigcup I$, where d is the number of patterns found from dataset D and I is the set of all items associated with D.

3.5.2 Finding integrated patterns using the patterns from multiple datasets

In the second step, we combine the individual patterns found from individual datasets to obtain the final integrated patterns. Suppose we have K heterogeneous datasets $\{D_k\}_{k=1}^{K}$ collected for the same set of n samples. Let d_k be the number of patterns found from dataset D_k in step 1, i.e., $|PS(D_k)| = d_k$. For each set of patterns found from the k^{th} dataset, we define a binary matrix A^k with dimensions $n \times d_k$ for ease of further discussion. Each entry of the binary matrix $\{A^k\}_{ij}$ represents whether the pattern $(P_j^{D_k})$ covers the sample i, for $i = 1...n$ and $j = 1...d_k$. Thus, each column of this matrix, (a_j^k) corresponds to the j^{th} pattern of k^{th} dataset, for $j \in \{1, ..., d_k\}$. We will look for the associations and interactions of these patterns across the datasets to obtain an *integrated pattern*,IP, of length l, which is defined by $IP = \bigcup_{t=1}^{l} P_j^{D_t}$, where $P_j^{D_t} \in PS(D_t)$ and $t \in [1...K]$. Note that an integrated pattern might only contain patterns from each of the datasets, so $2 \leq l \leq K$.

As discussed earlier, the first criteria of an integrated pattern to be considered as interesting is that it should be discriminative, i.e., $diffsup(IP) > \beta$. However, the $diffsup$ measure only provides a way for assessing the discrimination power of an IP, but not the discrimination power of the constituent patterns of that IP. To further explain this scenario, consider three IPs: $IP_{22} = \{P_2^X, P_2^Y\}$, $IP_{55} = \{P_5^X, P_5^Y\}$ and $IP_{66} = \{P_6^X, P_6^Y\}$ shown in Figure 3.1. All of these IPs will have a high diffsup score. Among them, the individual patterns $P_6^X = \{i13\}$ and $P_6^Y = \{i13\}$ of IP_{66} have highly

skewed discrimination power (i.e., 0 and 0.6, respectively). These types of IPs are not interesting because the discrimination power is coming mainly from one constituent pattern. In contrast, for IP_{22} and IP_{55}, the discrimination power of the constituent patterns are more balanced. Thus, we want to make sure that the discriminative power of each individual patterns of an integrated pattern is balanced, rather than skewed. This observation motivates our use of a heuristic measure called *balance score*, which is defined below.

Definition Balance score: For an integrated pattern $IP = \bigcup_{t=1}^{l} P_j^{D_t}$ of length l, where $P_j^{D_t} \in PS(D_t)$ and $t \in [1...K]$, we can represent the diffsups of each pattern $P_j^{D_t}$ as a ***diffsup vector***

$\vec{DV(IS)} =< diffsup(P_j^{D_1}), ..., diffsup(P_j^{D_t}), ..., diffsup(P_j^{D_l}) >$

The *balance score(bs)* is then defined as the cosine similarity ($\cos\theta$, where θ is the angle) between the perfectly balanced vector $\vec{I} =< 1, ..., 1 >$ of length l and $\vec{DV(IS)}$. More formally,

$$bs(IP) = \frac{\sum_{k=1}^{K} diffsup(a_j^k)}{\sqrt{n \times \sum_{k=1}^{K} (diffsup(a_j^k))^2}} \tag{3.4}$$

For any IP, $0 \leq bs \leq 1$. The larger the bs score, the more balanced the diffsups of the constituent patterns.

Example: The $IP_{66} =< P_6^X, P_6^Y >$ has a balance score = 0.7, while $bs(IP_{22}) = bs(IP_{44}) = 1$.

To summarize, we want the IP to be both discriminative and have balanced discrimination power in their constituent patterns. Hence, we use both diffsup and the balance score (bs), $DBS(IP) = diffsup(IP) * bs(IP)$ for assessing the discriminative power of final integrated patterns. Once we find all the discriminative IPs, we aim to further differentiate between the two types of IPs: *coherence-type* and *interaction-type* IP. In particular, we use the *improvement* and IS score to find interaction-type and coherence-type IPs, respectively.

We will use three different pruning criteria to search for such integrated patterns efficiently. This first pruning will be performed based on the anti-monotonicity property of IS measure. We generalize the original definition of the IS measure for integrated patterns from pairs to higher-order integrated pattern for $l \geq 2$, in such a way that the measure becomes anti-monotonic. More formally, the IS measure for an IP with length

$l > 2$ can be defined as the minimum of the IS measures of all pairwise subsets of IP. The anti-monotonicity property directly follows from the nature of the *min* function.

$$IS(IP) = \min_{r,s \in \{1,\ldots,l\}, i \in \{1,\ldots,d_r\}, j \in \{1,\ldots,d_s\}} IS(a_i^r, a_j^s) \tag{3.5}$$

Thus, if an IP of length 2 has $IS(IP) < \alpha$, then any superset of IP with $l > 2$ can be easily pruned.

Another level of pruning is done based on an alternative diffsup formulation suggested by Fang et al. [244], which makes the diffsup measure anti-monotonic. The last pruning criteria for making the algorithm efficient in this stage is to use the fact that diffsup is an upper bound of $DBS(IP)$.

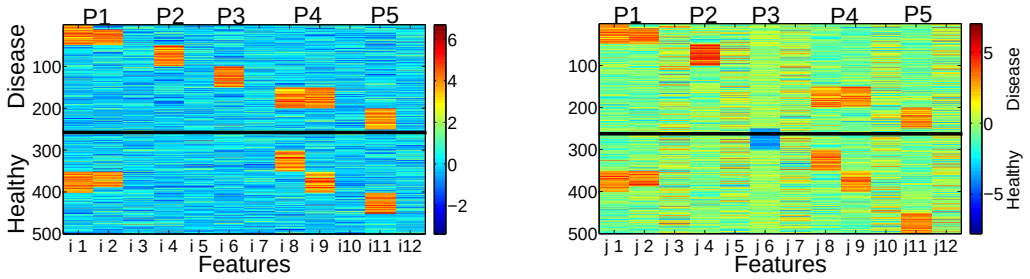
Lemma: For any IP, $DBS(IP) \leq diffsup(IP)$.

Proof: It follows from the definition of DBS and the observation that $0 \leq BS(IP) \leq 1$.

Thus, if $diffsup(IP) < \beta$, then we can prune the IP rather than calculating the DBS.

3.6 Experiments and Results

In this section, we present results for both synthetic datasets and real datasets. We will compare our proposed approach with CCA and its recent extension called DCCA, which finds discriminative components directly. One difficulty in comparing CCA with our approach is that CCA and DCCA combine original features into components by taking a linear combination of the features. Therefore, they do not provide a direct mapping of the discriminant components back to the original feature space and the true integrated patterns. We assess the activation level of each original feature for the discriminative canonical component by looking at the coefficients of component maps. More specifically, for each component, we compute the Z-score of the activation levels and then select the variables that have high z-scores [238]. DCCA finds only one discriminant component because of its restriction of number of components being less than number of classes.



(a) Dataset X (b) Dataset Y

Figure 3.3: Two datasets containing different types of patterns of interest(Best seen in color).

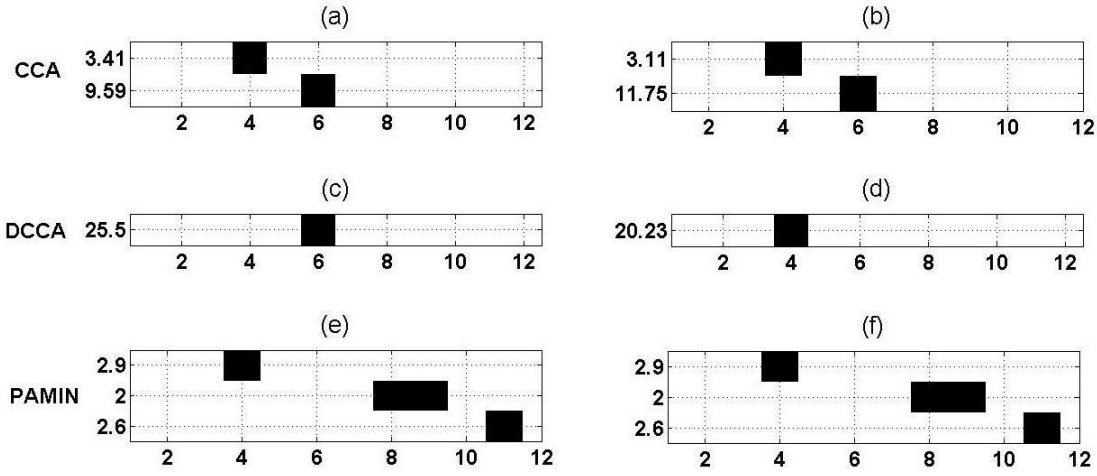


Figure 3.4: Comparison among CCA, DCCA and PAMIN based on the integrated patterns (IPs) from the dataset represented in Figure 3.3. Subfigures (a) and (b) represent the two IPs obtained by CCA from dataset X and Y, respectively. Similarly, subfigures (c-d) and (e-f) represent the IPs obtained by DCCA and PAMIN, respectively.

3.6.1 Synthetic datasets.

We generated two synthetic real-valued datasets X and Y (Figure 3.3) that contain features of the type similar to those shown in Figure 3.1 (a) and (b), respectively. For dataset X, we created real valued features $i_1 - i_{12}$ such that they reflect the nature of the features shown in Figure 3.1(a). Note that these features represent patterns $P_1^X - P_5^X$. When a feature i_l is present in a subject k, we generated the kl^{th} element

in matrix X (X_{kl}) from a normal distribution with $\mu = 2$ and $\sigma = 1$, and when i_l is not present in a subject k we generated the kl^{th} value in X from a normal distribution from $\mu = 0$ and $\sigma = 1$. We also added white noise to every element in the matrix using $X_{ij} = (0.8)X_{ij} + (0.2)n_{ij}$, where the value n_{ij} is the white noise generated from a normal distribution with $\mu = 0.1$ and $\sigma = 1$. We used a similar process for generating dataset Y, with an exception that for the feature, i_6, the values for subject in which the feature is present were generated from a normal distribution with $\mu = -2$ and $\sigma = 1$. Due to this, the correlation between i_6 in dataset X and i_6 in dataset Y is approximately 0.6. We will discuss the impact of this on the competing methods in the following paragraph. Of all the combinations of patterns imputed in dataset X and Y, the discriminative IPs (as discussed in section 3.3) are IP_{22}, IP_{44} and IP_{55}. Furthermore, IP_{22} and IP_{44} are coherence-type IPs and IP_{55} is an interaction-type IPs. An ideal integration framework for biomarker discovery is expected to discover all of the three discriminative IPs.

Figure 3.4 illustrates the components obtained from the dataset represented in Figure 3.3 using CCA, DCCA and PAMIN. In subfigures of Figure 3.4, each row represents a component and each column represents the variables in the original datasets, X and Y, respectively. Each entry of these component maps represents the activation profile or contribution of the variable to that component (binarized based on $Z - score >= 2$). The y-axis labels in these subfigures represent the discrimination power of these components in terms of their t-test (-log P value) on the modulation profile (See [238] for details).

We now compare the components and IPs discovered by the CCA, DCCA and PAMIN approaches. Figures 3.4(a) and 3.4(b) represent the discriminative components obtained from CCA from datasets X and Y, representatively. Each row of this figure denotes a component and each column denotes the features selected from the corresponding datasets. For example, the first row in 3.4(a) and (b) indicates that features (i_4 in X and i_4 in Y) pertaining to IP_{22} are selected. Similarly, the second row in 3.4(a) and (b) indicates that features pertaining to IP_{33} are selected. Therefore, CCA discovers one coherence-type IP and one non-discriminative IP. CCA is naturally expected to discover the coherence-type IP IP_{22} that has higher correlation across the two datasets. The other coherence type IP IP_{44} was not discovered by CCA because it was formed using within dataset interaction patterns (interaction between $< i_8, i_9 >$ ($< j_8, j_9 >$)

in dataset X (Y)) in individual datasets that have poor correlation between individual features. Owing to the same reason, the interaction-type IP IP_{55} was not discovered by CCA. On the other hand, CCA discovered IP_{33}, which is not a discriminative IP, because the features i_6 in X and i_6 in Y have a reasonably high correlation of **0.6**. Overall, CCA was able to discover one coherence-type IP(IP_{22}) and one non-discriminative IP (IP_{33}) as a false positive. However, it cannot find either within dataset interaction (IP_{44}) and across dataset interaction (IP_{55}).

Figure 3.4(c-d) represents the discriminative component obtained from DCCA. Similar to CCA, DCCA can find the coherence-type IP. Note that DCCA finds only one component, which is a combination of the features of IP_{22} and IP_{33}. On the other hand, our proposed PAMIN framework can find all coherence and interaction type IPs as shown in Figure 3.4(e-f).

Based on these observations, we conclude that CCA and DCCA are unable to find interaction-type IPs such as IP_{44} and IP_{55} and may find false positive coherence-type IPs, such as IP_{33}, where there is a correlation structure among the samples that does not support the corresponding patterns.

3.6.2 Neuroscience data.

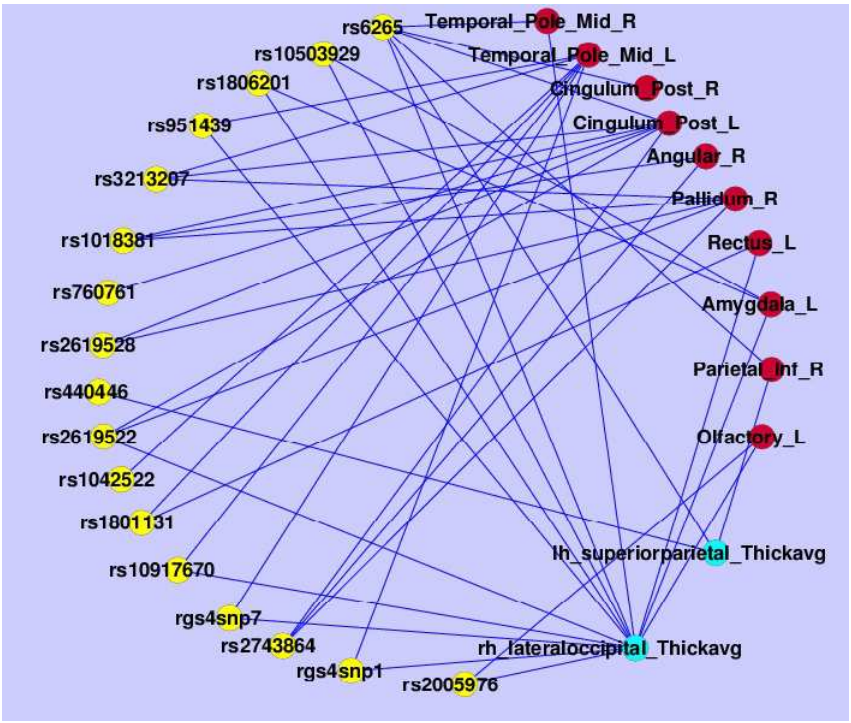


Figure 3.5: Coherence IPs obtained from three datasets: fMRI(red), SNP(yellow) and sMRI(green) for $IS > 0.6$ and $FDR < 0.1$

Here we demonstrate the applicability of our integration method on a data set with 76 schizophrenia cases and 94 controls. We integrated three different types of datasets: functional MRI (fMRI) [249], structural MRI (sMRI) cortical thickness and SNP data. fMRI data contains mean correlation for each of the 90 brain regions with respect to all of the other 89 brain regions [250]. The sMRI data contains thicknesses of 70 brain regions. The SNP data contains 162 preselected SNPs that are suspected to be associated with schizophrenia. For each of the fMRI and thickness features we create two binary features such that one feature has a value 1 when the original feature has a value in the highest 30% of its values and 0 otherwise; and the other binary feature has a value 1 when the original feature has a value in the lowest 30% of its values and 0 otherwise. We apply both PAMIN and DCCA on these datasets. DCCA finds one discriminative component for two class datasets. For PAMIN, we explored several

values of beta with [0.15, 0.2, 0.25, 0.3]. We then find two sub-types of IPs: coherence-type and interaction-type based on two parameters α and γ, respectively. The search space for pattern mining is exponential, and thus generates numerous hypothesis that can potentially lead to the increased type I error [251]. In particular, we randomized the class labels of each sample and then repeated the whole pattern mining procedure using the same parameter settings (α, β, γ) for 1000 runs, and then computed the false discovery rate of the DBS scores of the obtained discriminative integrated patterns in comparison to the randomized versions. We chose the thresholds of $\alpha = 0.6$, $\beta = 0.15$ and $\gamma = 0.1$, because the resultant IPs have an FDR value $<= 0.1$. After thresholding, we found 834 IPs of which 305 are coherence-type IPs and 5 are interaction-type IPs.

All the experiments presented here were run on a Linux machine with 8 Intel(R) Xeon(R) CPUs with 2.60 GHz and 16 GB memory. The algorithm took approximately 10 minutes for the neuroscience datasets with 758 variables from three different modalities. We used a high-performance parallel computing machine with 16 nodes to perform the 1000 random permutations for computing the statistical significance.

Coherence-type IPs: Figure 3.5 represents a three-partite graph depicting the global view of all 305 coherence integrated patterns (IPs) with $\alpha = 0.6$. Each node represents one feature and an edge between two features represents the presence of an coherence type pattern between them. The red, yellow and cyan nodes represent features from the fMRI, SNP and sMRI datasets, respectively. Most of the constituent associations are coming from SNP and fMRI regions. The topmost nodes in this network are the Temporal and Cingulum regions, which have been reported previously in [250] and [252], respectively. Note that, DCCA provides one discriminative component with the list of features selected from the three datasets only, but does not describe how those features are related among themselves. We found lots of overlap among the features selected by CCA based technique and PAMIN, which suggests that these features (spanning multiple datasets) are indeed correlated. However, PAMIN can discover fine-grained relationships among those features, which is a novel contribution of our proposed framework.

Interaction-type IPs: Table 3.1 presents 5 interaction-type IPs found by PAMIN. The first interaction-type IP has two features: Middle temporal pole from fMRI and rs6265 (also referred to as Val66Met, a SNP). Middle temporal pole was known to be

involved in the memory network along with hippocampus and other parietal regions including the precentral cortex [259]. Moreover, the functional connections from the temporal pole to other regions in the memory network were found to be of reduced strength in schizophrenic subjects [259]. Val66Met polymorphism is known to affect the cognitive function of the brain. It was also found to be associated with schizophrenia [260]. The fact that these two factors are found in an interaction type IP, where the discriminative power of the IP is much more than the two features, suggests that the reduction in the connection strength of connections originating at the temporal pole and a Val66Met polymorphism could potentially make schizophrenia symptoms worse. Similarly, the features from other interaction-type IPs, Superior Parietal, Parahippocampal, rs1018381, rs2743864, and rs3780103 were also found to be individually associated with schizophrenia [254, 257, 255, 256, 258]. As indicated above, the presence of the features in these interaction type IPs could result in an increase in severity of schizophrenia symptoms.

The strength of the pattern mining based integrative framework is that it captures the subjects covered by an IP. This is very useful in exploring biomarkers, as they can explain different subsets of the population. To illustrate this, in Figure 3.6 we show the subjects that are covered by the 5 interaction-type IPs shown in Table 3.1. From this figure, it can be seen that the first four IPs cover largely similar groups of subjects in cases and controls, while the last IP covers a very different group of subjects. This indicates that different IPs can explain different subgroups in the population. This relates to disease heterogeneity [261], which is often seen in the complex diseases, where different subgroups tend to have different biomarkers due to differences in ethnicity, biological underpinnings of the disease, demography, etc. PAMIN is more suitable for addressing this aspect of biomarker discovery, since it identifies the subgroups in the study that have different sets of biomarkers.

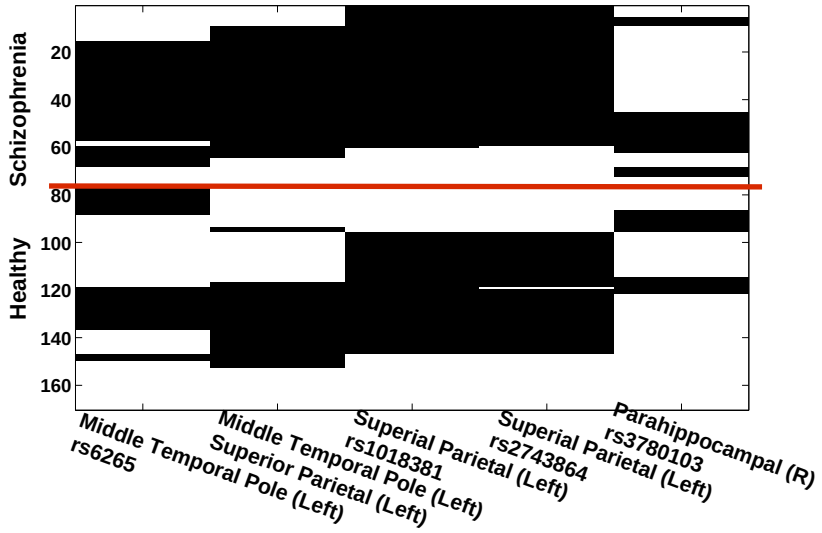


Figure 3.6: The subspace of diseased people covered by the interaction patterns in diseased and healthy people.

3.7 Future Work

The above mentioned discriminative pattern mining approach is mainly applicable for binary datasets. Here, finding relationships from continuous multi-source datasets is a much harder problem due to the lack of efficient techniques for searching the exponential search space. The discriminant pattern mining technique as described in the earlier section mostly work for the binary datasets, since discriminant pattern mining techniques are more efficient for binary datasets. Alternatively, direct approaches based on higher order statistics such as correlation has already been used for finding the relationship across datasets with continuous variables [228]. In this section, I will discusss a direct generic approach which can find both coherence and interaction type relationships from continuous-valued datasets. In particular, we used an optimization based technique to find such integrative patterns directly from the datasets.

I illustrate coherence type pattern for continuous datasets with the help of an example. Consider the three variables coming from three separated data sources being integrated as represented by three different colors in Figure 3.7 (x-axis represents the

activation level of each variable and y-axis represents samples coming from two groups: cases and controls). In Figure 3.7(a), we show the coherence-type pattern where each of the three variables are of similar discrimination power between two groups. Moreover, the variables are correlated overall across all the variables. Note that the discriminative power of the combination of features is approximately the same discriminative power as individual features. In contrast, each of the three features of Figure 3.7(b) coming from three different data sources has same activation level in both cases and controls, and so they do not explain the disease in question individually. However, these features together are correlated in case group, but not in the control group. Therefore, there is an inherenet relationship (or its absence), which leads potentially to disease progression. In this section, we aim at finding the such coherence and interaction type markers directly from diverse datasets.

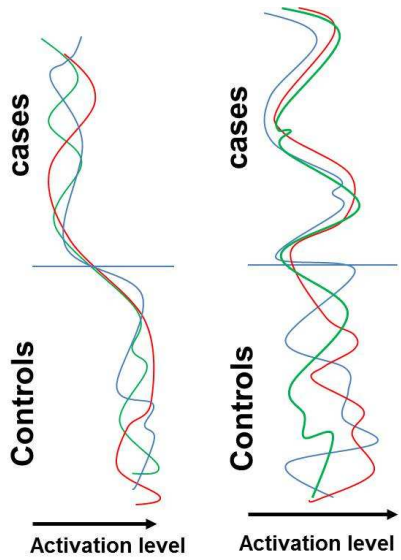


Figure 3.7: Example of tow-types of relationships present among three data sources. Each color represent markers from one data sources. a) Coherent-type relation and b) Interaction-type relation.

3.7.1 Problem Formulation

This approach finds a set of weight vectors, one for each dataset. Thus, instead of a biomarker being a collection binary weights (0 or 1) for each variable, variables can have

an arbitrary weight. This approach handles continuous variables without the need for binarization.

Consider two matrices X and Y representing case and control datasets, respectively. In each of these datasets, there are n columns (which represent n features) and $2m$ rows representing subjects. Among the $2m$ subjects, the first m subjects belong to healthy group and the last m subjects belong to diseased group. The sub matrices which only contain the healthy subjects are defined as $X1$ and $Y1$ and the sub matrices which contain the diseased group are called $X2$ and $Y2$. Intuitively, similar to feature-extraction based techniques, we aim to find the component vectors represented by two co-efficients w_a and w_b from the two datasets X and Y, respectively. Note that these componets are synonymous to the patterns described earlier.

3.7.2 Coherence-type patterns

The coherence-type of pattern of Figure 3.7(a) can be found by the following optimization formula. Here the first two terms want to maximize the correlation of the two components obtained from the two datasets X and Y within each class separately, similar to DCCA approach. The second term is for optimizing the discriminative power of the obtained components directly. In particular, we want to miniming the logistic loss function of the obtained prediction by $X_1 * w_a$ and class level C.

$$\max_{w_a,w_b} corr(X_1 w_a, Y_1 w_b) + corr(X_2 w_a, Y_2 w_b) - \lambda(logit(X w_a, C) - logit(Y w_b, C)) \quad (3.6)$$

Which equals to the following problem.

