DATA ANALYTICS AND STOCHASTIC OPTIMIZATION MODELS FOR DECISION SUPPORT IN CHRONIC DISEASE OPERATIONS MANAGEMENT

by

MOHAMMAD HESSAM OLYA

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: INDUSTRIAL ENGINEERING

Approved By:

Advisor Date

DEDICATION

To my beloved parents for their endless love, encouragement, and support.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere appreciation and gratitude to my advisor Prof. Kai Yang for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation. I would like to thank Dr. Yang for encouraging my research and for allowing me to grow as a research scientist. His advice on both research and my career have been priceless. Similar, profound gratitude goes to Mrs. Susan (Qian) Yu, Chief of Quality and Performance at the John D. Dingell VA Medical Center for her support and guidance. This study is a part of research supported by the National Science Foundation, Division of Civil, Mechanical, and Manufacturing Innovation (CMMI) under grant number 1233504.

Besides, I would like to thank the rest of my dissertation committee members: Dr. Kyoung-Yun Kim, Dr. Qingyu Yang, and Dr. Dongxiao Zhu for their comments that incented me to widen my research from various perspectives. In addition, my sincere appreciation goes to my fellow officemates in Healthcare System Engineering group at Wayne State University. Special mention goes to Hossein Badri and Milad Zafar Nezhad for useful comments, help, and advice.

Most importantly, I would like to thank my family from the deepest part of my heart. My parents and my siblings have always been there for me throughout my study here in the United States, even though they are thousands of miles away. I definitely would not make it this far, if it were not for their prayers, love, and support. Finally, but by no means least, thank you Sadaf, you have always been there for me through thick and thin.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

Chronic conditions constitute the leading causes of death and disability in the world. Chronic diseases are responsible for 60% of the global disease burden (WHO 2002). More than two-thirds of deaths in the United States are the result of chronic diseases according to Centers for Disease Control and Prevention (CDC). Heart disease, cancer, respiratory diseases, and stroke are the leading cause of death in the US. Studies show that diabetes is on the rise among Americans, and follows close behind as the seventh leading cause of death [1].

Recently the chronic disease pervasiveness has increased gradually among people of all ages. According to CDC, 133 million Americans had at least one chronic condition in 2005. In addition, it is projected that chronic disease affects 157 million Americans by 2020. Consequently, number of patients who have multiple chronic diseases is being increased in recent years. Studies show chronic diseases affected more than one in two adults and more than one in four children in the United States among them more than 25 percent live with multiple chronic conditions [2]. Multiple chronic conditions add complexity and cost to the health systems since they make the needs of patients with multiple chronic diseases more complicated compared to other patients. In addition, the rapidly aging population and a nationwide increase in risk factors for chronic disease can cause a significant increase in the number of patients with chronic conditions.

Chronic disease management is an integrated approach for managing illnesses that require coordinated patient care and treatment to improve the quality of care while reducing healthcare cost. Traditional formats of healthcare delivery are unable to address the needs of patients with chronic disease fully. The reason is that the complexity of the patient's clinical needs necessitates a pool of skills in the healthcare team. Based on prior systematic reviews of chronic disease management programs, Norris and colleagues defined chronic disease management in the clinical setting as "an organized, proactive, multi-component, patient-centered approach to healthcare delivery that involves all members of a defined population who have a specific disease entity (or a subpopulation with specific risk factors)"[3]. They believed that the care is integrated across the entire spectrum of the disease and its complications, the prevention of comorbid conditions, and the relevant aspects of the delivery system. They considered "identification of the population, implementation of clinical practice guidelines or other decision-making tools, implementation of additional patient-, provider-, or healthcare system-focused interventions, the use of clinical information systems, and the measurement and management of outcomes" as essential components of chronic disease management.

Chronic disease operations decision makers can prevent costly effects of chronic conditions with the help of numerous policies and programs. Access to comprehensive, quality health services is vital for everyone, but even more critical for those who have chronic conditions. Decision makers can enhance access to care by predicting the demand and workload of patients. Then they can optimize the use of healthcare resources. Thus, decision makers can improve prevention, detection, and treatment of chronic health conditions, yet many people face significant barriers for accessing to care.

More than 45 million Americans currently lack health insurance, many more are affected by the high cost of care, and others live in communities where services are difficult to access or unavailable.

Lack of insurance can adversely affect chronic disease operations management. One of the negative consequences of being uninsured is that people without insurance coverage are more likely to skip treatments, use emergency rooms, and be hospitalized due to the high cost of medical services. This issue is worse among seniors or patient with comorbidities since they need ongoing treatments and care. In addition, the mentioned issue may result in undiagnosed and uncontrolled chronic disease and eventually death.

Recent researches show that almost one-third of uninsured, working-age U.S. adults have at least one chronic condition. The primary goal of the 2010 Affordable Care Act (ACA) is to enhance access to health insurance through health insurance exchanges and optional Medicaid development. In addition, Federally Qualified Health Centers (FQHCs) can enhance access to healthcare. Many patients who lack access to care, including the uninsured patients, residents of rural and underserved areas can be provided with primary medical, dental, behavioral and social services in FQHCs. However, the cost of providing the aforementioned health services plays a significant role in the success and survival of the proposed interventions. Since many of governmental healthcare entities provide services regardless of an individual's ability to pay, there is a critical need to optimize the cost of these healthcare systems. Nowadays, debates and discussions about the cost-effectiveness and feasibility of these programs is a hot topic. Chronic disease operations management can help legislators in their decision-making by providing insight so that they can choose between many strategic policy options effectively. The purpose of chronic disease operations management is to provide a low cost and high-quality healthcare to patients. The reduced cost of providing healthcare can ensure access to a full range of quality health services for people with chronic diseases despite their location. Thus, even uninsured or underinsured patients can use healthcare benefits and seek care for chronic diseases. In addition, it ensures that the healthcare systems are well planned and cost-effective so that it can survive during a course of the strategic planning horizon.

Primary care plays a substantial role in healthcare delivery systems, in a way that it is considered as the primary resource for patients to get their consultation. Thus, improving primary care scope may result in a significant increase in patient’s satisfaction. The fundamental part of primary care is primary care physicians (PCP), and one of the essential attributes of PCP is patient panel list. A patient panel consists of the assignments of sets of patients to their providers. Typically, the patient panel size is predetermined and has a specific maximum size. In United States the average panel size is about 2,300 patients for one year [4]. One of the challenges in patient panel design is the quota size that is already predetermined, and it is not dynamic to balance the workload of each provider. Many factors can affect the PCP workload such as patient’s gender, age, insurance, and diagnostic code. These factors can change the workload amount based on their status so by assigning a certain number of patients to each PCP we cannot expect the same workload amount since a patient panel with a particular number of patients who are young and healthy can generate a different workload than the one with elderly patients with the chorionic disease. In the first case, PCP is underutilized, and it imposes a cost to the healthcare delivery system. In addition, in the second case, the PCP is doing excessive work that increases the risk of misdiagnosis as well as increasing waiting time, PCP switching number, the probability of visiting emergency department and cost caused by patient dissatisfaction and provider overtime. In addition, it is so important to maintain continuity of care, so patients and their PCP can be involved in increasing the quality of care over time. To achieve this goal, providers should be able to create a balance in their patient panel size so that patients are able to get a timely and regular appointment and see their physicians. This process leads to enhance the patient experience and the probability of being diagnosed accurately and taking medications correctly. Healthcare systems became more complicated and multi-level in recent years. An efficient healthcare system needs to have specific characteristics such as continuity of care, availability of care, comprehensiveness, coordination between providers, and focusing on all aspects of care that patient needs at a care period. One of the models that completely includes these factors is team-based care delivery.

With the complex need of patient with multiple chronic conditions and the advances in the treatment of chronic diseases, teamwork in the context of chronic diseases needs to be considered as an effective approach in order to cater the needs of patients and provide a high-quality healthcare service. A coordinated multidisciplinary care team is a crucial factor in successful chronic disease interventions [5]. A patient care team is a group of diverse clinicians who communicate with each other regularly about the care of a defined group of patients and participate in that care. A healthcare team that includes skilled clinicians contribute to chronic disease operations management effectiveness. Granting responsibilities to other team members rather than doing all the works by primary care physician can contribute to the success of chronic disease operations management since it ensures patients receive high-quality care. In order to provide team-based care effectively, one can add new disciplines and skills while defining some new roles for each member of the team, such as nurses and pharmacists. Team care increases the number and enhances the quality of available services. Therefore, it leads to improved health outcomes and reduced healthcare costs.

Recently, many healthcare entities used team-based care to improve the quality of care delivery. Team-based care aims to improve access to care and coordination of patient care. In team-based care delivery systems, a team of health care providers consisting of physicians, nurses, nutritionists, pharmacists, community health workers, and social workers all work together with focusing on a person’s overall health, in order to provide coordinated, comprehensive care for patients with multiple chronic conditions. Team-based care helps many healthcare entities to reduce costs and improve chronic disease care delivery. Since the healthcare delivery by a coordinated team of individuals benefits from the insights of different bodies of knowledge, and a wider range of skills.

Team-based care is beneficial for enhancing healthcare systems in various aspects. For example, it helps physicians to assign more time to diagnosis and treatment so the other healthcare team members can focus on different aspects of care. Therefore, since the healthcare team provides other care services, the cost of delivering healthcare decreases, and the system becomes more cost efficient. Moreover, team-based care leads to better chronic conditions management and enables more preventive screening. In addition, using team-based care reduces the waiting time of patients and helps to balance their required workload, due to providing care by teams and being patient-oriented instead of provider-oriented. The stated operational problems show that team-based care delivery systems have many potentials for improvement through using data analytics, operations research and optimization techniques.

As it is mentioned, one of the most effective ways to address chronic disease is through team-based care [8]. However, in some geographic areas, there is a shortage of necessary primary care providers for these teams. We project that the demand for primary care workforce increases due to the increase in the number of newly ensured Americans under the Affordable Care Act. Thus, considering the rate of enrollment, the need for more available providers to ensure that the new patients have adequate access to primary care becomes critical.

Many health centers and hospitals utilized the team-based care model lately. However, achieving its full potential in practice is a challenging task due to uncertain demand. Many studies evaluated the performance of team-based care in practice. Although the model performance is acceptable in some studies, there is room for improvements in team-based care operational modeling to achieve its potential fully. As we mentioned earlier, the patient panel adjustment can ensure a better quality and continuity of care in healthcare systems. Therefore, developing a systematic approach to predict the patient demand then assigning each patient to teams is critical for designing patient panels.

People with chronic conditions, and especially those with multiple chronic conditions, receive care from numerous providers in various settings. Therefore, the demand for healthcare is growing, so some policies need to be considered in order to make the workforce able to meet the healthcare demand.

Supply and demand realization in the traditional format of primary care and team-based care are different. In traditional primary care, supply is equal to the total available time of physician, and demand is total required workload. However, in team-based care model supply depends on all the available time of the members of each healthcare team. Therefore, it is a portfolio of the total available time of each healthcare provider in the team. Besides, in team-based care, we treat patient required workload as a stochastic variable that spreads through the healthcare team members based on their duties and professions. Therefore, it generates a portfolio of required workload that depends on different conditions and attributes of the patient. Thus, patient panel design has a more significant effect on the efficiency of team-based care model than traditional primary care model.

Recently, few studies focused on investigating the advantages of team-based delivery systems. The results of the studies show that using a multi-professional group including trained nurses and staffs who complement the physician in critical care functions, is associated with better outcomes, patient satisfaction improvement. In addition, it helps healthcare decision makers to cater the demand complexity issues occurring in the process of chronic disease healthcare management.

In order to achieve the above-mentioned objectives, we propose a comprehensive framework for chronic disease operations management in this study. This framework provides the modeling and solutions for optimizing chronic disease operations in three different management levels, i.e., strategic, tactical and operational.

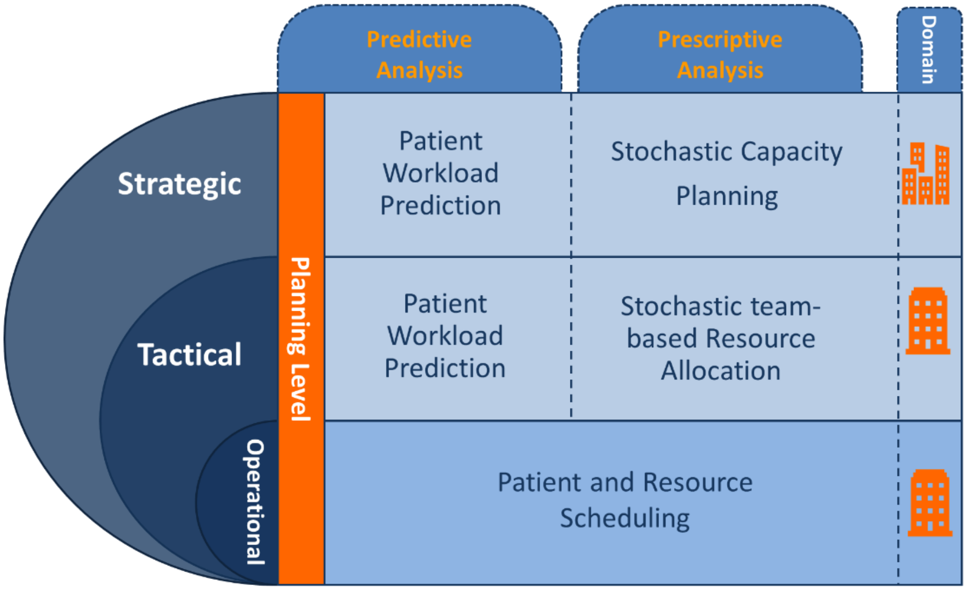
Predictive Analysis
Prescriptive Analysis
Domain
Strategic
Tactical
Operational
Planning Level
Patient Workload Prediction
Stochastic Capacity Planning
Patient Workload Prediction
Stochastic team-based Resource Allocation
Patient and Resource Scheduling

Figure 1.1: Chronic Disease Management Architecture

As it is shown in Figure 1.1, the proposed framework includes two phases: predictive and prescriptive analysis while the domain of action consists of two parts: inter-and intra-facility operations management.

In phase one; we present a predictive analysis that provides healthcare management boards with actionable insights based on data. Predictive analytics uses historical patient data, statistical models and prediction algorithms in order to provide patient's workload estimates and pattern identification. Therefore, we define two types of problems in this phase as below.

1. Modeling of the location of patients and estimating the required workload of patients for all facilities in the healthcare system

2. Predicting the required workload of patients based on their chronic disease features for each facility

After estimating the above-mentioned uncertain variables, we implement prescriptive analysis in phase two. Prescriptive analysis proposed in phase two assists decision makers to quantify and optimize the effect of future decisions and help them toward various strategic, tactical and operational solutions. Stochastic optimization helps decision makers to understand how they can achieve the best outcome and identify data uncertainties to make better decisions. Thus, we define three problems in this phase as below.

1. Stochastic capacity planning to determine the number of each healthcare provider for all facilities in strategic level

2. Stochastic recourse allocation, team workforce and workload optimization for determining the number of healthcare provider team types with different compositions within each facility in tactical level

3. Patient and resource operational planning within each facility in operational level

We discuss the current state of chronic disease operations management, its importance, and challenges, in the next part of this research

1.2. Research Motivation and Objectives

Chronic diseases are responsible for 7 in 10 deaths among Americans each year and the vast majority of health care costs. In addition to the serious consequences for the nations' health and health care systems, the increase in the number of patients with chronic conditions significantly contributed to health care costs [9]. Chronic disease treatment cost accounts for seventy-five percent of U.S. health care spending. Beneficiaries who have complex needs significantly influence healthcare costs and account for an even larger portion of spending [10].

Beyond the healthcare cost of chronic conditions, chronic diseases can adversely affect the economy and reduce economic productivity by increasing the rate of absenteeism and poor job performance. Studies show that chronic diseases cost the U.S. economy nearly $1.3 trillion annually, including $277 billion for treating chronic conditions and $1 trillion in lost productivity [11].

The main motivation of this research is to provide a solution for issues in chronic disease operations management in different planning levels by optimizing the healthcare operations. Managing the chronic disease operations leads to a reduced healthcare operation cost. This approach contributes in decreasing the cost of healthcare delivery while increasing the productivity of the chronic disease operations management systems. This approach leads to chronic disease operations cost reduction by using data analytics and operations research techniques in order to plan for required resources and allocate them to the patients based on their various demand while balancing workload of the resources. In addition, the proposed approach results in enhancing the access of patients with chronic conditions to a reliable and efficient team-based care. To the extent of our knowledge, there is no systematic analytic approach and framework defined in different management levels for team-based chronic disease operations management applied in inter-and intra-facility domains.

The remainder of this research is structured as follows: in the next chapter, we focus on the predictive part of this dissertation and present a Deep Multi-Task Learning (DMTL) approach for predicting the required workload of patients. In this chapter, after introducing the problem, we investigate the research articles focused on patient workload prediction. Afterwards, we explain the adopted prediction methods as well as different approaches for measuring patient workload. Subsequently, we propose our developed model for predicting patient workload in this chapter. Then, we present the results of our approach and compare the results with the earlier approaches. We conclude chapter two with elaborating more on the results of the proposed approach and analyzing the model performance. In chapter three, we mainly focus on the prescriptive part of this dissertation where we model and solve the aforementioned problems in strategical and tactical levels of decision-making in chronic disease operations management as it is shown in Figure 1.1. In addition, we review the existing literature about healthcare capacity planning and resource planning in team-based care. Then, we put forth the description of the proposed two-stage stochastic models as well as explanation of the related assumptions considered in developing the stochastic models. Afterward, we discuss results from two aspects of algorithm execution performance and cost optimization. Finally, we summarize the research and discuss the research contribution and novelty as well as future research directions in chapter four of this dissertation.

CHAPTER 2 PREDICTIVE ANALYTICS: DEEP MULTI-TASK PATIENT WORKLOAD PREDICTION

2.1. Problem Statement

One of the crucial factors affecting healthcare delivery systems is indeterminacy in patient’s demand and related workload. The variation in the required workload of the patient depends on the inherent diversity of individuals differing greatly in socioeconomic factors and health conditions. In this case, decision makers need to predict the patient workload and precisely estimate the needed healthcare resources to manage the workload of healthcare providers and related risks by widening their vision toward developing healthcare intervention strategies. Particularly, underestimating the demand influences the quality of provided care negatively, whereas overestimating the demand raises operating costs. Although many approaches are used to anticipate the patient demand recently, still there are rooms for improvement in healthcare demand prediction due to various types of uncertainty and patterns existing in these problems.

Recently, data analytics methods transformed the world of healthcare research on heterogeneous patient information significantly. Dependable datasets and suitable analytical approaches are two fundamental components for predicting the demand of patient and develop data-driven models. Electronic Health Records (EHR) is a valuable source of structured datasets, which help analysists to investigate the effects of different clinical features and the relationship between numerous types of diseases and comorbidities with the patient demand by applying several statistical analysis tools especially machine learning on clinical datasets. EHR plays the most important role as the primary source for analysis of healthcare datasets and strengthens research foundations in health systems. It can capture and integrate patient-specific information, provide a high dimensional dataset comprising diagnosis results (ICD codes), patient conditions, treatments, medications, laboratory test results, and imaging data, and eventually bill information, socioeconomic and demographic data such as age, gender, and employment status. As previously mentioned, EHR offers an important foundation for generating data-driven predictive clinical multivariate models using machine-learning methods. These models help decision makers to make inferences on patient demand primarily based on their different attributes and features. However, without using representative features the output of the model is not reliable.

Feature representation is a crucial task in healthcare demand prediction due to the presence of high dimensionality in healthcare datasets. In order to make predictive models more accurate when the number of features is large, the first step to do is extracting relevant and important features and transforming them into more explanatory features. Machine learning algorithms benefit from feature representation techniques in different ways. Feature representation enables machine-learning algorithms to make the training process faster, reduce the complexity of a model to make it easy to interpret, improve the model performance and reduce overfitting. Feature learning techniques are categorized into two major categories, namely feature selection and feature creation. In feature selection, a subset of features is extracted while removing redundant, irrelevant and noisy features. However, feature creation techniques map features to a new space and combine them to reduce dimensionality of the original input and capture the important information more effectively. Feature selection and feature creation algorithms include many techniques such as Principal component analysis (PCA), Independent component analysis (ICA), K-means clustering as well as deep leaning. The remarkable performance of deep learning techniques rises the popularity, acceptance, and utilization of deep learning methods with multiple hidden layers in healthcare data analytics for a variety of purposes. Feature representation by deep learning is distinct from classic feature learning techniques. Unsupervised deep learning approaches are considered as effective healthcare data abstraction methods in many recent studies. In addition, deep learning approaches are utilized as a predecessor for supervised learning. One advantage of using deep architecture for EHR feature representation is the capability of expressing different concept levels that are difficult to be expressed explicitly or formally in the problem domain. In addition, another advantage is simulating the complex procedure of human brains by storing the features as weights of connections between nodes when using deep learning models with multiple-layer networks [12].

Recently, patient-centered and team-based healthcare systems replaced disease-centered systems. This transformation necessitates the integration of patients, healthcare providers, and medical facilities. However, it challenges healthcare systems to quantify the overall performance and efficiency, measure the workload of healthcare providers, and define payment, billing and compensation procedures based on the actual workload done by providers. Among datasets provided by EHR, Relative Value Unit (RVU) is one of the most valuable information since it plays a key role in quantifying the patient workload assigned to each healthcare provider. Physician work RVU determines the relative measure of time to complete the service, technical skill, physical effort, the intensity of mental effort and training as well as judgment required to supply a specified health service. RVU is an essential component of numerous physician practices and serves as an effective assessment factor for amount of workload. Therefore, this unique methodology is employed as a basis for calculations of healthcare team compensation in the modern healthcare service industry.

RVUs are selected based on the Healthcare Common Procedure Coding System (HCPCS), an accredited healthcare procedure codes defined by the American Medical Association's Current Procedural Terminology (CPT) for reporting hospital procedures in different levels. In this system, every physician needs to submit a report including all the services and their associated codes for each patient. These codes indicate which performed a course of action, prescribed treatments, injected or delivered medication to the patient. The CPT codes must be written in accordance with condition of patient and physician's examination and diagnosis, which is represented by International Classification of Diseases (ICD) codes. Furthermore, RVU specifies the relativeness of values assigned to different healthcare service types or procedures. For this purpose, every single procedure type or service is determined by using a specific amount of RVU. The more complicated procedures need a larger amount of RVU. For instance, an invasive surgical procedure would have a higher RVU than a well patient visit. According to this relative scale, a medical practitioner visiting five complicated or high acuity patients per day accumulates far more RVUs than a physician who visits ten or more low acuity patients per day. As it is mentioned earlier, if a code has a higher RVU, it needs more time, intensity and technical skills. For example, an RVU of one can be allocated to a level one office visit, an RVU of two can be referred to a quality three office visit, and finally, an RVU of twenty may be assigned to a surgical treatment. Thus, different conditions of patients influence the workload of healthcare providers. However, similar patients in different locations can have different workloads due to efficiency of health systems, physical location of the facilities, quality of the provided care and the cost of providing the care. Therefore, developing a facility-based predictive model to estimate the workload is essential in order to enhance patient’s satisfaction and reduce the overall cost of healthcare systems.

In this research, we employ Multi-Task Learning (MTL) as a useful solution to approach datasets containing multiple related instances in EHR from many healthcare facilities. The final goal of MTL is to manage helpful information that into several similar tasks in order to improve the general performance of all learning tasks. For this purpose, MTL may be combined with some other learning models including semi-supervised learning and unsupervised learning. Machine learning approaches essentially require a large number of samples to learn a precise learner. However, satisfying this primary requirement might not be easy in healthcare systems analytics since there are some difficulties in gathering healthcare data. [13]. If each task has a limited number of samples and the majority of learning tasks are related, mutual learning of the tasks improves the training performance in comparison with the individual learning of them. MTL method categorizes the dataset based on particular tasks and considers a limited training dataset for every defined task. Then all tasks are learned jointly to use the shared representation. This process helps to improve the performance of learning other tasks by using what is learned for each task. Healthcare demand depends on many facility-dependent features that can be different across various facilities. Thus, we consider every facility as a task with its data in multi-task learning while learning them simultaneously to develop an accurate predictor.

As discussed above, the accuracy and dependability of the patient workload prediction is an important and critical issue in any healthcare system since it has a direct relationship with decreasing the overall costs and increasing patient’s satisfaction. In the present research, we implement a multi-task learning approach on a represented patients’ data to achieve an accurate workload prediction when the number of samples in the dataset is limited and the workload is facility-dependent. Then we compare the results of the proposed deep multi-task learning approach with that of other various predictive approaches in order to develop a more accurate patient workload predictive model and improve the prediction effectiveness in different ways.

2.2. Literature Review

We have categorized the previous works related to the present research into two groups. The first group predicts patient workload in healthcare issues by using RVU while the second one applies multi-task learning in patient workload prediction. In the following section, we have investigated conducted studies in the mentioned categories, separately.

2.2.1. Relative Value Units and Patient Workload Prediction

Relative value unit is a standardized measure for outpatient workload and a national standard for measuring resource allocation, productivity, budgeting, and cost benchmarking. RVU demonstrates the relative amount of needed physician work, expertise, and resources for healthcare services. Many researchers use RVU as the main source to assess the required workload associated with patient’s demand. Therefore, an appropriate prediction of future demand and reasonable estimation of the overall cost for allocating resources require an accurate workload predictor. In most of the conducted RVU prediction studies, the main limitation is descriptive analysis or regression-based modeling for anticipating the workload. In this section, we discuss some of these prominent studies that use RVU as a measure of workload.

Most of the studies are concentrated on finding the attributes that influence the workload in the early stages of studying RVU. In one study, Moniz [14] categorized patients based on their predisposing characteristics, i.e., gender, age, and social structure, then calculated the average RVU for each category, and finally analyzed the difference between the mean RVU and mean RVU per beneficiary by means of univariate analysis of variance. In other words, he used RVU to determine the relationship between workload and three different categories of gender, age and beneficiary type, using univariate analysis of variance. The results showed that age, gender and beneficiary category provide significant value for predicting the workload.