Note that here the parameter λ performs a trade-off between the coherence in terms of sample space and predictive power.

$$\begin{aligned} \max_{w_a,w_b} \quad & (w_a^T X_1^T Y_1 w_b + w_a^T X_2^T Y_2 w_b \\ + \quad & \lambda \textstyle\sum_{i=1}^{n} C_i \log\left(\frac{1}{1+e^{-X w_a}}\right) \\ + \quad & \lambda \textstyle\sum_{i=1}^{n} (1 - C_i) \log\left(1 - \frac{1}{1+e^{-X w_a}}\right)) \\ \text{subject to} \quad & ||w_a^T X_1^T X_1 w_a|| = ||w_b^T Y_1^T Y_1 w_b|| \\ = \quad & ||w_a^T X_2^T X_2 w_a|| = ||w_b^T Y_2^T Y_2 w_b|| = 1 \end{aligned} \quad (3.7)$$

Generally, X and Y are a matrix which include a few hundreds rows but thousands of columns. This can lead to often overfitting of the co-efficient vectors w_a and w_b. Therefore, we also impose sparsity on the co-efficient vectors such that they are not overfitted. Among several different ways to impose sparsity on the co-efficients as mentioned in [262], we used L_1 norm sparsity, as it can perform feature selection as well which will enhance the interpretability of the obtained set of biomarkers. Therefore, we aim to optimize the following formula.

$$\begin{aligned} \max_{w_a,w_b} \quad & (w_a^T X_1^T Y_1 w_b + w_a^T X_2^T Y_2 w_b \\ - \quad & \lambda \sum_{i=1}^{n} C_i \log\left(\frac{1}{1+e^{-Xw_a}}\right) \\ - \quad & \lambda \sum_{i=1}^{n} (1-C_i) \log\left(1 - \frac{1}{1+e^{-Xw_a}}\right)) \\ + \quad & b_1||w_a||_1 + b_2||w_b||_1 \\ \text{subject to} \quad & ||w_a^T X_1^T X_1 w_a|| = ||w_b^T Y_1^T Y_1 w_b|| \\ = \quad & ||w_a^T X_2^T X_2 w_a|| = ||w_b^T Y_2^T Y_2 w_b|| = 1 \end{aligned} \tag{3.8}$$

This problem is a convex problem but is not a smooth function. We aim to explore different optimization techniques including gradient-descent to find the solution of this formula. Once we find the first component w_a from a dataset, we will find the additional component such that it is orthogonal to the first component similar to the CCA approach.

3.7.3 Interaction-type Patterns

We aim to find interactions from more heterogeneous datasets such as real-valued data, data with mixed type. Consider the same problem set-up as described earlier for the two case-control datasets X and Y. In each of these dataset, there are n columns (which represent n features) and $2m$ rows representing subjects. Among the $2m$ subjects, the first m subjects belong to healthy group and the last m subjects belong to diseased group. The sub matrix which only contains the healthy subjects is defined as $X1$ and $Y1$ and the sub matrix which contains the diseased group is called $X2$ and $Y2$. Intuitively, similar to feature-extraction based techniques, we aim to find the component vectors represented by two co-efficients w_a and w_b from the two datasets X and Y, respectively.

In the scenerio of Figure 3.7(b), the correlation between dierent biomarkers can help

to distinguish the positive set and the negative set. In order to mine these patterns, we use the correlation as the discriminative measurement instead of linear combinations of "real" biomarkers. In details, we want to find the linear transformations of X and Y (using co-efficients w_a and w_b), which can maximize the correlation between the positive set (X1 and Y1) while minimize the absolute value of the correlation between the negative set (X2 and Y2). Thus, the problem is defined as below.

$$\max_{w_a,w_b} |corr(X_1 w_a, Y_1 w_b) - corr(X_2 w_a, Y_2 w_b)| \tag{3.9}$$

This is a non-convex function. Generally, X and Y are a at matrix which include a few hundreds rows but thousands of columns. Similar to the approach defined for finding coherence-type biomarker, we impose L_1 penalty on the canonical co-efficient vectors such as defined below.

$$\max_{w_a,w_b} corr(X_1 w_a, Y_1 w_b) - |corr(X_2 w_a, Y_2 w_b)| - b_1||w_a||_1 - b_2||w_b||_1 \tag{3.10}$$

which is identical to

$$\max_{w_a,w_b} w_a^T X_1^T Y_1 w_b - |w_a^T X_2^T Y_2 w_b| - b_1||w_a||_1 - b_2||w_b||_1 \tag{3.11}$$

subject to $||w_a^T X_1^T X_1 w_a|| = ||w_b^T Y_1^T Y_1 w^b|| = ||w_a^T X_2^T X_2 w_a|| = ||w_b^T Y_2^T Y_2 w_b|| = 1$

3.8 Conclusion

In this chapter, we pursue two approaches that can find both coherence and interaction patterns. They have different advantages and limitations as discussed below, but both represent a significant advance in the state of the art for finding integrative biomarkers.

- Pattern Mining Based Approach: We leverage the strength of discriminative pattern mining that has recently been used in the context of combinatorial biomarker discovery [263, 242] for its ability to ex plore the exponential combinatorial search space efficiently, enhanced interpretability of the obtained features and ability to identify the samples that are related with those features, a capability that can potentially handle disease heterogeneity. Moreover, pattern mining approach is a

non-parametric and non-linear approach to find higher-order interactions present among features in a particular dataset. A key innovation for this work is the creation of a two-step framework (described below) that en hances efficiency and naturally accommodates the differences between individual data sets in terms of dimensionality, noise level, etc.

- Optimization Based Approach: This approach finds a set of weight vectors, one for each dataset. Thus, instead of a biomarker being a collection of binary weights (0 or 1) for each variable, variables can have an arbitrary weight. This approach handles continuous variables without the need for binarization and is likely to be more computationally efficient for some formulations. This approach overcomes the limitations of current CCA based approaches that cannot find interaction and more generally, dont́ properly account for class labels.

Chapter 4

Incorporating Prior Knowledge into Biomarker Discovery

4.1 Introduction

Group1 ICD-9	Group1 survey features	Group2 ICD-9	Group2 survey features
250.61	Diabetes with neurological manifestation	401.1	Benign hypertension
294.20	Dementia	817	Multiple fractures of hand bones
272.4	hyperlipidemia	692.71	Sunburn

Table 4.1: Two groups of ICD-9 codes

Healthcare costs in the US are becoming unsustainable, reaching 18% of the gross domestic product (GDP) in 2011 and headed for 20% by 2020 [264]. The terabytes or even petabytes of health data available in EHRs present new opportunities and challenges for research that aims to effectively use these data to discover new knowledge to improve health-care. For example, half of the waste in healthcare spending (up to $425 billion) has been attributed to a failure of appropriate care delivery, a lack of coordination between different healthcare plans, and over-treatment [265, 266]. Mining significant patterns from EHRs can help elucidate such knowledge for potential new

care plans and enable more coordination between different healthcare plans.

We collected a large set of EHRs from 581 home healthcare (HHC) agencies for 270,068 patients. In particular, our data contains the diagnosis (ICD-9) codes during patients' admission into HHC. After admission, the patients received interventions designed to improve their health status. However, all patients are not equally likely to improve in their health status. For example, patients with poor memory are less likely to improve with respect to urinary incontinence. In general, the nursing interventions are designed mostly based on patients' initial health condition during their admission in the homecare agency. The ICD-9 diagnosis codes recorded during the admission into HHC can help to stratify patient groups for more customized homecare interventions and thus, an increased likelihood of improved health status. Finding the important groups of ICD-9 codes is also valuable for enhancing the interpretability of the final models. In this chapter, we aim to find the ICD-9 groups that help in improving health status as measured by urinary incontinence.

Unlike conventional predictive models which mainly focus on improving the predictive power of a target variable such as urinary incontinence, we are primarily interested in finding interpretable risk factors which can be used by the domain expert for further clinical purposes. Moreover, most classification approaches provide only one final set of biomarkers that are applicable for the overall population. Instead, we are interested in finding relatively homogeneous groups of ICD-9 diagnosis codes that are targeted to specific, homogeneous sub-populations. Indeed, this is the main goal of the this work. For example, Table 4.1 shows two groups of ICD-9 codes. The first group is more interpretable since they are more related and represent the patient group with diabetes and dementia. On the other hand, the second group of ICD-9 codes is not clinically interpretable, although it can be more predictive than the first group.

To find such homogeneous predictive groups of ICD-9 codes, we explored two distinct approaches: data-driven and prior knowledge driven. The data-driven models incorporate various clinical information such as demographic, behavioral, physiological, and psycho-social factors, which were collected as routine assessment in homecare EHRs through OASIS survey questions [1] . The goal of using such survey data is that the auxiliary information collected for same patient will provide more natural groupings of

[1] We will denote these assessments as survey data in rest of the chapter

ICD-9 codes. On the other hand, clinical classification software (CCS) provides systematic grouping of ICD-9 codes into a hierarchical tree structure using prior knowledge. We tried to incorporate such prior knowledge into the predictive models.

However, taking such diverse datasets into account creates a number of computational challenges. First, the three datasets (ICD-9 codes, survey questions, CCS prior knowledge) vary in terms of their innate properties such as type, format, and sparsity. Second, the relationships present between the ICD-9 codes and survey questions may be important, although not necessarily discriminative. Therefore, regular predictive models may overlook them. Third, there is a trade-off between data-driven grouping and prior-knowledge-driven groupings, which should be taken into account by the model.

In this chapter, we propose an integrative framework to address the above issues in a systematic way. Integration of multiple datasets for biomarker discovery techniques can be broadly classified into two groups: 1) Predictive models ([267] and [268] provide a good survey on several kernel fusion methods) and 2) Feature extraction based biomarker discovery techniques [233, 228]. The goal of the predictive model-based approaches is to build classification models with high accuracy, but often such techniques do not yield easily interpretable results. In contrast, biomarkers (that are constructed using a small number of features) can be directly useful in diagnosis, treatment or prevention, but equally as important, they can also provide insights into the underlying nature of the disease or related biomedical processes. Hence we focus only on such techniques in this chapter to find interpretable ICD-9 code groups.

Among the feature extraction based techniques, canonical correlation analysis (CCA) [269] is one of the most popular techniques for data integration because it can find natural grouping (by components) in each dataset and the potential relationships among those components is measured by correlation. It has been also shown that CCA has fewer model assumptions than other integration techniques [228]. Recently, CCA has been extended for handling high-dimensional data using different types of regularization including sparsity [262]. CCA has further been generalized to integrate more than two datasets [270]. However, none of these methods can take prior knowledge that is available from the CCS tree into account. Moreover, the existing CCA-based techniques are unable to handle datasets of different types because of their assumptions that all datasets have vector-based records which have been collected for the same set of samples.

Our proposed framework further extends the CCA to incorporate the prior knowledge available from the CCS tree into model development, which is different than the vector-based data format. Moreover, it can also trade off between the data-driven knowledge from survey data and prior-knowledge-driven CCS framework. The framework further builds a classification model to assess the predictive capability of the obtained components.

4.1.1 Contributions

Our goal is to find the groups of ICD-9 codes that are related with the improvement of urinary incontinence. Therefore, we want to build a model that is both predictive and interpretable. To enhance the interpretability of the predictive model, we incorporate several types of knowledge into the model development process as described below:

- To enhance the interpretability of the model, we use the clinical classification system (CCS) as prior knowledge in the predictive model (baseline model).
- We want to find the relationship between the ICD-9 codes and the clinical survey variables to enhance interpretability. In particular, we first use sparse-CCA to find the relationships present among the two datasets and then use them along with other useful discriminative features in a predictive model. We further develop a hybrid model called sparse hierarchical CCA (SHCCA), which can take both prior knowledge (CCS) and clinical survey data into account to enhance clinical interpretability.
- To assess the interpretability of the obtained ICD-9 features, we propose a novel metric called I-score based on searching PubMed articles. Our components are more interpretable than the individual ICD-9 and CCS codes while retaining similar prediction capability.
- SHCCA can extract relatively more homogeneous groups of ICD-9 codes, each representing a distinct subgroup of patients, in contrast to finding a global set of ICD-9 codes generated by the baseline predictive models.

4.2 Method

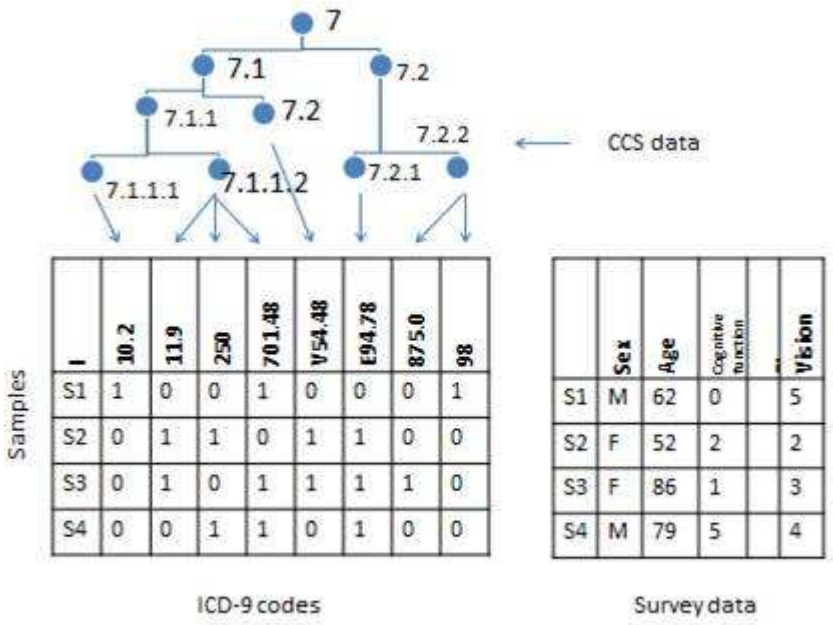


Figure 4.1: SHCCA framework containing three types of data: survey, ICD-9 codes and CCS hierarchy.

4.2.1 The Integrative Predictive Model Framework

The main goal of the chapter is to utilize as much information as possible from other clinical information to enhance the interpretability of the obtained ICD-9 groups without losing the baseline predictive power. In particular, we want to take two other types of information such as survey data and CCS tree into account during grouping ICD-9. However, integrating these three types of information poses some computational challenges. First, the datasets are of very different types. For example, ICD-9 codes contain binary data, while clinical factors contain binary, ordinal and numeric data. The ICD-9 codes are very sparse ($< 2\%$ density) compared to the dense clinical data. Second, there may be some relationships present among the two types of data. For example,

a subgroup of patients with mental disorders may have gone through the same set of interventions in the homecare. However, these factors may not be discriminative and thus will be missed by the traditional predictive models. Third, the CCS hierarchical tree provides a completely different type of information containing relationships present among ICD-9 codes. Moreover, they are not stored in traditional record-based datasets similar to the ICD-9 and survey data as shown in Figure 4.1. To address these three challenges, we will first describe how to leverage the survey data to group the ICD-9 codes and then finally take CCS prior knowledge into account.

4.2.1.1 Bringing survey data into grouping ICD-9 codes

The easiest way to integrate the ICD-9 and survey data is to concatenate the two datasets together and then build a predictive model such as LASSO as described earlier. However, this will not be able to handle the disparate nature of the two datasets as described earlier. Thus sparse ICD-9 data are more likely to be lost, since the coefficients of dense survey data will dominate the results. Also, such a predictive model will not be able to find relationships present among the types of features. Moreover, they only focus on providing one global set of biomarkers. Thus, it cannot provide information about disease heterogeneity, where different set of biomarkers affect different subsets of the population. We want to leverage canonical correlation analysis (CCA) based approaches to handle all these issues. Instead of merging the two datasets before performing the analysis, CCA finds components from each of the two datasets such that the components are maximally correlated. This correlation can help find relationships between two datasets. Moreover, each component can correspond to one homogeneous subgroup of the dataset. We used the sparse CCA (SCCA) approach for our analysis, because this will perform feature selection for both datasets as well, which will enhance the interpretability. We will describe the SCCA algorithm briefly as follows. Let $\mathbf{X}$ be a $n \times p$ matrix containing p sparse ICD-9 codes and $\mathbf{Y}$ be the $n \times q$ matrix containing q survey questions observed on same n observations. CCA tries to find the linear combination of $\mathbf{X}$ and $\mathbf{Y}$ such that they are maximally correlated. Therefore, we want to find coefficient vectors w_x and w_y from $\mathbf{X}$ and $\mathbf{Y}$ respectively such that

$$corr(w_x'X, w_y'Y) = \frac{w_x' C_{XY} w_y}{\sqrt{w_x' C_{XX} w_x}\sqrt{w_y' C_{YY} wy}} \tag{4.1}$$

is maximized where C_{XX},C_{XY} and C_{YY} are the variance matrix of X, covariance matrix for X and Y and variance matrix of Y, respectively. We can easily see that the correlation is invariant to the any arbitrary scaling of w_x and w_y (by replacing w_x by $a * w_x$). Therefore, equation 4.1 can be re-written as

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x' C_{XY} w_y \\ \text{subject to} \quad & w_x' C_{XX} w_x = 1, \\ & w_y' C_{YY} w_y = 1 \end{aligned} \tag{4.2}$$

Lets define a change of basis $u = C_{XX}^{1/2} w_X$ and $v = C_{YY}^{1/2} w_Y$. Substituting them in equation 4.2, we get

$$\max_{u,v} u' C_{XX}^{-1/2} C_{XY} C_{YY}^{-1/2} v \tag{4.3}$$

such that $u'u = v'v = 1$.

Among many such solutions of equation 4.3, we follow [271], where u and v can be computed using the singular valued decomposition of sample correlation matrix $K = C_{XX}^{-1/2} C_{XY} C_{YY}^{-1/2}$ and then used them back to get w_x and w_y. However, performing a linear combination on all the features of $\mathbf{X}$ and $\mathbf{Y}$ will lead to too many features which will lack biological interpretation. To perform feature selection along with finding the coefficient vectors, we perform sparse canonical analysis (SCCA) [262], where additional L_1 constraints were imposed on w_x and w_y as below.

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x' C_{XY} w_y \\ \text{subject to} \quad & w_x' C_{XX} w_x = 1, \\ & w_y' C_{YY} w_y = 1, \\ & \|w_x\|_1 \leq \lambda_x, \\ & \|w_y\|_1 \leq \lambda_y \end{aligned} \tag{4.4}$$

After the change of basis, the sparseness is imposed on the loading vectors u and v controlled by λ_x and λ_y which determines how many parameters will be selected. The SCCA solutions obtained by integrating ICD-9 dataset and the survey data will lead to

finding groups of ICD-9 that are not related among themselves but also correlated with the the selected survey data.

4.2.1.2 Taking CCS Prior Knowledge into Account

In this section, we will describe the utility of bringing prior CCS information to provide more interpretable solutions. Let us consider the example of Figure 4.1. To take the prior CCS tree structure into account, we need to penalize less for grouping the ICD-9 codes that are closer to each other in the CCS tree. Let H be a matrix that contains the similarity between each pair of ICD-9 codes in terms of the closeness in the CCS tree. Intuitively, this prior knowledge is parallel to the covariance matrix C_{XX} computed from the data as in equation 4.2. Intuitively, we want to tradeoff between these two matrices: the prior knowledge based similarity H and data-driven similarity matrix C_{XX}. The trade-off is imposed by introducing a new parameter $\lambda_h \in [01]$ in equation 4.4. When $\lambda_h = 0$ the solution is exactly equal to those of SCCA, while $\lambda_h = 1$ leads to ICD-9 codes that are purely similar based on the CCS tree H. We will call this Sparse Hierarchical CCA(SHCCA).

$$\begin{aligned} \max_{w_x, w_y} \quad & w_x' C_{XY} w_y \\ \text{subject to} \quad & w_x'[(1-\lambda_h)C_{XX} + \lambda_h H] w_x = 1, \\ & w_y' C_{YY} w_y = 1, \\ & \|w_x\|_1 \leq \lambda_x, \\ & \|w_y\|_1 \leq \lambda_y \end{aligned} \tag{4.5}$$

After the change of basis similar as described in equation 4.4, the solution can be obtained from the new sample correlation matrix $K_h = [(1-\lambda_h)C_{XX} + \lambda_h H]^{-1/2} C_{XY} C_{YY}^{-1/2}$. Note that matrices C_{XX}, C_{YY} and H have to be non-singular, which is ensured by computing them from the data with regularization if required, as mentioned in [272]. In this section we will first describe how to calculate the similarity matrix H from the CCS tree followed by the detailed algorithm for computing the solution of equation 4.5.

4.2.1.2.1 Computing CCS similarity: The similarity between any two ICD-9 codes was determined based on the depth of their lowest common ancestor(LCA) in

the tree. However, some of the ICD-9 codes are not labeled up to the 4th level in the tree(e.g., 875.0 and 95 in Figure 4.1). Therefore, we normalize that metric by the maximum depth of individual ICD-9 codes. Note that a similar type of edge-based similarity measure has also been applied in other biological ontologies such as the gene ontology [273]. More formally, it is defined as below:

$$H_{ij} = \frac{depth(LCA(X_i, X_j))}{\max(depth(X_i), depth(X_j))} \tag{4.6}$$

4.2.1.2.2 Finding the solution of SHCCA: Finding the solution of SHCCA relies on finding the SVD of the sample correlation matrix K_h approximated by the first singular vectors. We used the two parameter λ_x and λ_y as soft-thresholding parameters to perform feature selection on the datasets X and Y, which is similar to LASSO [274]. In addition, we have the third parameter λ_h which is used to incorporate the prior knowledge measured by H into account. We used an iterative soft-thresholding algorithm for performing SHCCA similar to [274]. This will lead to the first component of u_1 and v_1, where $u_1 = [(1 - \lambda_h)C_{XX} + \lambda_h H]^{1/2} w_X$ and $v_1 = C_{YY}^{1/2} w_Y$ from equation 4.5. The second canonical components, u_2 and v_2, can be computed such that they are orthogonal to the other components. This can be computed as below from the SVD solution of K_h.

$$K_h = \sum_{i=1}^{k} u_i * d_i * v_i'. \tag{4.7}$$

Therefore, the successive components of u_i and v_i can be computed as the SVD of the remaining sample correlation matrix $\{K_h\}_i = K_h - \sum_{i=1}^{k-1} d_i u_i v_i'$. The algorithm is given in the Appendix.

4.3 Experimental Setup

4.3.1 Dataset

We collected a large set of EHR data for 270,068 patients from 281 home healthcare (HHC) agencies. In particular, we collected 6,800 distinct ICD-9 diagnosis codes from HHC for those patients as shown in Figure 4.1. A clinician manually labeled each patient with at most twelve primary and secondary ICD-9 diagnosis codes during the

patient's admission in HHC. Moreover, the patients were assessed based on several survey questions related to their demographic, behavioral, physiological, and psycho-social factors during their admission and discharge in the homecare agencies. These survey questions were summarized into 184 variables guided by a domain expert and they were used as an auxiliary dataset for grouping the ICD-9 diagnosis codes. The class label was also created based on whether the urinary incontinence improved at discharge compared to the baseline level during admission in HHC. Furthermore, prior information is also available for the ICD-9 codes in the form of clinical classification software(CCS) [275]. CCS has been developed and maintained by Agency for Healthcare Research and Quality (AHRQ) to systematically manage the relationship among ICD-9 diagnosis codes as a multi-level hierarchical tree. The root contains very generic terms while leaves contain the most specific terms. Therefore, a CCS term is a summarization of several correlated ICD-9 codes. In this paper, we used a 4-level tree containing 15,073 CCS terms which finally contain all the 6,800 ICD-9 codes downloaded from [275].

A few preprocessing steps were performed on the datasets guided by domain experts. For example, the samples with no scope of improvement (highest urinary incontinence score during admission into homecare) were dropped from the analysis, which led to ultimately 121,956 samples. The very rare ICD-9 codes (occurrence in fewer than 10 samples) were removed leading to 2,705 ICD-9 codes. The categorical variables in the survey questions were converted into binary variables, each corresponding to one category. Finally, we ended up with 184 survey questions.

4.3.2 Evaluation

We evaluated the obtained ICD-9 codes using two metrics: the prediction power and the interpretability. The prediction power was assessed by the area under the ROC curve(AUC) score [276]. We first describe two baseline predictive models which were built on ICD-9 and CCS codes. Then, we describe how the components obtained from SHCCA were used to build the final predictive model. Finally, we discuss the techniques for assessing the interpretability of the ICD-9 codes.

4.3.2.1 Baseline Predictive Models

We created two baseline models for evaluating the prediction power of SHCCA. First, we only considered the basic ICD-9 diagnosis codes which are at the lowest level of granularity in the CCS hierarchy. Second, we used all internal nodes of the CCS hierarchy. CCS provides a systematic clustering of related ICD-9 codes and thus, provides a natural summarization of ICD-9 codes. Therefore, if we build the predictive model on the CCS terms, it can provide more correlated ICD-9 codes. We converted ICD-9 feature space into CCS feature space by taking the most conservative approach. In particular, we created a binary data set with 650 CCS codes (the internal nodes of CCS tree), where we denoted the presence of a CCS code for a particular patient if any of the ICD-9 codes belonging to the subtree rooted at that CCS node was present in that sample. Among different predictive models, we choose a LASSO based regularized model [277] because of its inbuilt feature selection technique using L_1 penalty on the coefficients of the solution. Selecting a few most important features in this way helps in the interpretation of the obtained features by a domain expert, which is the main goal of the paper. Furthermore, we used adaptive LASSO [278] to increase the stability of the obtained coefficients of both ICD-9 and CCS baseline models.

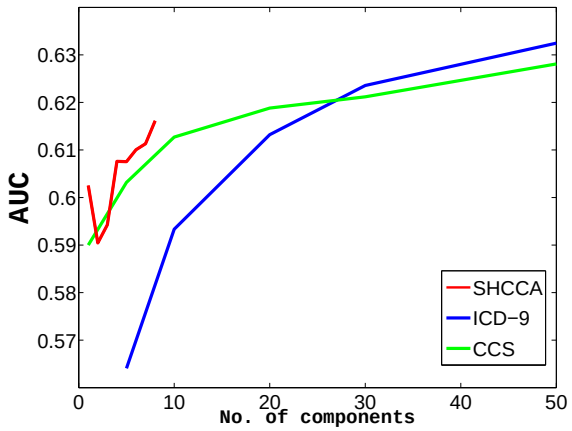


Figure 4.2: AUC scores for the three methods.

4.3.2.2 Building a Predictive Model on SHCCA Components

To assess the prediction power of the SHCCA components, we first transform the original ICD-9 data into the newly formed K components of ICD-9 codes. We multiply the original data matrix with component vector obtained from SHCCA creating a new matrix of size $n \times k$, where k is the number of components obtained from SHCCA. To evaluate the prediction power of the SHCCA method, we used two cross-validation (CV) frameworks. The external CV was used to find the prediction error of the predictive method built on the SHCCA components using a logistic regression model. For each of the training datasets, a 5-fold internal CV was further used to tune the parameters of SHCCA namely λ_u and λ_v (described later in the parameter selection section). We treated λ_h as an independent parameter since it is not related to the feature selection process from the two datasets.

4.3.2.3 Assessing Interpretability

The main goal of the paper was to improve the interpretability of the ICD-9 groups obtained from predictive models. Therefore, we evaluated the obtained ICD-9 codes rigorously based on their interpretability. First, the ICD-9 groups were analyzed by domain experts (also the co-author of the paper). The main evaluation criteria was whether the obtained groups of ICD-9 codes are coherent, representing similar types of pathology or disease symptoms. Second, we propose a novel measure called I-score to quantify the coherence of the obtained ICD-9 groups using the PubMed articles [279]. In particular, we searched each pair of the terms belonging to same group (or components of SHCCA) for their co-occurrence in the same article. Intuitively, the more frequently the terms co-occur in PubMed article, the more coherently they represent an underlying disease. Let t_i and t_j be two sets of PubMed articles containing the i-th and j-th ICD-9 terms, respectively. Then, a Jaccard similarity measure [276] is defined to assess the semantic similarity of the two terms based on the intersection and the union of two terms. Finally, all such semantic similarities between each possible pairs are summarized as the final similarity of the cluster. Note that the co-occurrence (thus the union of the two terms) of two terms is very rare and therefore, the I-score is low in general.

$$I - score(C) = \sum_i \sum_j |t_i \cap t_j| / |t_i \cup t_j| \tag{4.8}$$

4.4 Results

Initially, we report the predictive power of SHCCA components compared to the two baseline methods built on CCS and ICD-9 codes. Figure 4.2 represents the area under the ROC curve (AUC) of the three methods for different number of features (components for SHCCA) selected by the LASSO model. Among the three models, ICD-9 provides best overall prediction power. The prediction power of both ICD-9 and CCS models improves when the number of selected features increases. On the other hand, CCS provides better predictive power in the beginning, but saturates as the number of features increases. On the other hand, SHCCA (with best parameter of $\lambda_h = 0.8$, $\lambda_h = 0.16$, and $\lambda_v = 0.0016$) performs slightly better than both of the baseline methods. The AUC score of SHCCA does not vary too much on the λ_h value with a range between 0.59 and 0.62 (Appendix). It is quite natural for ICD-9 codes to have the best predictive power, because that is the lowest level of granularity in terms of feature selection and the LASSO model only picks the ICD-9 codes that have best predictive power. However, as the number of features go beyond 20, the interpretability of the ICD-9 codes becomes less since the ICD-9 codes are very disparate in nature (Appendix section for full list of ICD-9 codes). In contrast, each of the CCS and SHCCA components represents a cluster of ICD-9 codes, which may not be necessarily best predictive features. However, Figure 4.2 shows that even those groups are almost equally predictive as the raw ICD-9 codes. Note that the main purpose of this study is to group ICD-9 codes into more interpretable clusters rather than solely developing a predictive model.

The interpretability of the SHCCA method is greatly enhanced compared to the two baseline methods. Figure 4.3 represents the interpretability score (I-score) of SHCCA compared to two baseline methods for $\lambda_h = 0$, i.e., without bringing any prior information. The left subfigure of this figure shows the I-score of the two baseline methods when built by successively adding features into the model. On the other hand, the I-score of each component of SHCCA is shown on the right subfigure. The I-score is greatly enhanced by SHCCA (from 0.015 to 0.165), which shows the effectiveness of bringing

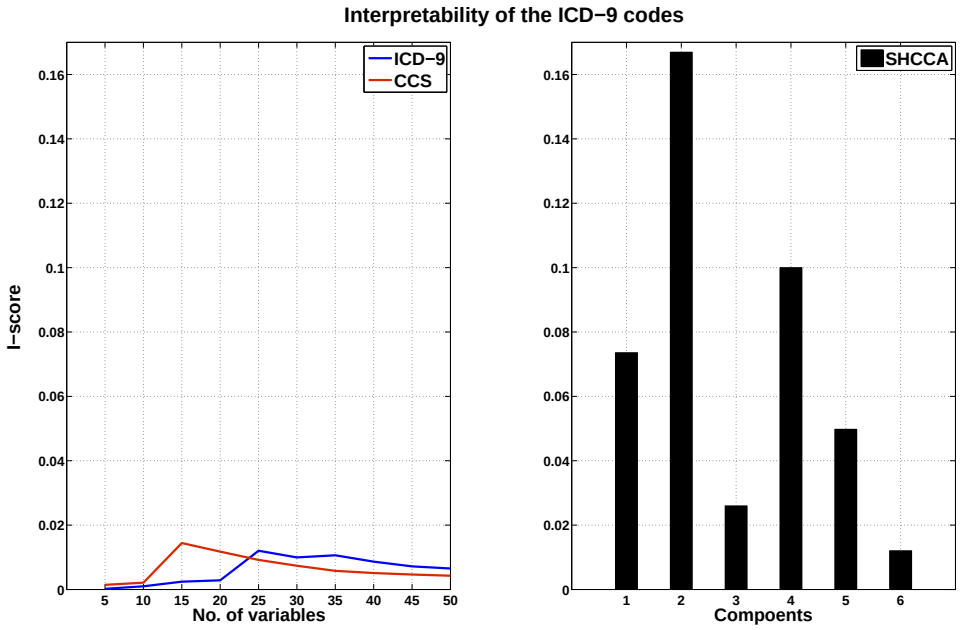


Figure 4.3: I-score of baseline methods

survey data into account. Note that, PubMed contains a large number of articles for any of the two terms being searched. However, finding co-occurrence of two disease codes (referred as co-morbidity in medical domain) is very rare. Therefore, even though the absolute value of the I-score is low compared to the perfect score of 1, the improvement of 0.15 is very significant. We also examined the ICD-9 groups selected by the two baseline methods (top 20 features with highest LASSO coefficients are shown for ICD-9 and CCS codes) and the SHCCA as shown in Table 4.2 and Table 4.3, respectively. Then, our domain experts evaluated the results obtained from three methods. It turned out that the components selected by SHCCA are more coherent, representing one underlying socio-psychological status of the patients. The ICD-9 codes shown in Table 4.2 represent codes from several diseases such as heart disease, radiological procedure, Alzheimerś disease, paralysis and so on. CCS terms are more interpretable in terms of representing only three major types of disease such as disease related to nervous system, several congenital anomalies, decubitus ulcer, and so on. On the other hand, Table 4.3 represents the three top components from ICD-9 codes and survey data

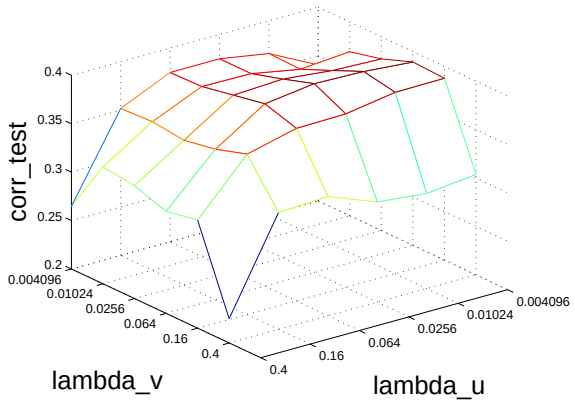


Figure 4.4: Effect of the sparseness parameter on the average test correlation.

both. The first components represent the ICD-9 codes that are only related to several neurological disorders. More interestingly, the corresponding survey features also exactly match with socio-psychological functions such as old age, poor cognitive function, speech, prior memory loss, memory deficiency and higher confusions. Similarly, the second component is more related to dysphagia, gastronomy, and blindness which lead to poor self-management skill. The third component consists of several aftercare therapies, which is confirmed by prior surgical wound observed in survey data.

We also studied the effect of bringing prior knowledge into SHCCA. Therefore, the λ_h was varied independently between the range $[0 : 0.2 : 1]$, with $\lambda_h = 0$ indicating no prior information included, while $\lambda_h = 1$ means only prior information is included. We found that $\lambda_h = 0.8$ provided the best predictive power as shown in Appendix Figure 2. We also checked how the interpretability varies with the increment of λ_h, but only considered the first component for this analysis. It turned out that the I-score remains almost same to 0.075 for all λ_h. However, the size of the components (number of the ICD-9 codes selected) becomes larger as more prior information is included. For example, 37 ICD-9 codes (Appendix section) were selected for $\lambda_h = 0.6$ in the first component as opposed to only four ICD-9 codes selected when no prior was included (Table 4.3). Actually, the 37 components comprise most of the ICD-9 codes represented by three subtrees rooted in three CCS level-3 codes representing dementia, transient mental disorders and persistent mental disorders, which are very related disorders. Note that, since we

computed the I-score using all of a component's pairwise I-score, a component is more likely to have lower I-score as the component becomes larger. In our case, using prior CCS information provides larger but very coherent ICD-9 without any loss of I-score. Therefore, using CCS prior information is important for both increasing the prediction power and interpretability of the SHCCA.

ICD-9 terms	CCS terms
Malignant hypertensive heart disease with heart failure	Delirium
Radiological procedure and radiotherapy	Congenital hip deformity
Alzheimer's disease	Other paralysis
Attention to Cystostomy	Decubitus ulcer
Other paralytic syndromes	Other congenital anomalies of urinary system
Neurogenic Bladder Nos	Benign neoplasm of uterus
Senile dementia with delusional or depressive features	Psychogenic disorders
Multiple sclerosis	Other lower gastrointestinal congenital anomalies
Other cerebral degenerations	Other nervous system congenital anomalies
Aftercare following surgery of the genitourinary system	Other aftercare

Table 4.2: Top 20 features selected by two baseline models based on ICD-9 and CCS terms.

4.4.1 Effect of the Parameters:

We also studied the effect of the sparseness parameters of the two methods. Since, we normalize the canonical vectors u and v in each step of the algorithm, the maximum value that any individual canonical coefficient can have is 1. Therefore, the maximum value of λ_X and λ_Y is 2. However, we found that if these parameters are set too high(≥ 0.5) no variable selection is performed. Therefore, we searched exponentially within the range of $[0, 0.4]$ to tune these parameters using the k-fold CV framework as mentioned earlier. For each CV run, SHCCA was computed for each combination of the two parameters on the training dataset and then, the obtained coefficients from the

training dataset were used to compute the correlation on the test dataset as defined below [262].

$$corr = \frac{1}{k}\sum_{j=1}^{k} |cor(X_j u^{-j}, Y_j v^{-j})| \tag{4.9}$$

Here X_j represents the j-th test set and the u^{-j} represents the canonical coefficients learnt from the corresponding training data. Finally, the test correlation was averaged over the k-fold CV steps and the parameters yielding the largest average correlation were used for building the final predictive model in the outer CV loop. The average correlation leads to a convex function. In most of the cases, the best correlation was obtained by the parameters $\lambda_u \in [0.0016, 0.16]$ for ICD-9 codes and $\lambda_v \in [0.0016, 0.1]$ for survey data.

4.5 Conclusion

In this chapter, we incorporated clinical information available from a survey survey data and prior information to group ICD-9 diagnosis codes into more coherent groups. In particular, we proposed a novel method to incorporate prior information into a sparse hierarchical canonical component analysis. The proposed method enhances the interpretability of ICD-9 codes greatly when assessed by both a novel score (I-score) based on search in PubMed articles and clinical interpretation by domain experts. The proposed SHCCA method can further be extended to take the class label into account during method development in our future work. A more systematic score can be also developed to search PubMed articles by mapping ICD-9 codes into Mesh terms for assessing interpretability.

SHCCA survey components-1	SHCCA ICD-9 terms-1	SHCCA survey components-2	SHCCA ICD-9 terms-2	SHCCA survey components-3	SHCCA ICD-9 terms-3
Age	Alzheimer's disease	Poor vision	Legal blindness	Fully granulating surgical wound	Aftercare for healing traumatic fracture of hip
Prior memory loss	Persistent mental disorders	Poor speech	Dysphagia, other	Missing surgical wound	Encounter for change or removal of surgical wound dressing
Poor Speech	Dementias	Worst Speech	Dysphagia		Knee joint replacement
Frequent Behavioral problem	Cerebral degenerations	Partially granulating surgical wound	Degeneration of macula and posterior pole		Hip joint replacement
Poor Cognitive Function		Not healing surgical wound	Non-healing surgical wound		Aftercare following surgery of the musculoskeletal system
Medium Confusion		Average feeding condition	Attention to gastrostomy		Aftercare following joint replacement
High Confusion		Poor feeding condition	Hemiplegia affecting dominant side		Aftercare following surgery for neoplasm
Highest Confusion		Worst feeding condition	Hemiplegia or hemiparesis		Aftercare following surgery of the circulatory system
Memory deficiency					

Table 4.3: The main three components of SHCCA with $\lambda_h = 0, \lambda_u = 0.3$, and $\lambda_v = 0.3$.

Chapter 5

Heterogeneity of the Biomarkers

5.1 Background and Motivation

One important but usually neglected issue in biomarker discovery is the heterogeneous nature of many diseases, i.e., different subsets of the population are known to have different biomarkers for the same disease [280], due to different pathways playing a role in the same disease, or due to the same pathway playing a different role in subjects from different ethnicities [32]. Some other potential reasons responsible for such disease heterogeneity are age [281], ethnicity and race [30, 31], or genetic predisposition [33]. All these factors may lead to different degrees of association between the other clinicogenomic factors and the disease. For example, consider Figure 5.1 which shows a subset of regions of interest (ROI) pairs selected from a fMRI dataset. These ROI pairs are categorized into two groups C1 and C2. From this figure, it can be seen that for a subset of healthy and schizophrenia subjects, correlation values of the ROI pairs in the group C1 are higher than the same for C2, whereas for most of the other subjects the converse is true. Hence, discriminating biomarkers for these two groups of subjects are likely to be different.

The issue of heterogeneity is even more important for integrative studies, because different types of data may measure very different aspects of disease phenomenon such as behavioral, environmental and biological factors. For example, genomic factors may have more effect on a particular group of patient while the clinical factors may affect in other group of patients. This concept can be further illustrated by the Figure 5.2,

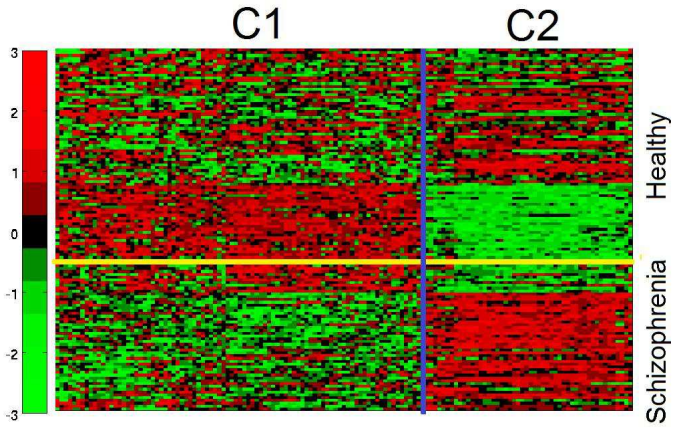


Figure 5.1: Subset of ROI pairs from VB fMRI dataset

which shows the features (represented by columns) coming from two different datasets: genomic and clinical data for the same set of samples (represented by rows) from two groups: healthy (control) and diseased (case). In this figure, each block represents a pattern consisting of a subset of features associated with a subgroup of samples. Among all these patterns, all patterns except B are discriminative because they are more representative in one group than the other and thus can act as biomarkers. Patterns A and E cover most of the samples and thus can be discovered by most existing techniques. Moreover, they are coherence type patterns as mentioned in Section 3. On the other hand, discriminative pattern D of the second dataset covers very few samples and thus it is missed by most techniques. However, the pattern D covers a different subset of diseased group of samples, which none of the other two patterns A or E cover. Therefore, it may be more interesting than E of the same dataset given the pattern A of first dataset.

Another issue that integrative studies provide is that the effects of one type of variables on a subspace can be explained by the other types of data being integrated, which will be treated as a confounding factor otherwise. For example, age, gender, ethnicity and race [30, 31] can be considered as the potential confounding factors. Alternatively, genomic markers can affect different subgroups defined by clinical variables in different way [33]. Without considering different kinds of data together, such underlying confounding factors driving the potential risk may remain unnoticed. Integrative studies thus can provide new opportunities to find such insights into disease heterogeneity.

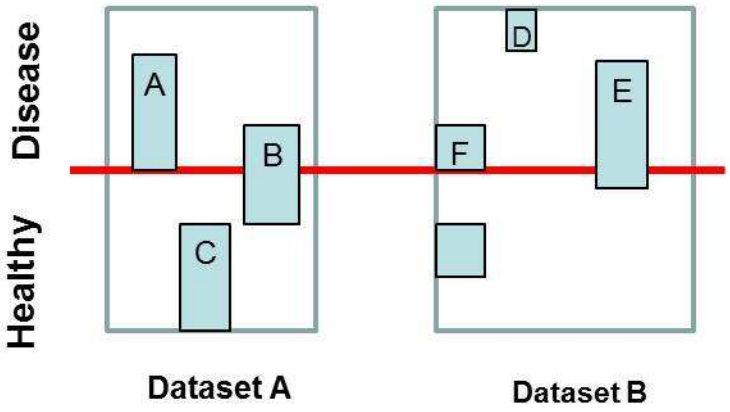


Figure 5.2: Integration of different types of patterns coming from two separate datasets

This creates the need for developing techniques that are able to find not only different types of biomarkers, but also the subgroups of patients or healthy population associated with each of those particular groups of markers. In this chapter, I will focus mainly on finding homogeneous groups of samples and the markers corresponding to each of these heterogeneous groups for integrative studies.

5.2 Related Work

Finding heterogeneous markers has been investigated by a few studies. The first type of studies analyzed how the heterogeneity of samples impacts the power of different algorithms to find meaningful biomarkers. For example, Camillo et al. [282] studied how the heterogeneity of samples affects the accuracy and stability of obtained biomarkers using several classification algorithms. Heterogeneity was produced by the intrinsic variability of the population by evolution of a pool of subjects. Moreover, the underlying gene regulatory mechanism was altered for those subgroups. Several classification algorithms were used to find accurate and stable biomarkers for several heterogeneous groups. It was observed that the stability and the accuracy of the obtained biomarkers decrease with the small sample size of the heterogeneous population. Among different classification algorithms, statistical analysis of microarray (SAM) which is a non-parametric version of the t-test and the bootstrapping algorithm performed the best. In addition to

assessing power of heterogeneous samples, Repsilber et al. [283] applied a matrix factorization based deconvolution approach for finding genes from heterogeneous tissues such as blood. However, these approaches do not aim to find the homogeneous population group directly from the dataset.

The second type of studies aim to find heterogeneous biomarkers along with the corresponding samples directly from the dataset. For example, Schwarz et al. [284] developed a two-step technique to find such heterogeneous groups of biomarkers. In the first step, they obtained all the biomarkers from a dataset using a uni-variate statistical test. In the second step, all these biomarkers were represented in a bi-partite graph with the biomarkers in one layer and the corresponding samples in another layer. Then, they used graph clustering approaches similar to [285] to find the homogeneous sample groups corresponding to those biomarkers. However, the two-step approach may miss the combinatorial markers each with weak effect on the corresponding samples. Most of the multi-variate integrative models such as canonical correlation analysis (CCA) and parallel ICA [228] are full-space models, i.e., they consider all samples to find combinatorial biomarkers. Therefore, they do not aim to find heterogeneous sample groups pertinent to a group of biomarkers directly. Some recent studies also tried to leverage the heterogeneous sample issue for integrating diverse data sources. For example, Obulkasim et. al. 2011 [286] developed a step-wise predictive model which determines automatically which samples will benefit the most by including molecular data in addition to the clinical data. First, they build two classifiers separately on clinical and molecular training data. Second, they determine the subgroup of test samples (using a re-classification score) that either lie in the decision boundary of the classifier built upon clinical variables (assuming that clinical variables are inexpensive and well-validated) or may improve the classification accuracy if molecular data is included. In particular, they project a test sample into the clinical space of training data and then, estimate the re-classification score based on how many training samples were correctly and wrongly classified in that local neighborhood. Finally, the samples with low score were reclassified using molecular data. However, this study aims at building a classification model rather than finding combinatorial biomarkers from diverse datasets as shown in Figure 5.2.

5.3 Generic Challenges

There are three main challenges for handling disease heterogeneity in integrative studies. First, finding biomarkers containing multiple types of features require searching the exponential number of combinations of features, which by itself is a computationally hard problem. The heterogeneous population groups corresponding to those combinations of features can also consist of any size and thus, finding such groups also requires exponential search on the sample space. This increases the computational complexity further. For example, in figure 5.2, the discriminative pattern D is more interesting than pattern C. Although pattern C is more discriminative than pattern D, the coverage of the pattern C is shared by the pattern F. On the other hand, pattern D covers new samples and thus is more interesting. Second, there may be both generic and specific characteristics pertinent to a subgroup. For example, knowing the overall risk factors for a particular subgroup is important for clinical decision making. On the other hand, the specific (local) characteristics that are only applicable to that particular subgroup can also be useful to understand the specific characteristics of that group, which are not applicable for other groups. Finding such types of global and local patterns pose further issues for developing computational methods. Third, each sample might belong to multiple homogeneous subgroups because of some shared phenomenon. This is very common in healthcare domain. For example, a few co-morbid patients can belong to multiple disease groups, sometime even containing diseases that are not related among themselves at all. This requires further adjustment of the methods for allowing samples to belong to multiple subgroups. Fourth, the number of samples available in each of the subgroup becomes much less than the original number of samples in the datasets. This may create a problem for the model development especially when datasets are high-dimensional. In one recent study, Karpatne et al. [287] imposed a regularized constrain on the parameters of generalized linear models (GLM) of two subgroups based on the similarity of the two subgroups computed by a cluster hierarchy. However, many times such cluster hierarchy may not be available or samples may belong to multiple subgroups simultaneously due to the co-morbidity of the patients. Note that this procedure is similar to risk adjusting variables such as mobility in a logistic regression model; however, logistic regression, even after adjustment for confounders, will provide a set of

global variables that apply to all patients, rather than finding variables that are specific only to one particular patient subgroup.

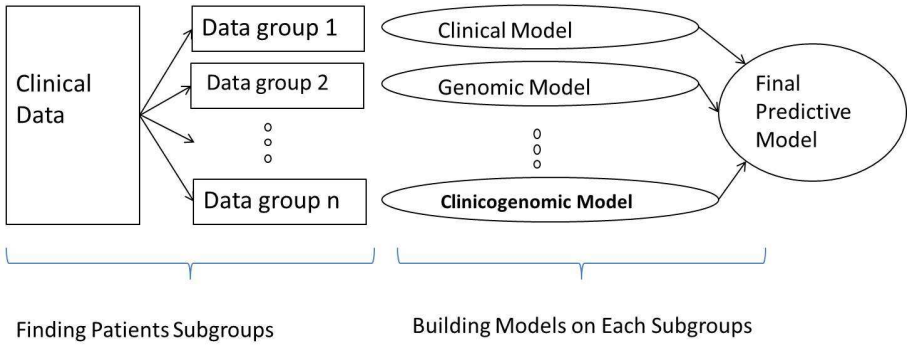


Figure 5.3: Schematic diagram for subspace models where clinical data is used first to identify the subspace. Steps for subspace models where clinical data is used first to identify the subspaces: use clinical data to stratify samples, use genomic or other left out clinical variables for each subgroups if required, and finally build a predictive model.