Many factors such as patient's age, gender, and diagnostic codes subject workload to variation. Østbye et al. [15] suggested that the type of chronic diseases affects the patient's visit frequency. Also, Naessens et al. [16] showed that clinical workload and related medical cost could be significantly varied based on the number of chronic conditions in a patient. Some studies used relative value units as predictor factors since they are among most important KPIs in order to measure workload. Turrentine et al., [17] studied the attributes of mostly elderly patients undergoing major operations in order to predict morbidity, mortality, and risk factors. They used stepwise logistic regression to predict mortality as an independent variable, and then investigated the effect of age on the outcome variables. In addition, they identified the risk factors predictive of morbidity and mortality associated with age groups.

Applying regression-based methods is considered as a straightforward solution to predict RVU in order to understand the demand of patients precisely. Murphy [18] used the surveys completed by PCMH team members in order to develop a demand-based forecast for RVU volume using multiple linear regression modeling. In this research, she studied the relationship between patient workload and per-encounter independent variables such as age, gender, beneficiary category, provider specialty, evaluation and management code, and appointment type, while assessing the relationship of each independent variable with the workload separately. Shah et al., [19] applied linear and multiple logistic regression and investigated the correlation of surgical procedures including measures of surgeon effort and RVU. They showed that there is a clear correlation between RVU and certain measures of surgeon work and patient’s attributes, such as the frequency of serious adverse events, and patient’s overall morbidity. Furthermore, they demonstrated that RVU is a compelling factor in predicting the operative time, length of stay, and serious adverse events.

We can forecast RVU by means of some other methods. For example, Barnes [20] developed two models in which RVU was used to anticipate the future patient’s demand for ten specialty practices. In the first model, he performed a time series model to find the most precise fitting forecast model. He evaluated the result of the model by least mean square error and used exponential smoothing method to minimize the error. In the second model, he used the past usage rate to anticipate the patient's demand; however, this method can be only applied in short-term prediction since in long-time horizon, irresponsible external factors may have some effects on the prediction.

As discussed above, RVUs are variables that represent healthcare demand. Etzioni et al., [21] estimated the effect of the aging population on the amount of surgical work. They also multiplied the age-specific surgery incidence rate for each procedure by the corresponding work RVU to anticipate the future workload required for surgical works. The results show a considerable increase in demand for surgical services due to the aging population. Therefore, it seems necessary to have a robust methodology to control the growth of workload as well as maintain the quality of care. Crane et al., [22] considered relative value units, RVU/h, and patients seen per hour as inputs, and proposed a task-based framework entitled “entropy”, then they evaluated the relationship between workload and the crowding in the emergency department. Their framework measures some aspects of the workload in emergency departments, such as efficiency and acuity. They regarded the workload as an operational complexity and defined that based on the total information collected from each task during observations in a certain period. Therefore, the entropy formula helped them in assigning an entropy value to different tasks, and finally estimating the workload of all tasks performed by healthcare providers. Chasan et al., [23] assessed workload and resources in eye care procedures using RVU and provided descriptive statistical analysis. Arndt et al. [24] applied another approach to assess the workload. They provided a survey of the perceived workload while both face to face and non-face to face encounters in order to assess the primary workload. Their results show that regardless of health status when routine primary care is not face-to-face, the total workload of panel management activities is more significant than the total workload associated with face-to-face encounters.

Since RVU represents the workload, some previous studies worked on optimizing the capacity considering RVUs. As an example, Bryce and Christensen [25] assumed a variable workload for the patient’s demand and found the mean and variance of workload during different time frames. Ultimately, they fitted normal distributions for demands and attempted to match the resource capacity efficiently to optimize staffing process. In addition, there are some other researches focusing on predicting the cost of operations. Such researches assumed RVU as an input helping the development of decision support systems for resource allocation process. Fulton et al., [26] incorporated data envelopment analysis of efficiency scores into a traditional logarithmic-linear cost function. They considered RVU as cost driver for hospital operations since it represents the workload volume and complexity.

2.2.2. Deep Feature Representation in Healthcare

In 2006, after Bengio’s review on deep learning and summarizing the prominent algorithms in deep learning, deep learning methods gained the attention of many researchers across the globe [8]. Deep predictive modeling and feature representation, are studied and applied to a variety of domains, including automatic text generation, image and speech recognition as well as biomedicine [27], [28]. The primary reasons for the broad use of deep learning approaches are the ability of modeling complex systems, generating a high-level representation of features, and enhancing the prediction accuracy.

Recent rapid growth in information gathering mechanisms provides voluminous, complex and high dimensional datasets with a huge number of features. Due to high dimensionality in these types of datasets, the traditional machine learning approaches such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Generalized Discriminant Analysis (GDA) may not be adequate to handle the dimensionality reduction and feature representation process. In contrary, deep learning approaches have shown an excellent performance in dealing with the challenge mentioned above. Pervasive sensing, medical imaging, medical informatics, bioinformatics, and public health are all examples of the application of deep learning in health informatics that require their specific input dataset including data from wearable devices, MRI/CT images, EHR, gene expression, social media data, respectively [29].

Feature learning and data representation are critical factors that can significantly affect the performance of clinical predictive models. There are many shortcomings such as inability in generalization and discovering new patterns as well as lack of scalability, associated with the traditional approaches for EHR dimensionality reduction where subject matter experts input is needed for identifying the patterns and important features [30]. Recently some research articles tried to address the mentioned limitations by suggesting data-driven approaches for EHR feature selection, and identifying risk factors to extract the correlation and the dependencies in the clinical data [31], [32]. However, these methods have limited performance in dealing with high-dimensional EHR. Therefore, recently some researchers started to utilize unsupervised learning and deep learning methods to tackle this issue. Recently, Miotto et al., used an unsupervised approach to represent patient's features by taking advantage of stacked autoencoders [30]. After extracting the patient features from EHRs through a deep learning approach, they applied a random forest method for patient's future disease prediction. They showed that using stacked autoencoders for deep feature representation before applying shallow models for prediction improves the prediction performance.

One example of applying deep learning method in the bioinformatics area can be found in [33]. In this study, the authors employed a two-phase approach consisting of PCA and a deep sparse encoding approach on gene expression data for feature dimension reduction and generating high-level abstractions in order to enhance the performance of cancer diagnosis and classification [33]. In order to capture the non-linear information on top of the linear transformation generated by the PCA, they took advantage of sparse auto-encoders in the second stage and achieved a higher classification performance. In another study, it is indicated that deep architectures such as Restricted Boltzmann Machine (RBM) tend to be more powerful than regular ones when they are utilized to assess the text information obtained from EHR data for discovering new patterns [34].

Furthermore, deep learning is applied extensively in the medical image processing and pattern recognition area. As in [35], authors utilized deep network with a RBM as a building block to find a latent hierarchical feature representation for image processing in order to diagnose Alzheimer's disease. Other researchers used various deep learning-based approaches to analyze medical images. For example, Hu et al. [36] proposed an autoencoder architecture customized by using a SoftMax output layer for image processing to predict Alzheimer's Disease (AD). In another study, in order to identify the progression stages of AD patients, Li et al. [37] proposed and applied an RBM approach on positron emission tomography and MRI scans. Unlike the previous studies, Suk et al. [38] considered the relations among the features in their study. They proposed a deep learning-based feature representation with a stacked autoencoder in order to discover complicated latent non-linear patterns in features. The results suggested that deep feature representation and using deep learning methods improve the model performance compared to conventional machine learning methods.

In a nutshell, despite the extensive research studies that we reviewed above, and broad utilization of deep learning approaches for feature representation and prediction in many areas in healthcare such as medical image and text processing and bioinformatics during the recent year, these techniques have not been used in the healthcare systems engineering area. Thus, there is still a significant research gap when it comes to predicting the patient demand and resource workload by using unsupervised feature representation to improve the healthcare system operations, and resource planning.

2.2.3. Multi-Task Learning

Multi-task learning is employable for many areas such as bioinformatics, natural language processing, computer vision, and healthcare informatics. However, there is no comprehensive study about the application of multi-task learning in health informatics in the literature. In this section, we aim to explore the studies on bioinformatics and healthcare informatics using MTL as an effective approach for workload prediction.

Widmer et al. used MTL method with various types of regularization terms to predict the sequence signals in genes finding [39]. In another study, in order to associate gene expression data with phenotypic signatures, the authors coupled multitask regression with co-clustering [40]. Under such circumstances, multi-task regression outperformed traditional Lasso and Ridge regression models. Liu et al. [41] ranked biological features based on the joint importance of siRNA and applied multi-task learning in predicting the efficiency of cross-platform siRNA. Mordelet and Vert [42] used multi-task learning and took advantage of shared information across disease genes to prioritize disease genes. In another example found in [43], the researchers applied multi-task learning in genetics in which genetic trait is predicted by multi-task learning and multiple output regression models instead of the linear regression model. In addition, they discovered a correlation between genetic markers in multiple populations while applying multi-task Lasso regression with L_1 and L_2 regularization. Moreover, Xu et al. captured the shared information among different organism by using the MTL method to predict subcellular protein location in another study [44].

Another application of MTL is the development of brain-computer interfaces by sharing a Gaussian prior on parameters of different tasks [45]. In another study, MTL is formulated as multiple kernel learning for MHC-I binding and splice-site prediction [46]. Zhou et al. [47] considered the prediction at each time point as a task for mini-mental state examination and Alzheimer’s disease assessment then used multi-task regression for forecasting. They used temporal group Lasso regularization term with two components including an $L_{2,1}$-norm penalty to ensure selection of a small subset of features, and a temporal smoothness term to have a small deviation between the two regression models at successive time points. In another research on Alzheimer’s disease prediction, Wan et al.

exploited the relationships between neuroimaging measures and cognitive scores by developing a sparse Bayesian multi-task learning algorithm [48]. Also, in some studies, researchers developed a model for Alzheimer’s disease progression prediction in which multi-task learning is integrated with time-series [49].

MTL is also able to solve other types of problems in healthcare informatics. As explained in [50], MTL is applied to analyze biological images where deep learning architectures such as convolutional neural networks are used for feature representation for improving the model performance. Another application of MTL can be observed in formulating survival analysis as a classification problem since they consist of multiple tasks related to survival prediction [51], [52]. As discussed, even though many extensive types of researches are conducted on MTL applications in various areas in recent years, there is still a research gap in developing patient workload prediction models using the MTL method. The limitations of previous studies in considering the relatedness between patient instances encouraged us to conduct the present research.

2.3. Methodology

We illustrate the framework of this study in Figure 2.1. The proposed approach consists of three phases, namely data pre-processing, unsupervised learning and supervised learning. We explain the steps in detail in the following sections.

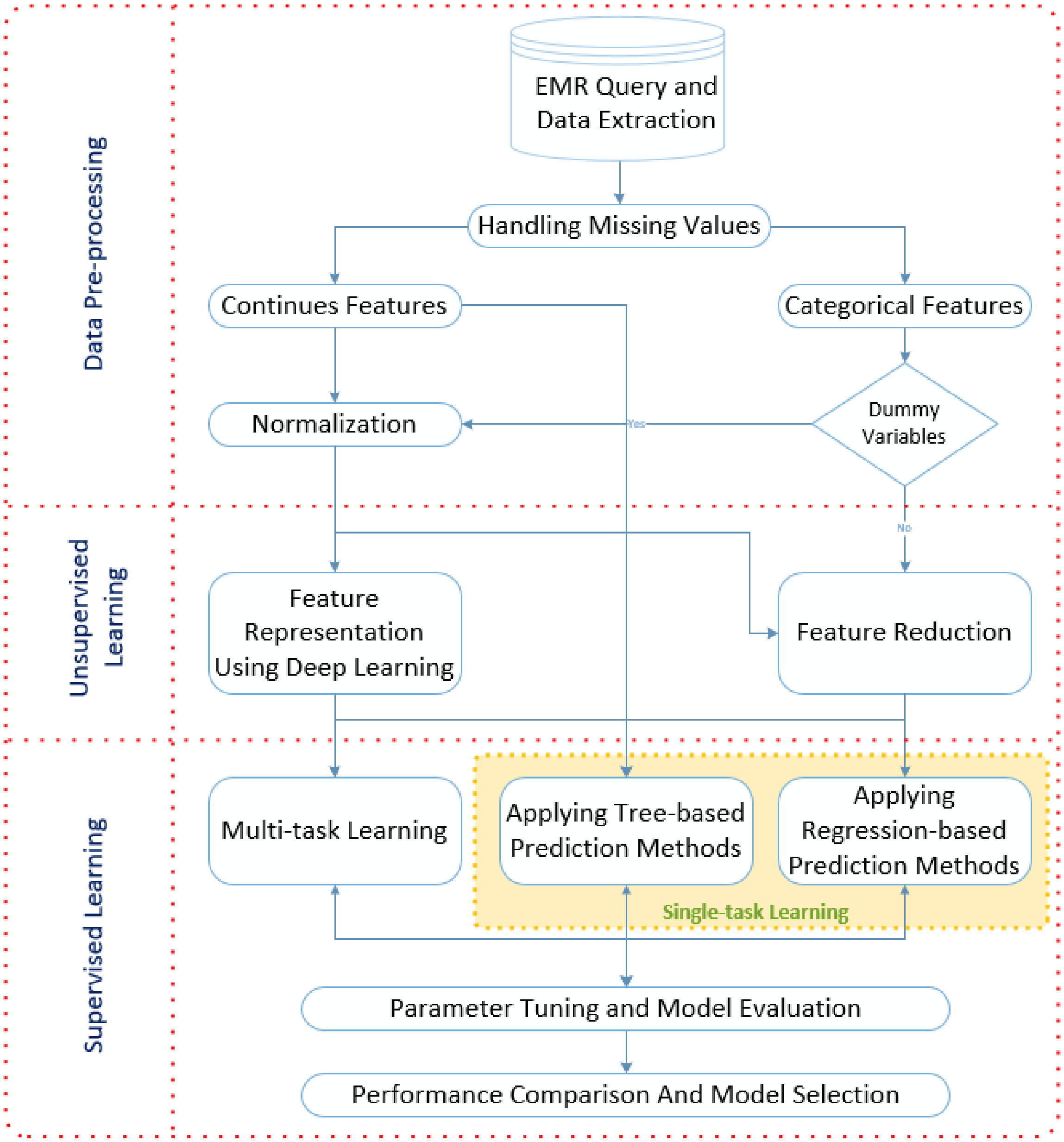
Data Pre-processing
Unsupervised Learning
Supervised Learning
EMR Query and Data Extraction
Handling Missing Values
Continues Features
Categorical Features
Dummy Variables
Yes
No
Normalization
Feature Representation Using Deep Learning
Feature Reduction
Multi-task Learning
Applying Tree-based Prediction Methods
Applying Regression-based Prediction Methods
Single-task Learning
Parameter Tuning and Model Evaluation
Performance Comparison And Model Selection

Figure 2.1: Predictive Analytics Framework

2.3.1. Data Pre-processing

The quality of data directly affects the quality of the prediction. Therefore, in the first stage of the research, after extracting the data, we cleaned the data by removing outliers and taking care of missing values. Based on the type of missing values, we either removed or imputed them by using the mean of all samples with the same class. Then, we transformed the categorical variables into numerical variables by generating dummy variables and one-hot encoding process. Afterward, we scaled the data by using min-max normalization method to reduce redundancy in the data.

2.3.2. Unsupervised Learning

2.3.2.1. Feature Reduction

The primary goal of dimension reduction is to transform the data into a dataset with lower dimension and ensure that the new dataset conveys similar information or the information with higher quality. Using dimension reduction results in many benefits such as tackling the multi-collinearity issue, reducing the computation time and noise reduction. The mentioned benefits lead to an enhanced prediction performance. In order to create different subsets of models containing important features that are required for building an accurate predictive model, many researchers used automatic feature selection methods.

Among many feature reduction techniques, we use Boruta algorithm [53] which is a wrapper built around the random forests algorithm to identify important features. There are many advantages in using random forests method such as a relatively high computation speed compare to other methods, no need for parameter tuning and generating a numerical output of feature importance. Random forests algorithm is an ensemble method that uses voting of multiple decision trees as unbiased weak classifiers for performing classification.

Random forest algorithm is developed based on a group of decision trees, which utilize a random sample of the original dataset. Thus, this model is able to remove the correlation existing between basic learners. Furthermore, every split created inside each tree uses only a random subset of sampled attributes. The number of attributes affects the balance between variance and bias of training. Classification tasks have a default value for the square root of the total number of attributes, giving us an immensely powerful way for selection. The random forest method is one of the most popular techniques since it is simply applicable in the domains of different regression and classification tasks. The advantages of using this method are not limited to the high quality of estimation. An additional advantage is the ability to determine the feature importance by means of measures of accuracy. Sometimes the prediction accuracy decreases if data for attributes are removed from the dataset.

As mentioned earlier, Boruta algorithm is a well-known method for ranking features and employs a random forest model to estimate feature relevance. The main reason to choose Boruta algorithm is the ability of algorithm in raking all the features by their relevance at the end of the iterations, in contrast with the conventional feature selection methods that remove some features during each iteration. Therefore, this process results in a small subset of the features at the end of all iterations. Furthermore, since this algorithm is developed based on random sampling of the original dataset, the correlation between basic learners are removed. Algorithm 2.1 describes Boruta algorithm in detail [54].

In this algorithm, randomizing the system and gathering results from groups of samples help us to reduce the mistakes resulting from random correlations and fluctuations. In this algorithm, joining copies of original attributes extends the original dataset. The values for the extended data are randomly rearranged based on learning cases in order to remove their connection with a decision attribute. This algorithm finds a statistically significant subset of features by incorporating randomness into the model via duplicating the variables and shuffling the duplicated copy of variables to create shadow features. Therefore, it results in a reduced misleading effect of correlation and random fluctuation.

The measure for identifying important features is obtained by mean decrease in accuracy and calculating the classification accuracy loss caused by the random permutation of the attribute. Then the algorithm calculates the Z-score by using the average and standard deviation of the accuracy loss for all trees in the forest that use the given attribute, separately. Afterward, the algorithm compares the importance of each feature with the best of its shadow by using Z-scores. If the Z-score of the variable is higher than the Z-score of its shadow, the algorithm records that variable in a vector called Hits. If the importance of attributes is significantly lower than Maximum Z-score of Shadow (MZS), the attributes are defined as irrelevant (rejectedSet). This procedure is iterated until the importance of all attributes is estimated. In the end, the number of recording times for each feature in Hits vector is calculated, and the features with higher frequencies are selected. The time complexity of the Boruta algorithm is $O(P.N)$ where p is the number of features and N represents the number of samples in the dataset.

Algorithm 2.1: Boruta algorithm for feature selection

```
Input: the input dataset: DFmain; the number of random forest execution: RFe
Output: the set of rejected and confirmed features: featureSet
$confirmedFeatures = \emptyset$
$rejectedFeatures = \emptyset$
for RFe do
    $predictorsMain \leftarrow DF(predictors)$
    $predictorsShadow \leftarrow permute(predictorMain)$
    $predictorsExt \leftarrow cbind(predictorsMain, predictorsShadow)$
    $DFext \leftarrow cbind(predictorsExt, DFmain(decisions))$
    $zScoreset \leftarrow randomForest(DFext)$
    $MZS \leftarrow max(zScoreset(predictorsShadow))$
    for a ∈ predictorsMain do
        if $zScoreset(a) > MZS$ then
        $Hits(a)$ ++
for $a \in predictorsMain$ do
    $significance(a) \leftarrow twoSidedTest(a)$
    if $significance(a) >> MZS$ then
        $confiremedFeatures \leftarrow confirmedFeatures \cup a$
    elseif if $significance(a) << MZS$ then
        $rejectedFeatures \leftarrow rejectedFeatures \cup a$
return $featureSet \leftarrow confiremedFeatures \cup rejectedFeatures$
```

2.3.2.2. Feature Representation by Using Deep Leaning

Representation learning methods transform the raw data into an abstracted form, which conveys the same information by discovering the various representations needed for classification or prediction automatically. Deep learning methods are categorized as representation learning methods that learn complex and non-linear modules at multiple transformation layers to transform the data into an abstract representation [55]. One of the most powerful properties of deep learning and neural networks is their flexibility that results in a great improvement in the predictive model performance. In order to get a higher-level abstraction of data, in this study, we implemented stacked autoencoders, which is one of the most popular approaches in deep learning.

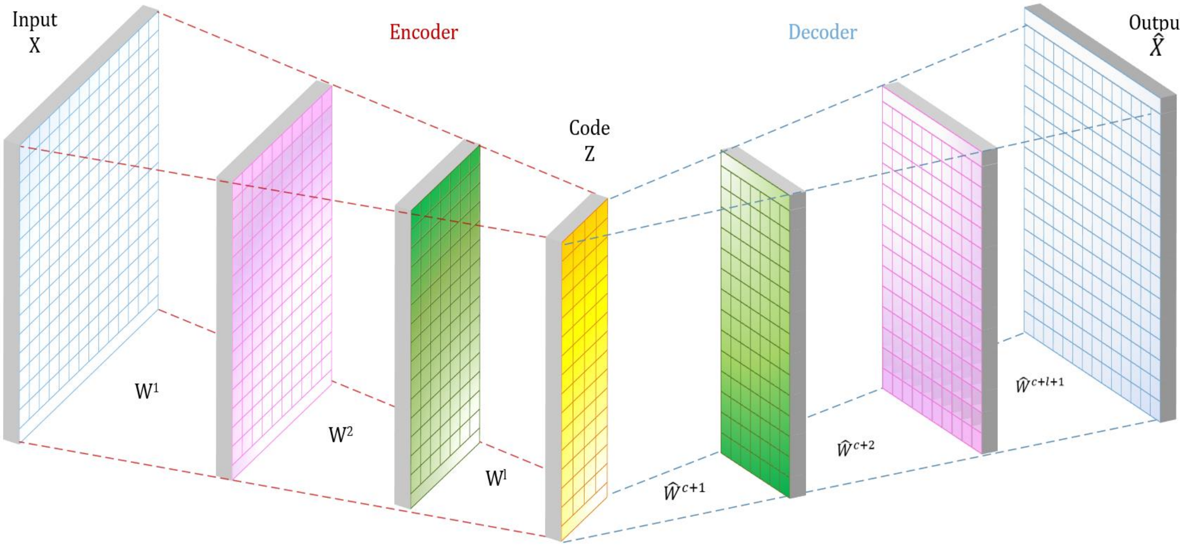

Figure 2.2: The Architecture of Stacked Autoencoders

Consider two sets *X* and *Z* that represent two datasets where *Z* has a lower dimension than *X*, and *Z* can reconstruct *X*.

$$X = \{x_{(1)}, x_{(2)}, \dots, x_{(n)}\} \tag{2.1}$$

$$Z = \{z_{(1)}, z_{(2)}, \dots, z_{(n)}\} \tag{2.2}$$

We are interested in mapping set *X* to set *Z* by reconstructing *X* via *Z*. We call the reconstructed set of *X* as $\hat{X} = \{\hat{x}_{(1)}, \hat{x}_{(2)}, \dots, \hat{x}_{(n)}\}$ since it is an estimation of elements of set *X*. In order to attain the mentioned objective, autoencoders follow two main steps, namely encoding at the first step and decoding at the second step. Consider l as a parameter that indicates the number of layers in the autoencoder system. Let us consider w and $\widehat{w}^l$ as the weights of the encoding and decoding processes for each layer of autoencoders, respectively. In addition, we show the bias for each layer with b^l and $\hat{b}^l$ in encoding and decoding steps, respectively. The mathematical formulation of the encoding and the decoding process for each layer of autoencoders is denoted as follows [56]. Let us start with a neural network with one hidden layer, and then we extend the mathematical formulation for stacked autoencoders with more than one hidden layer.

$$\hat{X}\left(X, W, \widehat{W}, b, \hat{b}\right) = \sum_{k=1}^{K} \mathbb{G}(\widehat{W}_{kj} \mathbb{F}\left(\sum_{i=1}^{D} W_{ki}\, X_i + \ b_i\right) + \hat{b}_k) \qquad j = 1, \ldots, D \qquad (2.3)$$

where $\mathbb{F}$ and $\mathbb{G}$ can be any activation function such as Tanh, Rectified linear, Sigmoid or Max-out.

As stated before, our main objective is to minimize the difference between the input and the output while reconstructing the input via the hidden layer. Therefore, we define the objective function as below.

$$\boldsymbol{\mathcal{L}}\left(\boldsymbol{W}, \widehat{\boldsymbol{W}}, \boldsymbol{b}, \hat{\boldsymbol{b}}\right) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} (\hat{\boldsymbol{x}}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{i}})^{\mathbf{2}} \qquad (2.4)$$

By incorporating the activation functions into the loss function, we restate the objective function as follows.

$$\underset{\boldsymbol{W}, \widehat{\boldsymbol{W}}, \boldsymbol{b}, \hat{\boldsymbol{b}}}{\boldsymbol{min}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \left\| \mathbb{G}\left(\widehat{\boldsymbol{W}\,\mathbb{F}}(\boldsymbol{W}\boldsymbol{X}_{\boldsymbol{i}} + \ \boldsymbol{b}_{\boldsymbol{i}}) + \hat{\boldsymbol{b}}_{\boldsymbol{i}}\right) - \boldsymbol{X}_{\boldsymbol{i}} \right\|_{\mathbf{2}}^{\mathbf{2}} \qquad (2.5)$$

Depending on the activation function, which can be a linear or non-linear function, one method to minimize expression 2.5 is to utilize stochastic gradient descent.