5.4 The Generic Framework

We propose a generic framework that can find both combinatorial markers and their corresponding sample groups directly from the data. Conceptually, given the two types of datasets, we use a two-step framework (Figure 5.3): we first identify the most dominating subgroups of samples using clinical data and then analyze each patient subgroup separately. Note that some studies reversed the order of these two steps. For example, a framework proposed by Schwarz et al. [284] first found the biomarkers from multiple datasets and then looked for the corresponding subgroups. However, such approach mainly focuses on finding the corresponding subsamples of the obtained markers, rather than finding all possible subgroups. Thus, they might miss the patterns (e.g. pattern D in Figure 5.2), which are marginally related to outcome, but which cover an interesting set of samples that are not covered by any other patterns. We aim to find all possible subgroups of patients and their corresponding factors from the data. These two steps are described below in detail.

Finding Subgroups of Patients: In this step, we aim to find all possible homogeneous subgroups of samples. The main idea is to use the putative confounding factors for

finding such confounding factors. The best option for selecting the features is to choose the features from domain knowledge. The multi-source datasets coming from rich EHRs provide a unique opportunity to obtain information about confounding factors such as age, ethnicity and gender. Note that in these cases, confounding factors are *measured* and selected by domain experts. Therefore, such domain information can be utilized for dividing the population into multiple groups. Although most of the clinicogenomic studies used only clinical information as the confounding factor, genomic data can also be used for subgrouping patients that have different degree of genetic predisposition towards disease.

Building Models on the Obtained Subgroups of Patients: Once patients have been assigned into multiple subgroups, each subgroup of samples can further be analyzed to find the potential biomarkers. Note that both predictive models and descriptive models for biomarker discovery can be designed for this purpose depending on the goal of the study. In the context of healthcare analytics, interpretability of the models is a much desired property and therefore we need to consider computational methods that are not only predictive of an outcome, but also interpretable enough to infer domain knowledge. In this chapter, we consider pattern mining techniques to achieve these goals. Moreover, we propose pattern mining methods to find both generic and specific characteristics pertinent to a subgroup.

In rest of the chapter, we will demonstrate the overall framework in the context of the large-scale Home Healthcare EHR data described earlier in Chapter 4. In particular, we will describe the problem definition, the specific details about the implementation of the methods, the obtained results and useful discussion regarding the obtained patterns.

5.5 Background and Problem Definition in Home Healthcare Application

In the context of Home Healthcare (HHC), *Mobility*, that is the ability to walk or use a wheelchair, is a strong determinant of an elderly individual's overall functional health status and ability to safely manage the home environment and personal health [288]. Mobility was defined as the ability to safely walk, once in a standing position (INDP), or with help of a device (DEVICE), or with supervision of others (SUPERV), or to

Score	Description
0 (INDP)*	Able to independently walk on even and uneven surfaces and climb stairs with or without railings (i.e., needs no human assistance or assistive device).
1 (DEVICE)	Requires use of a device (e.g., cane, walker) to walk alone or requires human supervision or assistance to negotiate stairs or steps or uneven surfaces.
2 (SUPERV)	Able to walk only with the supervision or assistance of another person at all times.
3 (CHAIR_I)	Chairfast, unable to ambulate but is able to wheel self independently.
4 (CHAIR_NI)	Chairfast, unable to ambulate and [not independent] to wheel self.
5 (BED)	Bedfast, unable to ambulate or be up in a chair.

Table 5.1: Mobility (M0700 Ambulation/ Locomotion) Score and Description and Inclusion for Outcome. The abbreviation in parenthesis will be used for referring them in rest of the paper. *INDP group does not have chance to improve, so was not used for analysis.

use a wheelchair either independently (CHAIR_I) or not independently (CHAIR_NI) [289]. Mobility was measured at admission and discharge using the OASIS question M0700: Ambulation/locomotion scored on a 6-point scale covering the aforementioned five categories and bedfast (BED) as shown in Table 5.1.

Mobility directly impacts fundamental performance of activities of daily living (ADLs), such as transferring, toileting, and bathing, as well as instrumental activities of daily living (IADLs) [290]. In 2010, 4.9 million Americans living in the community required another person in order to complete ADLs and 9.1 million Americans were unable to complete IADLs without assistance [291]. Impaired mobility may perpetuate a cycle of reduced activity, fear of falling, and social isolation [288]. Additionally, impaired mobility increases the risk of falls in the home which can result in hospitalization, periods of disability, and potential loss of independence [288, 292]. Hospitalization in itself is costly; moreover less than one-third of hospitalized older adults recover to pre-hospital function [293, 294, 295]. During home health care (HHC), only half the adults showed (46.9%) improved mobility (Agency for Healthcare Research and Quality, 2012); thus

there is a need to determine better ways to improve mobility. Improvement in mobility is a publicly reported outcome for comparing HHC quality by the Centers for Medicare and Medicaid Services (CMS) [296].

Mobility Score	Total (n = 262,035)		No Improvement Mobility Score=0 (n = 128,920)		Improvement Mobility Score=1 (n = 132,115)	
1	144,615	55.4%	99,119	68.5%	45,496	31.5%
2	89,860	34.4%	18,129	20.2%	71,731	79.8%
3	12,669	4.9%	5,322	42.0%	7,347	58.0%
4	11,339	4.3%	5,163	45.5%	6,176	54.5%
5	2,552	1.0%	1,187	46.5%	1,365	53.5%
All	261,035	100.0	128,920	49.4%	132,115	50.6%

Table 5.2: Number and Percent Patients by Mobility Score at Admission.

5.5.1 Selection of confounding factors based on domain knowledge

Although various studies found many factors related to mobility improvement in HHC patients [297, 298], the initial mobility status of patients at admission to HHC has

been the largest significant factor of improvement in mobility using Outcome Assessment Information Set (OASIS version B1) (OR = 5.96) [297]. However, mobility status alone only partially predicts mobility outcomes. Previous studies that included Omaha System interventions provide a unique contribution to understanding HHC outcomes [299, 300], but most HHC agencies do not use standardized intervention data. This observation also holds for our datasets as well. Indeed, the mobility score at admission was observed to be the strongest factor associated with improvement of mobility. Substantial heterogeneity in the outcome by subgroup was also observed: almost 80% of patients of in the SUPERV subgroup improved; a mere 31% of patients in the DEVICE subgroup improved; and, only 50-60% of the patients in other subgroups improved. Table 5.2 shows the distributions of the number of patients with improvement (mobility outcome = 1) vs. no improvement (mobility outcome = 0) across the five mobility scores at admission.

However, it is unknown if subgroups of patients based on their mobility score at admission vary in factors related with improvement and therefore it is not known if the types of interventions to improve mobility might vary by subgroup. Analyzing subgroups of patients and the variables associated with improvement of mobility outcome can help clinicians tailor interventions more effectively for improving outcomes.

Indeed, subgroup analysis of HHC patients by mobility status at admission may provide new insights for tailoring interventions or modifying the CMS risk adjustment models for more accurate comparison of outcomes across HHC agencies. Previous investigators identified differences in intervention effectiveness for community-dwelling older adults; however, there were conflicting results for types of interventions that were effective for frail elderly patients [301]. These studies support the need to examine subgroups of patients to better understand factors associated with mobility improvement.

Therefore, we use the initial mobility score as the confounding factor during grouping patient population. For this study, we took a straightforward approach of grouping all patients into five different subgroups based on the initial score of mobility at admission from 1-5. Samples with initial mobility score 0 were discarded since those patients had no chance of improvement. This stratified analysis allowed control for the strong effect of the admission mobility status and other confounding differences between subgroups.

Outcome Measure: The outcome of improvement in mobility was evaluated as

a change in mobility status from admission to discharge from HHC for patients with a $score > 0$ at admission (all except INDP subgroup). Improvement in mobility was measured as a binary variable of 1 for improvement or 0 for no improvement and was used as the only outcome of interest in rest of our paper. Figure 5.4 shows the distribution of improvement and no improvement outcome for each of the six groups of current ambulation score.

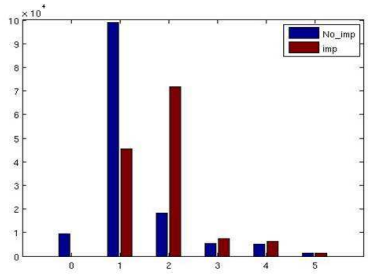


Figure 5.4: Number of samples for improvement and no improvement of mobility outcome belonging to each subgroup defined by cur ambulation score (0-5)

Potential variables associated with outcomes: OASIS data on the admission assessment includes agency information (pseudo identifier and geographical location from the previous study); demographic and patient history information (e.g., prior conditions, medical diagnoses, prognosis and life expectancy, and high risk factors affecting health, i.e., prior smoker; support systems (living arrangements and caregiver support); health status (e.g., sensory, integument, respiratory, emotional, cognitive, and behavioral); activities of daily living (ADLs) and instrumental activities of daily living (IADLs); medication and equipment management; and service utilization such as the need for therapy, emergent care, and discharge status. The complete OASIS assessment form and manual can be found on CMS's website (http://www.cms.hhs.gov/HomeHealthQualityInits/; last accessed 01/25/14). The OASIS data set was developed through 15 years of successive studies for reliable and valid data collection and outcome measurement in home care [302].

5.5.2 Analyzing each subgroup of patients

Since our goal of this study was understanding the disease mechanism better, interpretability of the obtained model was a much desired property. Therefore, we relied on finding patterns that are not only predictive of the mobility outcome, but also easily interpretable to domain experts. Another nice property of the pattern mining is that they can search for potential combinations of multiple variables and their corresponding samples simultaneously and efficiently. Most existing studies of mobility outcome assess single variables independently rather than analyzing the influence of multiple variables together. Often multiple variables are important to a particular subset of patients due to the heterogeneity of health and functional problems experienced by HHC patients. For example, two variables together, such as ability to dress and groom, may have different effects in a particular subset of patients than the overall effect for all patients. Thus, there may be several subsets of patients with different sets of variables (patterns) associated with outcomes. Finding these different patterns that are pertinent to different patient subsets may provide further knowledge about the relationships among the individual variables, and thus can potentially enhance clinical insights relevant for improving mobility. Therefore, we mainly considered a recent discriminative pattern mining based framework [305], since it is very effective to find the combinatorial patterns and the patterns are also easily interpretable.

Moreover, understanding the generic and specific patterns associated with mobility improvement for HHC patients (local patterns) within stratified subgroups using their mobility score at admission may infer new knowledge. In particular, the research questions addressed in this study are: 1) What are the patient and support system characteristics (generic pattern) associated with mobility improvement within the five subgroups of patients defined by their mobility score during admission (scores of 1 - 5)?, 2) Are there any local patterns specific to each of the five different subgroups, when contrasted with other subgroups?

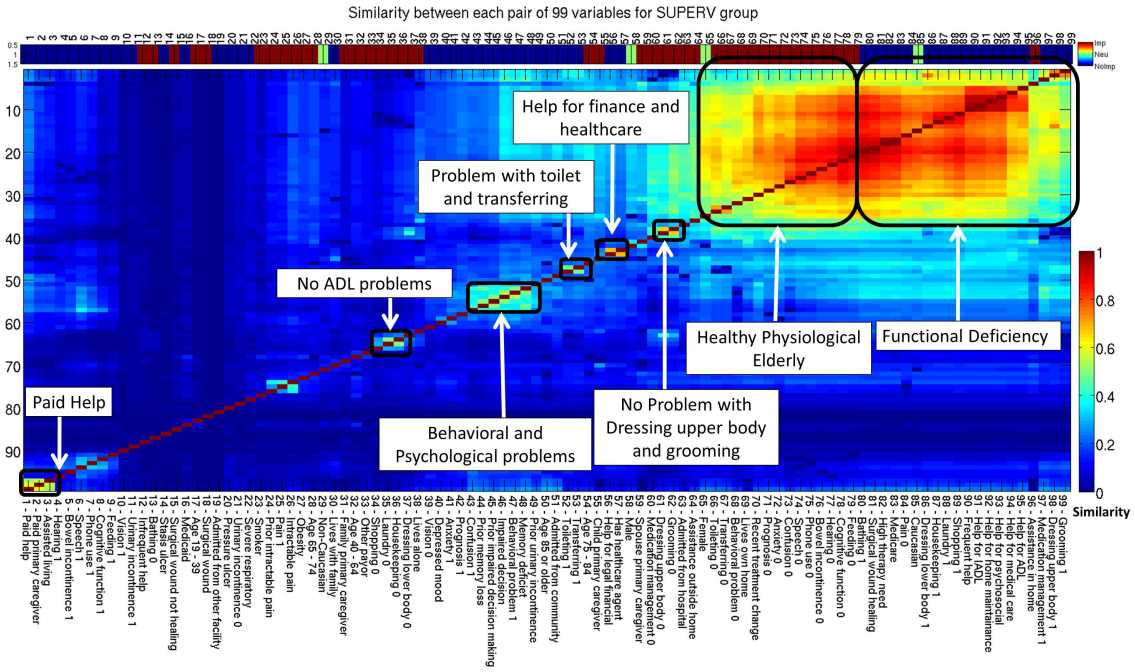


Figure 5.5: Coherence of any two variables out of 99 variables for the SUPERV group (mobility at admission = 2) is shown in the bottom panel, where x and y axes are OASIS variables. Each cell of this matrix represents the similarity between the corresponding two variables (represented by a row and a column), with red being highest similarity and blue being lowest. The top panel of the figure represents the association of each individual variable with mobility (red = improvement, blue = no improvement, and green = not significantly associated with OR close to 1).

5.5.3 Methods for Finding Generic Patterns

The pattern mining was performed in several steps for each of the five subgroups based on initial ambulation score for finding the generic patterns. First, the data were examined for variables associated with mobility improvement within each subgroup, and also evaluated for the consistency of those variables across the patient subgroups. Note that the variables associated with the improvement in mobility with odds ratios (ORs) < 1 were also analyzed in later steps and referred to as variables associated with no improvement of mobility outcome in rest of the paper. Second, the individual variables with significant ORs were grouped together into clusters (patterns) in order to expose

higher level, more interpretable relationships among individual factors.

Single-variables analysis: For each subgroup, the set of significantly predictive variables were identified using OR as a measure of discrimination (effect size). When the OR for a predictive variable associated with improvement was< 1.0, 1/OR was reported as an association with no improvement. Statistical significance was established by using a false discovery rate (FDR) of < 0.05 (FDR was computed to adjust for multiple hypothesis testing). In each patient subgroup, some sets of variables predictive of the outcome applied to roughly the same set of patients. We refer to this characteristic as coherence. (See Figure 5.5 which shows coherence of any two variables with each other based on whether they represent a common subset of patients as measured by the Jaccard coefficient. This coherence illustrates the existence of clusters (patterns) among the variables. For example, clustering the rows and columns resulted in nine clusters (variable numbers 1-3, 6-9, 24-26, 33-37, 55-56, 59-61, 64-78, 79-99).) As a result, grouping the individual variables can enhance clinical interpretability, since many of the variables represent a broader view of the patient's health and functional status. For example, behavioral problems, cognitive skills, and memory deficiency may be representative of a more general cognitive or psychological disorder in a patient. Thus, we further sought to group individual variables into patterns within each mobility subgroup based on their cohesiveness (whether they co-occur in the same patients) and whether they are predictive of the mobility outcome.

Multivariate pattern analysis. Grouping variables into patterns that are associated with improvement in mobility simultaneously is very challenging due to the exponential number of possible variable combinations that needs to be evaluated. Traditional clustering techniques [303] can group variables into patterns directly, however, they fail to discover all meaningful patterns and they also report incorrect patterns as they cannot take the outcome variable into account. In what follows, a technique for discovering such patterns efficiently is described.

Association analysis [304] was applied to identify sets of variables (patterns) associated with each other and also with the mobility outcome in each subgroup. More formally, a pattern is defined as a set of variables (characteristics) that have certain properties of interest. The most common property that any pattern should have is that if each binary variable represents the presence of a characteristic, e.g., memory deficit or

urinary incontinence, then all the variables in a pattern should co-occur in a sufficiently large number of patients. While exponentially many patterns can potentially be discovered, the choice of an association rule mining algorithm can often make the computation tractable by eliminating patterns that occurred in too few patients. Additionally, these patterns can be tested for further properties; cohesiveness and discriminative power. In this study, we were also interested in finding patterns that not only occur in a large number of patients (cohesiveness), but also can discriminate between improvement (mobility outcome =1) versus no improvement (mobility outcome =0) using a recent extension of association rule mining techniques called discriminative pattern mining [242, 305]. Patterns can be constructed using one or more variable(s), however, for brevity, in this paper we only report results for patterns consisting of at least two variables.

	Has Pattern (P=1)	Lacks Pattern (P=0)
Improved	n_{11}	n_{12}
Not Improved	n_{21}	n_{22}

Table 5.3: Contingency Table of a Pattern in relation to the mobility outcome. The second column represents the number of patients where all the variables of the pattern are present (has a value 1) and the second columns represents number of patients where at least one of the variables of the pattern is absent (value 0).

Discriminative pattern analysis was conducted within each subgroup of patients separately. The discovered patterns were evaluated based on two properties: their discriminative power and the cohesiveness of the two variables belonging to the pattern. The discriminative power of the patterns was assessed using the odds ratio (OR), which is calculated as $n_{11} * n_{22}/n_{12}/n_{21}$, where $n_{ij}$is defined by the 2x2 contingency table shown in Table 5.3. Specifically, n_{11} (and n_{21}) denote the number of patients who exhibit the pattern and had improvement or no improvement, respectively, in their mobility score. Analogously, n_{12} and n_{22} denote the number of patients who do not exhibit the pattern and had improvement or no improvement, respectively. High OR signifies high odds of improvement in mobility.

The cohesiveness of the patterns was measured by Jaccard similarity. It is defined as $f_{11}/(N - f_{00})$, where f_{11} is the number of patients where both variables in the pattern have a value of 1, f_{00} is the number of patients where both variables have a value of 0, and N is the total number of patients. Thus, the Jaccard coefficient indicates

how consistent the co-occurrences of the two variables are across the patients: a high value signals that the variables frequently co-occur, potentially suggesting that they may be related. Coincident absences of the variables are ignored. Moreover, we put an additional constraint on the cohesiveness of discriminative power of the variables such that the individual discriminative powers (ORs) of the constituent variables of a pattern are within a range.

An ideal pattern is highly discriminative (very predictive of improvement as measured by the OR) and highly cohesive (the two variables in the pattern frequently co-occur together within a subset of patients as measured by Jaccard). Discriminative pattern mining aims to discover patterns with precisely these properties. To visually study the relationships and to identify patterns of related variables, a graph was constructed. A graph is a mathematical construct consisting of nodes representing variables and edges that connect the nodes. An edge was drawn between two nodes (variables) if they co-occurred in a discriminative pattern and had odds ratio and Jaccard coefficient that exceeded predefined thresholds. The resultant graph was analyzed through the Cytoscape graph visualization software [306]. Of particular interest within a graph are the connected components. A connected component is a subset of nodes that are densely connected by edges but are not connected to any node outside the component. Connected components represent larger patterns where the constituent variables that are strongly pairwise related to each other and can therefore represent underlying patient conditions. These components are similar to clusters, but also take the discrimination power of the patterns into account. Note that these grouping in terms of patterns were performed among variables, not among patients; therefore there are patients who can belong to multiple patterns and there may be patients who do not belong to any pattern.

5.5.4 Finding local patterns that are specific to a particular subgroup

We also aimed to find all possible local patterns that are only pertinent to a specific subgroup of patients. Therefore, discriminative pattern mining not only has to consider the particular subgroup under consideration, but also contrast that subgroup with all other subgroups in the study so that the obtained patterns become applicable only to the subgroup under consideration. The motivation for finding such local pattern can be illustrated by Figure 5.6. In this figure, the blue pattern contains more specific local

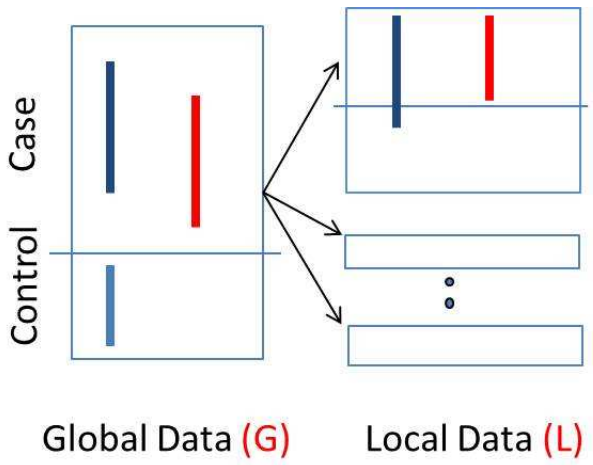


Figure 5.6: The schematic diagram containing the local patterns

information than red one, since the discrimination power increases substantially from global to local patterns.

In this section, we will describe methodologies to find such discriminative local patterns which are predictive of outcome for a particular subgroup, but distinct enough from other subgroups as well.

An association analysis based measure: Local discriminative patterns have to have two properties: 1) they are predictive of the disease outcome and 2) the local patterns are distinct enough from the other groups. For measuring the predictive power of a pattern, we used diffsup [305] because of its nice property of anti-monotonicity. The anti-monotonicity property guarantees that if a pattern is not discriminative enough then none of its superpatterns (i.e., patterns containing all the features of the original patterns and some additional features as well) is. The diffsup measure is defined below:

$DiffSup_P = Sup_p^+ - Sup_p^-$

To measure the the second criterion of locality of a pattern to a particular subgroup in contrast to other subgroups (global signal), we propose the following measure as shown below:

$LocalGainOriginal_P = DiffSup_p^L - DiffSup_p^G = (Sup_p^{L+} + Sup_p^{G-}) - (Sup_p^{L-} + Sup_p^{G+})$

Here, $DiffSup_p^L$ measures the discrimination power of the local pattern for the

corresponding subgroup under consideration and the $DiffSup_p^G$ measures the discrimination power of the same pattern P on other subgroups besides the one under consideration.

Note that this original formulation of LocalGainOriginal is not anti-monotonic, since the Diffsup of the local and global patterns can both increase and decrease from subpatterns to superpatterns. We used a simple trick to make this measure anti-monotonic as described below.

$LocalGain_P = (Sup_p^{L+} + Sup_p^{G-}) - max_{\alpha \subseteq P}(Sup_\alpha^{L-} + Sup_\alpha^{G+})$

Lemma: The measure $LocalGain$ is anti-monotonic.

Proof: Let Q be a subpattern of P. Then, the second term of the equation for a pattern P can only increase from its subpatterns Q, given the anti-monotonicity of the Sup measure. Now, the first term of P can only decrease from that of its subpattern Q. Therefore the LocalGain of P can only decrease from that of its subpattern Q.

This anti-monotonicity property will guarantee to find all possible local patterns if they are used in the apriori framework. However, note that this LocalGain measure is an approximation of the original LocalGainOriginal measure. We aim to explore the difference between the sets of patterns discovered by these two measures.

5.6 Results

Demographics of the sample and reasons for HHC are shown in Table 5.4. Patients were predominately white, older adults, with more females. Medicare was the most frequent payor and about half the patients were admitted to HHC following a hospitalization. The majority received home care for less than 60 days and about a third had care for less than 30 days. About two-thirds had a good prognosis and were expected to regain full or nearly full recovery and functionality. Many had chronic health conditions with the category of "Symptoms, Signs, and Ill-Defined Conditions" indicating that physical therapy and rehabilitation are the primary reasons for HHC. Overall, 49.4% of patients improved in mobility; while 50.6% did not improve. The percent of patients who improved or did not improve within each level of mobility status at admission is shown in Table 5.2

Results for Single Variables Associated with Outcomes: Single variables

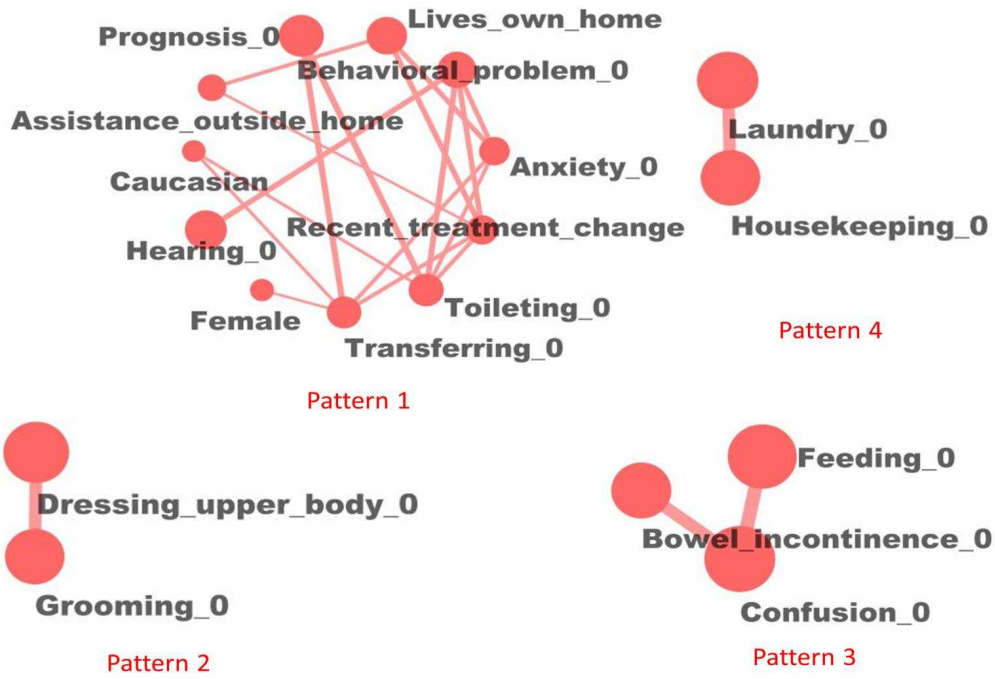


Figure 5.7: Patterns associated with IMPROVEMENT in the SUPERV group.

significantly associated with improvement (mobility outcome = 1) vs. no improvement (mobility outcome = 0) in mobility for all patients by the admission score are shown in Appendix Figure A.6 (with Odds Ratios (OR) for all Single Variables Significantly Associated with All Patients within each subgroup defined by the Admission Score for Mobility). There were no significant variables for patients in the BED subgroup, and therefore, these patients were dropped from further analysis. Many significant variables were found and the number of variables varied by patient subgroups (n = 27 - 60). In the DEVICE subgroup, the odds of improvement were highest for younger, post-surgical healthy patients needing infrequent help, who had adequate support. In contrast to this subgroup, the highest odds for no improvement represented frail chronically elderly with poor health, psychosocial, and functional status that needed considerable help. In the SUPERV subgroup, patients likely to improve were similar to DEVICE subgroup except they were slightly older and the ORs were more strongly associated with improvement if they had little or no problems with cognition. Patients in the SUPERV subgroup were

not likely to improve if they were chronically ill, very old with moderate to severe health, functional status, and psychosocial problems, and needed frequent help. In both the CHAIR_I and CHAIR_NI subgroups, the highest ORs included older age, chronic health problems, and difficulties with ADLs and IADLs. They differed in behavior problems. The highest ORs with no improvement for both CHAIR_I and CHAIR_NI subgroups were younger, chronically ill patients with ADL and IADL difficulties.

Significant variables that were common across at least 3 of 4 subgroups for mobility improvement include age 65 - 74, source of admission (hospital); prognosis (good); lives in own home; infrequent help needed; little or no problem with hearing, speech, bowel incontinence, behaviors, or confusion; and little or no problem with almost all ADLs and IADLs. Significant variables for no improvement in mobility common across at least 3 of 4 mobility levels include having paid help; needing help from primary caregiver for ADLs, power of attorney (financial, legal, and medical), and medical support; moderate to severe problems with vision, speech, urinary and bowel incontinence; poor cognitive functioning; having a pressure or decubitus ulcer or surgical wound not healing; cognitive and behavior problems; and moderate to severe problems with most ADLs and IADLs. The pattern for patients in the CHAIR_I subgroup was distinctly different than the DEVICE, SUPERV and CHAIR_NI subgroups.

Results for Patterns of Variables Associated with Outcomes: While patients in the CHAIR_I subgroup were the most different for both improvement and no improvement in mobility, further analysis was focused on the SUPERV subgroup because these patients had the most improvement. Figure 5.7 and 5.8 shows the patterns with $ORs > 1.4$, $Jaccard similarity \geq 0.4$, and with a range of ORs of the constituent variables of the patterns within 0.7 for patients in the SUPERV subgroup resulting from discriminative pattern mining analysis. Patterns 1 - 4 represent improvement (red circular nodes) vs. Patterns 5 - 9 represent no improvement in mobility (blue hexagonal nodes).

Patterns can be interpreted by visualizing both the circles and edges. The size of the circles represents the magnitude of individual ORs of each variable present in the patterns. The larger the circle / node in the pattern, the more likely the variable is associated with the outcome. The width of the edge represents the OR between two variables (nodes); the wider edge indicates a higher OR. Note that the joint OR of two

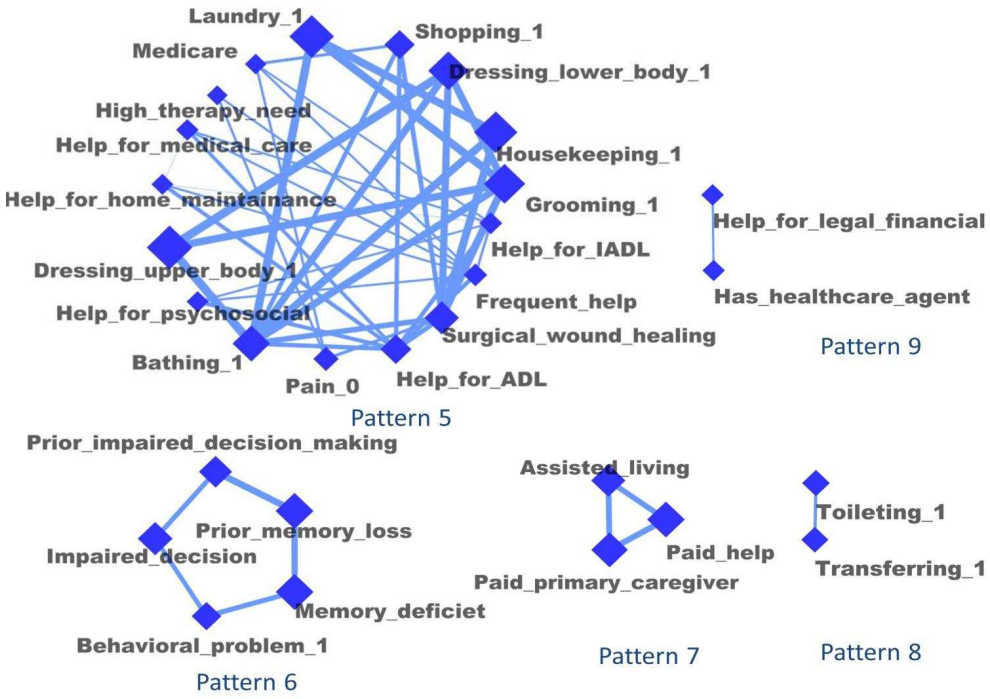


Figure 5.8: Patterns associated with NO IMPROVEMENT in the SUPERV group.

variables in a pattern may be different from the OR in SDC2 for individual variables. The notation after a variable name in the cluster is 0 = little or no problem and 1 = moderate to severe problems. A variable without a 0 or 1 indicates the presence of that variable such as High Therapy Need. There are four patterns for improvement and five for no improvement. Improvement patterns are associated with little or no problems, whereas no improvement problems are associated with moderate to severe problems, except for no pain in one of the patterns. The number of variables in a pattern is shown in parentheses and descriptions for improvement are as follows:

- Pattern 1: (p = 11) Independent with recent treatment change and good prognosis, as well as good functional and psychosocial status
- Pattern 2: (p= 2) Little or no problem with dressing upper body and grooming
- Pattern 3: (p = 2) Little or no problem with laundry and housekeeping

- Pattern 4: (p = 3) Little or no problem with confusion, feeding and bowel incontinence

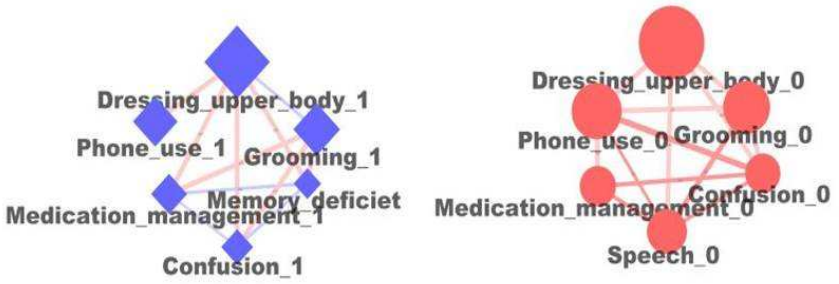


Figure 5.9: Local patterns associated with IMPROVEMENT and NO IMPROVEMENT in the SUPERV group. Patterns are interpreted by visualizing both the circles and edges.

Descriptions for patterns with no improvement in mobility are:

- Pattern 5: (p = 17) Frail requiring considerable help
- Pattern 6: (p = 5) Psychosocial impairment.
- Pattern 7: (p = 3) Requires paid help
- Pattern 8: (p = 2) Requires help with complex decision-making
- Pattern 9: (p = 2) Impaired mobility related ADLs

Other subgroups were also analyzed and many variables are consistent across subgroups. (Appendix Figures A.7 A.8, A.9, A.10,A.11, A.12 demonstrating the similarities and differences in patterns for patients by admission score for variables associated with improvement vs no improvement.). Examples of similar variables in all four subgroup patterns are functional and cognitive status, and the type and amount of help required.

The combination of variables within a pattern represents unique characteristics associated with improvement vs no improvement within each mobility level at admission; however, variables found in the CHAIR_I subgroup were quite different than those found in other subgroups. There were only six variables similar in the CHAIR_I subgroup patterns and with variables in any other subgroup for mobility improvement: Caucasian; female; and, little or no problem with speech, bowel incontinence, cognitive function, and use of a phone. Healing of a surgical wound was the only variable associated with no mobility improvement that was both in a cluster for CHAIR subgroup and other subgroups. Additionally, there were variables that uniquely appeared in the clusters of subgroup CHAIR_I. For improvement in mobility, these include: Medicare, needs frequent help, has a high need for therapy, and difficulty with dressing upper body, bathing, transferring, housekeeping, and shopping. In other subgroups, functional status variables were opposite with values of little or no problem when associated with mobility improvement. For no mobility improvement in the CHAIR_I subgroup, unique variables were admitted from the community (rather than the hospital or other facility) and needs assistance from a health care agent or medical power of attorney and a financial/ legal, power of attorney.

The local patterns: We also look for the patterns that are specific to a particular subgroup. To find such local specific patterns, we applied the proposed framework with multiple thresholds for diffsup (measuring the discrimination power) and LocalGain (measuring the local information for that particular group). Furthermore, we also validated the local patterns similar to the generic patterns as described earlier using the random permutation of the class label. Table 5.5 contains the number of patterns obtained for various thresholds of Diffsup and LocalGain. It was evident from the table that subgroup 2 (SUPERV) has the highest number of patterns. We also visualized these patterns to compare with the generic patterns obtained from the same group as shown in Figure 5.9. When compared with those obtained from generic pattern sets, the local patterns provided a few new insights. For example, confusion, medication management and phone use was not discovered in the generic patterns, and thus represents the special phenomenon that is only pertinent to this group SUPERV.

5.7 Significant Discoveries and Discussion

The obtained patterns resulted in a few interesting discovery from domain perspectives as well. Where recommendations with respect to home health care occur in the following section, they are recommendations from my collaborator, Bonnie Westra (Associate Professor and Director, Center for Nursing Informatics, University of Minnesota), which were based on the data mining results I generated. They are included here to show the potential usefulness of the data mining results.

This study confirms the high prevalence of mobility limitations in primarily older HHC patients, their low rate of improvement by discharge, and provides new information on the factors associated with both those who do or do not improve in mobility by level of mobility at admission. In this study, 97% of patients had at least some difficulty with mobility at the time they were admitted to HHC. In other studies, problems with functional status for older adults in the community were high; however, no studies were found specifically reporting the rate of difficulty with mobility and in particular for patients receiving HHC [291]. This study confirmed previous findings that about half the patients improve in mobility; AHRQ identified 46.9% of patients improved while this study found that 49.4% improved [307]. About three-fourths of patients were discharged in less than two months, with one-third receiving care for 30 days or less. In addition, more than 80% were age 65 or older.

This study also confirmed that mobility status at admission has the highest association for likelihood of mobility improvement by discharge from HHC. Using OASIS B1 data, CMS identified that the odds of improving in mobility increase by 5.96 for each increment of 1 point on the ambulation/ locomotion assessment question. Compared to CMS, this study showed the OR differed between each level of ambulation/ locomotion based on the admission score. Patients with a score of 1 at admission were less likely to improve (OR = 6.3) and those with a score of 2 were more likely to improve (OR = 7.3). Patients with a score of 3, 4, or 5 for mobility at admission had an OR of 1.28. Therefore, there is not a consistency in the odds of mobility improvement predicted across all mobility scores at admission.

Risk Adjustment: CMS calculates and reports risk-adjusted outcomes to HHC agencies to conduct outcome-based quality improvement and publicly compares home

care agencies on outcomes [297]. This study did not use the CMS risk adjustment methodology; however, it is interesting to note that many of the variables associated with mobility outcomes in this study differed from those found by CMS for OASIS B1. These differences may be due to the way in which variables were used (such as binary variables) as well as the rules applied when different data-driven methodologies were used. The findings in this study are novel in that significant factors associated with mobility outcomes were found that differed by the admission score. Therefore, it may be more accurate to report risk adjusted outcomes by subgroups of patients to HHC agencies and the public.

Global Outcomes: Overall, patients with a score of 1 (DEVICE subgroup) for mobility at admission (N=144,615) were less likely to improve (69%). In comparison to the DEVICE subgroup of patients, there were 89, 860 patients of SUPERV subgroup with 20% who did not improve; patients in CHAIR_I, CHAIR_NI and BED subgroups had approximately 45% of patients not improving. Therefore, it is likely there is a ceiling effect for measuring improvement using OASIS. This finding is consistent with other measures of ADL, such as the Katz ADL Scale [308]. In a previous study, methods were compared to determine an adverse event for functional decline. A floor effect was found; as a result a recommendation to CMS was made to change how to best calculate adverse events for functional status [309]. A similar analysis might be useful for improvement outcomes.

Patterns for Patients of SUPERV Subgroup: In the SUPERV subgroup, patients were most likely to improve in ambulation. Patterns of variables associated within this subgroup were further analyzed to better understand their risk for not improving. Furthermore, we also analyzed whether the obtained patterns are novel beyond the existing literature and whether this would provide actionable information. Tailoring of interventions for subgroups of patients could increase the percent of patients with improvement in mobility at discharge from home care.

Improvement in Mobility: Patients with mobility improvement were associated with four patterns representing overall patients who were healthier and had little or no functional impairment. Pattern 1 is consistent with results from other studies; healthier patients are more likely to improve in outcomes and regain their ability to function safely at home [310]. Pattern 2 included two ADL items and Pattern 3 included two IADL

items. Fortinsky et. al applied Rasch modeling to OASIS data to determine if ADL and IADL items represented a uni-dimensional scale and the order items in terms of difficulty [311]. ADL items were found to be easier and IADL more difficult. Our findings did not support a uni-dimensional relationship between functional status variables and mobility improvement; rather our study found unique patterns of functional status variables represented distinct patient subsets related to mobility improvement.

No Improvement in Mobility: There were five patterns associated with no improvement in mobility. Pattern 5 represents patients who might be considered "frail" and who are less likely to improve compared with non-frail patients. In previous studies frailty, poor life expectancy, or disease burden and chronic conditions that affect ambulation were associated with overall poorer functional status [312]. In another study, frail elderly patients were less likely to improve their mobility status with strength and balance exercises compared with less frail patients [310]. However, in our study, we observed novel patterns of poor ADL and IADLs requiring frequent help. It may be that an occupational therapist would be helpful since many variables for this pattern are related to ADLs and IADLs. Learning to use adaptive equipment may increase confidence and decrease fear of falling, thus improving mobility. Pattern 7 is composed of patients with paid services, including assisted living and lack of family to support their ongoing needs. Patients who can no longer live at home and have no or insufficient family help are on a functional decline and at risk for nursing home placement. While not surprising, this is a unique pattern requiring different strategies for supporting mobility and planning for the future. Nursing should consider a referral to social workers to determine additional financial resources to assure the adequacy of paid services. While patients in this clustering analysis received a score of 2 for mobility at admission to HHC, the poor score on additional functional status items, particularly those requiring mobility (toileting and transferring represented by pattern 9), places them at risk for falls as well as their ability to perform more difficult functions.

Two patterns focus more on cognitive rather than physical functioning. Pattern 8 demonstrates some challenges with requiring assistance with decisions about finances and healthcare and may reflect minimal to moderate cognitive impairment, which is distinctly different than pattern 6 which represents considerable cognitive ability that could influence daily living to remain safely at home. Cognitive decline is associated with

poor mobility in previous studies and risk of falling. When patients perceive they are at risk of falling, it further impairs their ability to improve in mobility [313]. Cognition can be improved through fitness, which in turns improves mobility [314]. Nurses should consider various ways of engaging HHC patients in fitness programs, beginning with therapy in the home and use of TV or video exercise programs, followed by engagement in community fitness programs after HHC discharge.

5.8 Future Work

Further research is needed both in domain side and computational method development.

Domain related future work: There were several limitations in this study. Patient care episodes were drawn from a convenience sample of HHC agencies. Data used for this study were collected for clinical documentation and not research; therefore, it is likely that there are inconsistencies in the way the data were recorded. The lack of information about interventions performed during HHC limits this study; the OASIS assessment data provided consistency, but which interventions were provided remain unclear. We did not analyze the degree of change, but rather the outcome was a binary variable of either improvement or no improvement. If the degree of change were represented in the outcome, it could change the patterns discovered. Furthermore, the patterns obtained for each subgroups need to be validated. This study further demonstrated the effective reuse of OASIS data from EHR across multiple agencies and software vendors to gain new insights. However, the addition of standardized terminology for nursing and other clinician interventions would increase the value of the EHR data for knowledge discovery in the future.

Future studies need to include process measures or interventions to understand the impact of clinical care in addition to patient and support characteristics that influence outcomes. OASIS C is the current version used by HHC agencies since January 1, 2010 which includes process measures. These process measures may explain more variance in outcomes. Additionally, state (http://www.health.state.mn.us/ehealth/advcommittee/index.html) and national efforts are in process to have sharable/

comparable nursing data in every healthcare setting (http://www.nursing.umn.edu/about/calendar-of-events/2014-events/big-data-2014/Agenda/index.htm). Standardization of intervention terms would facilitate the ability to better understand of the influence of care provided on patient outcomes.

Method related future work: In this context of OASIS data, several future directions can be pursued both steps of the generic framework, i.e., finding patient subgroups and developing models within each subgroup.

Often time the confounding factors are not measured properly, i.e., they cannot be determined from the domain knowledge completely. For example, patient's diagnosis codes can be a big confounding factor exhibiting different disease characteristics for different disease. In these cases of *unmeasured* confounding factors, we aim to find such groups by clustering the samples based on the features of interest. Note that the details about the clustering and the features on which the clustering will be performed depend on the particular applications. Since specific domain information about the confounding factors are not available here, we suggest using as many features as possible for clustering the patients. Also, overlapping clusters should be considered because some sample may belong to multiple clusters. Co-clustering based approaches can be more useful, since each possible subgroup is affected by only a subset of features. However, this poses computational challenge for integrative studies. Since the properties, types and formats of each of the datasets being integrated are very different, co-clustering techniques have to handle such disparate properties. An easier way will be to compute a separate distance metric for each type of datasets and then combine them to get the final distance metrics between samples. This co-clustering approach will also have the additional benefit of making the discovered subgroups more interpretable, i.e., each subgroup of sample can be characterized by the corresponding feature sets it belong to.

During model development strategy, although we analyzed the global and local patterns for each subgroups, the relationships between the global and local patters were not investigated directly. Alternatively, a predictive model can also be developed on each subgroups of patients. Moreover, mixed-effect statistical learning can be used to model each patient separately within each group. In such modeling the fixed-effect will measure the overall effects of the covariates and the random-effect models will measure the effect within each subgroup.

Finally, the two steps involved in the frameworks (i.e., grouping samples using clustering and building models for each group) can be combined in a single step. For example, clustering techniques in the first step can utilize the outcome variable especially in supervised setting to make the samples not only similar to each other within the same group, but also have similar outcome. This might increase the prediction power and interpretability of the overall model.

Demographic	Percent
Ethnicity	
White	83.2
Non-White	16.8
Gender	
Male	35.5
Female	64.5
Payer	
Medicare	89.7
Medicaid	5.3
Other	11.9
Location Prior to Admission	
Hospitalized in last 14 days	46.8
Admitted from other facility	20.5
Admitted from the community	35.7
Home care length of stay	
0 30 days	31.2
31 60 Days	41.0
¿ 60 days	37.8
Primary diagnosis (reason for admission)	
Symptom, signs, and ill-defined conditions	23.9
Circulatory system diseases	21.5
Injury and poisoning	10.9
Musculoskeletal system and connective tissue diseases	7.8
Respiratory	6.8
Endocrine, nutritional, and metabolic diseases and immunity disorders	6.2
Diseases of the nervous system and sense organs	6.1
All others	16.9

Table 5.4: Demographics and Reason for Admission to HHC

Ambulation	Diffsup	LocalGain	Singletons	Pairs	Triplets
Group 1 (DEVICE)	0.1	0.1	8	12	0
Group 1 (DEVICE)	0.1	0.2	0	0	0
Group 1 (DEVICE)	0.2	0.1	0	0	0
Group 2 (SUPERV)	0.1	0.1	37	272	1202
Group 2 (SUPERV)	0.1	0.2	14	33	42
Group 2 (SUPERV)	0.2	0.1	12	24	23
Group 2 (SUPERV)	0.2	0.2	10	17	12
Group 3 (CHAIR_I)	0.1	0.1	2	0	0
Group 3 (CHAIR_I)	0.1	0.2	0	0	0
Group 3 (CHAIR_NI)	0.1	0.1	16	52	93
Group 3 (CHAIR_NI)	0.1	0.2	6	6	2
Group 3 (CHAIR_NI)	0.2	0.1	7	9	5
Group 3 (CHAIR_NI)	0.2	0.2	6	6	2

Table 5.5: Number of patterns discovered for various threshold of Diffsup and LocalGain

Chapter 6

Conclusion and Future Work

6.1 Contribution

The thesis focused on reviewing and solving some of the key issues in the topic of integrating multiple biomedical datasets. In particular, it aims to make significant contributions to both computer science and the healthcare/biomedical domain.

6.1.1 Data Mining Contribution

The thesis developed novel techniques that can solve some of the domain related issues. First, the pattern mining based integration framework is a generic approach for integrating multiple types of data. In addition, it can handle the disparate dimensionalities of the data to some extent. Moreover, it can also find higher-order relationships such as interaction and coherence among diverse factors, unlike most of the prominent approaches for the integrative biomarker discovery, such as discriminative canonical correlation discriminative (dCCA) and related multi-variate techniques. Second, the Sparse Hierarchical Canonical Correlation Analysis (SHCCA), a multivariate integrative model, tries to enhance the interpretability of obtained integrative patterns, so that they can be evaluated by the domain experts for clinical decision making. Note that this technique is also a useful framework for incorporating relationships among features into model development for finding relationships and therefore, can be applied to many other domains where the relationships among features are known. Third, the thesis

aimed to find more localized patterns that are only applicable for a particular subgroup of patients. This kind of subgroup patterns can potentially unravel novel disease phenomenon, which many of the fullspace models will miss.

6.1.2 Domain Contribution

In the domain of biomedical Informatics, we tried to make also novel contributions. Since the research is interdisciplinary, I have collaborated with many experts in biomedical domain to properly understand the domain issues and for better deployment of the computational approaches in the domain. For example, I worked on a project to analyze the genetic, functional and structural activity of Schizophrenia patients collected from three different modalities ofdata such as functional MRI, SNP and structural MRI collected by the department of psychology, psychiatry, and neuroscience. Also, I actively worked on several health care projects, where data are collected from large-scale electronic health records (EHR). The techniques that was developed in this thesis was heavily applied on the above-mentioned real world datasets, each of which resulted in some novel discoveries in the domain which was described in the earlier chapters of the thesis.

6.2 Future Work

Since the topic of integration of diverse biomedical datasets is still in early stage of research, several future directions can be pursued. In this section, I will describe a few broad set of techniques to advance the domian of integration of biomedical datasets further.

6.2.1 Finding Relationships

In many biomedical applications, there may exist wide varieties of relationships between interaction and coherence, e.g., synergy, moderator, marginally interactive and coherence. Furthermore, finding causal relationship among the diverse factors is very much of interest to domain experts. Another important issue of integrative study, especially in mental health, is that often the disease endpoint is measured by several intermediate pathogenic phenotypes. Building a global disease progression model containing

relationships among diverse clinical and genomic factors leading to such intermediate phenotypes and finally to disease will be useful.