As it is illustrated in Figure 2.2, a stacked autoencoder is a deep autoencoder that includes more hidden layers compared to the neural network. The increase in the number of hidden layers makes the objective function minimization and training the deep neural networks difficult. There are some reasons for this issue such as the difference between the magnitude of gradients in the higher and lower layers, the high number of parameters that avoids a good generalization and the difficult landscape of objective function [56]. In order to avoid the mentioned issues, we use a greedy layer-wise approach proposed by Bengio et al. [57] for finding the parameters of the model. In this approach, neural networks with one hidden layer are considered shallow networks. Shallow autoencoders are trained one layer at a time by using unsupervised data in a greedy manner. So, the hidden layer of each shallow network is considered as an input to be reconstructed for creating the next hidden layer until the highest representation of the data with the desired number of hidden layers is created. Then the algorithm fine-tunes the network by using backpropagation. The mathematical formulation for encoding stacked autoencoders is expressed as follows.

$$D^l = \mathbb{F}(z^l) \qquad (2.6)$$

$$z^{l+1} = W^l D^l + b^l \qquad (2.7)$$

Also, the decoding formulation for stacked autoencoders can be stated as follows [58], where *c* represents the index of the central layer.

$$D^{c+l} = \mathbb{G}(z^{c+l}) \qquad (2.8)$$

$$z^{c+l} = \widehat{W}^{c+l} D^{c+l} + \hat{b}^{c+l} \qquad (2.9)$$

2.3.3. Supervised Learning

The last stage of the proposed framework consists comparing the performance of two different approaches (i.e., applying conventional prediction models, and multi-task learning method) in order to select the best model for patient workload prediction. There are parameter tuning and model evaluation steps for each of the approaches mentioned above. These two steps run iteratively until their performance no longer increases and the best parameters for every model are attained. Then, we select the best model based on the prediction accuracy and the performance. In the next sections, we explain the multi-task learning approach.

2.3.3.1. Multi-Task Learning Approach

In contrary to single-task learning, multi-task learning is a paradigm that takes advantage of the relatedness between samples to leverage the knowledge of other related tasks for learning a specific task. As mentioned in the earlier section, studies show that learning multiple tasks jointly rather than individually results in performance improvement.

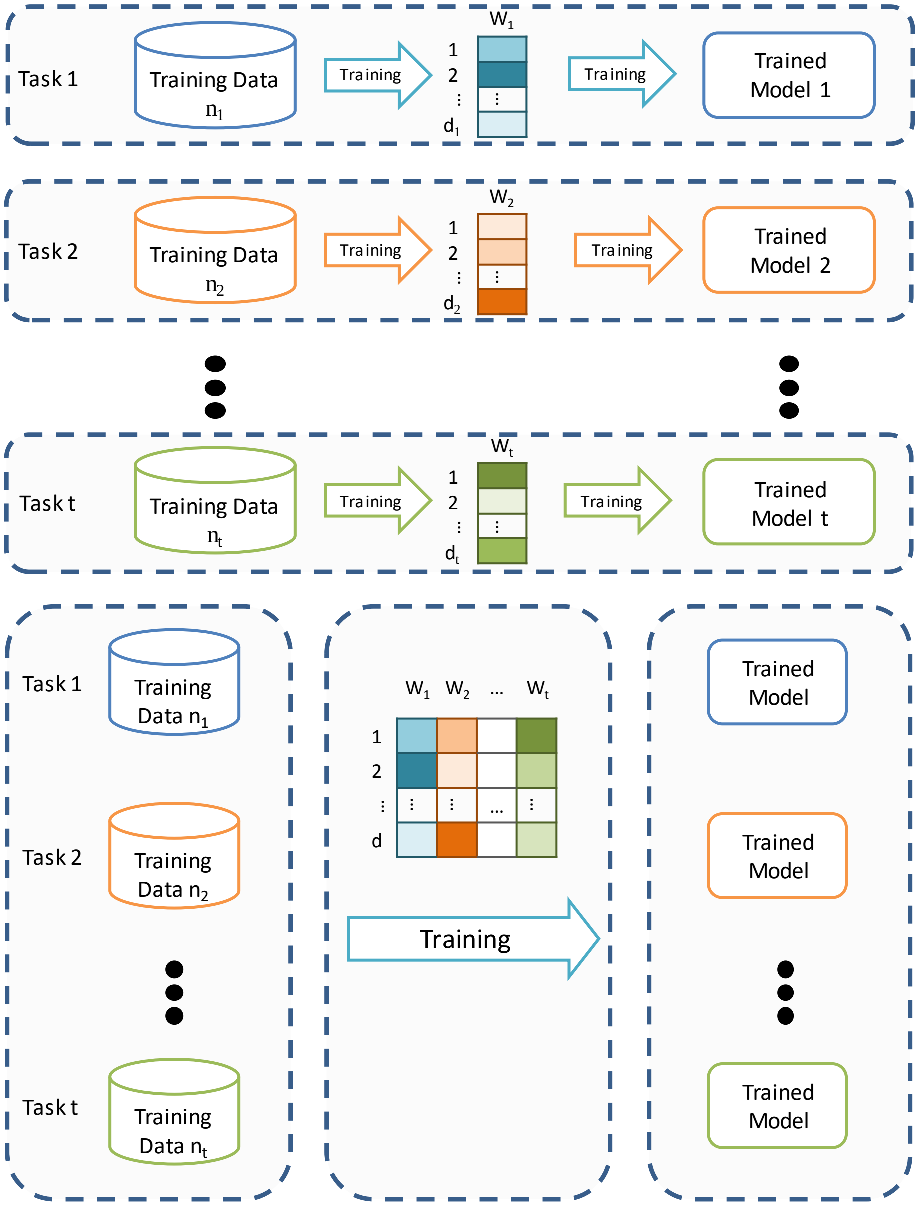

Figure 2.3: Comparison between Learning Schema of Single-Task Learning (Top) and Multi-Task Learning (Bottom)

Figure 2.3 illustrates the difference between multi-task learning and single-task learning. It shows that single-task learning trains each task individually (horizontal learning); however, in MTL the tasks are connected, so the hidden layer of information can be shared among the tasks (the vertical relationship between task). Therefore, tasks can affect each other. Furthermore, since in single-task learning each task uses its own data, when the number of data samples for each task is not enough, the problem of over-fitting may occur. In this case, MTL can be used to overcome over-fitting by sharing the data over different tasks and increasing the number of samples for each task. As it is shown in Figure 2.3, MTL is beneficial for better model training since tasks can affect each other and the knowledge between them can be transferred.

The objective function of MTL is to minimize the summation of the loss function and task regularization term that is defined as follows.

$$\min_{W} L(W) + \mathcal{R}(W) \tag{2.10}$$

In (2.10), W represents the collection of weight vectors learned for each task. In this study, T and D represent the number of tasks and attributes, respectively. So, the weight matrix is a $T \times D$ matrix ($W \in \mathbb{R}^{T\times D}$). $L(W)$ is the loss function and $\mathcal{R}(W)$ is the regularization term, which are expressed in (2.11) and (2.12), respectively [59].

$$L(W) = \frac{1}{2}\sum_{t=1}^{T}\|X_t W^T - \beta_t\|_F^2 \tag{2.11}$$

$$\mathcal{R}(W) = \|\mathrm{W}_{2,1}\| = \sum_{d=1}^{D}\sqrt{\sum_{t=1}^{T}|w_{td}|^2} \tag{2.12}$$

Where $X_t \in \mathbb{R}^{n_t\times D}$ and represents the input data for task t, n_t represents the number of samples for each task t and β_t is the response value corresponding to the samples in X_t. Therefore, we rewrite the objective function as below.

$$\min_{W} = \frac{1}{2}\sum_{t=1}^{T}\|X_t W^T - \beta_t\|_F^2 + \lambda \sum_{d=1}^{D}\sqrt{\sum_{t=1}^{T}|w_{td}|^2} \qquad (2.13)$$

Where λ ≥ 0 is defined as a tuning parameter that biases the data and controls the shrinkage of the model to make the model relativity simpler and sparser to reduce the complexity of the model.

2.4. Case Study and Results

We used the Veteran Health Administration (VHA) data from facilities across the nation. The dataset contains patient risk factors such as demographic, socioeconomic variables and number of visits as well as the number and type of chronic conditions for one year. The healthcare workload imposed on each provider is measured in relative value unit every year. Relative value units are a national standard used for measuring productivity, budgeting, allocating expenses, and cost benchmarking. RVU represents the relative amount of physician workload. Many government programs and private payers use the resource-based relative value scale and the relative value unit methodology as the basis for payment of many physician practices. This system is the foundation of medical group financial analysis and is unique to the medical service industry. The RVU schema is widely used for reimbursement in VHA and centers for Medicare and Medicaid services. In this schema, a value is assigned to every service as defined by a coding system called Current Procedural Terminology (CPT) rendered by a provider. The values are adjusted based on geographic regions.

As it is stated in the previous section, the first part of the proposed approach includes data preprocessing. We completed the pre-processing by performing four steps. First, we removed or imputed the missing values based on the type of the missing attribute. Then we identified and removed outliers. Afterward, depending on the method used for prediction, we converted categorical variables to a numerical variable by taking advantage of indicator variables. At last, we normalized the data by using the min-max normalization method.

According to Figure 2.1, the next phase of the framework includes unsupervised learning, where we applied feature reduction and feature representation techniques on the data seeking for a better and more robust representation of the data. This phase consists of feature reduction with Boruta algorithm and feature representation with deep learning.

Let us start with the feature reduction stage, we used Boruta algorithm for feature reduction. The preliminary result of the algorithm implementation is depicted in Figure 2.4. As it is shown in Figure 2.4, there are three possible decisions for each attribute including confirmed, rejected, and tentative represented by green, red, and yellow colors respectively. When the number of the random forest runs are not sufficient, the algorithm is unable to decide whether to reject or confirm the attribute, since the importance score of the attributes is close to their best shadow attributes. In this case, the algorithm labels the attribute as tentative. Figure 2.4 represents the importance of attributes over different runs.

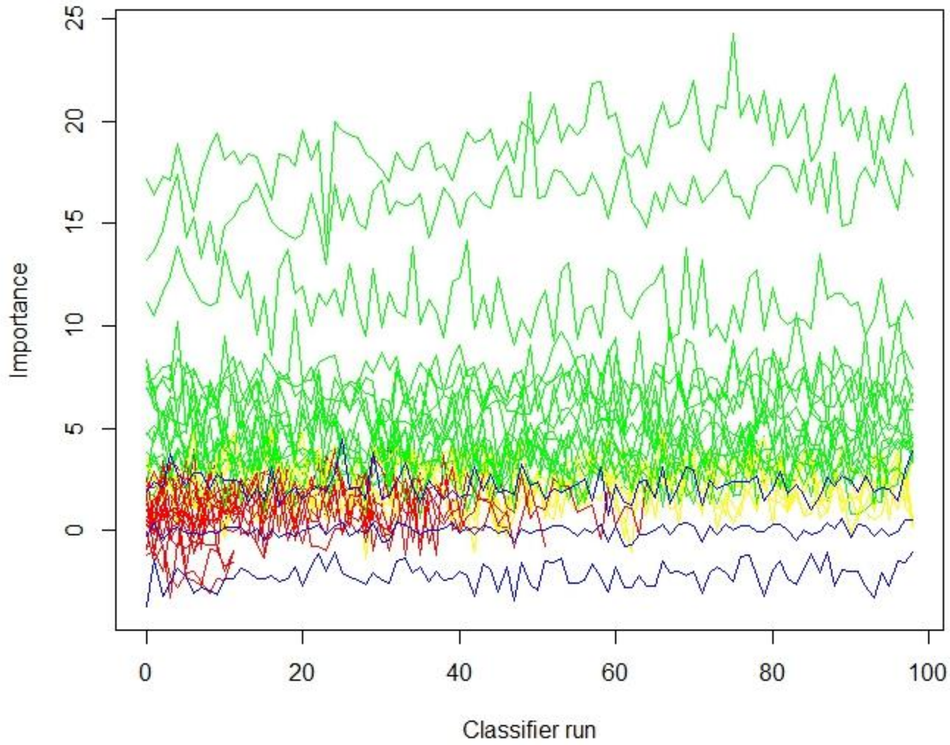

Figure 2.4: Importance of Attributes over Classifier Run

Thus, in order to make the final decision faster, we took advantage of a weaker test for deciding about the remained attributes while taking into the account that the number of tentative variables is not high. In this test, the median importance of each attribute is compared with the median importance of the maximal shadow attribute. So that the attributes that have higher median importance than those of their shadows are claimed as confirmed, and the rest of tentative attributes are rejected. The median of the test is counted over whole Boruta runs.

The results of Boruta algorithm are shown in Table 2.1 for each feature. This table contains the mean, median, maximum and minimum importance as well as the count of times that every attribute scored better than its shadow (hits) normalized by the number of performed runs. At the last column, we presented the final decision about the importance of every feature.

Table 2.1: The Results of Feature Analysis

	Mean Imp.	Median Imp.	Min Imp.	Max Imp.	Norm Hits	Decision
VISN	1.541	1.482	-0.449	4.388	0.091	Rejected
INCOME	3.092	3.230	0.615	5.507	0.788	Confirmed
SEX	2.776	2.755	0.819	4.657	0.697	Confirmed
MARITAL	0.795	0.992	-1.167	2.667	0.000	Rejected
EmpStatus	3.895	3.985	0.934	6.107	0.929	Confirmed
Priority	6.350	6.444	3.777	8.415	1.000	Confirmed
CanScore	19.105	19.025	13.974	24.294	1.000	Confirmed
Sta3n	2.221	2.221	-1.069	4.546	0.475	Confirmed
Ins	7.031	7.100	4.933	9.677	1.000	Confirmed
Age	16.168	16.277	12.978	21.403	1.000	Confirmed
ACC001	2.421	2.435	-0.356	4.922	0.596	Rejected
ACC002	1.026	1.118	-0.318	2.447	0.000	Rejected
ACC003	0.608	0.657	-0.982	2.273	0.000	Rejected
ACC004	7.695	7.676	4.986	10.791	1.000	Confirmed
ACC005	5.049	5.121	2.868	7.020	0.980	Confirmed
ACC006	-1.364	-1.362	-2.730	0.360	0.000	Rejected
ACC007	1.864	1.876	-1.330	4.471	0.404	Rejected
ACC008	6.825	6.853	3.604	10.225	1.000	Confirmed
ACC009	0.727	0.730	-1.199	2.132	0.000	Rejected
ACC010	3.360	3.445	0.712	6.033	0.798	Confirmed
ACC011	1.175	1.070	-1.178	3.999	0.081	Rejected
ACC012	1.379	1.507	-1.220	3.517	0.111	Rejected
ACC013	-1.723	-2.174	-3.104	0.395	0.000	Rejected
ACC014	1.033	0.872	-1.076	3.301	0.010	Rejected
ACC015	-1.055	-0.720	-3.293	-0.020	0.000	Rejected
ACC016	4.922	4.873	2.723	7.076	0.970	Confirmed
ACC017	1.481	1.510	-1.170	3.682	0.111	Rejected
ACC018	3.131	3.116	0.981	6.099	0.768	Confirmed
ACC019	2.523	2.502	-0.036	5.028	0.626	Confirmed
ACC020	1.862	1.830	-1.052	4.211	0.394	Rejected
ACC021	0.850	0.810	-0.222	2.040	0.010	Rejected
ACC022	0.188	0.368	-1.854	2.523	0.010	Rejected
ACC023	1.212	1.042	-0.995	3.094	0.131	Rejected
ACC024	1.463	1.432	-0.726	3.959	0.182	Rejected
ACC025	3.914	3.852	1.703	6.340	0.909	Confirmed
ACC026	3.161	3.225	0.917	5.728	0.747	Confirmed
ACC027	5.111	5.087	2.345	7.982	0.990	Confirmed
ACC029	0.498	0.210	-1.020	2.853	0.051	Rejected

ACC030	11.178	11.125	8.554	14.149	1.000	Confirmed

Figure 2.5 represents the final decision regarding the attributes. Blue boxplots represent minimal, average and maximum Z-score of a shadow attribute. Red and green boxplots represent Z-scores of rejected and confirmed attributes, respectively.

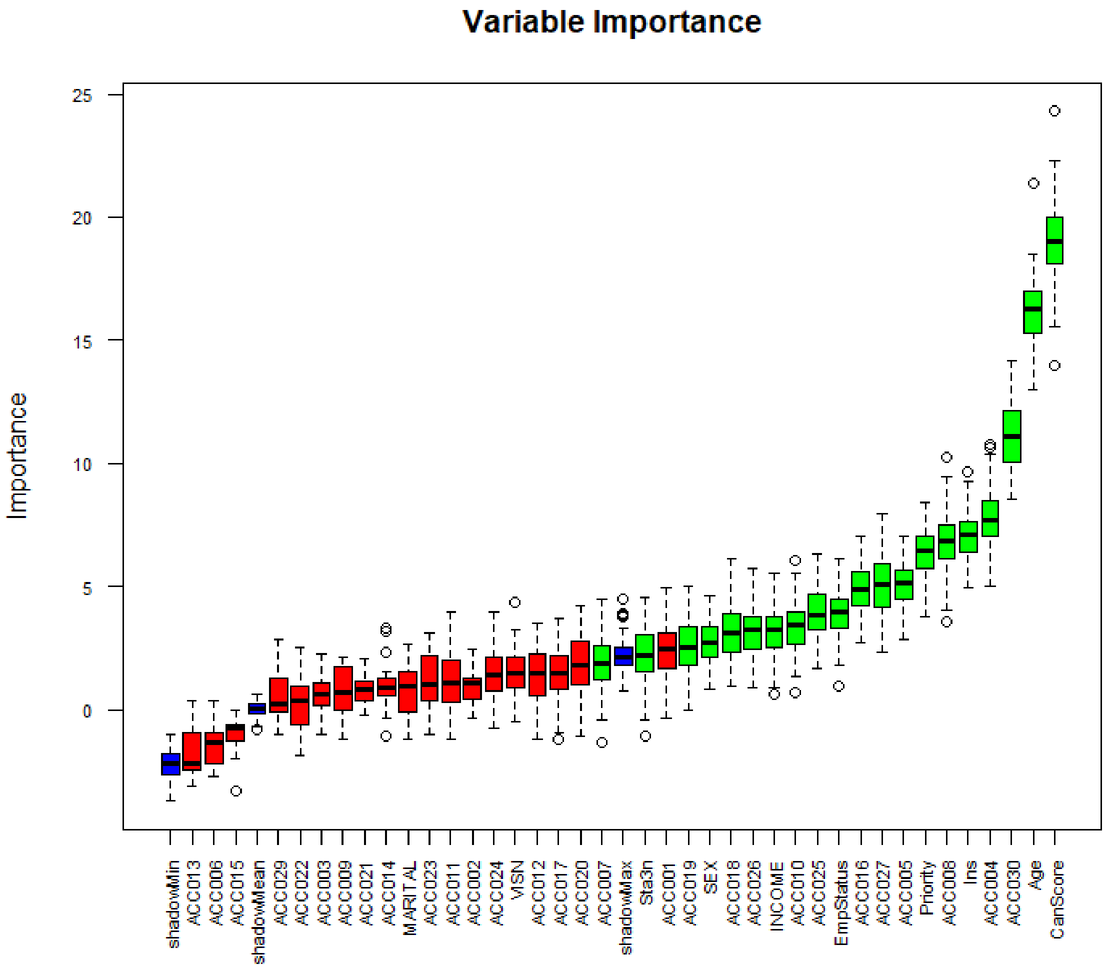

Figure 2.5: Final Decision for Features

In the second stage of unsupervised learning, features are represented using deep learning. We used stacked autoencoders due to the existence of binary variables in the dataset, the ability of the method to reduce the noise, and high computation speed of the method. We used package H2O in R for implementing stacked autoencoders. The proposed approach is applied for deeply stacked autoencoders with different number of nodes in the hidden layer.

The results suggest that using stacked autoencoders for feature representation improves the accuracy of the prediction compared to using the original data or feature reduction. The improvement in the performance happens since by using deep learning, the data is transformed into a data that is more robust. In addition, the results show that changing the number of nodes in the hidden layers leads to different prediction performances. As it is indicated in the Table 2.2, reducing the number of nodes in the hidden layer does not negatively affect the MSE. According to Table 2.2 by reducing the number of nodes, the MSE either remains unchanged or improves for all prediction methods.

The last phase of the framework is related to supervised learning where two different types of single-task and multi-task learning approaches are adapted to predict the RVU of each patient. In this research, we used three well-known machine-learning methods for workload prediction (i.e., Random Forests, Regression Tree and Lasso Regression) along with multi-task learning. In order to compare the performance of different methods, we considered Mean Squared Error (MSE) as a performance metrics for performance evaluation. MSE indicates the squared average deviation of the estimated value. The result of single-task learning is presented in Table 2.2.

For the single-task learning method, we took advantage of regression-based and tree-based methods. We tried to consider both linear and nonlinear predictors to compare the prediction performance. Lasso regression has been widely used in for prediction in high dimensional datasets. Although Lasso performs the variable selection, the results show that it does not improve the performance. Therefore, we conclude that removing features does not necessarily improve the prediction performance. Based on the performance of feature representation by stacked autoencoder discussed before, we conclude that transforming features into a higher-level representation not only reduces the complexity of the problem but also enhances the prediction performance. The advantages of feature representation are generating features that are more robust; also, it transforms the data into a less sparse and less noisy data while conveying the same information as the original dataset. The results show that using random forests as a non-linear predictor leads to better performances compare to Lasso regression.

Furthermore, for every predictive model, the parameters of the model are fine-tuned in order to identify the best parameter that generates the lowest MSE. As an example, the tuning process for finding the best value of λ (the penalty coefficient) that corresponds to the lowest MSE is depicted in Figure 2.6. As it is shown, the best penalty coefficient for the model is equal to 0.0172. Figures 2.7 and 2.8 represent the value of MSE for different numbers of trees and different numbers of predictors at every split of random forests model. In Figure 2.9, the optimal number of splits for the regression tree is found by using the relative error. As it is shown the relative error decreases by increasing the number of splits up to nine splits, after that point the relative error becomes ascending.

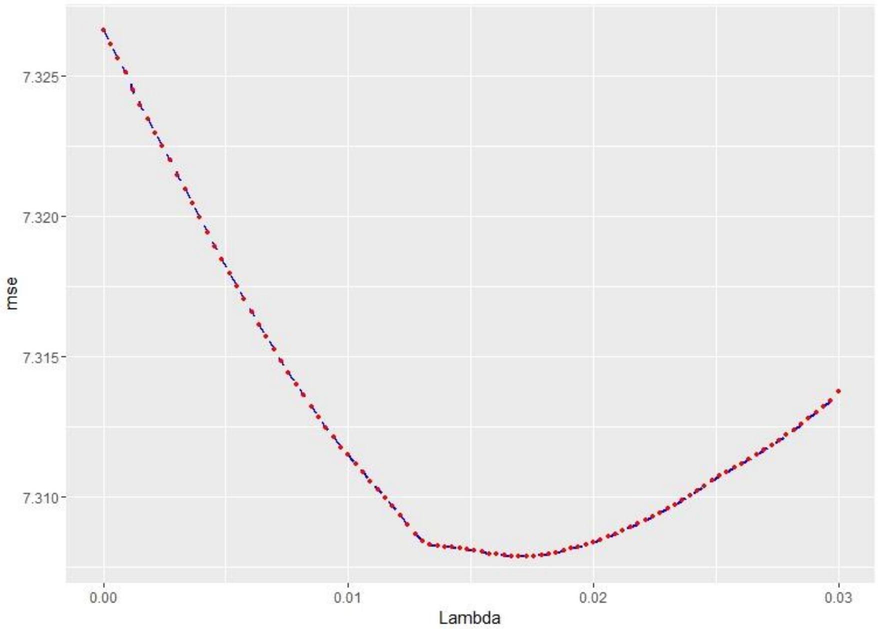

Figure 2.6: Lasso Parameter Tuning for SAE with 10 Nodes in the Hidden Layer

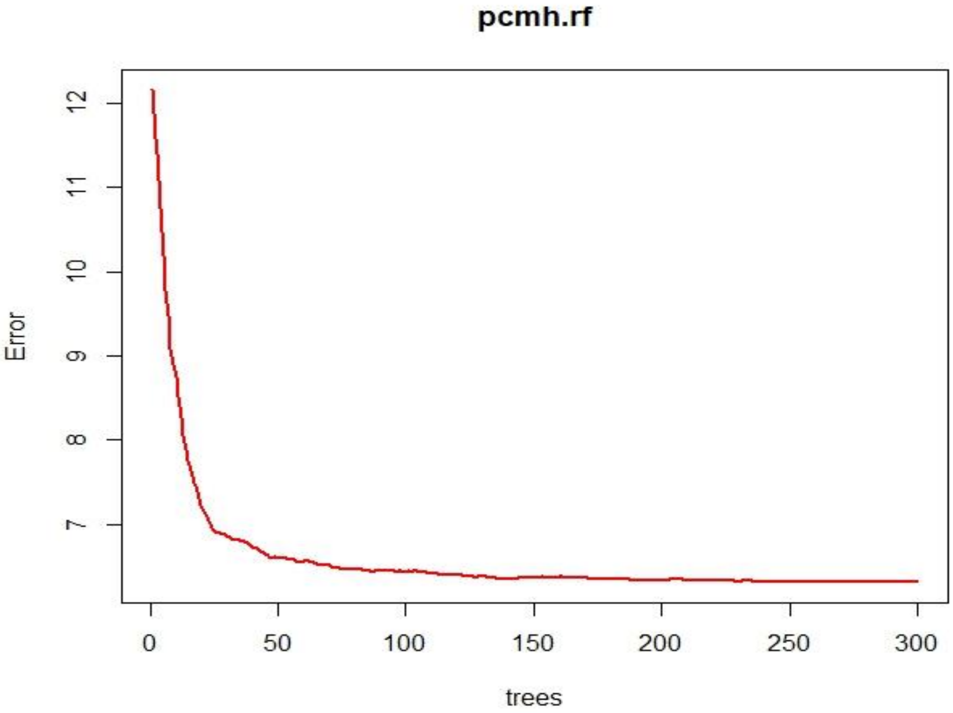

Figure 2.7: Number of Random Forest Trees vs. MSE for SAE with Five Nodes in the Hidden Layer

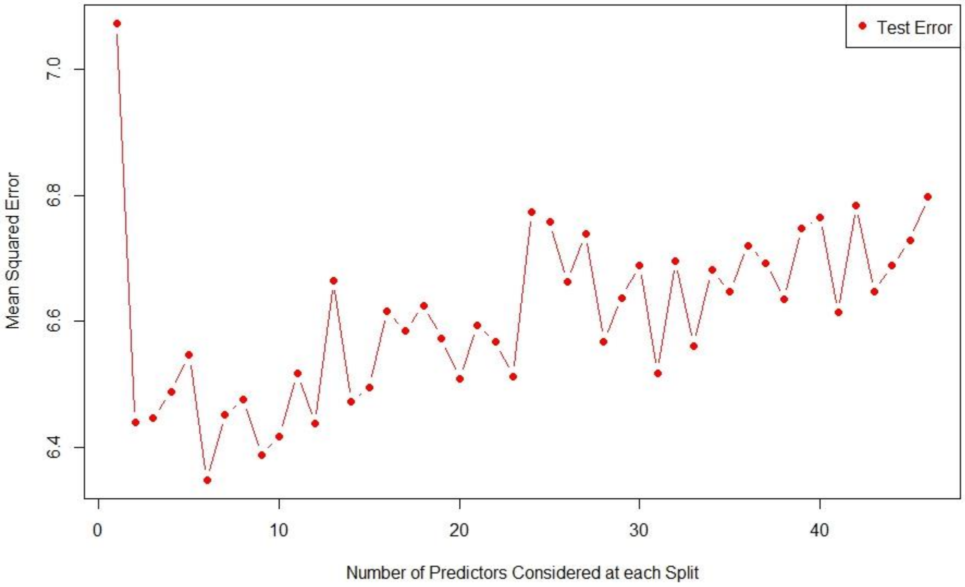

Figure 2.8: Number of Random Forest Predictors vs. MSE for SAE with Five Nodes in the Hidden Layer

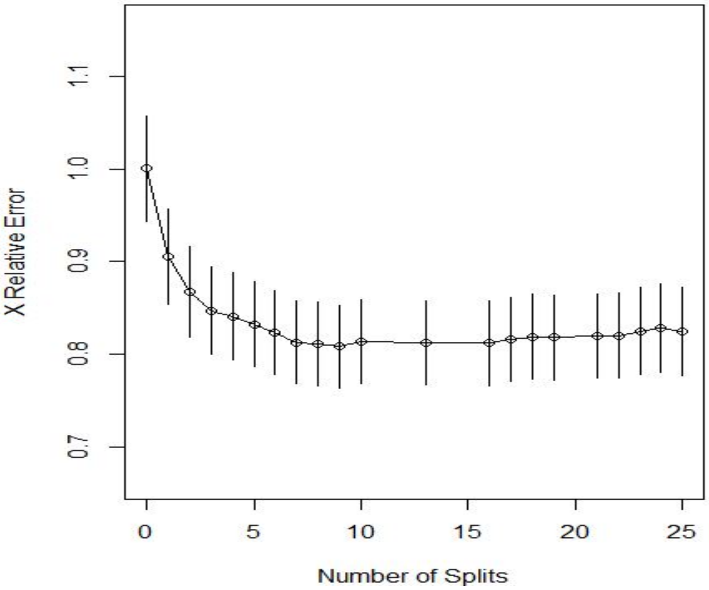

Figure 2.9: Relative Error vs. Number of the Splits for Regression Tree for the Original Data

In the last stage of the supervised learning phase, we implemented multi-task learning. So far, in the approaches mentioned above, the information that is contained between tasks were ignored. However, in multi-task learning, we considered a shared hidden layer for the related tasks. That is why multi-task learning has the best performance among other approaches according to the results of Table 2.2. We performed multi-task learning on the data for predicting the workload of patients with 10-fold cross-validation for the training and testing process [60]. We categorized the data into 130 tasks, which indicate the number of healthcare facilities in the dataset.

Table 2.2: Performance of the Proposed Method (MSE)

	Lasso Regression	Regression Tree	Random Forest	Multi-Task Learning
Original data	7.61	6.74	6.67	4.83
Boruta	7.42	6.61	6.52	4.74
SAE 15 Nodes	7.38	6.52	6.41	4.71
SAE 10 Noes	7.34	6.43	6.38	4.33
SAE 5 Nodes	7.18	6.43	6.35	3.32

Implementing multi-task learning on the represented and unrepresented data shows that this approach outperforms other conventional approaches in predicting RVU. In other words, the prediction accuracy of the proposed approach where stacked autoencoders represent the data with five nodes in the hidden layer, and multi-task learning is used as a predictor is significantly higher than other approaches.

Few reasons can be considered as the reasons that multi-task learning outperforms other approaches. One reason is the high number of binary variables existing in our dataset. The conventional approaches, specifically speaking the random forests approach, tend to bias towards numeric variables that have a greater number of distinct values over binary variables. The other reason is that our data like most EHR data contain both patient-level and facility-level information. Therefore, there is a possibility that many patients have the same attributes but different workloads. The source of this variability can be traced in the facility-level information where the quality and the cost of providing healthcare service, as well as the diagnosis of PCP may differ in different facilities. For example, the RVU for a service with CPT code 99213 (refers to office/other outpatient services) performed in San Francisco is higher compared to the service performed in Detroit. The other reason that multi-task learning outperforms the rest of approaches is that MTL takes advantage of pooling the sample across the related tasks. Therefore, MTL increases the number of samples for each task in the case that the data is not enough. Thus, not only it avoids overfitting and increases the ability to fit random noise by introducing inductive bias (regulating the model), but also helps the model to achieve better representation and generalization compare to single-task learning approaches as the results suggest in Table 2.2.

2.5. Discussion and Conclusion

In this chapter of the dissertation, we developed a novel predictive approach for patient workload prediction that takes advantage of integrating deep feature representation and multi-task learning. The proposed framework consists of three phases of data preprocessing, unsupervised learning and supervised learning. We analyzed the performance of different predictive approaches where the original dataset and the represented data are used.

There are many methods for feature engineering. The traditional approach is the supervised approach where subject matter experts use their knowledge for choosing important variables. This approach is not efficient due to the required amount of engineering skills and work, especially when the data is high dimensional. Another approach is to use unsupervised feature learning methods that do not need explicit labels, such as deep learning where features are represented automatically by using a general-purpose learning procedure. Deep learning approaches have a multilayer structure that consists of stacks of simple modules. By taking advantage of multiple modules that transform their inputs, the system can distinguish between irrelevant and relevant feature more precisely. As it is evident in the result of this study, this characteristic of deep learning feature representation leads to an enhanced selectivity and invariance of the representation.

We considered three data types in this research. For the first type, we used the original pre-processed data where all the features are considered in each predictive model. Applying Boruta feature selection method on the original data generates the second type of data. Finally, we transformed the original data by using deep feature representation as the third type of input data. In addition, we considered three different number of nodes in the hidden layers of the represented data. We considered 15, 10, and 5 nodes in the hidden layers for each represented dataset. The results underline that the choice of the number of nodes in the SAE hidden layer affects the prediction accuracy.

After preparing various data types as inputs for the RVU prediction models, we applied two types of predictive approaches (i.e., single-task learning and multi-task learning) to predict the RVU. We utilized three well-known single-task learning approaches consisting of Lasso regression, random forests and regression tree to predict RVU. The parameters of the predictive models are all fine-tuned, and the parameters corresponding to the lowest MSE are chosen. The tuning process ensures that the model is at its best accuracy level. The results suggest that using the appropriate represented data improves the performance of the predictive models. On top of that, it suggests that using multi-task learning instead of single-task learning for predicting the RVU enhances the prediction accuracy for all types of input data. The results also show that the integration of SAE with five nodes in the hidden layer with multi-task learning outperforms the rest of the combinations.

We believe that to the best of our knowledge, this study is the first study in patient workload prediction that provides a systematic framework for RVU prediction and uses the RVU as a quantitative measure of workload in team-based care. In addition, the proposed deep multi-task approach is the first attempt to use deep feature representation and multi-task learning for patient workload prediction with multiple chronic diseases in team-based care systems. In this study, we considered patient-level and facility-level information as well as the hidden information that is shared between different tasks. This method can be applied to other healthcare problems as well as any prediction problem in other industries such as supply chain, transportation, and manufacturing where the number of binary variables is high, or the training samples are limited.

The future steps of this research can be stated as using other types of deep learning approaches and applying the proposed approach on more datasets of various types to make the prediction model more robust. On the management point of view, this approach can be a basis for resource allocation and optimization for chronic disease operations management. The more accurate workload prediction contributes to enhancing the quality of the decision-making process. Healthcare decision makers can use the result of this model as an input for resource planning in their systems. In addition, by using the results of the proposed approach, the patients can be assigned to proper healthcare providers. This leads to an increased patient satisfaction and a balanced workload of healthcare providers.

CHAPTER 3 PRESCRIPTIVE ANALYTICS: STOCHASTIC OPTIMIZATION MODELS

3.1. Problem Statement

With a size of $2.9 trillion spend in 2015, and a rapid expected employment growth rate of 18 percent for years 2016 to 2026, the healthcare industry accounts for the largest sector of the economy in the United States (US) [61]. Despite advances in medical technology and, thereby, the increasing use of medical diagnostic, monitoring, and treatment equipment, the health care industry is highly labor-intensive. According to the US Department of Labor, the health care industry provided 13.5 million jobs in 2004, out of which 13.1 million jobs are for wage and salary workers and about four hundred eleven thousand are for the self-employed workers [62]. It follows that human resource wages and salaries account for a substantial part of the total expenditures for any health care facility. For instance, hospitals spend on average about 54 percent of all expenditures on wages and salaries. Hence, health care personnel planning, i.e., determining the proper mix of health care personnel needed to provide safe, effective, timely, and cost-efficient services to patients is an important problem.

Chronic diseases are among the most common and costly health conditions in the United States. Almost half of Americans suffer from at least one chronic condition, and unfortunately, despite all the efforts, the number is growing. There are various definitions for chronic diseases, but most of the researchers consider cancer, diabetes, hypertension, stroke, heart disease, respiratory diseases, arthritis, obesity, and oral diseases as major chronic diseases that are also the nation’s leading causes of death and disability. These diseases can lead to hospitalization, long-term disability, reduced quality of life and death [63].

In order to prevent and control chronic diseases and their risk factors, the National Center for Chronic Disease Prevention and Health Promotion's (NCCDPHP) budget for the fiscal year 2016 is $1.17 billion. According to Center for Chronic Disease Prevention (CDC), a 13% reduction in the number of people with uncontrolled hypertension (about 4.7 million people) would save the health care system $25.3 billion per year in averted disease costs.

The mentioned facts and information show us that there is a critical need to reduce the cost of chronic disease operations, and to improve the quality of health services. In fact, reducing the chronic disease operations management cost leads to a reduced cost of healthcare delivery and an improved efficiency of the system so more people can benefit from quality healthcare.

Healthcare decision makers can benefit from the integration of multidisciplinary teams and clinical information decision support systems in order to manage chronic disease operations efficiently. Team-based care is recognized as a model for transforming the structure and delivery of primary care in the US that provides high quality, accessible and efficient healthcare in the US. Team-based care consists of teams of healthcare professionals including primary care provider, and the care team, which consists of specialist, nurse, nutritionist, pharmacist, social worker, and other professionals working together toward improving patient care. These entities are in collaboration with each other and use shared patient records to focus on various patient needs and to deliver continuous, coordinated, accessible, safe and high-quality care healthcare services to patients through the healthcare system.

Designing an efficient mechanism for establishing a trade-off between supply and demand in team-based care model is a challenging task. The reason is that one must consider the experience and expertise of staffs, as well as different combinations and numbers of providers in order to have a well-balanced workload. In addition, the workload of patient is a random value, which depends on demographic, diagnostic, and health conditions of patients. To design the model for allocating patients to teams, one should consider spreading the workload on the teams, and utilizing the care providers evenly. In this way, the healthcare service can be delivered in a timely manner by minimizing provider's idle time and the overtime generated by excessive patient's demand.

Demand uncertainty is an integral part of stochastic programming [67], [68]. Thus, a good demand prediction that truly reflects the actual workload required for patients can increase the accuracy of the model, and consequently results in a better patient panel workload planning and assignment. Let us explain the necessity of having a good workload prediction and efficient resource allocation with an example. Consider two healthcare providers. The first provider has a certain number of young and healthy patients in his/her patient panel. On the other hand, the second provider has the same number of patients who are older and have multiple chronic disease. One can conclude empirically that the first provider is underused, while the other one experiences excessive workload. This issue causes an increase in the patient waiting time and may force patients to switch their PCP. But the main challenge in patient assignment is that the actual required workload of a patient is not typically known at the time of assignment. Hence, in order to design patient allocation procedures efficiently, and to minimize the hiring cost and total workload of providers at a given time, some adjustments such as paying overtime are needed to compensate provider overloading [65], [66].

The primary mission in any industry is to cater to the needs of the customer while taking the resources efficiency into the account [64]. For developing a comprehensive model for team-based care, we study two important problems for chronic disease operations management in this dissertation. The scope of the first problem is strategical planning level decision-making where the demand of patients is unknown. The main objective of the first problem is to determine the number of providers for each facility in a multi-facility healthcare system. In order to minimize the hiring cost, we offer some incentives to eligible patients for traveling between facilities in this problem. The main scope of the second problem is in tactical planning. The objective of the second problem is to minimize the number of teams while balancing their workload for multiple facilities when demand is uncertain.

3.2. Literature Review

3.2.1. Resource Planning in Team-Based Care

Team-based healthcare delivery model is introduced as an important enabler of U.S. healthcare transformation [69], [70]. Some principles such as coordinated care, person-oriented system, physician-directed medical practice, continuity of care, having a personal physician, emphasis on quality, safety, and a proper payment mechanism, which are complement one another in ideal situations build the team-based care systems. These principles remained abstract [71], and their implementations in everyday practice need to be investigated [72].

In this study, we used a model for predicting clinical workload, which considers different features such as different types of disease, and distance between the location of patient and the assigned facility in mathematical modeling. Many factors such as patient’s age, gender, and diagnostic codes can affect the workload. Østbye et al. [15] suggested that patient’s visit frequency is affected by the type of chronic disease. Also, Naessens et al. [16] showed that the number of chronic conditions of a patient significantly affects the clinical workload and medical cost. By combining the demand of patients with different characteristics and predicted workload, it becomes more likely that a high demand from one patient balances out by a low demand from another in the aggregated patient panel.

A good approach to tackle problems of this type can be using a two-stage stochastic program. Stochastic programming is widely used in various settings when demand is unknown. Recently, researchers focused on staffing and workforce planning in healthcare using different stochastic models. Kao and Queyranne [73] suggested a two-stage stochastic program for budgeting the nurses cost where it determines nurses working hours, a the first step and then it determines their overtime in the second step. Some surveys of two-stage stochastic integer programming with mixed-integer recourse are presented in [74], [75]. A two-stage stochastic integer programming model for assigning nurses to the patient is proposed in [76], where in the first stage assigns each nurse to patients then balances the nurse workload in the second stage. Zhu and Sherali [77] considered a continuous workload variable for every worker at the second-stage decision of their two-stage stochastic workforce planning model. Bodur, Merve, and Luedtke [78] used two-stage stochastic programming with continuous variables in the second stage to propose an integrated staffing and scheduling model for service system. However, they did not consider adjustment decisions as a recourse to different demands.

Staffing decisions can be integrated into stochastic models by defining binary variables [79]. Many researchers benefited from L-shaped method [76], [79] which is based on Benders' decomposition [80]. Benders' decomposition method is capable of solving large-scale problems in a reasonable time [81]. Laporte and Louveaux [82] proposed an integer L-shaped method where integer variable can be used in both first and second stage to ensure the optimality by branch and bound method. It also creates a finite number of subspaces to ensure finiteness of method [83]. They also proposed a multi-cut approach in addition to single cut approach. The multi-cut approach may decrease the number of iterations significantly by keeping the second stage cut information separately. Another advantage of the multi-cut approach is the possibility of scenario aggregation to reduce the number of cuts [84].

One of the recent works on staffing and scheduling problems under demand uncertainty is conducted by Kim and Mehrotra [85]. They considered two decision stages, that is, initial staffing and initial schedule adjustment, based on the demand realization. They formulated the problem as a two-stage stochastic integer program with mixed-integer recourse. They solved the problem by a modified multi-cut approach in L-shaped method and achieved improvements in computational efficiency of the algorithm.

Although, designing patient panel in team-based care is a new topic, there are some important works on workload planning and allocation in terms of task assignment and scheduling which are sub-problems of workforce planning. The framework of workload planning and staff rostering can be found in [86]. Also, some extensive literature review and surveys on workforce planning methods and models in healthcare can be found in [87], [88]. Operations research and resource allocation models with several applications inside and outside of healthcare scope have a broad literature (see [79], [89]–[97]). For instance, Cardoen et al. [98] conducted an extensive literature review on operation room planning and scheduling. One of the researches on patient allocation in operation rooms is studied in [99]. The authors developed a non-linear stochastic programming model for allocating surgical specialties to operating rooms and minimize the total expected cost by considering a penalty in the model for any patient who is not allocated to a provider, and over and under usage of the operating room. Jebali and Diabat [100] used a two-stage stochastic program to solve the operation room planning problem considering surgery uncertainties and hospital capacity constraints. There are other studies on different aspects of staff allocation. As an example, studies about workforce allocation with respect to the effects of cross training are conducted in [101], [102]. In order to allocate and schedule cross-trained workers in a multi-department setting where the demand is uncertain, Campbell [103] suggested a two-stage stochastic program. Chalabi, Epstein, McKenna, and Claxton, [104] conducted a study to allocate resources in the presence of budget constraints and uncertain healthcare variables using two-stage stochastic programming approach. Liang and Turkcan [105] addressed the problem of nurse assignment in oncology clinics. They developed a multi-objective optimization model to minimize patient waiting time as well as nurse overtime.

Lanzarone and Matta [106] studied reference nurse assignment to each patient concerning continuity of care concept in-home care context taking into account the uncertainty in both new and assigned patients. They only considered one nurse at the time of the assignment. However, in this study, we considered the problem of assigning a team of care providers to every patient. Villarreal and Keskinocak [107] presented a model for nurse and surgical staff planning in which staff can be assigned to one service line and switched to another one while considering the forecasted demand. The study tried to find the number of employees and number of staffs assigned to each service line while considering overtime and penalizing deviations from assignments by using a two-stage model.

Task assignment and allocation to teams can be a challenging problem [108]. Balasubramanian, et al. [109] developed a stochastic dynamic program to allocate patients with same-day appointments to physicians in a multi-physician environment with uncertain demand. They also assumed a certain duration for their appointment slots. Zhen [110] suggested splitting tasks into deterministic and uncertain parts considering their workload to assign tasks to teams. Unlike the proposed method by Zhen [110] on task scheduling, which split the task, we proposed a direct method to make the assignments. However, we focused on patient panel assignment in team-based care with respect to healthcare demand of patient, which depends on the patient’s condition. Finding a balance between supply and demand of care service is the key factor in health delivery. Although in most of the healthcare environment supply is deterministic and is calculated by using the number of staff and working hours, finding the demand and patient assignment are challenging tasks, specifically in team-based care where a team of different healthcare providers is assigned to a patient. We believe that there is no literature on allocating healthcare providers to patients in team-based care setting when demand is unknown. In addition, estimating the clinical workload portfolio based on key features of patient and provider and then stochastic task assignment in a team-based health delivery system would be a knowledge contribution to researches on team-based care environment data and operations analytics.

3.2.2. Healthcare Capacity Planning

It has been a decade that researchers study health care capacity planning to address strategic planning, medium-term staffing and short-term scheduling decisions. Since one of the major parts of healthcare system costs is associated with staffing cost, among the conducted researches in the field of capacity planning for healthcare systems, most of the researches are focused on short-term planning, which includes personnel planning problems such as staffing, provider scheduling and nurse rostering problems [73]. State of the art in nurse staffing can be found in a review conducted by Burke et al. [87].

Optimization methods are popular among researchers to address staffing problems. Many of them benefit from operations research techniques and mathematical modeling to address personnel planning problems [111]. There is a wide range of using operations research techniques to model the capacity-planning problem in healthcare systems. These techniques includes using linear programming for resource planning [112], resource shortage modeling [113], using simulation-based optimization modeling [114], [115] and using stochastic multi-objective models as developed by Abdelaziz and Masmoudi [116] in order to determine the number of beds that need to be assigned to hospital departments for satisfying the random demand. One of the studies that investigated healthcare capacity planning in the tactical planning level was conducted by Dellaert et al. [117]. They considered the creation of tactical planning for patient surgeries and the resource utilization of healthcare facilities in order to increase hospital efficiency. They determined the number of patients in each category that need surgery on a daily basis.

Most of the works reviewed above, focus on minimizing the cost of healthcare resources or maximizing the provider utilization, however there is a need to address other different problems from the patient's perspective such as minimizing the patient waiting time and time-span in a healthcare system, as well as maximizing the availability of the resources in the system. Queuing theory is a method that is useful in order to address the mentioned problems. Fomundam et, al. conducted an extensive survey of queuing theory applications in healthcare [118]. Keller and Laughhunn focused on a different perspective in the queuing theory applications in healthcare. They minimized the cost of the required capacity in healthcare facility via minimizing the costs of a healthcare queuing system while considering the capacity of the servers working in the system [119].

Foregoing is another scope that researchers focused on in the literature. Foregoing is defined as the situation that patients decide to leave a system since they do not wish to wait any longer in the queue. Most of the researchers assumed customer arrival rate as a constant value. For instance, Fiems et al. studied the association between emergency requests on the waiting time of scheduled patients with deterministic processing times [120]. However, in many healthcare systems the arrival rate is variable rather than a deterministic value. Also, it is known that increasing the capacity of service cannot significantly affect the queue length since the arrival rate increases when patients realize that the service time is reduced [121]. Broyles and Cochran [122] calculated the revenue loss results from the patients who leave an emergency department without getting help by using the arrival rate, service rate, utilization, and capacity.

Limiting the queue length results in blocking in a queuing system. Koizumi et al. [123] determined that blocking in a chain of extended care, residential and assisted housing facilities leads to an increase in the time that patients spend in facilities. In addition, they investigated the effect of increasing the capacity in downstream facilities on the queue length and waiting time of patients. Every healthcare system has a queuing discipline. Most of the healthcare systems define patient priorities in order to prioritize the patients in the queue. This process can be defined based on the first-in-first-out system or priority group classifications. Generally, patients who are classified in a low-priority group do not receive the healthcare service until the high-priority patients receive the service.

The effect of using the emergency department by primary care patients on patient waiting times was analyzed by Siddhartan et al., [124]. In their study, they proposed a priority discipline for different patient categories and then the first-in-first-out discipline for each category. By analyzing the relationship between the composition of prioritized queues and the number of nurses responding to inpatient demands, Haussmann [125] realized that a slight increase in the number of patients assigned to a patient mix with more high priority demands results in a significant increase in waiting time of low priority patients. Taylor et al. [126] investigated the probability that a patient would have to wait more than a certain amount of time to receive the healthcare service by modeling an emergency department operating with priority queuing discipline. Fiems et al. [120] studied the relationship between emergency requests and the waiting time of scheduled patients with deterministic processing times.

3.3. Methodology and Problem Formulation

As it is indicated in Figure 1.1, this research includes two major planning phases of predictive and prescriptive analysis. In this chapter, we focus on the second part of the dissertation that is prescriptive analytics. In this study, we discuss two optimization models in the strategical and tactical levels of management, namely Strategic Chronic Disease Decision Optimization (SCDDO) and Tactical Chronic Disease Decision Optimization (TCDDO). In the SCDDO model, we propose a model and solution for incentive-based capacity management with patient transportation for multiple facilities under uncertainty. This problem is formulated as a two-stage stochastic optimization problem with recourse. In TCDDO, we propose an integrated team-based workforce and workload stochastic optimization model for each facility when the workload is unknown. For solving the problems mentioned above, we made some assumptions that are discussed in the next section.

3.3.1. Assumptions

The assumptions that are used in order to model SCDDO and TCDDO problems are listed in this section.

Assumption 1: The decision-making process happens in two stages.

In SCDDO problem, the goal of the first stage is to determine the number of providers in order to minimize the human resource cost well ahead in time. Then in the second stage, the solution determines the assignment of patients to facilities while minimizing the expected service coverage cost. We follow the same decision-making procedure for TCDDO problem too. In the first stage, the model determines the number of each team well ahead in time. Then in the second stage, we adjust our decision by taking recourse actions and minimizing the expected cost of overtime by assigning patients to teams.

Assumption 2: We consider the capacity of the providers as a constant value.

We use the maximum capacity of each provider and assume that providers work on their full capacity. In addition, we assume that the capacity of providers is independent of their performance, quality of the provided care, and the required workload of the patient. We simply calculate the total available capacity of teams based on the headcount and the duration of the provider availability in a certain planning period.

Assumption 3: Required workload of the patient is not known.

We define the required workload of the patient as an unknown variable that contains the stochastic required workload of the patient from providers in chronic disease operations management model. We measure the required workload in terms of relative value unit, which accounts for the time, technical and mental effort as well as judgment, and stress for providing specific healthcare service to the patient. Therefore, to estimate the required workload of the patient, we use patient chronic disease information and MTL approach.

Assumption 4: The required workload of each patient is correlated with the patient attributes.

Assumption 5: The assignment proportion of the required workload for each provider is known. Therefore, the quota of each provider from total RVU is an input of the model.

Assumption 6: The efficiency of similar providers is not different. Thus, identical providers provide identical services in terms of quality and efficiency.

Assumption 7: There is no coordination among identical providers.

We assume that the patient-provider assignment is one to one. Therefore, identical providers cannot split their tasks to provide a certain service to a patient.

Assumption 8: Pay-for-travel service can be offered to patients only after a certain distance threshold.

In order to provide a high-quality service and maintain continuity of care, we design the SCDDO model to be able to assign patients to the nearest facility that has all the required resources available. In addition, we determine a service coverage threshold in the proposed model. In case the model fails to assign patients to the nearest facility, we offer transportation services to the patient to a proper facility at no cost for the patient. However, we penalize the model by imposing service coverage cost that is calculated based on the distance of patients from facilities.

Assumption 9: We incur a cost when care providers are required to work more than their available capacity.

Assumption 10: The composition of teams in TCDDO model is known. Thus, the decision makers know the number of providers from each type for a certain team.

3.3.2. SCDDO Problem Statement

In the first phase of prescriptive analytics for chronic disease operations management, an incentive-based capacity optimization problem with patient transportation under uncertainty is investigated for multiple facilities in the strategical planning level. We consider the distance of patients from the facilities to formulate the problem by a two-stage problem with recourse where the patient’s demand and location are uncertain.