6.2.2 Handling Disparate Data

Often, the natures of diverse clinical and genomic datasets are very different. In particular, there are differences in data formats, types, properties, dimensionalities, and the amount of noise present in these data due to the differences between the experimental design and data collection protocols. Kernel-based predictive models have been applied in such cases, where kernels are learnt separately for each dataset respecting the individual nature. However, several issues remain unanswered. Examples include: how to choose kernels appropriately and how to merge them. Since kernels transform the data into another feature space, interpreting the results of the predictive models for domain expert can be difficult. These issues require exploring new data mining approaches. For example, alternative models such as graphical models can be used for this purpose.

6.2.3 Handling Large-scale Data and Dimensionality

As more and more diverse clinical and genomic factors are collected for the same patients, the overall dimensionality increases quite rapidly, while the number of samples remains the same. Thus, the issue of finding statistically significant factors from integrative models is even more challenging. Applying sparse machine learning approaches can be useful for such cases. However, different degree of sparsity may be needed for each dataset depending on the amount of co-linearity present in each dataset which requires further research. Moreover, all features are not equally important in the domain. Sometimes, there are underlying relations among the features, e.g., among genes and proteins through pathways. I plan to incorporate this information into the model development to further constrain the learning process. This will not only increase the stability of the model but also their interpretability. Moreover, large-scale EHRs are expected to be generated by 2017 leading to terabytes of patient records. This will increase the statistical power of the models greatly. The advantage of techniques coming from b́ig dataśolutions such as Hadoop and Spark can be applied on those datasets.

6.2.4 Temporal Modeling

Much of the large-scale EHR data are longitudinal in nature where diagnosis, lab results, interventions and the after-care complications are collected over a long period of time. However, there are challenges unique to longitudinal EHR data, which traditional temporal models cannot handle. First, the time-stamps are highly irregular in nature, since the data collection depends on a patients visit to hospital which can vary randomly. Second, longitudinal EHR data are collected for secondary analysis rather than in a controlled study to account for experimental bias. Often, medical guidelines and follow- up treatments are applied to patients throughout their longitudinal history, which can ultimately alter the health status of the patient, diagnosis and the lab tests. This makes population cohorts highly heterogeneous in nature at any point of time. Second, before performing any analysis, it is necessary to find the links among diverse health factors and handle the heterogeneity of the patients dynamically. Third, patients admitted to hospital often have different prior risks depending on their genetic mark-up, age, prior health status. These factors may act as confounding factors when analyzing the effectiveness of treatment or therapeutics.

In summary, integrating diverse biomedical data provides considerable opportunity for developing new data mining techniques that can farther advance the state-of-the-art in data integration topic. Moreover, these techniques can also be applicable to many other data domains such as image processing, social network and climate science, which possess issues and challenges similar to biomedical integration.

Bibliography

[1] P Pavlidis, J Weston, J Cai, and WN Grundy. Gene functional classification from heterogeneous data. page 255. ACM, 2001.

[2] EE Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, 2009.

[3] J McClellan and MC King. Genetic heterogeneity in human disease. *Cell*, 141(2):210–217, 2010.

[4] Qing Zhao, Xingjie Shi, Yang Xie, Jian Huang, BenChang Shia, and Shuangge Ma. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from tcga. *Briefings in bioinformatics*, page bbu003, 2014.

[5] Lynda Chin, Jannik N Andersen, and P Andrew Futreal. Cancer genomics: from discovery science to personalized medicine. *Nature medicine*, 17(3):297–303, 2011.

[6] David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnr. Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, 8(Suppl 2):I1, 2014. Gomez-Cabrero, David Abugessaisa, Imad Maier, Dieter Teschendorff, Andrew Merkenschlager, Matthias Gisel, Andreas Ballestar, Esteban Bongcam-Rudloff, Erik Conesa, Ana Tegner, Jesper eng England 2014/07/18 06:00 BMC Syst Biol. 2014 Mar 13;8 Suppl 2:I1. doi: 10.1186/1752-0509-8-S2-I1. Epub 2014 Mar 13.

[7] S Dey, R Gupta, M Steinbach, and V Kumar. Integration of clinical and genomic data: a methodological survey. 2013.

[8] P.N. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

[9] Anneleen Daemen, Dirk Timmerman, Thierry Van den Bosch, Cecilia Bottomley, Emma Kirk, Caroline Van Holsbeke, Lil Valentin, Tom Bourne, and Bart De Moor. Improved modeling of clinical data with kernel methods. *Artificial intelligence in medicine*, 54(2):103–114, 2012.

[10] John PA Ioannidis. Microarrays and molecular research: noise discovery? *The Lancet*, 365(9458):454–455, 2005.

[11] AL Boulesteix, C Porzelius, and M Daumer. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698, 2008.

[12] C Truntzer, D Maucort-Boulch, and P Roy. Comparative optimism in models involving both classical clinical and gene expression information. *BMC bioinformatics*, 9(1):434, 2008.

[13] A Obulkasim, GA Meijer, and MA van de Wiel. Stepwise classification of cancer samples using clinical and molecular data. *BMC bioinformatics*, 12(1):422, 2011.

[14] L Ein-Dor, I Kela, G Getz, D Givol, and E Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171, 2005.

[15] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. National Acad Sciences.

[16] A Naderi, AE Teschendorff, NL Barbosa-Morais, SE Pinder, AR Green, DG Powe, JFR Robertson, S Aparicio, IO Ellis, and JD Brenton. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507–1516, 2006.

[17] HY Chuang, E Lee, YT Liu, D Lee, and T Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(1), 2007.

[18] Peter Hall, JS Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

[19] L Zhang and X Lin. Some considerations of classification for high dimension low-sample size data. *Statistical methods in medical research*, 22(5):537–550, 2013.

[20] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*, 2013.

[21] AL Boulesteix and W Sauerbrei. Added predictive value of high-throughput molecular data to clinical data, and its validation. *Breifings in bioinformatics*, 2011.

[22] H Binder and M Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC bioinformatics*, 9(1):14, 2008.

[23] Riccardo De Bin, Willi Sauerbrei, and Anne-Laure Boulesteix. Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in medicine*, 33(30):5310–5329, 2014.

[24] J Pittman, E Huang, H Dressman, CF Horng, SH Cheng, MH Tsou, CM Chen, A Bild, ES Iversen, and AT Huang. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8431, 2004.

[25] T Hastie, R Tibshirani, and JH Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Verlag, 2009.

[26] I Jolliffe. Principal component analysis. 2002.

[27] AL Boulesteix and K Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32, 2007.

[28] I Ulitsky, R Karp, and R Shamir. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. pages 347–359. Springer, 2008.

[29] G Fang, R Kuang, G Pandey, M STEINBACH, CL MYERS, and V KUMAR. Subspace differential coexpression analysis: problem definition and a general approach. *Pac Sympos Biocomput*, 15:145–56, 2010.

[30] W Zhou, G Liu, DP Miller, SW Thurston, LL Xu, JC Wain, TJ Lynch, L Su, and DC Christiani. Gene-environment interaction for the ercc2 polymorphisms and cumulative cigarette smoking exposure in lung cancer. *Cancer research*, 62(5):1377–1381, 2002.

[31] David J Hunter. Geneenvironment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287–298, 2005.

[32] Ryan Kelley and Trey Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature biotechnology*, 23(5):561–566, 2005.

[33] Casey S Greene, Nadia M Penrod, Scott M Williams, and Jason H Moore. Failure to replicate a genetic association may provide important clues about genetic architecture. *PloS one*, 4(6):e5639, 2009.

[34] Sara R Jaffee, Avshalom Caspi, Terrie E Moffitt, Kenneth A Dodge, Michael Rutter, Alan Taylor, and Lucy A Tully. Nature nurture: Genetic vulnerabilities interact with physical maltreatment to promote conduct problems. *Development and psychopathology*, 17(01):67–84, 2005.

[35] Avshalom Caspi and Terrie E Moffitt. Geneenvironment interactions in psychiatry: joining forces with neuroscience. *Nature Reviews Neuroscience*, 7(7):583–590, 2006.

[36] Anne-Kathrin Wermter, Manfred Laucht, Benno G Schimmelmann, Tobias Banaschweski, Edmund JS Sonuga-Barke, Marcella Rietschel, and Katja Becker. From nature versus nurture, via nature and nurture, to gene environment interaction in mental disorders. *European child and adolescent psychiatry*, 19(3):199–210, 2010.

[37] J Loscalzo, I Kohane, and AL Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular systems biology*, 3(1), 2007.

[38] Stephen B Manuck and Jeanne M McCaffery. Gene-environment interaction. *Annual review of psychology*, 65:41–70, 2014.

[39] Anneleen Daemen, Olivier Gevaert, and Bart De Moor. Integration of clinical and microarray data with kernel methods. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 5411–5415. IEEE, 2007.

[40] M Gnen and E Alpaydn. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[41] NM Correa, T Adali, YO Li, and VD Calhoun. Canonical correlation analysis for data fusion and group inferences. *Signal Processing Magazine, IEEE*, 27(4):39–50, 2010.

[42] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[43] K Kammers, M Lang, JG Hengstler, M Schmidt, and J Rahnenfhrer. Survival models with preclustered gene groups as covariates. *BMC bioinformatics*, 12(1):478, 2011.

[44] Sanjoy Dey, Gyorgy Simon, Bonnie Westra, Michael Steinbach, and Vipin Kumar. Mining interpretable and predictive diagnosis codes from multi-source electronic health records. 2014.

[45] M Szklo. Population-based cohort studies. *Epidemiologic reviews*, 20(1):81, 1998.

[46] MH Galea, RW Blamey, CE Elston, and IO Ellis. The nottingham prognostic index in primary breast cancer. *Breast cancer research and treatment*, 22(3):207–219, 1992.

[47] A Goldhirsch, AS Coates, RD Gelber, JH Glick, B Thrlimann, and HJ Senn. First-select the target: better choice of adjuvant treatments for breast cancer patients. *Annals of Oncology*, 17(12):1772, 2006.

[48] D Thomas. Geneenvironment-wide association studies: emerging approaches. *Nature Reviews Genetics*, 11(4):259–272, 2010.

[49] JW Blumberg. Pda applications for physicians. *ASCO News*, 16:S4–S6, 2004.

[50] C Sotiriou and MJ Piccart. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature Reviews Cancer*, 7(7):545–553, 2007.

[51] K Driouch, T Landemaine, S Sin, SX Wang, and R Lidereau. Gene arrays for diagnosis, prognosis and treatment of breast cancer metastasis. *Clinical and Experimental Metastasis*, 24(8):575–585, 2007.

[52] A Potti, S Mukherjee, R Petersen, HK Dressman, A Bild, J Koontz, R Kratzke, MA Watson, M Kelley, and GS Ginsburg. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *New England Journal of Medicine*, 355(6):570, 2006.

[53] ME Garber, OG Troyanskaya, K Schluens, S Petersen, Z Thaesler, M Pacyna-Gengelbach, M Van De Rijn, GD Rosen, CM Perou, and RI Whyte. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13784, 2001.

[54] D Anastassiou. Computational analysis of the synergy among multiple interacting genes. *Molecular systems biology*, 3(1), 2007.

[55] TR Golub, DK Slonim, P Tamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, and MA Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531, 1999.

[56] AA Alizadeh, MB Eisen, RE Davis, C Ma, IS Lossos, A Rosenwald, JC Boldrick, H Sabet, T Tran, and X Yu. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[57] A Bhattacharjee, WG Richards, J Staunton, C Li, S Monti, P Vasa, C Ladd, J Beheshti, R Bueno, and M Gillette. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13790, 2001.

[58] AJ Stephenson, A Smith, MW Kattan, J Satagopan, VE Reuter, PT Scardino, and WL Gerald. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, 104(2):290, 2005.

[59] P Edn, C Ritz, C Rose, M Fern, and C Peterson. good old clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *European journal of cancer*, 40(12):1837–1841, 2004.

[60] AE Teschendorff, A Naderi, NL Barbosa-Morais, SE Pinder, IO Ellis, S Aparicio, JD Brenton, and C Caldas. A consensus prognostic gene expression classifier for er positive breast cancer. *Genome Biol*, 7(10):R101, 2006.

[61] DE Redmond Jr. Tobacco and cancer: the first clinical report, 1761. *New England Journal of Medicine*, 282(1):18–23, 1970.

[62] RA Weinberg. *The biology of cancer*. Garland Science, 2007.

[63] M West, GS Ginsburg, AT Huang, and JR Nevins. Embracing the complexity of genomic data for personalized medicine. *Genome research*, 16(5):559, 2006.

[64] PN Tan, M Steinbach, and V Kumar. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

[65] L Li. Survival prediction of diffuse large-b-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*, 22(4):466, 2006.

[66] K Shedden, JMG Taylor, SA Enkemann, MS Tsao, TJ Yeatman, WL Gerald, S Eschrich, I Jurisica, TJ Giordano, and DE Misek. Gene expressionbased survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827, 2008.

[67] Y Sun, S Goodison, J Li, L Liu, and W Farmerie. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23(1):30, 2007.

[68] L Li, L Chen, D Goldgof, F George, Z Chen, A Rao, J Cragun, R Sutphen, and JM Lancaster. Integration of clinical information and gene expression profiles for prediction of chemo-response for ovarian cancer. 2005.

[69] J Beane, P Sebastiani, TH Whitfield, K Steiling, YM Dumas, ME Lenburg, and A Spira. A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prevention Research*, 1(1):56, 2008.

[70] JR Nevins, ES Huang, H Dressman, J Pittman, AT Huang, and M West. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human molecular genetics*, 12(Review Issue 2):R153, 2003.

[71] J Clarke and M West. Bayesian weibull tree models for survival analysis of clinico-genomic data. *Statistical methodology*, 5(3):238–262, 2008.

[72] KA Le Cao, E Meugnier, and GJ McLachlan. Integrative mixture of experts to combine clinical factors and gene markers. *Bioinformatics*, 26(9):1192–1198, 2010.

[73] R Tibshirani and B Efron. *Pre-validation and inference in microarrays*. Stanford University, Department of Biostatistics, 2002.

[74] H Hofling and R Tibshirani. A study of pre-validation. *Annals*, 2(2):643–664, 2008.

[75] CR Acharya, DS Hsu, CK Anders, A Anguiano, KH Salter, KS Walters, RC Redman, SA Tuchman, CA Moylan, and S Mukherjee. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. *Jama*, 299(13):1574, 2008.

[76] SM Wang, LLPJ Ooi, and KM Hui. Identification and validation of a novel gene signature associated with the recurrence of human hepatocellular carcinoma. *Clinical cancer research*, 13(21):6275, 2007.

[77] A Daemen, O Gevaert, and B De Moor. Integration of clinical and microarray data with kernel methods. pages 5411–5415, 2007.

[78] O Gevaert, FD Smet, D Timmerman, Y Moreau, and BD Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22(14):e184, 2006.

[79] M Campone, L Campion, H Roch, W Gouraud, C Charbonnel, F Magrangeas, S Minvielle, J Genve, AL Martin, and R Bataille. Prediction of metastatic relapse in node-positive breast cancer: establishment of a clinicogenomic model after fec100 adjuvant regimen. *Breast cancer research and treatment*, 109(3):491–501, 2008.

[80] J Silhava and P Smrz. Additional predictive value of microarray data compared to clinical variables. *4th IAPR International Conference on Pattern Recognition in Bioinformatics*, 2009.

[81] ME Futschik, M Sullivan, A Reeve, and N Kasabov. Prediction of clinical behaviour and treatment for cancers. *Applied Bioinformatics*, 2:53–58, 2003.

[82] H Bovelstad, S Nygard, and O Borgan. Survival prediction from clinico-genomic models-a comparative study. *BMC bioinformatics*, 10, 2009.

[83] S Ma and J Huang. Combining clinical and genomic covariates via cov-tgdr. *Cancer Informatics*, 3:371, 2007.

[84] AL Boulesteix and T Hothorn. Testing the additional predictive value of high-dimensional molecular data. *BMC bioinformatics*, 11(1):78, 2010.

[85] Jing Sui, Tlay Adali, Qingbao Yu, Jiayu Chen, and Vince D Calhoun. A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of neuroscience methods*, 204(1):68–81, 2012.

[86] JS Hamid, P Hu, NM Roslin, V Ling, CMT Greenwood, and J Beyene. Data integration in genetics and genomics: Methods and challenges. *Human Genomics*, 2009.

[87] Georgia Tsiliki and Sophia Kossida. Fusion methodologies for biomedical data. *Journal of proteomics*, 74(12):2774–2785, 2011.

[88] Gurkan Bebek, Mehmet Koyutrk, Nathan D Price, and Mark R Chance. Network biology methods integrating biological data for translational science. *Briefings in bioinformatics*, page bbr075, 2012.

[89] Shawn N Murphy, Michael E Mendis, David A Berkowitz, Isaac Kohane, and Henry C Chueh. Integration of clinical and genetic data in the i2b2 architecture. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1040. American Medical Informatics Association.

[90] Georgia Tsiliki, Michalis Zervakis, Marina Ioannou, Elias Sanidas, Efstathios Stathopoulos, George Potamias, Manolis Tsiknakis, and Dimitris Kafetzopoulos. Multi-platform data integration in microarray analysis. *Information Technology in Biomedicine, IEEE Transactions on*, 15(6):806–812, 2011.

[91] Stuart Watt, Wei Jiao, Andrew MK Brown, Teresa Petrocelli, Ben Tran, Tong Zhang, John D McPherson, Suzanne Kamel-Reid, Philippe L Bedard, and Nicole Onetto. Clinical genomics information management software linking cancer genome sequence and clinical decisions. *Genomics*, 102(3):140–147, 2013.

[92] Giuseppe Tradigo, Claudia Veneziano, Sergio Greco, and Pierangelo Veltri. An architecture for integrating genetic and clinical data. *Procedia Computer Science*, 29:1959–1969, 2014.

[93] Christine Desmedt, Fanny Piette, Sherene Loi, Yixin Wang, Franoise Lallemand, Benjamin Haibe-Kains, Giuseppe Viale, Mauro Delorenzi, Yi Zhang, and Mahasti Saghatchian d'Assignies. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, 13(11):3207–3214, 2007.

[94] Isaac S Kohane. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6):417–428, 2011.

[95] JP Ioannidis. Microarrays and molecular research: noise discovery? *Lancet*, 365(9458):454, 2005.

[96] L Ein-Dor, O Zuk, and E Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. National Acad Sciences, 2006.

[97] VD Calhoun, T Adali, GD Pearlson, and KA Kiehl. Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data. *Neuroimage*, 30(2):544–553, 2006.

[98] Tom Eichele, Vince D Calhoun, Matthias Moosmann, Karsten Specht, Marijtje LA Jongsma, Rodrigo Quian Quiroga, Helge Nordby, and Kenneth Hugdahl. Unmixing concurrent eeg-fmri with parallel independent component analysis. *International Journal of Psychophysiology*, 67(3):222–234, 2008.

[99] Vince D Calhoun, Jingyu Liu, and Tlay Adali. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1):S163–S172, 2009.

[100] P Buhlmann and T Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[101] NK Kasabov. On-line learning, reasoning, rule extraction and aggregation in locally optimized evolving fuzzy neural networks. *Neurocomputing*, 41(1-4):25–45, 2001.

[102] TM Cover and JA Thomas. *Elements of information theory*. wiley, 2006.

[103] Jing Gao, Wei Fan, Yizhou Sun, and Jiawei Han. Heterogeneous source consensus learning via decision propagation and negotiation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 339–348. ACM, 2009.

[104] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226–239, 1998.

[105] Mathias Fuchs, Tim Beibarth, Edgar Wingender, and Klaus Jung. Connecting high-dimensional mrna and mirna expression data for binary medical classification problems. *Computer methods and programs in biomedicine*, 111(3):592–601, 2013.

[106] Ary Noviyanto and Ito Wasito. Evaluation of data integration strategies based on kernel method of clinical and microarray data. *Bioinformation*, 8(3):147, 2012.

[107] Minta Thomas, Kris D Brabanter, Johan AK Suykens, and Bart D Moor. Predicting breast cancer using an expression values weighted clinical classifier. *BMC bioinformatics*, 15(1):6603, 2014.

[108] Ito Wasito, Aulia N Istiqlal, Mujiono Sadikin, and Indra Budi. Empirical evaluation of integration of biological data model using kernel based approach. *Journal of Convergence Information Technology*, 9(1), 2014.

[109] Shi Yu, Lon-Charles Tranchevent, Bart De Moor, and Yves Moreau. *Weighted Multiple Kernel Canonical Correlation*, pages 173–190. Springer, 2011.

[110] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48, 2003.

[111] Ferran Reverter, Esteban Vegas, and Josep M Oller. Kernel-pca data integration with enhanced interpretability. *BMC Systems Biology*, 8(Suppl 2):S6, 2014.

[112] Itziar Irigoien and Concepcin Arenas. Diagnosis using clinical/pathological and molecular information. *Statistical methods in medical research*, page 0962280214534410, 2014.

[113] Aapo Hyvrinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[114] JP Klein and ML Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Verlag, 2003.

[115] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. volume 14, pages 1137–1145. Citeseer.

[116] G Tutz and H Binder. Boosting ridge regression. *Computational Statistics and Data Analysis*, 51(12):6044–6059, 2007.

[117] J Friedman and BE Popescu. Gradient directed regularization for linear regression and classification, 2004.

[118] M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, and JT Eppig. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[119] G Michal and D Schomburg. *Biochemical pathways: an atlas of biochemistry and molecular biology*. John Wiley and Sons, 2013.

[120] A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, and ES Lander. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545, 2005.

[121] B Efron and R Tibshirani. On testing the significance of sets of genes. *The Annals of Applied Statistics*, 1(1):107–129, 2007.

[122] X Yan and F Sun. Testing gene set enrichment for subset of genes: Sub-gse. *BMC bioinformatics*, 9(1):362, 2008.

[123] MA Harris, JI Deegan, J Lomax, M Ashburner, S Tweedie, S Carbon, S Lewis, C Mungall, J Day-Richter, and K Eilbeck. The gene ontology project in 2008. *Nucleic Acids Res*, 36:D440–D444, 2008.

[124] C Fan, DS Oh, L Wessels, B Weigelt, DSA Nuyten, AB Nobel, LJ van't Veer, and CM Perou. Concordance among gene-expressionbased predictors for breast cancer. *New England Journal of Medicine*, 355(6):560–569, 2006.

[125] Y Wang, JGM Klijn, Y Zhang, AM Sieuwerts, MP Look, F Yang, D Talantov, M Timmermans, ME Meijer-van Gelder, and J Yu. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

[126] P Royston and W Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in medicine*, 23(5):723–748, 2004.

[127] KK Dobbin, DG Beer, M Meyerson, TJ Yeatman, WL Gerald, JW Jacobson, B Conley, KH Buetow, M Heiskanen, and RM Simon. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clinical cancer research*, 11(2):565, 2005.

[128] AE Teschendorff, A Miremadi, SE Pinder, IO Ellis, and C Caldas. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biol*, 8(8):R157, 2007.

[129] Paola Sebastiani, Ross Lazarus, Scott T Weiss, Louis M Kunkel, Isaac S Kohane, and Marco F Ramoni. Minimal haplotype tagging. *Proceedings of the National Academy of Sciences*, 100(17):9900–9905, 2003.

[130] OG Troyanskaya, K Dolinski, AB Owen, RB Altman, and D Botstein. A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8348, 2003.

[131] PA Konstantinopoulos, SA Cannistra, H Fountzilas, A Culhane, K Pillay, B Rueda, D Cramer, M Seiden, M Birrer, and G Coukos. Integrated analysis of multiple microarray datasets identifies a reproducible survival predictor in ovarian cancer. *PloS one*, 6(3):e18202, 2011.

[132] Dokyoon Kim, Hyunjung Shin, Young Soo Song, and Ju Han Kim. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *Journal of biomedical informatics*, 45(6):1191–1198, 2012.

[133] R Bellazzi and B Zupan. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, 77(2):81–97, 2008.

[134] R Simon, MD Radmacher, K Dobbin, and LM McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1):14, 2003.

[135] E Bair, T Hastie, D Paul, and R Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

[136] P Smialowski, D Frishman, and S Kramer. Pitfalls of supervised feature selection. *Bioinformatics*, 26(3):440, 2010.

[137] Daniela Dunkler, Stefan Michiels, and Michael Schemper. Gene expression profiling: does it add predictive accuracy to clinical characteristics in cancer prognosis? *European Journal of Cancer*, 43(4):745–751, 2007.

[138] TM Therneau and PM Grambsch. *Modeling survival data: extending the Cox model*. Springer Verlag, 2000.

[139] KC Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[140] MA Shipp, DP Harrington, JR Anderson, JO Armitage, G Bonadonna, G Brittinger, F Cabanillas, GP Canellos, B Coiffier, and JM Connors. A predictive model for aggressive non-hodgkins lymphoma the international non-hodgkins lymphoma prognostic factors project. *N Engl J Med*, 329(14):987–994, 1993.

[141] CM Bishop and SpringerLink. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006. (Online service).

[142] A Goldhirsch, WC Wood, RD Gelber, AS Coates, B Thurlimann, and HJ Senn. Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. *Journal of Clinical Oncology*, page 200304576, 2003.