In two-stage stochastic programing, decisions are made in two stages, and the decision in each stage is made based on the available information at the time of the decision. At the first stage of the problem, the decision maker must decide before the realization of the uncertain workload and location of the patient. In the second stage when the realization of the patient’s random attributes becomes available, the second stage recourse decision is made. We consider the second stage decision as an optimization problem describing the optimal behavior where the uncertain data are realized or as a recourse action. Since the recourse actions are costlier than the first stage decisions, the goal is to minimize the sum of the first stage decision costs and the expected cost of the second stage decisions where problem instances are random with a known probability distribution.

The main objective of this SCDDO model is to minimize the cost of hiring healthcare providers in the facilities while taking the workload demand and distance of the patients into account. We formulate the stochastic optimization model for this problem as below. To solve this problem, we use the sample average approximation method.

Indices

$\rho \in$ P: index for patients

$l \in L$: index for facility location

$h \ \in H$: index for healthcare providers

$\omega \in \Omega$: index for scenarios, Ω denotes the sample space

Parameters

c_h : fixed cost of hiring healthcare provider h

ψ_h: capacity of each provider

$\bar{d}$: service coverage upper-bound

T: coverage violation cost

$$b_{hl} = \begin{cases} 1 & \text{if provider } h\ can\ be\ hired\ in\ facility\ l \\ 0 & \qquad\qquad\qquad otherwise \end{cases}$$

Scenario Dependent Parameters

$D^{\omega}{}_{\rho h}$: required workload of patient ρ from provider h under scenario ω

$d^{\omega}{}_{\rho l}$: the distance of patient ρ from facility *l* under scenario ω

$$R_{\rho l}^{\omega} = \begin{cases} d^{\omega}{}_{\rho l} \times T, & if\ d^{\omega}{}_{\rho l} \geq \ \bar{d} \\ \quad 0 & otherwise \end{cases} \quad \textit{the service coverage cost of patient p to facility l}$$

Decision Variables

x_{hl}: number of providers of type h in facility *l*

$y_{\rho l}$: binary variable representing if patient ρ is assigned to facility *l*

$$\min \sum_{h} \sum_{l} c_h x_{hl} + \mathbb{E}_p[Q\ (Y,\omega)] \tag{3.1}$$

s.t. (3.2)

$$Q\ (Y,\omega) = \sum_{\rho\in\rho^{\omega}} \sum_{l} R^{\omega}_{\rho l} y_{\rho l}$$

$$\sum_{\rho\in\rho^{\omega}} D^{\omega}{}_{\rho h} y_{\rho l} \leq x_{hl}\, \psi_h \ , \quad \forall h,l \tag{3.3}$$

$$x_{hl} \leq M.b_{hl} \quad , \ \forall h,l \tag{3.4}$$

$$\sum_{l} y_{\rho l} = 1, \qquad \forall \rho \in \rho^{\omega} \tag{3.5}$$

$$y_{\rho l} \in \{0,1\} \ , \ \forall \rho \in \rho^{\omega}, l \tag{3.6}$$

$$x_{hl} \in \mathbb{Z}^{+}, \qquad \forall h,l \tag{3.7}$$

In the above formulation, x represents the first stage decision that includes determining the number of providers of each type for every facility and y denotes the second stage decisions. Also, ω corresponds to uncertain random data with known distribution. The symbol $\mathbb{E}$ denotes mathematical expectation. In this problem, $Q\ (Y,\omega)$ corresponds to the total service cost of each patient. Thus, the objective function is to minimize the sum of hiring cost of providers and the expected patient service cost as it is indicated in (3.1) and (3.2). Constraint (3.3) ensures that the total demand of patients from providers cannot exceed the total available capacity of the providers in each facility. Constraint (3.4) determines whether it is possible to hire a certain healthcare provider in a certain facility. In addition, in the case that the provider can be assigned to the facility, it determines the number of assigned providers to each facility. Constraint (3.5) shows that every patient is only assigned to one facility. Constraints (3.6) indicates that the variable $y_{\rho l}$ is a binary variable. In constraint (3.7), we show that the variable x_{hl} belongs to non-negative integer set.

3.3.3. TCDDO Problem Statement

In the second phase of the prescriptive analytics part of the proposed framework, we discuss the Integrated Team-based Workforce and Workload stochastic optimization (ITWWSO) model in order to improve the quality of the decisions for chronic disease operations management in the tactical level. Integrated team-based workforce and workload stochastic optimization problem is formulated as a large-scale two-stage stochastic optimization model. The objective is to minimize the overall number of healthcare teams plus the expected overloading cost of healthcare providers in each team while balancing the workload of the teams. As discussed before, we use the estimated workload for each disease category generated by the deep multi-task learning predictive model as an input for developing different workload scenarios in stochastic capacity planning model. Afterward, the TCDDO model determines the number of different team types while assigning patients to certain teams based on the patient required workload. We model this problem as a MIP problem. To solve this problem, we use the sample average approximation method.

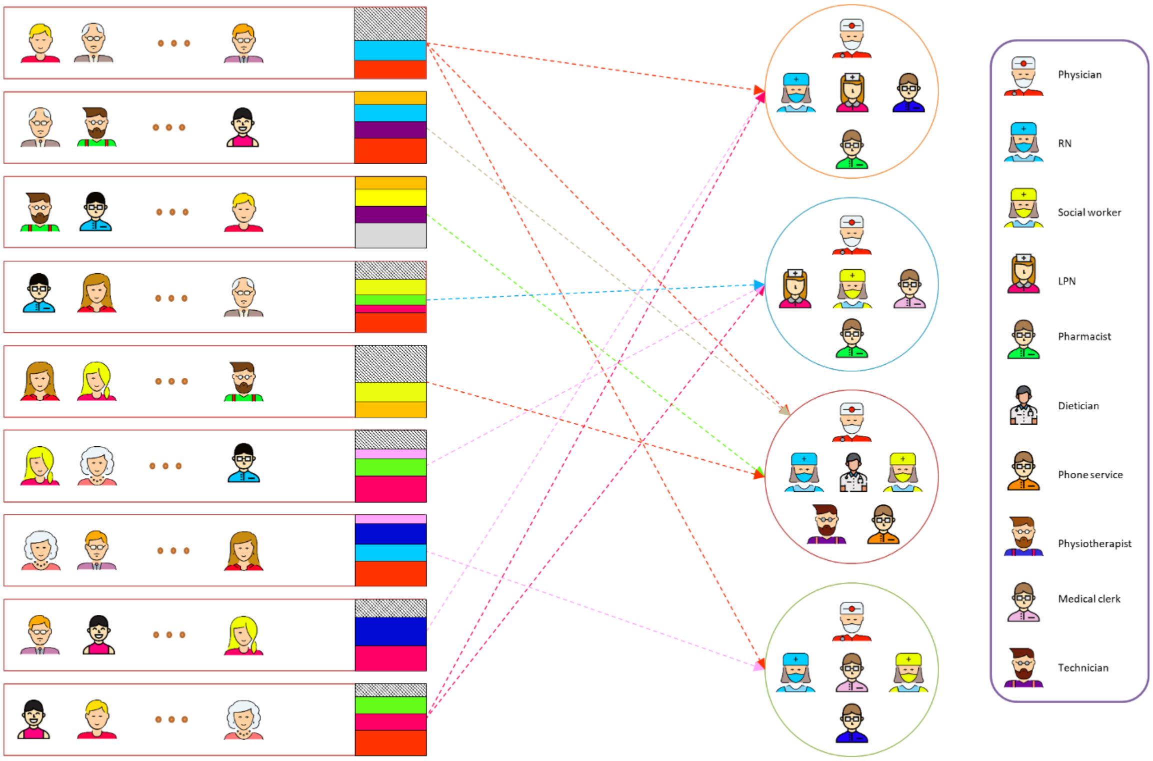

Figure 3.1: Integrated Team-based Workforce and Workload stochastic optimization Architecture

The model consists of three essential elements: chronic disease attributes of the patient, workload portfolio of the patient and various providers as it is shown in Figure 3.1. The model makes sure that patients are categorized based on their chronic condition type and are assigned to only one team that can provide all the necessary services to satisfy the workload while balancing the workload of providers.

Our proposed model aims to find the optimal number of available team types such that each patient is assigned exactly to one team, subject to resource constraints limiting the workload capacity of providers. We model the problem as a two-stage stochastic program with mixed 0-1 recourse. The first stage decisions involve minimizing the number of teams and their associated cost. In the second stage of the stochastic optimization problem, after getting closer to the actual workload realizations, the model minimizes the expected cost of overtime along with the difference between the workload of each team and the average workload of the teams in order to balance the total workload of providers by reassigning patients to the underutilized teams. Therefore, it assists the decision makers in estimating the required resources for the healthcare system while considering the patient assignment. This provides the decision makers with an efficient patient assignment procedure. We define α as a weight factor that determines the importance of each objective in the second stage of decision-making. We designed the mathematical formulation of TCDDO problem as follows.

Indices

$\rho \in$ P: index for patients

$\tau \in$ T: index for the team among the set of teams

$h \ \in H$: index for healthcare provider

$\omega \in \Omega$: index for scenarios, Ω denotes the sample space

Model Parameters

c_τ : fixed cost of team τ

ψ_h: capacity of each provider h

ξ_h: cost of overtime for each provider

υ_h: amount of overtime upper bound for each provider

$$a_{\rho\tau} = \begin{cases} 1 & \text{if patient } \rho \ can\ be\ assigned\ to\ team\ of\ type\ \tau \\ 0 & otherwise \end{cases}$$

W_τ: nominal standard capacity of a team of type τ

α_1: weight of overload

α_2: weight of any deviation from the available capacity of the team

Scenario Dependent Model Parameters

${D^{\omega}}_{\rho h}$: demand of patient ρ from provider h under (demand) scenario ω

Decision Variables

x_τ: number of teams of type τ

$y_{\rho\tau}$: binary variable representing if patient ρ is assigned to team τ

$\varphi_{h\tau}$: the amount of overload for each provider in each team τ

$$\min \sum_{\tau} c_\tau x_\tau + \mathbb{E}_p[Q\ (Y,\omega)] \tag{3.8}$$

Where

$$Q\ (Y,\omega) = \alpha_1 \sum_{\tau,h} \xi_h\, \varphi_{h\tau} + \alpha_2 \sum_{\tau} \left| \sum_{h} \sum_{\rho\in\rho^\omega} {D^{\omega}}_{\rho h} y_{\rho\tau} - x_\tau W_\tau \right| \tag{3.9}$$

$$\sum_{\rho\in\rho^\omega} {D^{\omega}}_{\rho h} y_{\rho\tau} \leq\ x_\tau \psi_h + \varphi_{h\tau}\ ,\ \ \forall\ h,\tau,\omega \tag{3.10}$$

$$\varphi_{h\tau} \leq M.x_\tau\ \ ,\forall h,\tau \tag{3.11}$$

$$\varphi_{h\tau} \leq\ \upsilon_{h\tau}\, ,\ \ \forall h,\tau \tag{3.12}$$

$$y_{\rho\tau} \leq\ a_{\rho\tau}, \qquad \forall\ \tau,\rho \in \rho^\omega \tag{3.13}$$

$$\sum_{\tau} Y_{\rho\tau} = 1, \qquad \forall \rho \in \rho^\omega \tag{3.14}$$

$$y_{\rho\tau} \in \{0,1\}, \forall \rho \in \rho^\omega, \tau \tag{3.15}$$

$$x_\tau \in \mathbb{Z}^+, \qquad \forall \tau \tag{3.16}$$

$$\varphi_h \geq 0, \qquad \forall h \tag{3.17}$$

In order to convert the TCDDO minimization model to a linear stochastic optimization model, we transform the expression in the absolute value symbol into two constraints (3.20) and (3.21).

$$\min \sum_{\tau} c_\tau x_\tau + \mathbb{E}_p[Q\ (Y, \omega)] \tag{3.18}$$

Where

$$Q\ (Y, \omega) = \alpha_1 \sum_{\tau,h} \xi_h\ \varphi_{h\tau} + \alpha_2 \sum_{\tau} \sigma_\tau \tag{3.19}$$

$$\sum_{h} \sum_{\rho \in \rho^\omega} D^{\omega}{}_{\rho h} y_{\rho\tau} - x_\tau W_\tau \leq\ \sigma_\tau, \qquad \forall\ \tau \tag{3.20}$$

$$x_\tau W_\tau - \sum_{h} \sum_{\rho \in \rho^\omega} D^{\omega}{}_{\rho h} y_{\rho\tau} \leq\ \sigma_\tau, \qquad \forall\ \tau \tag{3.21}$$

$$\sum_{\rho \in \rho^\omega} D^{\omega}{}_{\rho h} y_{\rho\tau} \leq\ x_\tau \psi_{h\tau} + \varphi_{h\tau}\ , \qquad \forall\ h, \tau, \omega \tag{3.22}$$

$$\varphi_{h\tau} \leq M.x_\tau, \qquad \forall h, \tau \tag{3.23}$$

$$\varphi_{h\tau} \leq\ \upsilon_{h\tau}\,, \qquad \forall h, \tau \tag{3.24}$$

$$y_{\rho\tau} \leq\ a_{\rho\tau}, \qquad \forall\ \tau, \rho \in \rho^\omega \tag{3.25}$$

$$\sum_{\tau} Y_{\rho\tau} = 1, \qquad \forall \rho \in \rho \tag{3.26}$$

$$y_{\rho\tau} \in \{0,1\}, \qquad \forall \rho \in \rho^\omega, \tau \tag{3.27}$$

$$x_\tau \in \mathbb{Z}^+, \qquad \forall \tau \tag{3.28}$$

$$\varphi_h \geq 0, \qquad \forall h \tag{3.29}$$

The TCDDO model is transformed into a standard form of the stochastic optimization problem. As is indicated in TCDDO, the objective is to minimize the total hiring cost of teams. Note that $\mathbb{E}$ stands for mathematical expectation. In the proposed model, $Q\ (Y,\omega)$ is equal to the total capacity violation cost. The objective function then minimizes the sum of team construction cost and the expected overloading cost as well as the deviation of the total workload between the teams as it is indicated in (3.18) and (3.19). Constraints (3.20) and (3.21) ensure that the workload is spread out evenly in order to balance the total required workload of the patients from providers in each team. Constraint (3.22) indicates that workload can exceed the capacity of the provider by φ_h unit at the cost of ξ_h per unit within each team. Constraints (3.23) and (3.24) limits the overload of providers in team. Constraint (3.23) ensures that the overload is equal to zero when there is no team available. Constraint (3.24) ensures that the overload is lower than the overload upper bound and limits the amount of overload. Constraint (3.25) prevents the assignment of the patient to the teams, which do not have the required providers who can provide chronic disease-specific healthcare services. Constraint (3.26) ensures that the patients are assigned to only one team where there is at least one team of type τ. Constraint (3.27) indicates that the variable $y_{\rho\tau}$ is a binary variable. We show that that x_τ and φ_h belong to non-negative integer and real number sets respectively in constraints (3.28) and (3.29).

Now that we defined SCDDO and TCDDO models, we go over the solution approach adapted in this research in the next section.

3.3.4. Sample Average Approximation

Proper estimation of expected recourse function and optimizing the expected recourse function over the first stage of stochastic optimization are among the main challenges in solving them. One approach to address the above-mentioned issues efficiently is Sample Average Approximation (SAA). SAA algorithm is considered as an efficient method for solving large-scale stochastic problems. SAA algorithm is based on Monte Carlo simulation, and it is capable of solving discrete optimization problems. This approach is generally used when obtaining the optimal solution by considering all the possible scenarios is not feasible in a reasonable amount of time. This method takes advantage of a random sampling of all possible scenarios to approximate the solution. Also, SAA has other advantages such as ease of numerical implementation, good convergence properties, a better approach for parallel computations, well developed statistical inference, and easy adaption to variance reduction techniques [127].

In order to explain the SAA algorithm, let us denote the number of replications and the number of scenarios in the sampled problem by M and N, respectively. Also N' represents the sample size used to estimate $c^T x + \mathbb{E}[\mathcal{Q}(\bar{x}, \xi)]$ for a given feasible solution $\bar{x}$. The necessary steps for implementing the SAA algorithm is explained below [128].

We put forth the SAA algorithm in the following. In addition, we depict the flowchart of SAA algorithm implementation in Figure 3.2 [129].

- **SAA Algorithm:**

1. Repeat the following steps for *m = 1,...,M.*

1.1. Generate a random sample of scenarios with N realizations (i.e. $\{\xi^1, \dots, \xi^N\}$).

1.2. Solve the following problem and record the solution and the optimal objective value in vectors $\hat{x}_N^m$ and $\hat{v}_N^m$, respectively.

$$\min_{x \in X}\{c^T x + N^{-1}[Q(\bar{x}, \xi^n)]\} \tag{3.30}$$

1.3. Evaluate the upper bound of the true optimal solution value $\hat{g}_{N'}\,(\hat{x}_N^m)$ and the estimate of the variance $S^2{}_{\hat{g}_{N'}\,(\hat{x}_N^m)}$ by generating N' independent random samples $\{\xi^1, \dots, \xi^{N'}\}$ using the following formulas.

$$\hat{g}_{N'}\,(\bar{x}) = c^T\bar{x} + N'^{-1}\sum_{n=1}^{N'} Q(\bar{x}, \xi^n) \tag{3.31}$$

$$S^2{}_{\hat{g}_{N'}\,(\bar{x})} := [N'(1 - N')]^{-1}\sum_{n=1}^{N'}[c^T\bar{x} + Q(\bar{x}, \xi^n) - \hat{g}_{N'}\,(\bar{x})]^2 \tag{3.32}$$

2. By using the following equations, find an unbiased estimator of $\mathbb{E}[\hat{v}_N]$ which is considered as the lower bound to v^* and its estimate of the variance $S^2_{\bar{v}_N^M}$.

$$\bar{v}_N^M = \frac{1}{M}\sum_{m=1}^{M}\hat{v}_N^m \tag{3.33}$$

$$S^2_{\bar{v}_N^M} = \frac{1}{M(M-1)}\sum_{m=1}^{M}(\hat{v}_N^m - \bar{v}_N^M)^2 \tag{3.34}$$

3. For every solution $\hat{x}_N^m$, *m = 1,...,M*, estimate the optimality gap by $\hat{g}_{N'}\,(\hat{x}_N^m) - \bar{v}_N^M$, along with an estimated variance of $S^2_{\bar{v}_N^M} + S^2{}_{\hat{g}_{N'}\,(\hat{x}_N^m)}$. Choose one of the M candidate

solutions based on the pre-defined criteria such as the least estimated objective value, or the smallest estimated gap.

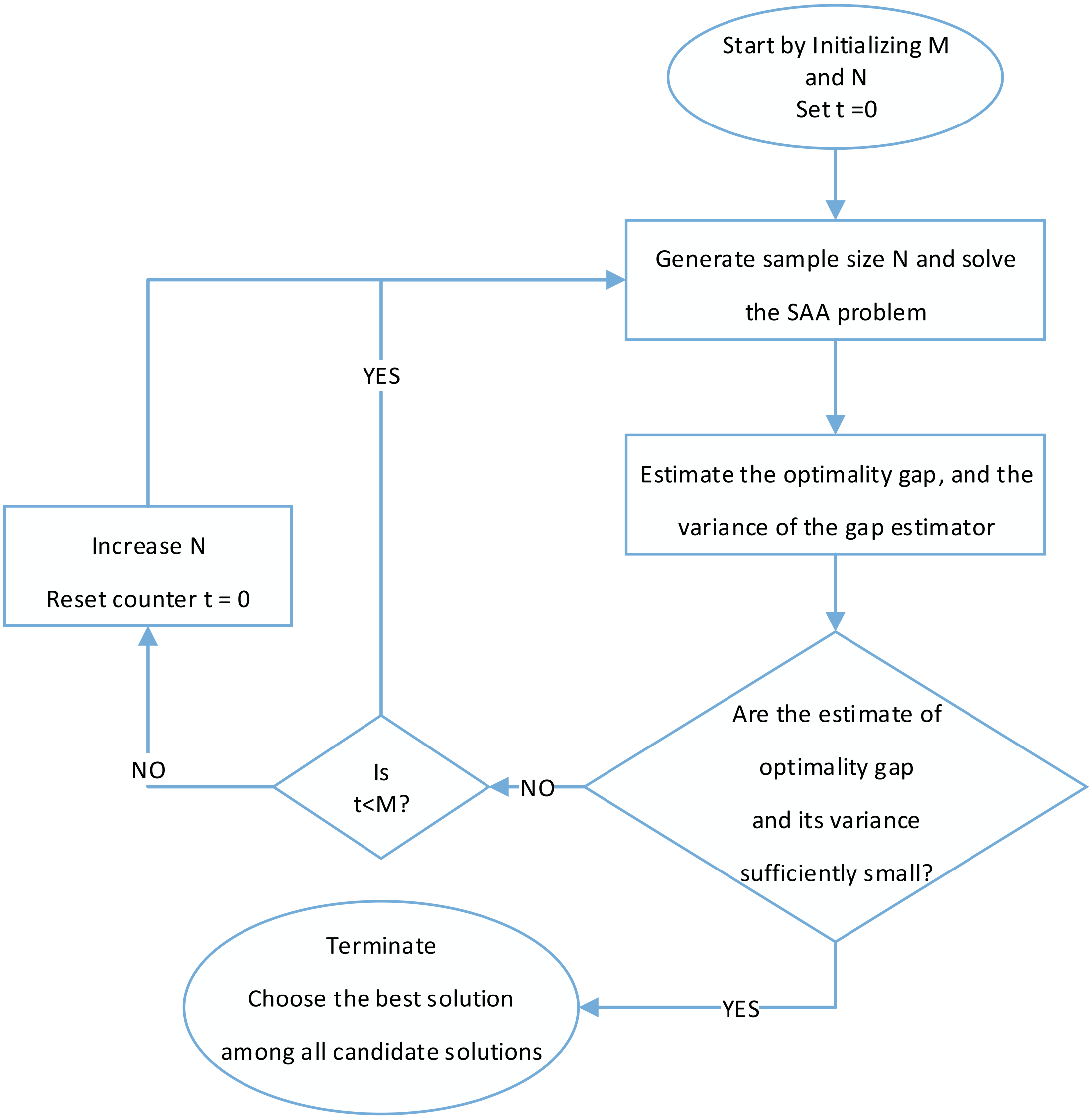

Figure 3.2: Flowchart of SAA Algorithm

It is proven that the optimal value of SAA problem converges to the optimal value of the true problem when the sample size tends to infinity [130]. However, selecting a larger number of samples increases the complexity and the computation time of the model. In order to tackle this issue, the SAA problem is solved several times with smaller independent and identically distributed samples rather than using a large number of samples. The quality of SAA solution depends on several factors such as the size of the sample, the convergence rate, and the algorithm termination criteria (see [131]).

We use the Value of Stochastic solution (VSS) to measure the performance of the SAA algorithm in this research. VSS justifies the significance of using stochastic approaches over the expected value solution. Due to the complexity of solving stochastic recourse problem, there is a tendency toward replacing random variables of the model by their expected value and solving the mean value problem to find the Expectation of the Expected Value Problem (EEVP). However, EEVP is not necessarily close to the solution of Recourse Problem (RP) unless the optimal solution of the expected value problem is independent of random variables realizations. Thus, we use VSS to determine the usefulness of the model. VSS measures the cost saving where the stochastic solution is used rather than the mean value solution. Therefore, VSS represents the possible gain from solving the model by considering randomness. The value of the stochastic solution is defined as VSS = EEVP – RP.

3.4. Computational Study

In this section, we perform and discuss an extensive experimental analysis of proposed models.

To assess and evaluate the computational performance of the proposed method, we compare the amount of change in the objective function for each problem for various problem sizes. Afterward, we establish the models by using the results of the comparison.

In addition, in this section, in order to find the optimal parameters of the model, we evaluate the computation time complexity of the models for different parameters. We present the value of the objective function for the stochastic problem, and the expected value problem in this analysis. Also, we evaluate the benefit of using randomness for modeling both problems, separately. We implemented the proposed model by using C++ programming language. To report the numerical results, a personal machine with a quad-core 2.4 GHz processor and 16 GB RAM is used. Due to the large execution time, we set the maximum CPLEX run time and the algorithm gap tolerance to 3600 seconds and one percent, respectively. The optimality gap instructs CPLEX to stop once it finds a feasible integer solution that is within one percent of optimal.

3.4.1. SCDDO Result Discussion

As mentioned before, the main objective of the SCDDO problem is to determine the number of providers that are necessary for each facility while minimizing the hiring and service coverage cost. In order to determine the model parameters, we evaluate the problem size in different stages. We consider four problem instances in this study that include 200, 300, 400, and 500 patients during a particular strategic planning horizon. The problem scenarios are generated based on the number of patients, their random location, and the distribution of different attributes such as age and comorbidities. We uniformly assign patients to a physical location in a pre-defined location grid designed for specifying the distance of the patient from every available facility. The Manhattan distance approach is used to measure patient distance from available facilities. In the first step of the SAA algorithm, we generate M samples, each including N independent scenarios. Afterward, we calculate the probability distribution of different attributes of the patient. Then, we generate different scenarios based on probability distribution of the patient's attributes. Solving deterministic equivalent of the stochastic model for samples results in M candidate solutions. Then, SAA generates an evaluation scenario sample with N' samples to evaluate the candidate solutions. Finally, SAA selects the optimal solution with the smallest gap among the evaluated candidate solution

In Figure 3.3, the effect of scenario sample size on the quality of solution for different numbers of patients is shown. The number of sampling replications for instances is considered as 10. In order to determine the proper scenario sample size, we consider the second stage estimation sample size as 60 to be sufficiently greater than the scenario sample size. This way we compare the objective change for different values of scenario sample size N. In the following figures, we compare the change rate in the value of the objective function for a given scenario sample size to that of the previous size. As an example, in the first chart of Figure 3.3, the objective value of the problem with sample size 5 is compared to that of a problem with scenario sample size 10. Then the objective value of the problem with sample size 10 is compared to that of a problem with scenario sample size 15. This procedure is followed until we reach the point that the change rate of the objective value for a given scenario sample size is not significantly higher than the objective change rate of its preceding scenario sample size. We set the threshold for change ratio to 5%.

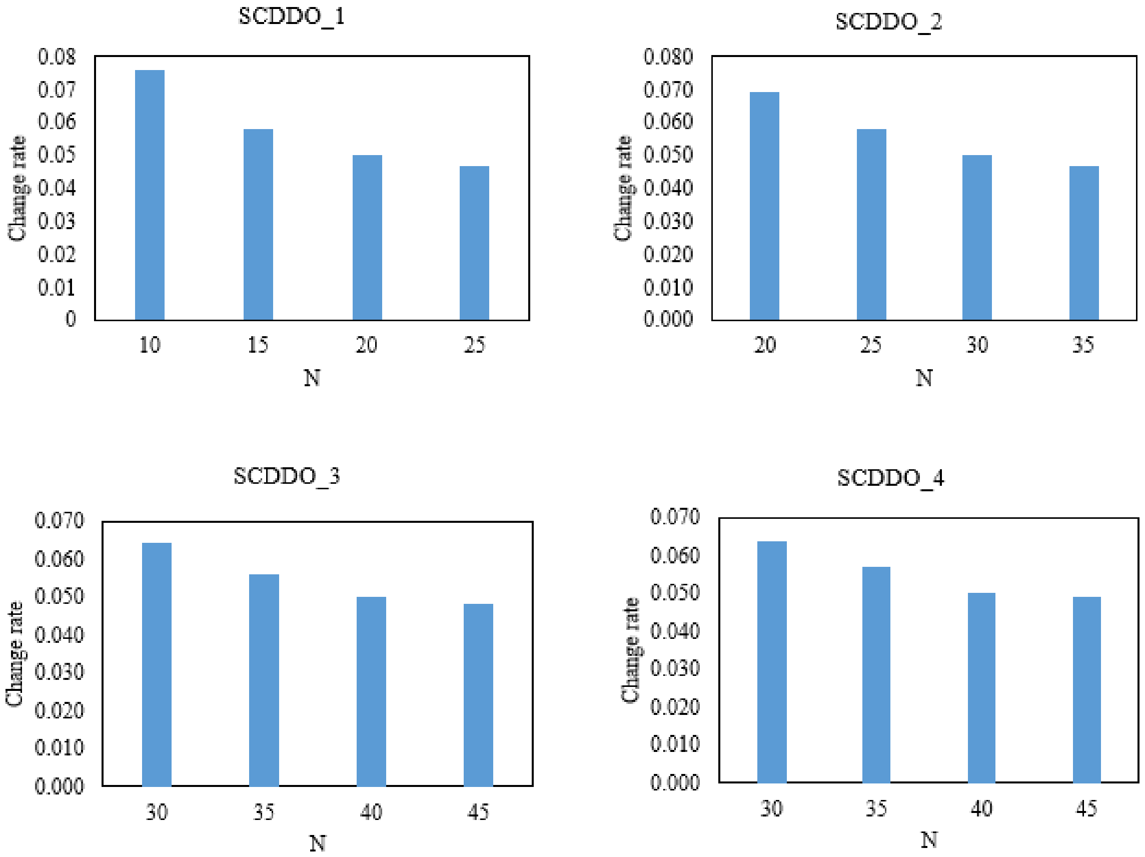

Figure 3.3: Analyzing Scenario Sample Size vs. Objective Change for SCDDO

Figure 3.3 shows that by increasing the number of samples, the quality of the solution improves. This happens since more possible scenarios are considered, so the problem gets closer to reality. However, as it is indicated in this figure, by increasing the number of scenario sample size, the amount of improvement is reduced until it gets to the point that there is no significant change. This means that although increasing the scenario sample size enhances the quality of the solution, the relative rate of improvement is reduced.

Now that the optimal number of N is determined in the previous step, the number of scenario samples is fixed for evaluating the best N' for each problem in the next step.

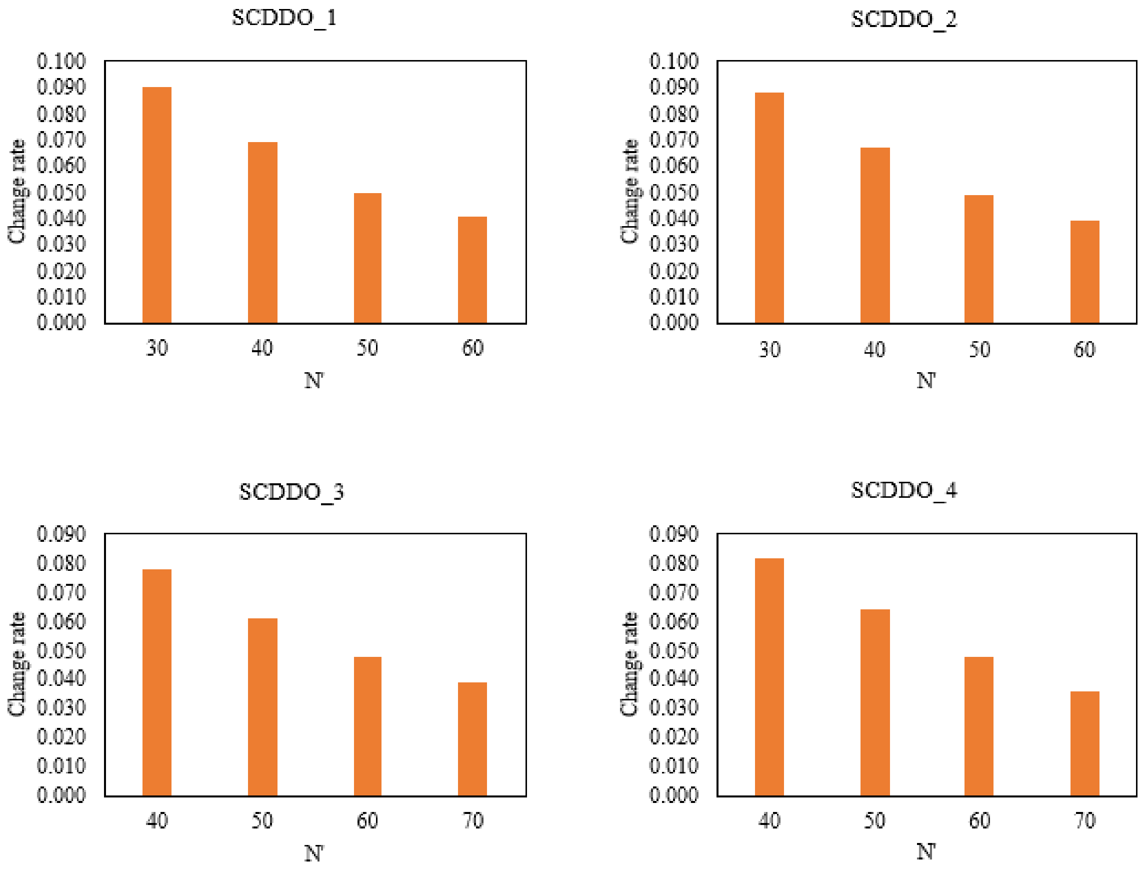

Figure 3.4: Analyzing Estimation Sample Size vs. Objective Change for SCDDO

Figure 3.4 shows the effect of the second stage evaluation sample size on the amount of improvement in the objective function. As it is shown, we use the optimal number of scenarios that are determined in the previous step for different number of patients to find the best estimation sample size. It is clear that by increasing the number of estimation sample size, the quality of the solution improves. However, the rate of improvement decreases while N' increases. Thus, we select N' as the point where there is no significant improvement in the value of the objective function after that point.

After selecting the parameters of the stochastic programming model, we solve the proposed problem and summarize the results in Table 3.1. We used two different methods to solve the problem. First, we solved the problem by using the SAA method explained in this research. Then we solved the expected value problem whereby all the stochastic variables are replaced by their expected value. Afterward, we compared the solution of these two approaches. We expressed the absolute value of the difference between the RP and EEVP solutions in Table 3.1. As it is explained in the methodology of this chapter, we use the value of the stochastic solution to measure the usefulness of the model. The absolute value of the VSS represents the amount that decision makers can save if they use stochastic solution with random scenarios instead of only using the mean value scenario.

Table 3.1: The Results of Stochastic Optimization for SCDDO Problem

Instance No.	# of Patients	# of Replications (M)	Sample Size (N)	Evaluation Sample Size (N')	EEVP	RP	Abs. VSS
SCDDO_1	200	10	20	50	104,553.82	93,351.63	11,202.20
SCDDO_2	300	10	30	50	158,573.02	142,858.58	15,714.44
SCDDO_3	400	10	40	60	200,032.54	177,019.95	23,012.59
SCDDO_4	500	10	40	60	283,542.38	246,558.59	36,983.79

As the results suggest, the VSS improves by increasing the number of patients. However, we must be aware that the number of scenarios is increased along with the problem size. So, one reason affecting the VSS is that by increasing the number of scenarios and evaluation sample size, the model gets closer to reality which means that the algorithm benefits from more possible scenarios and can generate better candidate solutions compared to the solution of the mean value problem.

Please note that we set the MIP gap tolerance to one percent in this study. This value indicates that CPLEX stops when it finds a feasible integer solution that is within one percent of optimality. Since the objective function of the proposed model amounts to a hundred thousand, we avoid further processing and stop at one percent. However, one can proceed with a tighter optimality gap where more accuracy is required, or the execution time is not costly to avoid any chance of missing the best possible solution.

3.4.2. TCDDO Result Discussion

In the second study, we determine the number of each type of teams for each facility while balancing the workload among the teams within the facility. We designed different problem instances in this study. We solved the problem by considering many types of team compositions with different number and type of providers. Also, we considered patients with different chronic diseases along with their associated workload. Each team consists of various numbers of individuals including primary care physician (PCP), registered nurse (RN) which include psychiatric-mental health nurse, licensed practical nurse (LPN), nutritionist, pharmacist, and clerical assistant. As a best practice, some resources can be shared between care provider teams in the team-based care delivery systems. To tackle this matter, we divide the total available capacity of each shared resource by the number of teams using that particular resource and consider that resource as a standalone resource but with a divided capacity between the teams.

We calculate the capacity of full-time employees based on their number of working days and the maximum RVU value per hour for each provider. The available capacity of the providers is also calculated based on their Full-Time Equivalent (FTE). The FTE represents a scale to compare the hours that a part-time employee works to that of a full-time employee. Thus, the FTE for a full-time employee is equal to one, and the FTE for part-time employees is determined as a proportion of the FTE of full-time employees with respect to their working hours. We consider 1920 available working hours in each year for every full-time employee considering 20 days off due to vacation. Afterward, we use the maximum RVU values per hour for each provider, which was obtained by the CPT codes performed and the RVU scale schema [132] to convert and express the available capacity of providers based on RVU. For example, the maximum RVU/hour for a physician in Detroit VA medical center is equal to 12, so the provider can deliver 20304 RVU during a year. We consider four types of teams in this study each of which consists of various types of healthcare providers.

In order to generate the scenarios, we consider different problem sizes and generate random scenarios based on the probability distribution of the features in the dataset. Then, the required workload of patients is predicted for every scenario. In order to have a robust estimation, we repeat the sampling process for multiple times and pick the best candidate solution. Then, we run the SAA algorithm five times and take the average of the solutions to remove any variances in the solution that may happen due to the machine performance.

In order to approximate the recourse function of the stochastic optimization model, we use the SAA algorithm. The algorithm takes independent and identically distributed samples where each sample has a constant number of scenarios. The main goal of the first part of this analysis is to determine the optimal number of scenarios in the sampled problem. After determining the number of sampling replications and the number of scenarios, we solve the stochastic model and evaluate the results.

In the first stage of the analysis, M samples are generated; each of these samples consists of N independent scenarios to be utilized for estimating the recourse function of the stochastic optimization model. Then, according to SAA algorithm, we solve the deterministic equivalent of the TCDDO problem independently for every sample M, which results in M candidate solutions. In order to evaluate the candidate solutions, we generate a scenario sample set with a much larger number of scenarios compared to the number of the scenarios in the initial scenario sample set N ($N' >> N$). Then, we compare each candidate solution to the solution of the evaluation problem and choose the best candidate solution. The reason that we take one evaluation sample is to establish a fair comparison between solutions where the chance of being chosen is equal for all the candidate solutions. The described analysis procedure makes evident that determining the optimal number of samples and scenarios is critical to our analytic approach in terms of solution quality and computational efficiency. Accordingly, we analyze the improvement in the value of the objective function for different sample and scenario sizes subsequently. This analysis gives us the capability to determine the optimal combination of the number of samples and scenarios. In order to determine the best sample and scenario sizes, we follow the procedure of comparing the improvement in the value of the objective function until there is no significant improvement in the objective function of a given sample size compared to the objective function value associated with its preceding sample size.

The TCDDO problem is solved by considering different problem sizes. We consider six instances with 40, 60, 80, 100, 120 and 140 patient sample size over a particular tactical planning horizon. We design the scenarios by considering different attributes of patients and their associated workload. We analyze the sensitivity of the model in terms of two different aspects which are objective function change and time complexity. Let us begin with the objective function change analysis by explaining the procedure that is followed to identify the optimal sample and scenario sizes. As it is shown in Figure 3.5, the number of sampling replication (M) is set to 10. Then we determine scenario sample size (N) where the probability of each scenario is equally likely. In addition, we consider the second stage estimation sample size N' equal to 60 to be sufficiently large for problem evaluation.

Afterward, we investigate the effect of the scenario sample size on the quality of the solution in terms of the change in the objective function for each problem size. As it is depicted in Figure 3.5, the procedure followed here is to compare the objective function for each sample size with the value of the objective function of its preceding sample size then calculating their difference as the change in the value of the objective function. As an example, in the chart on the top left of Figure 3.5, we consider the basis value for the sample scenario size as 5 and compare its associated value of the objective function to that of a problem with scenario sample size equal to 10. Then we follow this procedure to compare the value of the objective function associated with samples with 10 and 15 scenarios. We continue the procedure until reaching the point that there is no significant improvement in the objective function value associated with the current sample size compared to that of a problem with the succeeding sample scenario size.

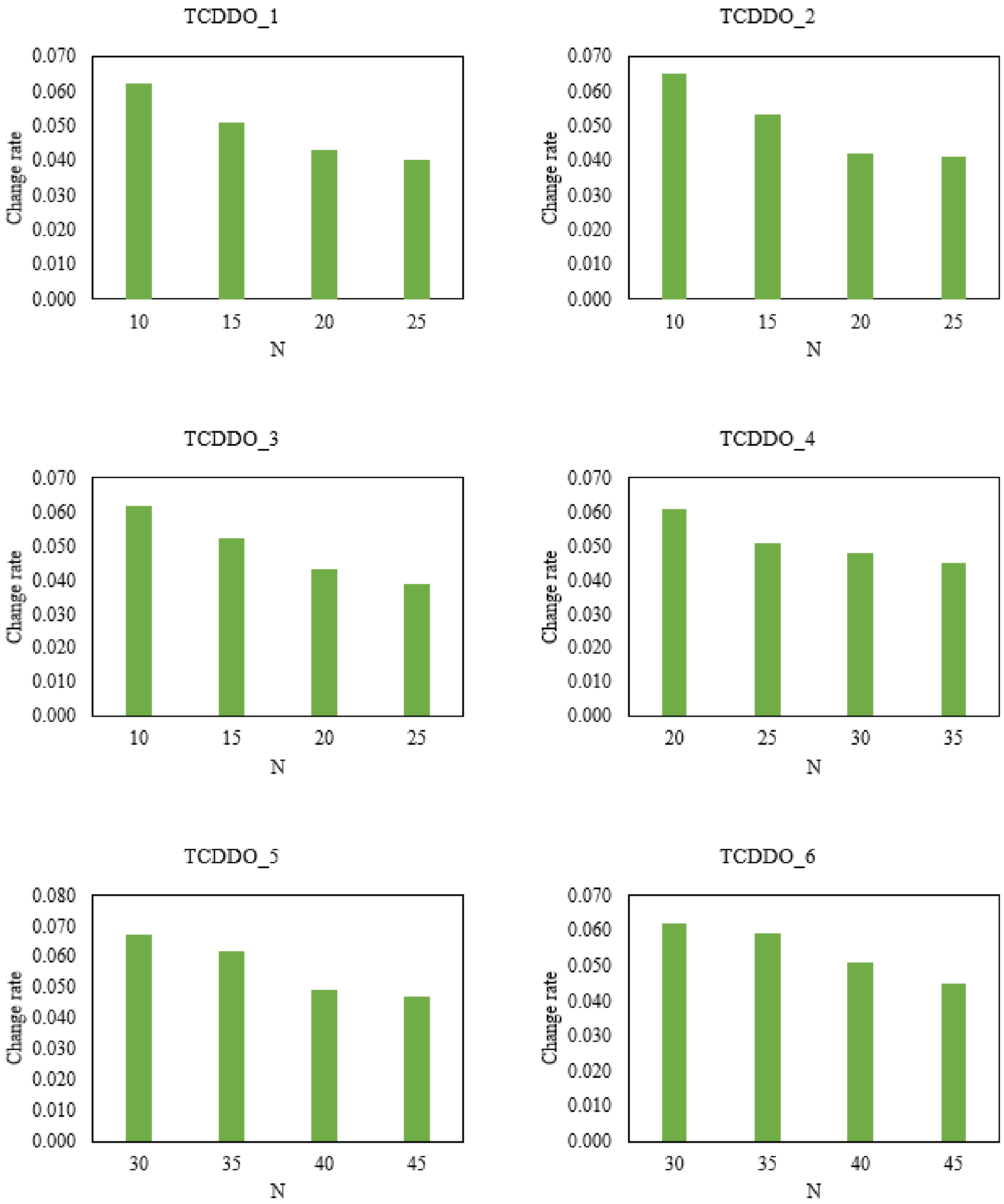

Figure 3.5: Analyzing Scenario Sample Size vs. Objective Change for TCDDO

As the results in Figure 3.5 shows, for each problem from top left to bottom right the optimal number of scenarios are 20, 20, 20, 30, and 40, respectively.

As the second set of running time analysis, we focused on analyzing the effect of changing the number of scenario samples on the performance of the algorithm. We consider the number of sampling replication as 10 for this analysis. Also, we run each instance for five times to reduce the effect of CPU performance variation on the problem solution. As it is displayed in Figure 3.6, by increasing the scenario sample size, the running time increases exponentially. One possible reason is that the algorithm must solve both the first and second stage for each instance.

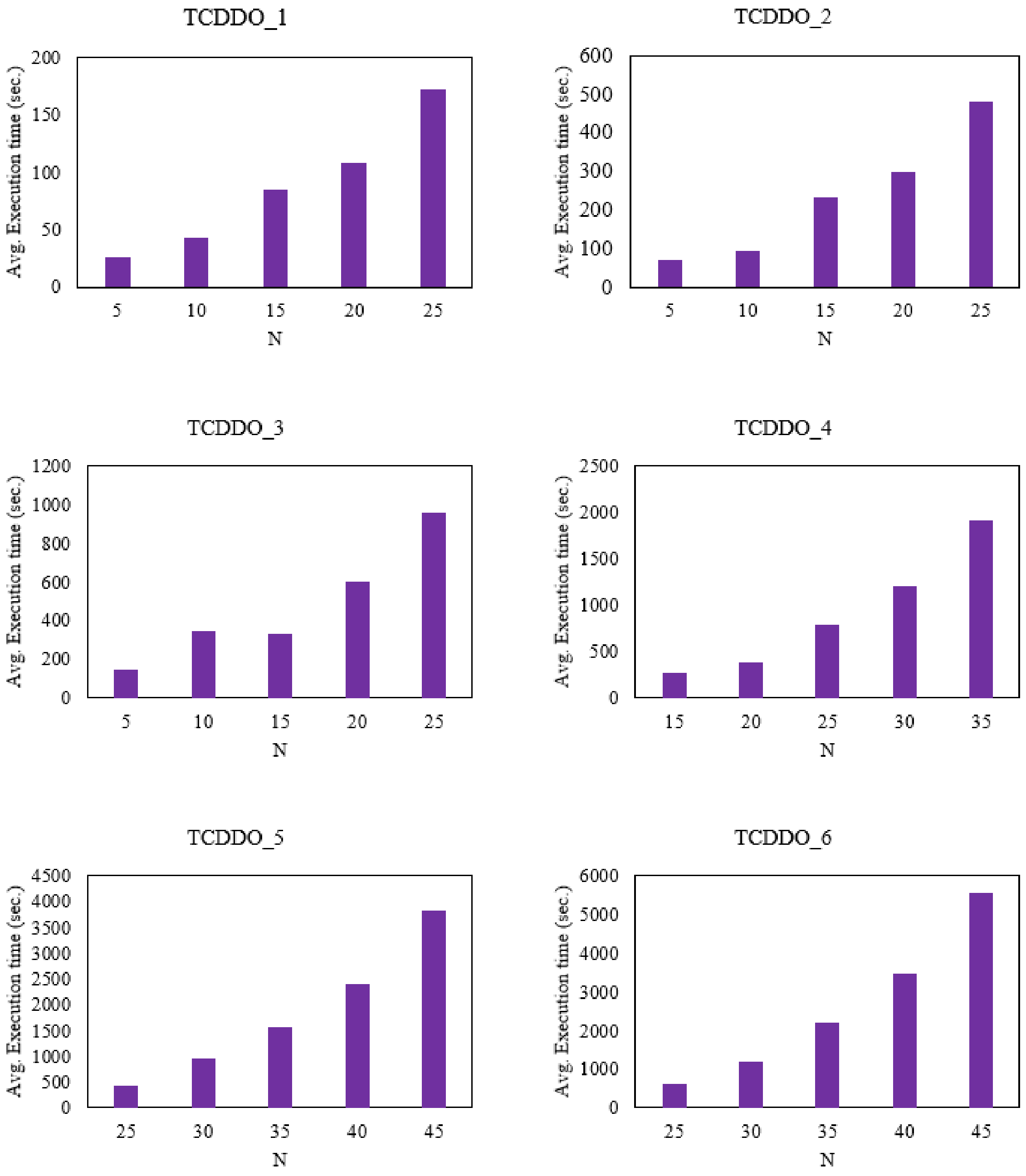

Figure 3.6: Analyzing Scenario Sample Size vs. Execution Time for TCDDO

We use the scenario sample sizes from the previous analysis as inputs for their associated problem to perform the analysis on the evaluation scenario sample size. The primary goal of the analysis depicted in Figure 3.7 is to find the optimal number of evaluation sample size N' by fixing the optimal sample scenario size for each problem. We try different evaluation sample sizes and compare their associated change in the objective function. We follow the same procedure as described earlier in order to determine the evaluation sample size. As it is demonstrated in the Figure 3.7, the quality of the solution increases when the evaluation sample size is increased. However, the rate of change plunges significantly. As we are only interested in the amount of change that is significantly higher than that of its preceding, we choose the evaluation sample size for which the change in the objective function is relatively significant compared to its following sample size. Therefore, we defined the threshold of the change as 5% and choose the scenario sample size that reaches the threshold first. Thus, the optimal number of evaluation sample size for problem instances are equal to 50, 50, 50, 50, 60, and 60 from the top right to the bottom left chart, respectively.

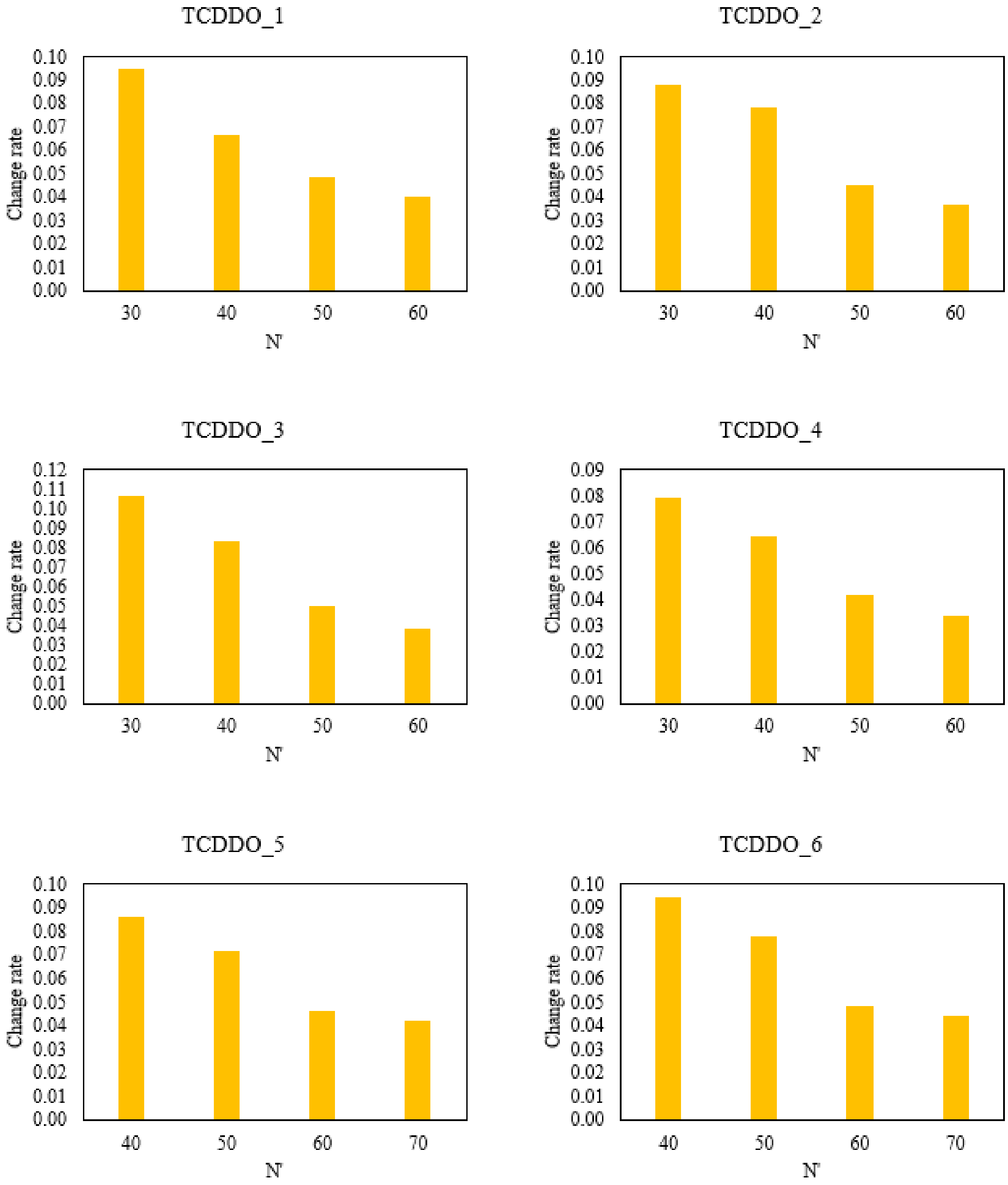

Figure 3.7: Analyzing Estimation Sample Size vs. Objective Change for TCDDO

Now that we determined the number of evaluation sample size, in this part of the analysis, we illustrate how the average running time of the algorithm changes concerning changes in the evaluation scenario sample size. SAA algorithm requires a large number of scenarios for accurately estimating the second stage of the problem. So, we consider a relatively large number of evaluation sample size in this analysis. As the results of the execution time displayed in Figure 3.8 suggests, the running time of the algorithm increases by increasing the number of the evaluation samples. The running time increases exponentially, however, the rate of the increase resulted from increasing the number of evaluation samples, is less than the rate of increase in the running time that is discussed earlier when changing the number of scenario samples.