[143] P Eifel, JA Axelson, J Costa, J Crowley, WJ Curran Jr, A Deshler, S Fulton, CB Hendricks, M Kemeny, and AB Kornblith. National institutes of health consensus development conference statement: adjuvant therapy for breast cancer, november 1-3, 2000. *Journal of the National Cancer Institute*, 93(13):979, 2001.

[144] RO Duda, PE Hart, and DG Stork. *Pattern classification*, volume 2. Citeseer, 2001.

[145] H Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[146] A Spira, JE Beane, V Shah, K Steiling, G Liu, F Schembri, S Gilman, YM Dumas, P Calner, and P Sebastiani. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature medicine*, 13(3):361–366, 2007.

[147] FE Harrell Jr, RM Califf, DB Pryor, KL Lee, and RA Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543, 1982.

[148] AW Partin, JL Mohler, S Piantadosi, CB Brendler, MG Sanda, PC Walsh, and JI Epstein. Selection of men at high risk for disease recurrence for experimental adjuvant therapy following radical prostatectomy*. *Urology*, 45(5):831–838, 1995.

[149] MW Kattan, TM Wheeler, and PT Scardino. Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *Journal of Clinical Oncology*, 17(5):1499, 1999.

[150] ML Blute, EJ Bergstralh, A Iocca, B Scherer, and H Zincke. Use of gleason score, prostate specific antigen, seminal vesicle and margin status to predict biochemical failure after radical prostatectomy. *The Journal of urology*, 165(1):119–125, 2001.

[151] M Graefen, PI Karakiewicz, I Cagiannos, E Klein, PA Kupelian, DI Quinn, SM Henshall, JJ Grygiel, RL Sutherland, and PD Stricker. Validation study of the accuracy of a postoperative nomogram for recurrence after radical prostatectomy for localized prostate cancer. *Journal of Clinical Oncology*, 20(4):951, 2002.

[152] VN Vapnik. *The nature of statistical learning theory*. Springer Verlag, 2000.

[153] CR Williams-DeVane, DM Reif, EC Hubal, PR Bushel, EE Hudgens, JE Gallagher, and SW Edwards. Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes. *BMC systems biology*, 7(1):119, 2013.

[154] T Hothorn, P Buhlmann, S Dudoit, A Molinaro, and MJ Van Der Laan. Survival ensembles. *Biostatistics*, 2005.

[155] J Kittler. Combining classifiers: A theoretical framework. *Pattern Analysis and Applications*, 1(1):18–27, 1998.

[156] JJ Oliver and DJ Hand. On pruning and averaging decision trees. page 430437. Citeseer, 1995.

[157] AE Raftery, D Madigan, and JA Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[158] JA Hoeting, D Madigan, AE Raftery, and CT Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999.

[159] M Calnan. Clinical uncertainty: is it a problem in the doctor-patient relationship? *Sociology of Health and Illness*, 6(1):74–85, 2008.

[160] Lei Xu, Michael I Jordan, and Geoffrey E Hinton. An alternative model for mixtures of experts. *Advances in neural information processing systems*, pages 633–640, 1995.

[161] SS Dave, G Wright, B Tan, A Rosenwald, RD Gascoyne, WC Chan, RI Fisher, RM Braziel, LM Rimsza, and TM Grogan. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *New England Journal of Medicine*, 351(21):2159, 2004.

[162] LJ Van't Veer, H Dai, MJ Van de Vijver, YD He, AAM Hart, M Mao, HL Peterse, K van der Kooy, MJ Marton, AT Witteveen, GJ Schreiber, RM Kerkhoven, and C Roberts. Gene expression profiling predicts clinical outcome of breast cancer. 2002.

[163] MH van Vliet, HM Horlings, MJ van de Vijver, MJT Reinders, and LFA Wessels. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PloS one*, 7(7), 2012.

[164] JT Chi, Z Wang, DSA Nuyten, EH Rodriguez, ME Schaner, A Salim, Y Wang, GB Kristensen, A Helland, and AL Brresen-Dale. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS medicine*, 3(3):e47, 2006.

[165] C Sotiriou, P Wirapati, S Loi, A Harris, S Fox, J Smeds, H Nordgren, P Farmer, V Praz, and B Haibe-Kains. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *JNCI Cancer Spectrum*, 98(4):262, 2006.

[166] R Liu, X Wang, GY Chen, P Dalerba, A Gurney, T Hoey, G Sherlock, J Lewicki, K Shedden, and MF Clarke. The prognostic role of a gene signature from tumorigenic breast-cancer cells. *New England Journal of Medicine*, 356(3):217–226, 2007.

[167] Riccardo De Bin, Tobias Herold, and Anne-Laure Boulesteix. Added predictive value of omics data: specific issues related to validation illustrated by two case studies. *BMC medical research methodology*, 14(1):117, 2014.

[168] H Wold. Partial least squares. 1985.

[169] L Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[170] James William Hardin, Joseph M Hilbe, and Joseph Hilbe. *Generalized linear models and extensions*. Stata Press, 2007.

[171] Margret-Ruth Oelker and Anne-Laure Boulesteix. *On the Simultaneous Analysis of Clinical and Omics Data: A Comparison of Globalboosttest and Pre-validation Techniques*, pages 259–267. Springer, 2013.

[172] WHO. Biomarkers in risk assessment: Validity and validation, 2001.

[173] Kyle Strimbu and Jorge A Tavel. What are biomarkers? *Current opinion in HIV and AIDS*, 5(6):463, 2010.

[174] Gyorgy J Simon, John Schrom, M Regina Castro, Peter W Li, and Pedro J Caraballo. Survival association rule mining towards type 2 diabetes risk assessment. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1293. American Medical Informatics Association.

[175] KI Goh, ME Cusick, D Valle, B Childs, M Vidal, and AL Barabsi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685, 2007.

[176] E Schwarz, FM Leweke, S Bahn, and P Li. Clinical bioinformatics for complex disorders: a schizophrenia case study. *BMC bioinformatics*, 10(Suppl 12):S6, 2009.

[177] S Van Dongen. A cluster algorithm for graphs. *Report-Information systems*, (10):1–40, 2000.

[178] R. Agrawal, T. Imieli ski, and A. Swami. Mining association rules between sets of items in large databases. volume 22, pages 207–216. ACM.

[179] M Berlingerio, F Bonchi, M Curcio, F Giannotti, and F Turini. Mining clinical, immunological, and genetic data of solid organ transplantation. *Biomedical Data and Applications*, pages 211–236, 2009.

[180] H Dai, L van't Veer, J Lamb, YD He, M Mao, BM Fine, R Bernards, M van de Vijver, P Deutsch, and A Sachs. A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer research*, 65(10):4059–4066, 2005.

[181] MJ Sillanp. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity*, 106(4):511–519, 2011.

[182] John R Schrom, Pedro J Caraballo, M Regina Castro, and Gyrgy J Simon. Quantifying the effect of statin use in pre-diabetic phenotypes discovered through association rule mining. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1249. American Medical Informatics Association.

[183] G Fang, G Pandey, W Wang, M Gupta, M Steinbach, and V Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 2010.

[184] Anne-Laure Boulesteix, Silke Janitza, Alexander Hapfelmeier, Kristel Van Steen, and Carolin Strobl. Letter to the editor: On the term interactionand related phrases in the literature on random forests. *Briefings in bioinformatics*, page bbu012, 2014.

[185] D Singh, PG Febbo, K Ross, DG Jackson, J Manola, C Ladd, P Tamayo, AA Renshaw, AV D'Amico, and JP Richie. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

[186] U Kumar, SI Grigorakis, HL Watt, R Sasi, L Snell, P Watson, and S Chaudhari. Somatostatin receptors in primary human breast cancer: quantitative analysis of mrna for subtypes 15 and correlation with receptor protein expression and tumor pathology. *Breast cancer research and treatment*, 92(2):175–186, 2005.

[187] A Delorme and S Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[188] L Xu, G Pearlson, and VD Calhoun. Joint source based morphometry identifies linked gray and white matter group differences. *Neuroimage*, 44(3):777–789, 2009.

[189] BB Biswal and JL Ulmer. Blind source separation of multiple signal sources of fmri data sets using independent component analysis. *Journal of computer assisted tomography*, 23(2):265, 1999.

[190] A Cichocki, R Zdunek, AH Phan, and S Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[191] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[192] Adrian R Groves, Christian F Beckmann, Steve M Smith, and Mark W Woolrich. Linked independent component analysis for multimodal data fusion. *Neuroimage*, 54(3):2198–2217, 2011.

[193] Harold Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[194] J Liu, G Pearlson, A Windemuth, G Ruano, NI Perrone, and V Calhoun. Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica. *Human brain mapping*, 30(1):241–255, 2009.

[195] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[196] Yi-Ou Li, Tlay Adali, Wei Wang, and Vince D Calhoun. Joint blind source separation by multiset canonical correlation analysis. *Signal Processing, IEEE Transactions on*, 57(10):3918–3929, 2009.

[197] Nicolle M Correa, Yi-Ou Li, Tlay Adali, and Vince D Calhoun. Fusion of fmri, smri, and eeg data using canonical correlation analysis. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 385–388. IEEE.

[198] Ignacio Gonzlez, Sbastien Djean, Pascal GP Martin, Olivier Gonalves, Philippe Besse, and Alain Baccini. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17(02):173–199, 2009.

[199] K.A. L Cao, P.G.P. Martin, C. Robert-Grani, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, 10(1):34, 2009.

[200] Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, 2011.

[201] Dongdong Lin, Hongbao Cao, Vince D Calhoun, and Yu-Ping Wang. Sparse models for correlative and integrative analysis of imaging and genetic data. *Journal of neuroscience methods*, 237:69–78, 2014.

[202] Tingkai Sun, Songcan Chen, Jingyu Yang, and Pengfei Shi. A novel method of combined feature extraction for recognition. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 1043–1048. IEEE.

[203] J. Sui, T. Adali, G. Pearlson, H. Yang, S.R. Sponheim, T. White, and V.D. Calhoun. A cca+ ica based model for multi-task brain imaging data fusion and its application to schizophrenia. *Neuroimage*, 51(1):123–134, 2010.

[204] Jing Sui, Godfrey Pearlson, Arvind Caprihan, Tlay Adali, Kent A Kiehl, Jingyu Liu, Jeremy Yamamoto, and Vince D Calhoun. Discriminating schizophrenia and bipolar disorder by fusing fmri and dti in a multimodal cca+ joint ica model. *Neuroimage*, 57(3):839–855, 2011.

[205] Brett A McKinney, David M Reif, Marylyn D Ritchie, and Jason H Moore. Machine learning for detecting gene-gene interactions. *Applied bioinformatics*, 5(2):77–88, 2006.

[206] Heather J Cordell. Detecting genegene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

[207] K Van Steen. Travelling the world of genegene interactions. *Briefings in bioinformatics*, 13(1):1–19, 2012.

[208] Xuefeng Wang, Robert C Elston, and Xiaofeng Zhu. The meaning of interaction. *Human heredity*, 70(4):269, 2011.

[209] Sophia J Docherty, Yulia Kovas, and Robert Plomin. Gene-environment interaction in the etiology of mathematical ability using snp sets. *Behavior genetics*, 41(1):141–154, 2011.

[210] Jinying Zhao, Li Jin, and Momiao Xiong. Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845, 2006.

[211] Matthew B Lanktree and Robert A Hegele. Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease. *Genome Med*, 1(2):28, 2009.

[212] Gang Fang, Majda Haznadar, Wen Wang, Haoyu Yu, Michael Steinbach, Timothy R Church, William S Oetting, Brian Van Ness, and Vipin Kumar. High-order snp combinations associated with complex diseases: efficient discovery, statistical power and functional interactions. *PloS one*, 7(4):e33531, 2012.

[213] Sanjay Purushotham, Martin Renqiang Min, C-C Jay Kuo, and Rachel Ostroff. Factorized sparse learning models with interpretable high order feature interactions. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 552–561. ACM.

[214] S Dey, K Lim, G Atluri, A MacDonald III, M Steinbach, and V Kumar. A pattern mining based integrative framework for biomarker discovery. In *Proceedings of the*

ACM Conference on Bioinformatics, Computational Biology and Biomedicine, pages 498–505. ACM, 2012.

[215] K.A. L Cao, E. Meugnier, and G.J. McLachlan. Integrative mixture of experts to combine clinical factors and gene markers. *Bioinformatics*, 26(9):1192–1198, 2010.

[216] Benjamin R Jefferys, Iheanyi Nwankwo, Elias Neri, David CW Chang, Lev Shamardin, Stefanie Hnold, Norbert Graf, Nikolaus Forg, and Peter Coveney. Navigating legal constraints in clinical data warehousing: a case study in personalized medicine. *Interface focus*, 3(2):20120088, 2013.

[217] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, and Ken Aldape. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[218] Hal Caswell. Prospective and retrospective perturbation analyses: their roles in conservation biology. *Ecology*, 81(3):619–627, 2000.

[219] David L Sackett. *Evidence-based medicine*. Wiley Online Library, 2000.

[220] Qing Zhao, Xingjie Shi, Yang Xie, Jian Huang, BenChang Shia, and Shuangge Ma. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from tcga. *Briefings in bioinformatics*, page bbu003, 2014.

[221] J. Loscalzo, I. Kohane, and A.L. Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular systems biology*, 3(1), 2007.

[222] W. Zhou, G. Liu, D.P. Miller, S.W. Thurston, L.L. Xu, J.C. Wain, T.J. Lynch, L. Su, and D.C. Christiani. Gene-environment interaction for the ercc2 polymorphisms and cumulative cigarette smoking exposure in lung cancer. *Cancer research*, 62(5):1377–1381, 2002.

[223] Stephen B Manuck and Jeanne M McCaffery. Gene-environment interaction. *Annual review of psychology*, 65:41–70, 2014.

[224] Anne-Laure Boulesteix, Silke Janitza, Alexander Hapfelmeier, Kristel Van Steen, and Carolin Strobl. Letter to the editor: On the term interactionand related phrases in the literature on random forests. *Briefings in bioinformatics*, page bbu012, 2014.

[225] D. Singh et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

[226] C. Sotiriou and M.J. Piccart. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nature Reviews Cancer*, 7(7):545–553, 2007.

[227] U Kumar, SI Grigorakis, HL Watt, R Sasi, L Snell, P Watson, and S Chaudhari. Somatostatin receptors in primary human breast cancer: quantitative analysis of mrna for subtypes 1–5 and correlation with receptor protein expression and tumor pathology. *Breast cancer research and treatment*, 92(2):175–186, 2005.

[228] N.M. Correa, T. Adali, Y.O. Li, and V.D. Calhoun. Canonical correlation analysis for data fusion and group inferences. *Signal Processing Magazine, IEEE*, 27(4):39–50, 2010.

[229] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[230] Bharat B Biswal and John L Ulmer. Blind source separation of multiple signal sources of fmri data sets using independent component analysis. *Journal of computer assisted tomography*, 23(2):265–271, 1999.

[231] Lai Xu, Godfrey Pearlson, and Vince D Calhoun. Joint source based morphometry identifies linked gray and white matter group differences. *Neuroimage*, 44(3):777–789, 2009.

[232] David J Hunter. Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287–298, 2005.

[233] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N.I. Perrone-Bizzozero, and V. Calhoun. Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica. *Human brain mapping*, 30(1):241–255, 2009.

[234] VD Calhoun, T. Adali, GD Pearlson, and KA Kiehl. Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data. *Neuroimage*, 30(2):544–553, 2006.

[235] V.D. Calhoun, J. Liu, and T. AdalI. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1):S163–S172, 2009.

[236] T. Sun et al. A novel method of combined feature extraction for recognition. In *ICDM*, pages 1043–1048. IEEE, 2008.

[237] K.A. Lê Cao, I. González, and S. Déjean. integromics: an r package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855, 2009.

[238] N.M. Correa, Y.O. Li, T. Adali, and V.D. Calhoun. Fusion of fmri, smri, and eeg data using canonical correlation analysis. In *ICASSP*, pages 385–388. IEEE, 2009.

[239] H.J. Cordell. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 2009.

[240] D. Thomas. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, 11(4):259–272, 2010.

[241] L.W. Hahn, M.D. Ritchie, and J.H. Moore. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, 2003.

[242] Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *The Journal of Machine Learning Research*, 10:377–403, 2009.

[243] S.D. Bay and M.J. Pazzani. Detecting group differences: Mining contrast sets. *DMKD*, 2001.

[244] Gang Fang, Gaurav Pandey, Wen Wang, Manish Gupta, Michael Steinbach, and Vipin Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. *IEEE TKDE.*, 2012.

[245] E.E. Schadt et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 2005.

[246] P.N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*. ACM, 2002.

[247] G. Fang, W. Wang, B. Oatley, B. Van Ness, M. Steinbach, and V. Kumar. Characterizing discriminative patterns. *Arxiv preprint arXiv:1102.4104*, 2011.

[248] Steinbach M. Fang, G. and, V. Kumar, , et al. High-order snp combinations associated with complex diseases: Efficient discovery, statistical power and functional interactions. *PLoS ONE*, To appear, 2012.

[249] D.S. Bassett, K.O. Lim, et al. Altered resting state complexity in schizophrenia. *NeuroImage*, 2011.

[250] M.E. Lynall et al. Functional connectivity and brain networks in schizophrenia. *The Journal of Neuroscience*, 2010.

[251] S. Hanhijärvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila. Tell me something i don't know: Randomization strategies for iterative data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–388. ACM, 2009.

[252] Y. Zhou, N. Shu, Y. Liu, M. Song, Y. Hao, H. Liu, C. Yu, Z. Liu, and T. Jiang. Altered resting-state functional connectivity and anatomical connectivity of hippocampus in schizophrenia. *Schizophrenia research*, 100(1-3):120–132, 2008.

[253] O.R. Phillips et al. Fiber tractography reveals disruption of temporal lobe white matter tracts in schizophrenia. *Schizophrenia research*, 2009.

[254] G.R. Kuperberg et al. Regionally localized thinning of the cerebral cortex in schizophrenia. *Archives of general psychiatry*, 60(9):878, 2003.

[255] M. Fatjó-Vilas et al. Dysbindin-1 gene contributes differentially to early-and adult-onset forms of functional psychosis. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 2011.

[256] L. Zuo et al. Association study of dtnbp1 with schizophrenia in a us sample. *Psychiatric genetics*, 2009.

[257] N.E.M. van Haren, H.G. Schnack, W. Cahn, M.P. van den Heuvel, C. Lepage, L. Collins, A.C. Evans, H.E.H. Pol, and R.S. Kahn. Changes in cortical thickness during the course of illness in schizophrenia. *Archives of general psychiatry*, 68(9):871, 2011.

[258] H. Gurling et al. Genetic association and brain morphology studies and the chromosome 8p22 pericentriolar material 1 (pcm1) gene in susceptibility to schizophrenia. *Archives of general psychiatry*, 63(8):844, 2006.

[259] K. Kasai et al. Differences and similarities in insular and temporal pole mri gray matter volume abnormalities in first-episode schizophrenia and affective psychosis. *Archives of general psychiatry*, 2003.

[260] M. Gratacòs et al. Brain-derived neurotrophic factor val66met and psychiatric disorders: meta-analysis of case-control studies confirm association to substance-related disorders, eating disorders, and schizophrenia. *Biological psychiatry*, 61(7):911–922, 2007.

[261] M. Dunoyer. Accelerating access to treatments for rare diseases. *Nature Reviews Drug Discovery*, 10(7):475–476, 2011.

[262] Daniela M Witten and Robert J Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

[263] G. Fang et al. Mining low-support discriminative patterns from dense and high-dimensional data. *IEEE TKDE*, 2010.

[264] Sean P Keehan, Andrea M Sisko, Christopher J Truffer, John A Poisal, Gigi A Cuckler, Andrew J Madison, Joseph M Lizonitz, and Sheila D Smith. National health spending projections through 2020: economic recovery and reform drive faster spending growth. *Health Affairs*, 30(8):1594–1605, 2011.

[265] Donald M Berwick and Andrew D Hackbarth. Eliminating waste in us health care. *JAMA: the journal of the American Medical Association*, 307(14):1513–1516, 2012.

[266] Robert Steinbrook. Health care and the american recovery and reinvestment act. *New England Journal of Medicine*, 360(11):1057–1060, 2009.

[267] T. Diethe, D. Hardoon, and J. Shawe-Taylor. Constructing nonlinear discriminants from multiple data views. *Machine Learning and Knowledge Discovery in Databases*, pages 328–343, 2010.

[268] M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[269] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[270] J.R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[271] Kantilal Varichand Mardia, John T Kent, and John M Bibby. Multivariate analysis. 1980.

[272] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[273] Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443, 2009.

[274] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

[275] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ccs/.

[276] Pang-Ning Tan et al. *Introduction to data mining*. Pearson Education India, 2007.

[277] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[278] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[279] http://www.ncbi.nlm.nih.gov/pubmed/advanced.

[280] J. McClellan and M.C. King. Genetic heterogeneity in human disease. *Cell*, 141(2):210–217, 2010.

[281] Gyorgy J Simon, John Schrom, M Regina Castro, Peter W Li, and Pedro J Caraballo. Survival association rule mining towards type 2 diabetes risk assessment. In *AMIA Annual Symposium Proceedings*, volume 2013, page 1293. American Medical Informatics Association, 2013.

[282] B. Di Camillo, T. Sanavia, M. Martini, G. Jurman, F. Sambo, A. Barla, M. Squillario, C. Furlanello, G. Toffolo, and C. Cobelli. Effect of size and heterogeneity of samples on biomarker discovery: Synthetic and real data assessment. *PloS one*, 7(3):e32200, 2012.

[283] D. Repsilber, S. Kern, A. Telaar, G. Walzl, G.F. Black, J. Selbig, S.K. Parida, S.H.E. Kaufmann, and M. Jacobsen. Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC bioinformatics*, 11(1):27, 2010.

[284] E. Schwarz, F.M. Leweke, S. Bahn, and P. Liò. Clinical bioinformatics for complex disorders: a schizophrenia case study. *BMC bioinformatics*, 10(Suppl 12):S6, 2009.

[285] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, and A.L. Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[286] A. Obulkasim, G.A. Meijer, and M.A. van de Wiel. Stepwise classification of cancer samples using clinical and molecular data. *BMC bioinformatics*, 12(1):422, 2011.

[287] Anuj Karpatne, Ankush Khandelwal, Shyam Boriah, and Vipin Kumar. Predictive learning in the presence of heterogeneity and limited training data.

[288] Judith Maria Mathea Meijers, RJG Halfens, Jacques CL Neyens, YC Luiking, G Verlaan, and JMGA Schols. Predicting falls in elderly receiving home care: the role of malnutrition and impaired mobility. *The journal of nutrition, health & aging*, 16(7):654–658, 2012.

[289] Outcome and assessment information set implementation manual: Implementing oasis at a home health agency to improve patient outcomes, 2008. Retrieved from http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/Downloads/HHQIArchivedOASISInformation.zip.

[290] Robert E Schlenker, Martha C Powell, and Glenn K Goodrich. Initial home health outcomes under prospective payment. *Health Services Research*, 40(1):177–193, 2005.

[291] PE Adams, ME Martinez, JL Vickerie, and WK Kirzinger. Summary health statistics for the u.s. population: National health interview survey, 2010. *Vital and Health Statistics, Series 10, Data From the National Health Survey*, 251:1–117, 2011.

[292] Lindy Clemson, Lynette Mackenzie, Claire Ballinger, Jacqueline CT Close, and Robert G Cumming. Environmental interventions to prevent falls in community-dwelling older people a meta-analysis of randomized trials. *Journal of Aging and Health*, 20(8):954–971, 2008.

[293] Cheryl Chia-Hui Chen, Charlotte Wang, and Guan-Hua Huang. Functional trajectory 6 months posthospitalization: a cohort study of older hospitalized patients in taiwan. *Nursing research*, 57(2):93–100, 2008.

[294] Thomas M Gill, Heather G Allore, Evelyne A Gahbauer, and Terrence E Murphy. Change in disability after hospitalization or restricted activity in older persons. *JAMA*, 304(17):1919–1928, 2010.

[295] Susan E Hardy and Thomas M Gill. Recovery from disability among community-dwelling older persons. *Jama*, 291(13):1596–1602, 2004.

[296] Medicare: Home health compare. `https://www.medicare.gov/homehealthcompare/`, 2014.

[297] Documentation of prediction models used for risk adjustment of home health agency outcomes reported on the cms home health compare web site, 2005. Retrieved from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HomeHealthQualityInits/downloads/HHQIOASISOBQIOverviewRADocumentation.p

[298] William A Satariano, Jack M Guralnik, Richard J Jackson, Richard A Marottoli, Elizabeth A Phelan, and Thomas R Prohaska. Mobility and aging: new directions for public health action. *American Journal of Public Health*, 102(8):1508–1515, 2012.

[299] Karen A Monsen, Bonnie L Westra, S Cristina Oancea, Fang Yu, and Madeleine J Kerr. Linking home care interventions and hospitalization outcomes for frail and non-frail elderly patients. *Research in nursing & health*, 34(2):160–168, 2011.