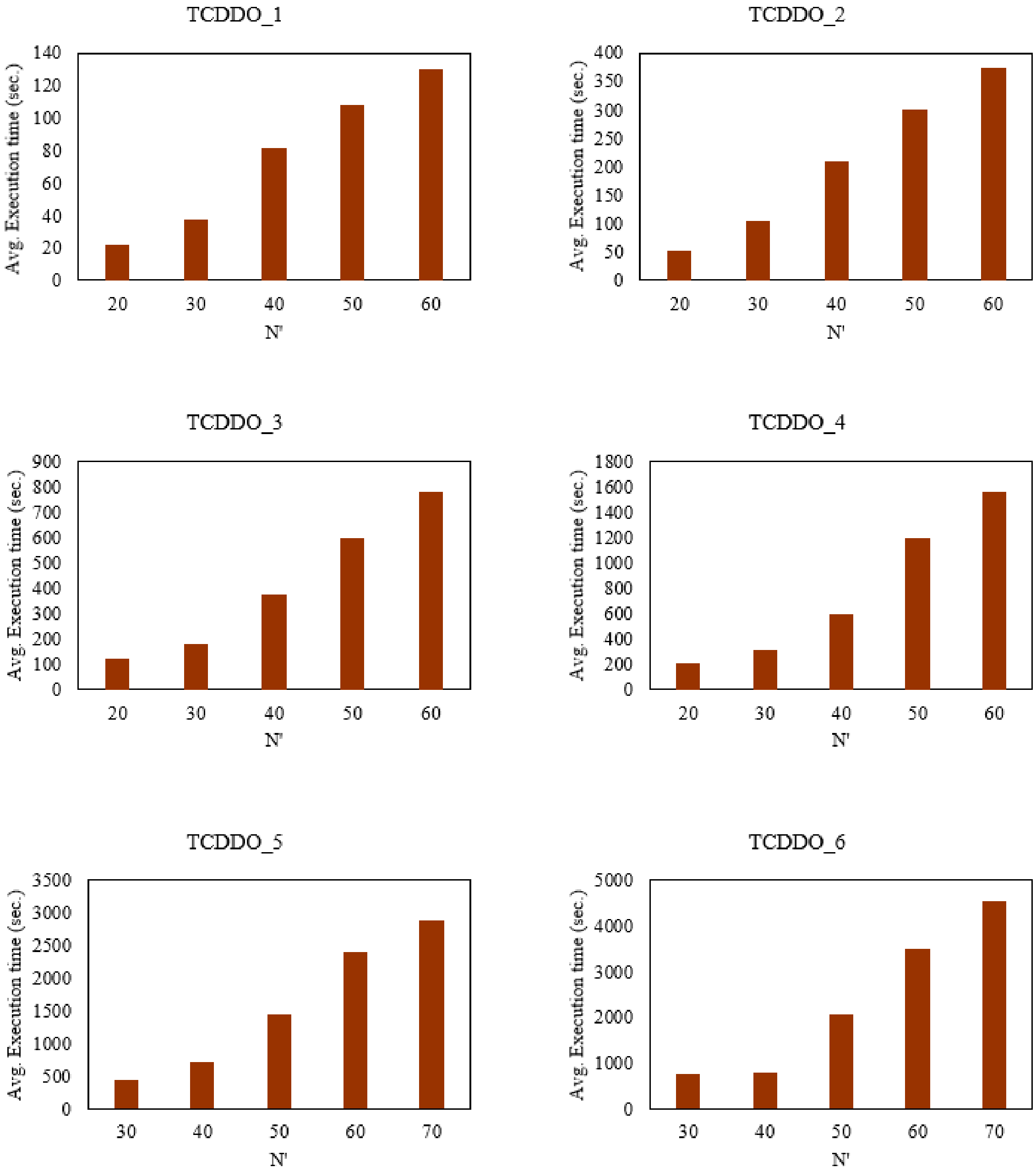

Figure 3.8: Analyzing Estimation Sample Size vs. Execution Time for TCDDO

To summarize, we illustrated the execution running time for each problem instance in Figure 3.9. We executed the algorithms for five times for each problem size to reduce the effect of machine performance variation. In this figure, we analyze the effect of increasing the problem size in our model. The results suggest that the average execution time of the algorithm increases exponentially when the problem size is increased. From the definition of SAA algorithm explained in the previous section, this is a reasonable outcome. Since by increasing the problem size, the algorithm must solve the first and the second stages of the stochastic optimization problem for every sample. Hence, due to the increase in the sample size, the model requires more scenarios.

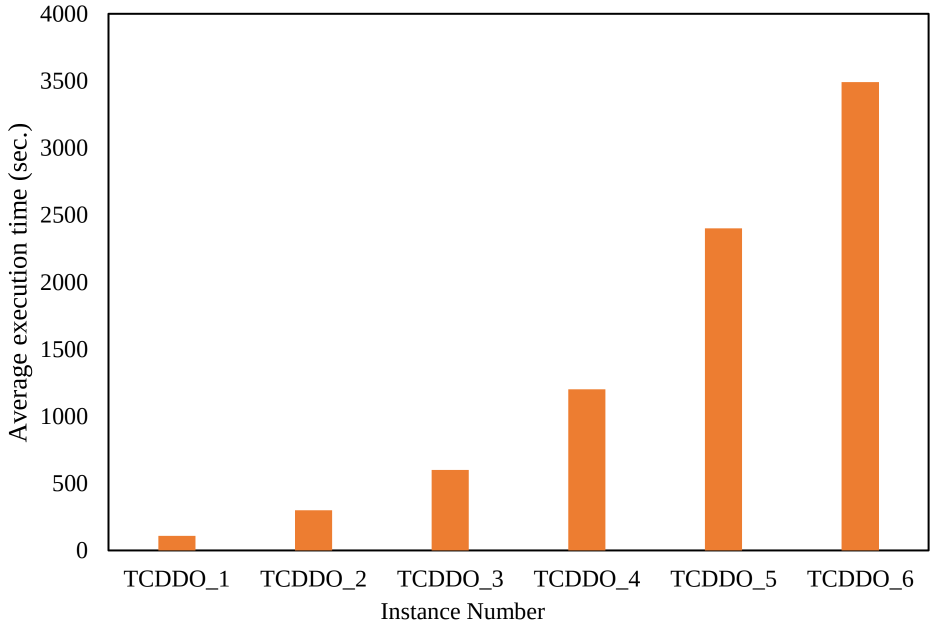

Figure 3.9: Execution Time Analysis for TCDDO Problem Instances

The established running time and objective change analysis enable the decision makers to optimally fine-tune the model parameters. Now that optimal parameters of the TCDDO problem are chosen, we solve the problem by using two solution approaches. Initially, we solve the problem by using the expected value problem where the random variables are replaced by their expected value. To provide the second set of solutions, we solve the stochastic problem by using SAA algorithm. Finally, we present and compare the results of both approaches in Table 3.2. In addition, we present the value of stochastic solution so that the decision makers are able to measure the performance of the stochastic solution. VSS represents the impact of considering the variable stochasticity into the solution method and is considered as a measure for performance of the stochastic solution approach.

Table 3.2: The Results of Stochastic Optimization for TCDDO Problem

Instance No.	# of Patients	# of Replications (M)	Scenario Sample Size (N)	Evaluation Sample Size (N')	EEVP	RP	Abs. VSS
TCDDO_1	40	10	20	50	49,021.60	43,381.95	5,639.65
TCDDO_2	60	10	20	50	68,159.42	60,318.07	7,841.35
TCDDO_3	80	10	20	50	74,655.85	66,067.13	8,588.73
TCDDO_4	100	10	30	50	86,349.53	76,415.51	9,934.02
TCDDO_5	120	10	40	60	144,497.96	127,874.30	16,623.7
TCDDO_6	140	10	40	60	154,933.75	137,109.52	17,824.2

We solved the stochastic problem and its expected value problem five times for every instance and reported the average of the objective function in Table 3.2. Please note that the parameters of the model are tuned for each instance. As we observe in the results of Table 3.2, generally the VSS increases when the problem size increases. The results also suggest VSS rises suddenly for problem instance TCDDO_5. This happens because of the increase in the evaluation scenario sample size. Since increasing N', results in a more realistic estimation of the candidate solution due to the greater number of covered scenarios and more coverage of the reality. Moreover, we know that by increasing N, we provide the model with more information about the possible scenarios so by using a greater number of scenario samples, we can improve the quality of solution and generate better solutions. In this way, it is evident that using stochastic optimization techniques can significantly contribute to improve the solution and deliver value by using more information and scenarios from reality rather than only using average scenario to solve the problem.

3.5. Discussion and Conclusion

In this chapter of the dissertation, we focused on the prescriptive analysis. After determining the required workload of the patients, the main issue is to provide efficient healthcare systems to the patient to improve the quality of the service and reduce the cost of providing the care. We discussed two different management level decision-making problems in this chapter. First, we focused on the strategic decision-making level where the decision maker deals with multiple healthcare facilities. Facilities play a significant role where it comes to strategic capacity management. We formulated this problem as a two-stage problem with recourse where the patient’s demand and location are uncertain. The main objective of the strategic chronic disease decision optimization (SCDDO) problem is to manage the required capacity of the healthcare providers by optimizing the number of providers in the first stage of the decision-making process. The objectives of SCDDO in the second stage are assigning patients to the healthcare facilities, offering incentives to patients by transferring them between the facilities, and then adjusting the hiring decisions of the first stage of the problem.

The scope of the second model discussed in this chapter lies in the tactical planning for chronic disease operations management. Unlike the previous model which focused on higher-level capacity management for multiple facilities, we proposed an integrated model for the team-based workforce and workload optimization within one single facility where the required workload of the patients is stochastic. Team-based healthcare delivery system is proven as a successful model for reducing cost and improving the quality of the care. With more healthcare organizations establishing team-based delivery systems, some challenges emerged to be critical for the efficiency of these systems. One of the most important challenges is establishing a systematic approach for assigning patients to the teams by considering the limitation of the teams in terms of their capacity. In order to develop more sustainable teams with a balanced workload, we developed a stochastic optimization model for a patient-team assignment where the required workload of the patient is not known.

In team-based healthcare delivery systems, the demand is considered as a stochastic variable which can be spread through the healthcare team based on the specialty and responsibility of healthcare providers. The stochasticity of the demand results in a portfolio of demand which depends on different conditions and attributes of the patient. In the previous chapter, we used the patient attributes to model the workload. Then in this chapter, we modeled the tactical chronic disease decision optimization problem as a two-stage stochastic optimization model which aims to minimize the number of teams and balance the workload between teams. As the results suggest, the stochastic optimization provided us with a more realistic solution. However, we must consider the increase in time complexity of the approach as the number of scenarios and problem size increase. We discussed the scalability of the problem and analyzed the solution performance with respect to the running time of the algorithm. In addition, we compared the change in the objective function for a various number of scenarios and chose the optimal scenario sample size for every instance to optimize the efficiency of the proposed solution.

Furthermore, we provided some insights about the value of considering uncertainty into the model. As the result of this study suggests, considering randomness can help to reduce the cost of team-based healthcare delivery. This study provides comprehensive modeling and solution for capacity planning in a team-based healthcare delivery system which is an essential step and has a prominent impact on improving recent health delivery systems and leads to improvements in patient satisfaction as well as maintaining continuity of care.

Modeling the real-world problem without using any assumption is not reasonable. As the future steps, the proposed model can be improved by relaxing some preliminary constraints and assumptions used in this research. Such modifications include considering stochastic utilization rate, different efficiency rate for team members and different quality of provided care by each care provider as well as taking into account the possibility of switching teams by patients and coordination between care providers of teams.

CHAPTER 4 CONCLUSION

4.1. Summary and Contribution

Healthcare delivery system strategic planning is a critical contributor to the effectiveness of chronic disease operations management and quality improvement of provided services to patients. With considering aging of the population, there is more need for coordinated care teams with team members who communicate regularly to make sure that patients with chronic disease receive appropriate services in a timely manner and in an efficient way. As it is discussed, the efficient design and development of the multidisciplinary teams are essential for better management of the chronic disease. This is important due to the fact that well-structured health professional teams with a clear division of responsibilities and well-balanced workload offer a wide range of skills and insight coming from different individuals. Consequently, patients with chronic disease receive high-quality care.

As it is discussed in the previous chapters of this dissertation, some researchers investigated the importance of building multidisciplinary teams for chronic disease operations management and focused on examining its impact on chronic disease interventions. However, two critical questions needed to be answered. The first question was how the required workload of patient should be estimated and measured, and the second question was how decision makers should estimate, plan and optimize the required capacity of teams and their assignment to patients to be able to satisfy the patient’s needs efficiently and minimize the cost of the healthcare systems.

The goal of this dissertation is to assist healthcare decision makers in chronic disease operations decision-making by providing a systematic and scientific approach to answer the aforementioned questions. For this reason, we proposed a methodological and conceptual framework for understanding the essential elements of chronic disease operations management for different decision-making levels in the first chapter of this dissertation. The proposed framework covers two important decision-making stages, namely predictive and prescriptive analytics. We focused on predictive analysis part of the research in the second chapter of this dissertation. The developed model predicts the required workload of patients based on their various features and characteristics such as age and different types of chronic diseases. Afterward, we used the output of the predictive analysis to develop the strategic and tactical capacity planning models for multiple and single facilities in chapter three. Finally, we elaborated on the summary, contribution and potential future directions of this dissertation in chapter four.

Since demand of the patients from the healthcare system is not always known, we developed different predictive approaches in the predictive analysis phase of this dissertation. The performance of the patient workload prediction is a key factor for decision-making in chronic disease operations management. Thus, we compared the performance of the developed methods and chose the most accurate approach for predicting the patient required workload. In order to achieve a better prediction performance, we developed a deep multi-task learning approach. We used stacked autoencoders for transforming the data representation. Then, we used multi-task learning instead of single-task learning in order to be able to consider the similar characteristics of the patients and distinguish between the workload of the patients in each facility as well as improve the performance of the prediction. In terms of performance, which is defined by the mean squared error of prediction, the deep multi-task learning approach outperformed many machine-learning techniques. In general, a large set of samples that fully represents the targeted statistical population is an essential input for almost any machine learning technique. However, sometimes this requirement is not satisfied due to the data gathering issues. Thus, the burden of dealing with the prediction accuracy and training a precise learner is on researchers and data scientists. They must choose the appropriate machine learning approach in order to assist the decision makers in their subsequent management decisions for capacity planning and process optimization. Therefore, in this dissertation, we took advantage of a special property of multi-task learning where the samples are trained jointly. In the proposed approach, after transforming the data by using stacked autoencoders, the data is categorized based on the healthcare facilities with a limited dataset for each category. We showed that transforming the data and using the multi-task learning approach improve the performance of the patient workload prediction compared to the performance of utilizing feature selection and bagging techniques.

As discussed above, the accuracy and dependability of patient workload prediction is an important and critical aspect to be considered in the chronic disease decision-making process since it has a direct relationship with decreasing the overall cost of healthcare systems and increasing patient satisfaction. After obtaining an accurate healthcare workload prediction as an input for the prescriptive analytics phase of this research, we studied healthcare capacity and resource planning in team-based chronic disease operations management systems in this dissertation.

The prescriptive analysis phase of the proposed framework consists of two optimization problems discussed in chapter three of this dissertation. In the first part, we focused on strategic healthcare capacity planning that includes capacity planning for multiple facilities. For many healthcare delivery organizations, facility is a very important entity and many long-term capacity management tasks such as demand estimation and capacity management are facility-based. For this reason, we developed a two-stage stochastic optimization model in order to minimize the number of each specific healthcare provider. We considered the location of patients to model the problem. This model is capable of minimizing the required service coverage cost of patient transportation between the facilities in its second stage.

The goal of the second problem discussed in chapter three is to assist decision makers in tactical decision-making within each healthcare facility. The results of this model provide useful insights for resource allocation. This model is analyzed under several scenarios. The result determines the number of every specialty team and the patient allocation for each of them while minimizing the amount of overtime for each team. After finding the optimal parameters for every model with different problem sizes, the performance of the model is examined by using the objective function value improvement and the time complexity metrics. In addition, the use of stochastic optimization is justified by evaluating the value of the stochastic solution for problems with a various number of patients who require healthcare services throughout a course of certain planning horizon.

We believe this dissertation can contribute toward the advancement of the research and knowledge in chronic disease operations management in many aspects. We believe that to the best of our knowledge there is no existing literature on designing an analytical and comprehensive framework for chronic disease operations management in different management levels when the workload is unknown. In this research, we suggested RVU as a quantitative measure for the required workload of the patients. This research contributes to the existing literature by adopting multi-task learning approach for forecasting the patient workload for the first time. We developed a statistical approach that categorizes the instances based on their facility-dependent features while training the instances simultaneously. We addressed the issue of limitation in training samples by integrating the relatedness of tasks for training the model. Moreover, we developed a deep multi-task learning approach to improve the accuracy of the prediction by feature representation. Besides, we provided a comprehensive performance comparison to evaluate the accuracy of our proposed approaches compared to well-known prediction techniques. As this research continued with optimization models, we defined and proposed two novel decision-making problems in chronic disease operations management, namely SCDDO and TCDDO.

The scope of decision-making in SCDDO is at strategic management level. We developed a model that provides a mathematical and systematic solution for allocating healthcare providers to patients who have different chronic conditions in a team-based healthcare delivery system when the required workload of patients is unknown. We developed a model that takes the distance of patient from facilities into account. Decision makers can use the proposed model to provide service coverage incentives in terms of transportation cost reimbursement to specific patients in order to reduce the overall healthcare cost.

TCDDO model developed in this research accounts for the tactical decision-making for chronic disease operations management focused on planning for one facility unlike the SCDDO model, which was focused on multiple facilities. We developed a novel stochastic capacity planning model to determine the number of every type of healthcare teams with different compositions within each facility in the tactical level. This problem is modeled as a two-stage stochastic optimization model, which provides insights about three critical tactical decisions. The solution of the model helps decision makers to determine the number of required teams in each facility and the assignment of the patients to teams when balancing the workload of teams where the workload of the patient is not deterministic. So stochastic capacity management for team-based health delivery systems would be a knowledge contribution to researches on chronic disease operations management.

To summarize, this research provides a comprehensive modeling and solution for decision-making in chronic disease operations management under uncertainty, which is essential for designing effective healthcare delivery interventions. We believe that research in data analytics, operations research and management aspects of chronic disease strategic planning is limited. As discussed before, there are lots of opportunities yet unanswered questions in the context of chronic disease operations management. The need and urgency for more research in this context become clearer when one investigates the cost burden of ignoring the necessity of having a systematic design for chronic disease care delivery. That is to say, due to the aging population, growing the number of people with chronic conditions in recent years, the complexity of the comorbidity treatment, the need of continued care and regular treatments, as well as the shortage in skilled resources, there should be a special attention to use systematic and scientific approaches for designing chronic disease operations management systems.

This research is an attempt to make the process of healthcare decisions making more structured and transparent. We tried to develop a mechanism to reduce the redundant operations and costs in healthcare delivery, which is a prominent element of the overall cost of providing healthcare services. We hope that this research can contribute to the research society and consequently affect the healthcare delivery systems positively so that high-quality and low-cost healthcare becomes available to everyone in the globe.

4.2. Future Works

In this section, we discuss some of the possible future work and potential research opportunities. As it is suggested in the proposed framework of chronic disease operations management, operational planning is the last level of decision-making in chronic disease operations management where the focus is on the operations within each facility. This includes many types of optimization problems such as resource scheduling to find an optimal way to assign healthcare providers to patients in a timely manner, shift assignment, as well as admission and bed scheduling.

In this study, we proposed a deep multi-task learning approach for predicting the required patient workload. Technically, using any other prediction method that can improve the accuracy of the prediction is an improvement of the predictive part of this dissertation. In this research, we tried stacked autoencoders for feature representation. One can use other deep learning approaches. However, the interpretation of the represented features remains as an unknown and needs an extensive amount of research. Furthermore, as it is discussed in this research, due to the similarities between some patients, we can use unsupervised learning methods specifically clustering algorithms to group similar patients together based on their various features such as patient location or comorbidity that may significantly contribute to defining the similarity.

As the number of trips increases with the number of patients, developing automated vehicle assignment, scheduling and dispatching optimization model for vehicles in the patient transportation network can be considered as an essential improvement for SCDDO problem. This way the service providers can enhance their efficiency by minimizing the cost of operation. In addition, we can investigate the vehicle routing problem in the future steps of this research to choose the best route for vehicles in the network and ensure that the riders take the route with minimum cost and can get to their destinations in a timely fashion. Furthermore, developing routing algorithms that can respond to road and weather conditions as well as vehicle breakdowns can be a direction for future research. Moreover, a combination of vehicle routing optimization and ride sharing between patients can be a worthwhile attempt for enhancing this research.

We believe that there are some potential improvements associated with TCDDO problem. We tried to make reasonable assumptions to solve the problem; however, to make the model closer to reality, some constraints can be lifted. Such modifications of assumptions include considering stochastic utilization rate, different efficiency rate for providers in the team and different quality of provided care by each care provider. In addition, one can consider the possibility of switching teams by patients as well as the existence of coordination between care providers among teams. Moreover, developing a dynamic model for patient migration and dynamically updating the patient allocation, and enter and dropout from the system can be helpful to make the model closer to the reality. Furthermore, patient scheduling within the teams can be considered as future research in this area.

As the results of this dissertation suggest, the execution time of the SAA algorithm changes dramatically when the number of scenarios increases. Therefore, we can conclude that an efficient scenario optimization method can be used to reduce the number of scenarios when covering an acceptable range of events. In addition, one can use different optimization techniques to solve the problem such as L-shaped method, progressive hedging algorithm, and meta-heuristic approaches. In addition, by adding various stages of decision-making, the problem can be formulated as a multi-stage stochastic optimization problem. Since the scenario sample size significantly affects the performance of the solution method, one can use parallel algorithms to speed-up the solving process. In that case, due to the quick response of the algorithm, the approach can be more responsive so that decision makers can be provided with real-time solutions for short-term decision-making.

REFERENCES

[1] D. L. Hoyert and J. Xu, “Deaths: preliminary data for 2011,” *Natl Vital Stat Rep*, vol. 61, no. 6, pp. 1–51, 2012.

[2] N. C. of S. Legislatures, “Health Care Safety-Net Toolkit for Legislators.” [Online]. Available: http://www.ncsl.org/documents/health/chronicdtk13.pdf. [Accessed: 20-Jan-2017].

[3] S. L. Norris, R. E. Glasgow, M. M. Engelgau, P. J. Os’Connor, and D. McCulloch, “Chronic disease management,” *Dis. Manag. Heal. Outcomes*, vol. 11, no. 8, pp. 477–488, 2003.

[4] G. Caleb Alexander, J. Kurlander, and M. K. Wynia, “Physicians in retainer (‘concierge’) practice: a national survey of physician, patient, and practice characteristics,” *J. Gen. Intern. Med.*, vol. 20, no. 12, pp. 1079–1083, 2005.

[5] E. H. Wagner, C. Davis, J. Schaefer, M. Von Korff, and B. Austin, “A survey of leading chronic disease management programs: are they consistent with the literature?,” *J. Nurs. Care Qual.*, vol. 16, no. 2, pp. 67–80, 2002.

[6] J. M. Pines, V. Keyes, M. van Hasselt, and N. McCall, “Emergency department and inpatient hospital use by medicare beneficiaries in patient-centered medical homes,” *Ann. Emerg. Med.*, vol. 65, no. 6, pp. 652–660, 2015.

[7] M. Hasselt, N. McCall, V. Keyes, S. G. Wensky, and K. W. Smith, “Total Cost of Care Lower among Medicare Fee-for-Service Beneficiaries Receiving Care from Patient-Centered Medical Homes,” *Health Serv. Res.*, vol. 50, no. 1, pp. 253–272, 2015.

[8] T. Bodenheimer, E. Chen, and H. D. Bennett, “Confronting the growing burden of chronic disease: can the US health care workforce do the job?,” *Health Aff.*, vol. 28, no. 1, pp. 64–74, 2009.

[9] B. W. Ward, J. S. Schiller, and R. A. Goodman, “Peer reviewed: Multiple chronic conditions among us adults: A 2012 update,” *Prev. Chronic Dis.*, vol. 11, 2014.

[10] G. F. Anderson, *Chronic care: making the case for ongoing care*. Robert Wood Johnson Foundation, 2010.

[11] R. DeVol *et al.*, “An unhealthy America: The economic burden of chronic disease,” 2007.

[12] Z. Liang, G. Zhang, J. X. Huang, and Q. V. Hu, “Deep learning for healthcare decision making with EMRs,” in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, 2014, pp. 556–559.

[13] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *arXiv Prepr. arXiv1707.08114*, 2017.

[14] C. Moniz, “Outpatient workload (RVU) predictors: age, gender & beneficiary category,” JOHNS HOPKINS UNIV BALTIMORE MD, 2008.

[15] T. Østbye, K. S. H. Yarnall, K. M. Krause, K. I. Pollak, M. Gradison, and J. L. Michener, “Is there time for management of patients with chronic diseases in primary care?,” *Ann. Fam. Med.*, vol. 3, no. 3, pp. 209–214, 2005.

[16] J. M. Naessens *et al.*, “Effect of multiple chronic conditions among working-age adults,” *Am. J. Manag. Care*, vol. 17, no. 2, pp. 118–122, 2011.

[17] F. E. Turrentine, H. Wang, V. B. Simpson, and R. S. Jones, “Surgical risk factors,

morbidity, and mortality in elderly patients," *J. Am. Coll. Surg.*, vol. 203, no. 6, pp. 865–877, 2006.

[18] R. G. Murphy, "A Primary Care Workload Production Model for Estimating Relative Value Unit Output," DTIC Document, 2011.

[19] D. R. Shah, R. J. Bold, A. D. Yang, V. P. Khatri, S. R. Martinez, and R. J. Canter, "Relative value units poorly correlate with measures of surgical effort and complexity," *J. Surg. Res.*, vol. 190, no. 2, pp. 465–470, 2014.

[20] T. D. Barnes, "Demand Analysis for Proposed Medical Services at the Future Naval Health Clinic Charleston, South Carolina: A Graduate Management Project," DTIC Document, 2006.

[21] D. A. Etzioni, J. H. Liu, M. A. Maggard, and C. Y. Ko, "The aging population and its impact on the surgery workforce," *Ann. Surg.*, vol. 238, no. 2, pp. 170–177, 2003.

[22] P. W. Crane, Y. Zhou, Y. Sun, L. Lin, and S. M. Schneider, "Entropy: A conceptual approach to measuring situation-level workload within emergency care and its relationship to emergency department crowding," *J. Emerg. Med.*, vol. 46, no. 4, pp. 551–559, 2014.

[23] J. E. Chasan, B. Delaune, A. Y. Maa, and M. G. Lynch, "Effect of a teleretinal screening program on eye care use and resources," *JAMA Ophthalmol.*, vol. 132, no. 9, pp. 1045–1051, 2014.

[24] B. Arndt, W.-J. Tuan, J. White, and J. Schumacher, "Panel Workload Assessment in US Primary Care: Accounting for Non–Face-to-Face Panel Management

Activities," *J. Am. Board Fam. Med.*, vol. 27, no. 4, pp. 530–537, 2014.

[25] D. J. Bryce and T. J. Christensen, "Finding the sweet spot: how to get the right staffing for variable workloads: a simulation tool can help hospitals uncover hidden opportunities to reduce costs by optimizing staffing in a way that best reflects demand," *Healthc. Financ. Manag.*, vol. 65, no. 3, pp. 54–61, 2011.

[26] L. Fulton, L. S. Lasdon, and R. R. McDaniel, "Cost drivers and resource allocation in military health care systems," *Mil. Med.*, vol. 172, no. 3, pp. 244–249, 2007.

[27] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Mol. Pharm.*, vol. 13, no. 5, pp. 1445–1454, 2016.

[28] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[29] D. Ravì *et al.*, "Deep learning for health informatics," *IEEE J. Biomed. Heal. informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[30] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, p. 26094, 2016.

[31] J. Sun *et al.*, "Combining knowledge and data driven insights for identifying risk factors using electronic health records," in *AMIA Annual Symposium Proceedings*, 2012, vol. 2012, p. 901.

[32] T. T. Van Vleck and N. Elhadad, "Corpus-based problem selection for EHR note summarization," in *AMIA Annual Symposium Proceedings*, 2010, vol. 2010, p.

817.

[33] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, “Using deep learning to enhance cancer diagnosis and classification,” in *Proceedings of the International Conference on Machine Learning*, 2013, vol. 28.

[34] Z. Liang, G. Zhang, Z. Li, J. Yin, and W. Fu, “Deep learning for acupuncture point selection patterns based on veteran doctor experience of chinese medicine,” in *Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on*, 2012, pp. 396–401.

[35] H.-I. Suk, S.-W. Lee, D. Shen, and A. D. N. Initiative, “Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis,” *Neuroimage*, vol. 101, pp. 569–582, 2014.

[36] C. Hu, R. Ju, Y. Shen, P. Zhou, and Q. Li, “Clinical decision support for Alzheimer’s disease based on deep learning and brain network,” in *Communications (ICC), 2016 IEEE International Conference on*, 2016, pp. 1–6.

[37] F. Li, L. Tran, K.-H. Thung, S. Ji, D. Shen, and J. Li, “A robust deep model for improved classification of AD/MCI patients,” *IEEE J. Biomed. Heal. informatics*, vol. 19, no. 5, pp. 1610–1616, 2015.

[38] H.-I. Suk, S.-W. Lee, D. Shen, and A. D. N. Initiative, “Latent feature representation with stacked auto-encoder for AD/MCI diagnosis,” *Brain Struct. Funct.*, vol. 220, no. 2, pp. 841–859, 2015.

[39] C. Widmer, J. Leiva, Y. Altun, and G. Rätsch, “Leveraging sequence classification by taxonomy-based multitask learning,” in *Annual International Conference on*

Research in Computational Molecular Biology, 2010, pp. 522–534.

[40] K. Zhang, J. W. Gray, and B. Parvin, “Sparse multitask regression for identifying common mechanism of response to therapeutic targets,” *Bioinformatics*, vol. 26, no. 12, pp. i97–i105, 2010.

[41] Q. Liu, Q. Xu, V. W. Zheng, H. Xue, Z. Cao, and Q. Yang, “Multi-task learning for cross-platform siRNA efficacy prediction: an in-silico study,” *BMC Bioinformatics*, vol. 11, no. 1, p. 181, 2010.

[42] F. Mordelet and J.-P. Vert, “ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples,” *BMC Bioinformatics*, vol. 12, no. 1, p. 389, 2011.

[43] D. He, D. Kuhn, and L. Parida, “Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction,” *Bioinformatics*, vol. 32, no. 12, pp. i37–i43, 2016.

[44] Q. Xu, S. J. Pan, H. H. Xue, and Q. Yang, “Multitask learning for protein subcellular location prediction,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 8, no. 3, pp. 748–759, 2011.

[45] M. Alamgir, M. Grosse–Wentrup, and Y. Altun, “Multitask learning for brain-computer interfaces,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 17–24.

[46] C. Widmer, N. C. Toussaint, Y. Altun, and G. Rätsch, “Inferring latent task structure for multitask learning by multiple kernel learning,” *BMC Bioinformatics*, vol. 11, no. 8, p. S5, 2010.

[47] J. Zhou, L. Yuan, J. Liu, and J. Ye, “A multi-task learning formulation for predicting disease progression,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 814–822.

[48] J. Wan *et al.*, “Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer’s disease,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 940–947.

[49] H. Wang *et al.*, “High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer’s disease progression prediction,” in *Advances in Neural Information Processing Systems*, 2012, pp. 1277–1285.

[50] W. Zhang *et al.*, “Deep model based transfer and multi-task learning for biological image analysis,” *IEEE Trans. Big Data*, 2016.

[51] L. Wang, Y. Li, J. Zhou, D. Zhu, and J. Ye, “Multi-task Survival Analysis,” in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 485–494.

[52] Y. Li, J. Wang, J. Ye, and C. K. Reddy, “A multi-task learning formulation for survival analysis,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1715–1724.

[53] M. B. Kursa and W. R. Rudnicki, “Feature selection with the Boruta package,” *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.

[54] U. Stańczyk, B. Zielosko, and L. C. Jain, *Advances in Feature Selection for Data and Pattern Recognition*. Springer, 2018.

[55] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, p. 436, 2015.

[56] Q. V Le, “A tutorial on deep learning part 2: autoencoders, convolutional neural networks and recurrent neural networks,” *Google Brain*, pp. 1–20, 2015.

[57] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in *Advances in neural information processing systems*, 2007, pp. 153–160.

[58] A. Coates, A. Ng, and H. Lee, “An analysis of single-layer networks in unsupervised feature learning,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[59] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. R. Stat. Soc. Ser. B*, pp. 267–288, 1996.

[60] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Ijcai*, 1995, vol. 14, no. 2, pp. 1137–1145.

[61] A. Hogan and B. Roberts, “Occupational employment projections to 2024,” *Mon. Lab. Rev.*, vol. 138, p. 1, 2015.

[62] H. R. D. Gordon, *The history and growth of career and technical education in America*. Waveland press, 2014.

[63] C. for D. C. and Prevention, “The power of prevention: Chronic disease... the public health challenge of the 21st century,” *Atlanta, GA Natl. Cent. Chronic Dis. Prev. Heal. Promot. Centers Dis. Control Prev.*, 2009.

[64] C. Aguwa, M. H. Olya, and L. Monplaisir, “Modeling of fuzzy-based voice of

customer for business decision analytics,” *Knowledge-Based Syst.*, vol. 125, pp. 136–145, Jun. 2017.

[65] J. F. Bard and H. W. Purnomo, “Short-term nurse scheduling in response to daily fluctuations in supply and demand,” *Health Care Manag. Sci.*, vol. 8, no. 4, pp. 315–324, 2005.

[66] J. F. Bard and H. W. Purnomo, “Hospital-wide reactive scheduling of nurses with preference considerations,” *Iie Trans.*, vol. 37, no. 7, pp. 589–608, 2005.

[67] J. F. Bard, D. P. Morton, and Y. M. Wang, “Workforce planning at USPS mail processing and distribution centers using stochastic optimization,” *Ann. Oper. Res.*, vol. 155, no. 1, pp. 51–78, 2007.

[68] M. J. Fry, M. J. Magazine, and U. S. Rao, “Firefighter staffing including temporary absences and wastage,” *Oper. Res.*, vol. 54, no. 2, pp. 353–365, 2006.

[69] K. Davis, M. Abrams, and K. Stremikis, “How the affordable care act will strengthen the nation’s primary care foundation,” *J. Gen. Intern. Med.*, vol. 26, no. 10, pp. 1201–1203, 2011.

[70] D. R. Rittenhouse and S. M. Shortell, “The patient-centered medical home: will it stand the test of health reform?,” *JAMA*, vol. 301, no. 19, pp. 2038–2040, 2009.

[71] M. W. Friedberg, D. J. Lai, P. S. Hussey, and E. C. Schneider, “A guide to the medical home as a practice-level intervention.,” *Am. J. Manag. Care*, vol. 15, no. 10 Suppl, pp. S291-9, 2009.

[72] T. Hoff, W. Weller, and M. DePuccio, “The Patient-Centered Medical Home: A Review of Recent Research,” *Med. Care Res. Rev.*, vol. 69, no. 6, pp. 619–644,

2012.

[73] E. P. C. Kao and M. Queyranne, “Budgeting costs of nursing in a hospital,” *Manage. Sci.*, vol. 31, no. 5, pp. 608–621, 1985.

[74] F. V Louveaux and R. Schultz, “Stochastic integer programming,” *Handbooks Oper. Res. Manag. Sci.*, vol. 10, pp. 213–266, 2003.

[75] S. Sen, “Algorithms for stochastic mixed-integer programming models,” *Handbooks Oper. Res. Manag. Sci.*, vol. 12, pp. 515–558, 2005.

[76] P. Punnakitikashem, J. M. Rosenberger, and D. B. Behan, “Stochastic programming for nurse assignment,” *Comput. Optim. Appl.*, vol. 40, no. 3, pp. 321–349, 2008.

[77] X. Zhu and H. D. Sherali, “Two-stage workforce planning under demand fluctuations and uncertainty,” *J. Oper. Res. Soc.*, vol. 60, no. 1, pp. 94–103, 2009.

[78] Bodur, Merve, and J. Luedtke, “Integrated Service System Staffing and Scheduling via Stochastic Integer Programming,” 2014.

[79] P. Punnakitikashem, J. M. Rosenberber, and D. F. Buckley-Behan, “A stochastic programming approach for integrated nurse staffing and assignment,” *Iie Trans.*, vol. 45, no. 10, pp. 1059–1076, 2013.

[80] J. F. Benders, “Partitioning procedures for solving mixed-variables programming problems,” *Numer. Math.*, vol. 4, no. 1, pp. 238–252, 1962.

[81] M. Khatami, M. Mahootchi, and R. Z. Farahani, “Benders’ decomposition for concurrent redesign of forward and closed-loop supply chain network with demand and return uncertainties,” *Transp. Res. Part E Logist. Transp. Rev.*, vol. 79, pp. 1–

21, 2015.

[82] G. Laporte and F. V Louveaux, “The integer L-shaped method for stochastic integer programs with complete recourse,” *Oper. Res. Lett.*, vol. 13, no. 3, pp. 133–142, 1993.

[83] J. R. Birge and F. V. Louveaux, “A multicut algorithm for two-stage stochastic linear programs,” *Eur. J. Oper. Res.*, vol. 34, no. 3, pp. 384–392, 1988.

[84] S. Trukhanov, L. Ntaimo, and A. Schaefer, “Adaptive multicut aggregation for two-stage stochastic linear programs with recourse,” *Eur. J. Oper. Res.*, vol. 206, no. 2, pp. 395–406, 2010.

[85] K. Kim and S. Mehrotra, “A Two-Stage Stochastic Integer Programming Approach to Integrated Staffing and Scheduling with Application to Nurse Management A Two-Stage Stochastic Integer Programming Approach to Integrated Staffing and Scheduling with Application to Nurse Management,” *Oper. Res.*, vol. 63, no. 6, pp. 1431–1451, 2015.

[86] A. T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier, “Staff scheduling and rostering: A review of applications, methods and models,” *Eur. J. Oper. Res.*, vol. 153, no. 1, pp. 3–27, 2004.

[87] E. K. Burke, P. De Causmaecker, G. Vanden Berghe, and H. Van Landeghem, “The state of the art of nurse rostering,” *J. Sched.*, vol. 7, no. 6, pp. 441–499, 2004.

[88] B. Cheang, H. Li, A. Lim, and B. Rodrigues, “Nurse rostering problems––a bibliographic survey,” *Eur. J. Oper. Res.*, vol. 151, no. 3, pp. 447–460, 2003.

[89] O. EL-Rifai, T. Garaix, V. Augusto, and X. Xie, “A stochastic optimization model

for shift scheduling in emergency departments," *Health Care Manag. Sci.*, vol. 18, no. 3, pp. 289–302, 2015.

[90] H. Fazlollahtabar and M. H. Olya, "A cross-entropy heuristic statistical modeling for determining total stochastic material handling time," *Int. J. Adv. Manuf. Technol.*, vol. 67, no. 5–8, pp. 1631–1641, Jul. 2013.

[91] P. J. H. Hulshof, N. Kortbeek, R. J. Boucherie, E. W. Hans, and P. J. M. Bakker, "Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS," *Heal. Syst.*, vol. 1, no. 2, pp. 129–175, 2012.

[92] S. Nickel, M. Reuter-Oppermann, and F. Saldanha-da-Gama, "Ambulance location under stochastic demand: A sampling approach," *Oper. Res. Heal. Care*, vol. 8, pp. 24–32, 2016.

[93] M. H. Olya, "Finding shortest path in a combined exponential-gamma probability distribution arc length," *Int. J. Oper. Res.*, vol. 21, no. 1, p. 25, 2014.

[94] M. H. Olya, "Applying Dijkstra's algorithm for general shortest path problem with normal probability distribution arc length," *Int. J. Oper. Res.*, vol. 21, no. 2, p. 143, 2014.

[95] P. M. Pardalos *et al.*, "Springer Optimization and Its Applications," 2013.

[96] J. L. Vile, J. W. Gillard, P. R. Harper, and V. A. Knight, "Time-dependent stochastic methods for managing and scheduling Emergency Medical Services," *Oper. Res. Heal. Care*, vol. 8, pp. 42–52, 2016.

[97] M. H. Olya and H. Fazlollahtabar, "Finding Shortest Path in a Combined Exponential-Gamma-Normal Probability Distribution Arc Length," *Adv. Ind. Eng.*

Manag., vol. 3, no. 4, pp. 35–44, 2014.

[98] B. Cardoen, E. Demeulemeester, and J. Beliën, “Operating room planning and scheduling: A literature review,” *Eur. J. Oper. Res.*, vol. 201, no. 3, pp. 921–932, 2010.

[99] S. Choi and W. E. Wilhelm, “On capacity allocation for operating rooms,” *Comput. Oper. Res.*, vol. 44, pp. 174–184, 2014.

[100] A. Jebali and A. Diabat, “A stochastic model for operating room planning under capacity constraints,” *Int. J. Prod. Res.*, vol. 53, no. 24, pp. 7252–7270, 2015.

[101] L. L. X. Li and B. E. King, “A healthcare staff decision model considering the effects of staff cross-training,” *Health Care Manag. Sci.*, vol. 2, no. 1, pp. 53–61, 1999.

[102] B. Maenhout and M. Vanhoucke, “An integrated nurse staffing and scheduling analysis for longer-term nursing staff allocation problems,” *Omega*, vol. 41, no. 2, pp. 485–499, 2013.

[103] G. M. Campbell, “A two-stage stochastic program for scheduling and allocating cross-trained workers,” *J. Oper. Res. Soc.*, vol. 62, no. 6, pp. 1038–1047, 2011.

[104] Z. Chalabi, D. Epstein, C. McKenna, and K. Claxton, “Uncertainty and value of information when allocating resources within and between healthcare programmes,” *Eur. J. Oper. Res.*, vol. 191, no. 2, pp. 529–538, 2008.

[105] B. Liang and A. Turkcan, “Acuity-based nurse assignment and patient scheduling in oncology clinics,” *Health Care Manag. Sci.*, vol. 19, no. 3, pp. 207–226, 2016.

[106] E. Lanzarone and A. Matta, “Robust nurse-to-patient assignment in home care

services to minimize overtimes under continuity of care," *Oper. Res. Heal. Care*, vol. 3, no. 2, pp. 48–58, 2014.

[107] M. C. Villarreal and P. Keskinocak, "Staff planning for operating rooms with different surgical services lines," *Health Care Manag. Sci.*, vol. 19, no. 2, pp. 144–169, 2016.

[108] A. Hassanzadeh, M. Rasti-Barzoki, and H. Khosroshahi, "Two new meta-heuristics for a bi-objective supply chain scheduling problem in flow-shop environment," *Appl. Soft Comput.*, vol. 49, pp. 335–351, 2016.

[109] H. Balasubramanian, S. Biehl, L. Dai, and A. Muriel, "Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments," *Health Care Manag. Sci.*, vol. 17, no. 1, pp. 31–48, 2014.

[110] L. Zhen, "Task assignment under uncertainty: stochastic programming and robust optimisation approaches," *Int. J. Prod. Res.*, vol. 53, no. 5, pp. 1487–1502, 2014.

[111] D. P. Schneider and K. E. Kilpatrick, "An optimum manpower utilization model for health maintenance organizations," *Oper. Res.*, vol. 23, no. 5, pp. 869–889, 1975.

[112] E. P. C. Kao and G. G. Tung, "Aggregate nursing requirement planning in a public health care delivery system," *Socioecon. Plann. Sci.*, vol. 15, no. 3, pp. 119–127, 1981.

[113] M. J. Brusco and M. J. Showalter, "Constrained nurse staffing analysis," *Omega*, vol. 21, no. 2, pp. 175–186, 1993.

[114] D. H. Kropp and R. C. Carlson, “Recursive modeling of ambulatory health care settings,” *J. Med. Syst.*, vol. 1, no. 2, pp. 123–135, 1977.

[115] J. B. Jun, S. H. Jacobson, and J. R. Swisher, “Application of discrete-event simulation in health care clinics: A survey,” *J. Oper. Res. Soc.*, pp. 109–123, 1999.

[116] F. Ben Abdelaziz and M. Masmoudi, “A multiobjective stochastic program for hospital bed planning,” *J. Oper. Res. Soc.*, vol. 63, no. 4, pp. 530–538, 2012.

[117] N. Dellaert, E. Cayiroglu, and J. Jeunet, “Assessing and controlling the impact of hospital capacity planning on the waiting time,” *Int. J. Prod. Res.*, vol. 54, no. 8, pp. 2203–2214, 2016.

[118] S. Fomundam and J. W. Herrmann, “A survey of queuing theory applications in healthcare,” 2007.

[119] T. F. Keller and D. J. Laughhunn, “An application of queuing theory to a congestion problem in an outpatient clinic,” *Decis. Sci.*, vol. 4, no. 3, pp. 379–394, 1973.

[120] D. Fiems, G. Koole, and P. Nain, “Waiting times of scheduled patients in the presence of emergency requests,” *Tech. Rapp. URL http//www. math. vu. nl/koole/articles/report05a/art. pdf,(Accessed 18/12/2012)*, pp. 1–19, 2007.

[121] D. Worthington, “Hospital waiting list management models,” *J. Oper. Res. Soc.*, pp. 833–843, 1991.

[122] J. R. Broyles, “Estimating business loss to a hospital emergency department from patient reneging by queuing-based regression,” in *IIE Annual Conference. Proceedings*, 2007, p. 613.

[123] N. Koizumi, E. Kuno, and T. E. Smith, “Modeling patient flows using a queuing network with blocking,” *Health Care Manag. Sci.*, vol. 8, no. 1, pp. 49–60, 2005.

[124] K. Siddharthan, W. J. Jones, and J. A. Johnson, “A priority queuing model to reduce waiting times in emergency care,” *Int. J. Health Care Qual. Assur.*, vol. 9, no. 5, pp. 10–16, 1996.

[125] R. K. D. Haussmann, “Waiting time as an index of quality of nursing care,” *Health Serv. Res.*, vol. 5, no. 2, p. 92, 1970.

[126] T. H. Taylor *et al.*, “A study of anaesthetic emergency work,” *BJA Br. J. Anaesth.*, vol. 41, no. 1, pp. 70–75, 1969.

[127] A. Shapiro, “Complexity of two and multi-stage stochastic programming problems,” *Tutor. Notes Sch. Ind. Syst. Eng. Atlanta, Georg.*, pp. 30205–30332, 2005.

[128] S. Ahmed, A. Shapiro, and E. Shapiro, “The sample average approximation method for stochastic programs with integer recourse,” *Submitt. Publ.*, pp. 1–24, 2002.

[129] A. Schaefer, “An Overview of Sampling Methods in Stochastic Programming,” 2008.

[130] R. Schultz, “Rates of convergence in stochastic programs with complete integer recourse,” *SIAM J. Optim.*, vol. 6, no. 4, pp. 1138–1152, 1996.

[131] G. Bayraksan and D. P. Morton, “Assessing solution quality in stochastic programs via sampling,” *Tutorials Oper. Res.*, vol. 5, pp. 102–122, 2009.

[132] W. C. Hsiao, P. Braun, E. R. Becker, and S. R. Thomas, “The resource-based

relative value scale: toward the development of an alternative physician payment system,” *JAMA*, vol. 258, no. 6, pp. 799–802, 1987.

ABSTRACT

DATA ANALYTICS AND STOCHASTIC OPTIMIZATION MODELS FOR DECISION SUPPORT IN CHRONIC DISEASE OPERATIONS MANAGEMENT

by

MOHAMMAD HESSAM OLYA

August 2019

Advisor: Dr. Kai Yang

Major: Industrial Engineering

Degree: Doctor of Philosophy

Meeting the complex needs of patients with chronic illness is the single greatest challenge in medical practices. Chronic disease is a prevalent and high-cost issue in the United States healthcare systems. Efficient spending of healthcare funds and better management of healthcare operation costs lead to an enhanced access to high-quality healthcare services and reduces the overall healthcare cost. Thus, in this research, we have proposed a comprehensive framework for chronic disease operations management. Due to uncertainty in patient demand and workload, this framework consists of two predictive and prescriptive analysis phases. In the first phase, we have proposed a deep multi-task learning approach for predicting the required workload of patients. Then in the second phase, we have developed two stochastic optimization models for capacity planning and resource allocation for decision-making in strategic and tactical management levels where the scope of decision-making includes single and multiple facilities, respectively. One of the drawbacks of earlier studies in workload prediction is that the problem is not investigated for multiple facilities where the quality of provided services, equipment and resources used

for provided services as well as diagnosis and treatment procedures may differ even for patients with similar conditions. Besides, the sparsity of chronic disease data is another challenge in workload prediction. To tackle the mentioned issues, we have considered patient-dependent and facility-dependent attributes as well as the relation between them into the proposed model and trained multiple related tasks simultaneously. In addition, we have transformed the data using multiple non-linear transformations through several hidden layers to capture data complexity and sparsity for providing a robust abstraction. The results of this study show that feature representation and training related instances jointly increase the performance of patient workload prediction. Moreover, we have addressed two critical issues in team-based healthcare strategic and tactical planning. The first issue is to determine the optimal number of providers for multiple facilities and eligible patients for pay-to-travel incentives where the demand and location of patients are unknown. The second issue is to minimize the number of different healthcare teams and balance their workload within every single facility. We have developed a stochastic workforce and workload optimization model under various scenarios to address this issue. The result of prescriptive analysis suggests considering the randomness rather than replacing the stochastic variables by their expected value significantly contributes in reducing the overall cost of healthcare and practically enhancing access to care.

AUTOBIOGRAPHICAL STATEMENT

Mohammad Hessam Olya is a Ph.D. candidate at Wayne State University, Michigan. He started his Ph.D. program in Industrial engineering at the department of Industrial and systems engineering at Wayne State University in 2014. He received his bachelor's and master's degrees in industrial engineering from Iran in 2010 and 2012, respectively. Also, he received his second master's degree in computer science from Wayne State University in 2018.

His research interests include data analytics, machine learning, optimization methods and their applications in decision support system development, healthcare, supply chain, and transportation. During his career in WSU, he served as a graduate research assistant and graduate teaching assistant for several courses such as probability and statistics, and operations research.

He is a member of many well-known professional societies and associations in industrial engineering community such as Institute for Operations Research and the Management Sciences (INFORMS), Institute of Industrial and Systems Engineers (IISE). Also, he is the vice president and treasurer of WSUInforms.

He published and submitted articles in well-known journals such as Knowledge-Based Systems, The International Journal of Advanced Manufacturing Technology, and Expert Systems with Applications. Also, his research works are presented in well-known international conferences.