[300] Bonnie L Westra, Kay Savik, Cristina Oancea, Lynn Choromanski, John H Holmes, and Donna Bliss. Predicting improvement in urinary and bowel incontinence for home health patients using electronic health record data. *Journal of wound, ostomy, and continence nursing: official publication of The Wound, Ostomy and Continence Nurses Society/WOCN*, 38(1):77, 2011.

[301] Thomas M Gill, Dorothy I Baker, Margaret Gottschalk, Peter N Peduzzi, Heather Allore, and Amy Byers. A program to prevent functional decline in physically

frail, elderly persons who live at home. *New England Journal of Medicine*, 347(14):1068–1074, 2002.

[302] Robert J Rosati. The history of quality measurement in home health care. *Clinics in geriatric medicine*, 25(1):121–134, 2009.

[303] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[304] Steinbach-M. Kumar V. Tan, P.N. *Introduction to data mining*. Pearson Addison-Wesley, Boston, MA, 2006.

[305] Gang Fang, Gaurav Pandey, Wen Wang, Manish Gupta, Michael Steinbach, and Vipin Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):279–294, 2012.

[306] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[307] Agency for Healthcare Research and Quality. *National Healthcare Report, 2011*. U.S. Department of Health and Human Services, 540 Gaither RoadRockville, MD 20850, 3 2012. AHRQ Publication No. 12-0005.

[308] Robert A Fieo, Elizabeth J Austin, John M Starr, and Ian J Deary. Calibrating adl-iadl scales to improve measurement accuracy and to extend the disability construct into the preclinical range: a systematic review. *BMC geriatrics*, 11(1):42, 2011.

[309] Tanya P Scharpf, Natalie Colabianchi, Elizabeth A Madigan, Duncan Neuhauser, Timothy Peng, Penny H Feldman, and John FP Bridges. Functional status decline as a measure of adverse events in home health care: an observational study. *BMC health services research*, 6(1):162, 2006.

[310] Marjan J Faber, Ruud J Bosscher, Marijke J Chin A Paw, and Piet C van Wieringen. Effects of exercise programs on falls and mobility in frail and pre-frail older adults: a multicenter randomized controlled trial. *Archives of physical medicine and rehabilitation*, 87(7):885–896, 2006.

[311] Richard H Fortinsky, Ramon I Garcia, T Joseph Sheehan, Elizabeth A Madigan, and Susan Tullai-McGuinness. Measuring disability in medicare home care patients: application of rasch modeling to the outcome and assessment information set. *Medical care*, pages 601–615, 2003.

[312] Kenneth E Covinsky, Catherine Eng, Li-Yung Lui, Laura P Sands, and Kristine Yaffe. The last 2 years of life: functional trajectories of frail older people. *Journal of the American Geriatrics Society*, 51(4):492–498, 2003.

[313] Sherry A Greenberg. Analysis of measurement tools of fear of falling for high-risk, community-dwelling older adults. *Clinical Nursing Research*, 21(1):113–130, 2012.

[314] Stanley Colcombe and Arthur F Kramer. Fitness effects on the cognitive function of older adults a meta-analytic study. *Psychological science*, 14(2):125–130, 2003.

Appendix A

Glossary and Acronyms

A.1 Supplementary Figures and Tables

Variables	All	1	2	3	4	All	1	2	3	4
	Improved					No Improved				
Age										
Age 19 – 39	1.37	3.10	1.11						3.29	1.63
Age 40 – 64	1.18	1.65	1.54						1.74	
Age 65 – 74	1.26	1.50	1.58	1.23						
Age 75 – 84			1.08	1.41			1.11			
Age 85 or older				1.37		1.42	2.59	1.76		
Gender										
Male	1.10	1.38							1.50	
Female				1.50		1.10	1.38			
Ethnic										
White (vs non-White)	1.04			1.32			1.06			
Non-White		1.06				1.04			1.32	
Payor for homecare										
Medicare				1.40		1.13	1.56			
Medicaid		1.35						1.09	1.88	
Other payor	1.14	1.32	1.21							
Admitted from										
Hospital	1.45	1.79	1.88	1.31	1.43					
Community						1.40	1.46	1.82	1.59	1.74
Other facility	1.13			2.32			1.30	1.24		
Medical and treatment regimen										
Recent Treatment Change	1.31	1.40	1.49							

Figure A.1: Odds Ratios for Single Variables Significantly Associated with All Patients and by Admission Score for Mobility (Continued)

Prognosis scale										
Good prognosis	1.30	1.61	2.26	1.17	1.85					
Moderate to poor prognosis						1.28	1.64	2.23		1.88
Current residence										
Assisted living						1.51	3.07	2.30		
Lives with family							1.16	1.29		
Lives in own home	1.24	1.74	2.00		1.31					
Lives alone			2.00		1.70	1.19	1.07			
Assisting person										
Person residing in the home	1.29	1.25	1.12							
Paid help						1.47	2.73	2.57	1.45	1.54
Relative/ friend outside the home			1.31							
Primary caregiver										
Child			1.10	1.52		1.07	1.27			
Other family	1.05	1.23	1.28						1.39	1.28
Paid help						1.47	2.78	2.49	1.36	1.33
Spouse	1.31	1.45	1.32							
Frequency of primary assistance										
1-2 times per week		1.03	2.26		1.64	1.30				
3-6 times per week	1.08						1.12	1.48		

Figure A.2: Odds Ratios for Single Variables Significantly Associated with All Patients and by Admission Score for Mobility(Continued)

Type of primary caregiver assistance										
Activities of daily living	1.14						1.47	2.04		1.35
Environmental support	1.05						1.17	1.35		
Financial/ legal, power of attorney				1.17		1.07	1.18	1.38		1.35
Health care agent, medical power of attorney				1.17		1.06	1.16	1.36		1.33
Instrumental activities of daily living	1.03						1.25	1.42		
Medical care support							1.18	1.39		1.33
Psychosocial support	1.03						1.14	1.35		
Vision										
Little or no problem				1.22		1.03	1.47	1.47		
Moderate to severe						1.31	2.75	2.16		1.84
Hearing										
Little or no problem	1.15	1.82	2.08		1.49				1.33	
Moderate to severe				1.33		1.15	1.82	2.08		1.49
Speech										
Little or no problem	1.33	1.72	3.87		2.51					
Moderate to severe						1.31	1.79	3.83		2.11
Pain frequency										
Little or no problem		1.10				1.32		1.61	1.16	2.17
Moderate to severe	1.32		1.61	1.16	2.17		1.10			
Intractable pain	1.23		1.39		1.84		1.12			
Prior Intractable pain	1.27		1.71		1.74		1.07			
Urinary incontinence frequency										
Prior urinary incontinence						1.11	1.84	1.79		1.32
Little or no problem					1.42	1.18	1.47	1.22		
Constantly	1.04			1.55	2.06		1.49	1.11		

Figure A.3: Odds Ratios for Single Variables Significantly Associated with All Patients and by Admission Score for Mobility(Continued)

Bowel incontinence										
None or rarely	1.25	1.22	3.01	2.18	3.05					
Frequently						1.34	2.08	3.32	1.96	2.98
Respiratory Status										
Respiratory status - moderate to severe				1.3			1.45	1.19		
Wounds										
Has pressure ulcer						1.49	2.06	2.3	2.77	2.06
Has stasis ulcer						1.64	1.84	1.39	1.57	
Has surgical wound	1.61	1.83	3.70	1.58	2.31					
Status of surgical wound										
Healing						1.60	2.19	2.31		1.78
Not healing	1.61	2.19	2.32		1.83					
Prior risk factors										
Obesity			1.31		1.36		1.16			
Smoking	1.07	1.31	1.29		1.91				1.32	
Anxiety frequency										
No anxiety or only in new situations		1.07	1.43						1.19	
Anxiety daily or all the time	1.03			1.19			1.07	1.42		
Depressed										
Depressed mood							1.1	1.22		
Behavior problems										
No or infrequent behavior problems		1.39	1.91		1.36	1.03			1.24	
Frequent behavior problems	1.03			1.24			1.39	1.91		1.36
Behaviors										
Prior impaired decision making				1.19		1.06	1.51	2.18		1.34
Current impaired decision making				1.25			1.52	2.3		1.44
Prior memory loss				1.24		1.12	2.01	2.53		1.45
Current memory deficit				1.37		1.11	1.83	2.53		1.68

Figure A.4: Odds Ratios for Single Variables Significantly Associated with All Patients and by Admission Score for Mobility(Continued)

Cognitive function										
Little or no problem	1.69	2.22	5.94		2.81					
Moderate to severe						1.69	2.22	5.94		2.81
When confused										
Never or new situations only	1.32	1.99	3.62		2.21				1.16	
Upon waking to constantly				1.16		1.31	2.00	3.61		1.98
Current grooming										
Little or no problem		1.45	2.84		2.98	1.31			1.27	
Moderate to severe	1.31			1.27			1.45	2.84		2.98
Ability to dress upper body										
Little or no problem		1.38	3.17		4.23	1.42			1.21	
Moderate to severe	1.42			1.21			1.38	3.17		4.23
Ability to dress lower body										
Little or no problem		1.53	2.76		5.30	1.51			1.26	
Moderate to severe	1.51			1.26			1.53	2.76		5.30
Ability to bath										
Little or no problem		1.68	2.55		4.19	1.36			1.58	
Moderate to severe	1.36			1.58			1.68	2.55		4.19
Ability to toilet										
Little or no problem		1.61	1.69		2.65	1.62				
Moderate to severe	1.62						1.61	1.69		2.65
Ability to transfer										
Little or no problem		1.50	1.65		4.58	2.02			1.22	
Moderate to severe	2.02			1.22			1.50	1.65		4.58
Mobility										
1 - Uses device/ supervision for steps or uneven surfaces						6.33				

Figure A.5: Odds Ratios for Single Variables Significantly Associated with All Patients and by Admission Score for Mobility(Continued)

2 - Walks only with device or supervision at all times	7.26									
3-5 - Chairfast or bed fast	1.28									
Ability to eat										
Little or no problem	1.15		3.51		3.06		1.24			
Moderate to severe		1.24				1.15		3.51		3.06
Ability to do laundry										
Little or no problem		1.68	3.09		4.24	1.26			1.53	
Moderate to severe	1.26			1.53			1.68	3.09		4.24
Ability to do housekeeping										
Little or no problem		1.62	2.99		3.81	1.26			1.64	
Moderate to severe	1.26			1.65			1.62	2.99		3.81
Ability to shop										
Little or no problem		1.53	1.99		2.27	1.25			2.00	
Moderate to severe	1.25			2.00			1.53	1.99		2.29
Ability to use phone										
Little or no problem	1.41	2.32	4.71		3.20					
Moderate to severe						1.39	2.39	4.79		2.83
Ability to manage medications										
Little or no problem		1.66	3.29		2.26				1.18	
Moderate to severe				1.18			1.66	3.29		2.25
High Therapy Need										
Yes				1.52		1.09	1.43			
Total Number Variables	46	38	42	36	32	44	60	51	27	37

Figure A.6: Odds Ratios for Single Variables Significantly Associated with All Patients and by Admission Score for Mobility(End)

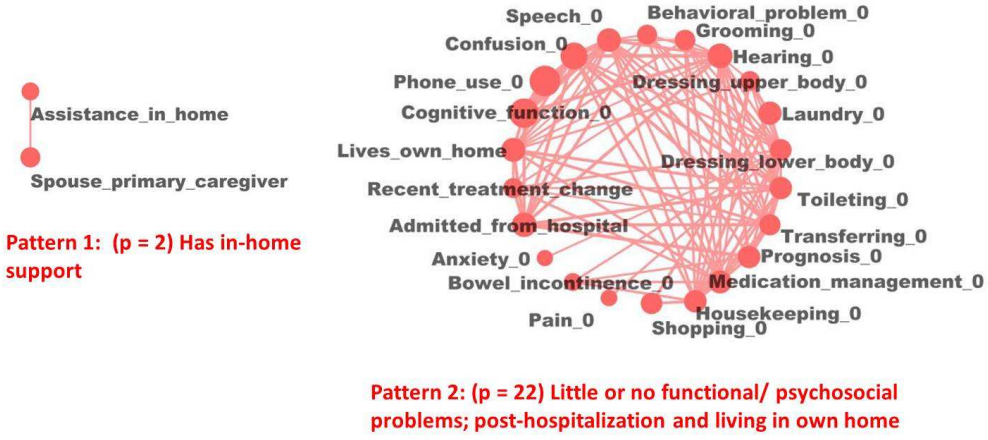


Figure A.7: Patterns associated with IMPROVEMENT in the DEVICE group.

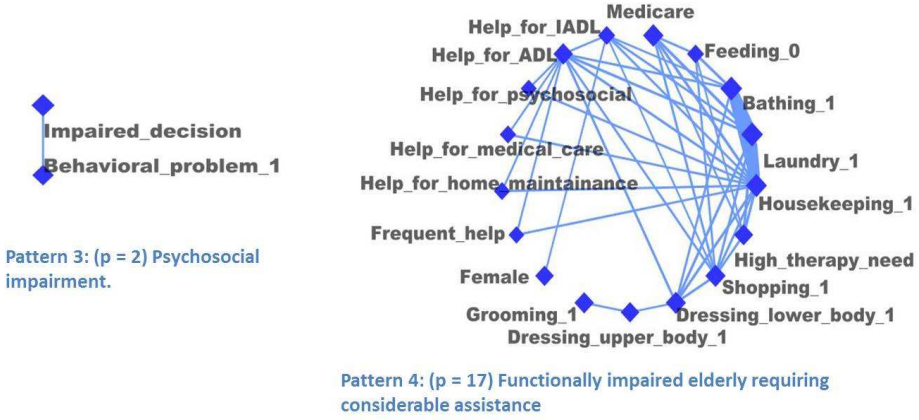


Figure A.8: Patterns associated with NO IMPROVEMENT in the DEVICE group.

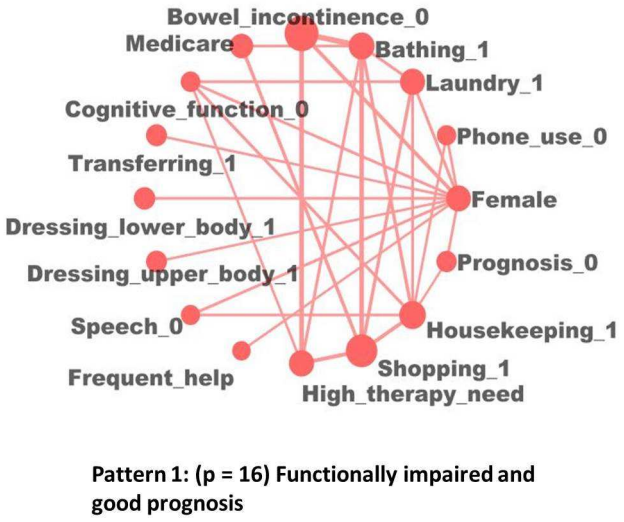


Figure A.9: Patterns associated with IMPROVEMENT in the CHAIR_I group.

Figure A.10: Patterns associated with NO IMPROVEMENT in the CHAIR_I group.

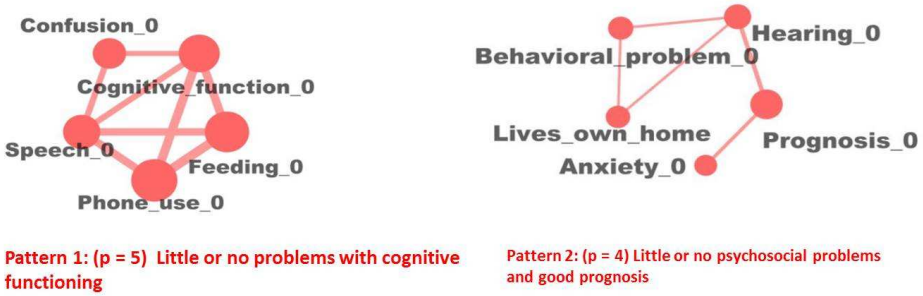


Figure A.11: Patterns associated with IMPROVEMENT in the CHAIR_NI group.

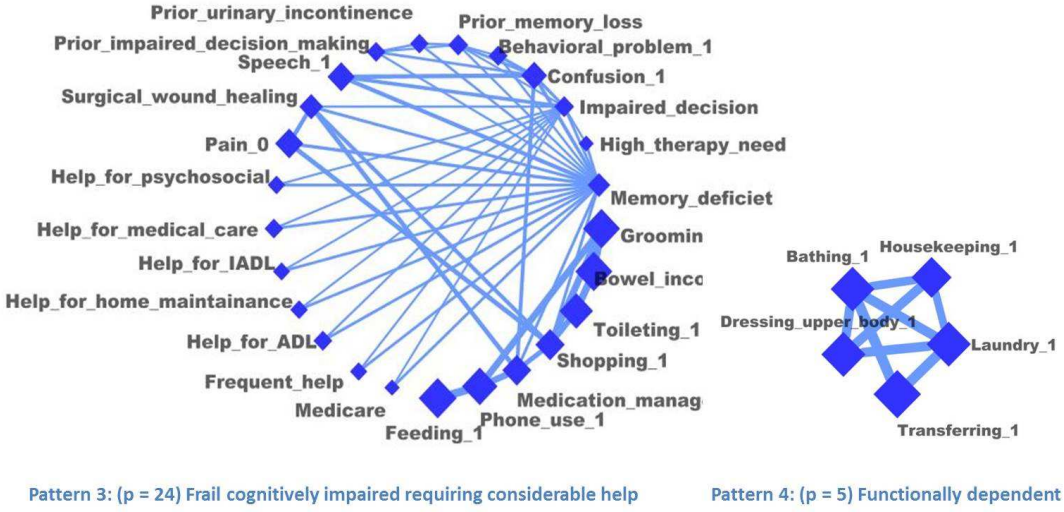


Figure A.12: Patterns associated with NO IMPROVEMENT in the CHAIR_NI group.

Table A.1: A summary of predictive models

Clinico-genomic Study	Stage of integration	Stage of dimensionality reduction	Full-space/ Sub model	Dimensionality reduction technique	Methods	Clinical Endpoint	Disease Phenotype	Goal of the study
Li 2006 [65]	Early	2-step	Full	Two-step feature extraction: PCA, SIR	Cox hazard model	Survival time after chemotherapy	large-B-Cell lymphoma (DLBCL)	Generic Predictive Modeling
Stephenson et al. 2005 [58]	Early	2-step	Full	Ranking using statistical test	Logistic regression	Recurrance after Radical Prostatectomy	Prostate cancer	Generic Predictive Modeling
Sun et al., 2007 [67]	Early	2-step	Full	Wrapper Model	Linear Discriminant analysis	Survival prediction (Two class)	Breast cancer	Generic Predictive Modeling
Li et al, 2005 [68]	Early	2-step	Full	Gene selection based on t-test	SVM	Response to platinum-based based Chemotherapy (Survival)	Ovarian cancer	Generic Predictive Modeling
Beane et al. 2008 [69]	Early	2-step	Full	Ranking using Statistical test	Logistic regression	Development of metastasis after pathology	Lung cancer	Generic Predictive Modeling

Table A.2: A summary of predictive models(cont.)

Pittman et al. 2004 [24]	Early	2-step	Full	Feature creation by Cluster and PCA	Tree	Survival prediction(Two class metastasis development)	Breast Cancer	Generic Predictive Modeling
Clarke et al. 2008 [71]	Early	2-step	Full	Clustered based metagene	Tree	Survival time after primary chemotherapy/ disease relapse	Ovarian cancer	Generic Predictive Modeling
Cao et al. 2010 [72]	Early	2-step	Full	Three dimensionality reduction methods	Mixture of Experts	Binary outcome	Breast cancer, Prostate cancer, Medulloblastomas	Generic Predictive Modeling
Binder et al. 2008 [22]	Early	1-step	Full	Regularization Techniques	Cox based prior model	Survival data	DLBCL	Generic Predictive Modeling
Ma et al. 2007 [83]	Early	1-step	Full	Statistic test and regularization	Penalized logistic and Cox regression	Binary class (metastasis), Survival analysis	Breast cancer, Follicular lymphoma	Generic Predictive Modeling
Bovelstad et al. 2009 [82]	Early	1-step	Full	Regularization method	Cox regression	Survival prediction	Breast cancer, DLBCL, Neuroblastoma	Generic Predictive Modeling

Table A.3: A summary of predictive models(cont.)

Berlingerio et al. 2009 [179]	Early	2-step	Sub	Only HLA alleles corresponding to six loci are considered	Frequent Pattern Mining	Liver transplant VS. normal	Liver diseases leading to liver transplantation	Generic Predictive Modeling
Schwarz et al. 2009 [176]	Early	2-step	Sub	Domain guided	Network based framework	Case-control	Schizophrenia	Generic Predictive Modeling
Jana Silhava et al. 2009[80]	Late	2-step	Full	Filtering	Logit, Bionomial boosting	Recurrance vs. Not recurrence	Breast Cancer	Generic Predictive Modeling
Campone et al. 2008 [79]	Late	2-step	Full	Filtering by univariant cox regression+PCA	Multivariate Cox regression Aanalysis	Metastasis free survival	Breast cancer	Generic Predictive Modeling
Futschik et al. 2003 [81]	Late	2-step	Full	Filtering using statistical test	Bayesian network & ANN	Two class Survival after 5-yrs.	DLBCL	Generic Predictive Modeling
Daemen et al. 2007 [9]	Intermediate	2-step	Full	Ranking using statistical test	SVM	Metastasis	Breast cancer	Generic Predictive Modeling
Gevaert et al. 2006 [78]	Intermediate	2-step	Full	Gene Filtering	Bayesian network	Metastasis	Breast cancer	Generic Predictive Modeling

Table A.4: A summary of predictive models(cont.)

Noviyanto et al. 2012 [106]	Inte- rme- di- ate	2- step	Full	Gene Filter- ing	LS- SVM	Two- class	Breast cancer	Generic Pre- dictive Model- ing
Wasito et al. 2014 [108]	Inte- rme- di- ate	2- step	Full	Kernel Dimen- sion- ality Reduc- tion	SVM	Two- class	Lymp- homa cancer	Generic Pre- dictive Model- ing
Irigoien et al. [112]	Early, inte- rme- di- ate	2- step	Full- space	Distance metrics	Clusters, dis- crim- inant analy- sis	Two- class	Synthetic data	Generic Pre- dictive Model- ing
Kammers et al.[43]	Early	2- step	Full- space	Prior knowl- edge (Path- way based)	Cox Model	Survival	Breast Can- cer	Generic Pre- dictive Model- ing
van Vilet et al.[163]	Early, Inte- rme- di- ate, Late	Both of them	Full- space	T-test or Chi- squared test	Many	Binary class	Breast Can- cer	Generic Pre- dictive Model- ing
Zhao et al. 2015 [210]	Early	Both	Full	PCA, PLS, LASSO	Cox Regres- sion	Survival analysis	Several Can- cer from TCGA	Generic Pre- dictive Model
Nevins et al. 2003 [70]	Early	2- step	Full	Feature cre- ation by Clus- tering and PCA	Tree	Survival pre- diction (Two class metas- tasis develop- ment)	Breast Can- cer	Generic Pre- dictive Model- ing

Table A.5: A summary of predictive models(cont.)

Wang et al. 2007 [76]	Early	2-step	Full	Stepwise logistic regression for gene selection	SVM, SLD, KNN	Recurrent vs. Non-recurrent	Human Hepatocellular Carcinoma	Additional power of genes
Tibshirani et al. 2002 [73]	Early	2-step	Full	Filtering approach based on p-value of fold change	Logistic regression	Binary class (metastasis vs. normal)	Breast cancer	Additional power of genes
Hofling et al. 2008 [74]	Early	2-step	Full	Filtering approach based on p-value of fold change	Logistic regression	Binary class (metastasis vs. nomral)	Breast cancer	Additional power of genes
Boulesteix et al. 2008 [11]	Early	2-step	Full	Supervise feature extraction, PLS	Tree based method	Binary class (metastasis vs. nomral)	Breast and Colorectal cancer	Additional power of genes
Boulesteix et al. 2010 [84]	Early	1-step	Full	Regularization based technique	Logistic regression with boosting	Binary class (metastasis vs. normal & remission vs. no-remission)	Breast cancer, Leucemia	Additional power of genes
Oelker et al. 2013 [171]	Early	1-step	Full	Regularization based technique	Regression with boosting	Binary class	Breast cancer, Synthetic data	Additional power of genes

Table A.6: A summary of predictive models(cont.)

De Bin et al. 2014a [167]	Early	2-step	Full	Several techniques	Cox Regression	Survival analysis	Breast cancer and Neuroblastoma	Additional power of genes
De Bin et al. 2014b [23]	Early	2-step	Full and Sub	Any technique to generate a genomic score	Cox Regression	Survival analysis	Acute Myeloid leukemia, Chronic lymphoid leukemia	Additional power of genes
Dunkler et al. 2014b [137]	Early	2-step	Full	Previous Studies	Cox Regression	Survival analysis	Lymphoma, Breast, Head and Neck Cancer	Additional power of genes