Depression Management Using Electronic Health Record:

Individual Progression Prediction

Weiwei Shang

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2016

Committee:

Shan Liu, Chair

Shuai Huang

W. Art Chaovalitwongse

Program Authorized to Offer Degree:

Industrial and Systems Engineering
University of Washington

Abstract

Depression Management Using Electronic Health Record: Individual Progression Prediction

Weiwei Shang

Chair of the Supervisory Committee:
Assistant Professor Shan Liu
Department of Industrial & Systems Engineering

Mitigating depression has become a national health priority and is the most common mental illness seen in primary care. Due to the complex dynamics of individual's depression trajectory, how to predict the progression of an individual patient's depression has long been an open problem. In this thesis, by using the electronic Patient Health Questionnaire (PHQ)-9 data, a new nature-history model is proposed to provide individual depression prediction, based on which the PHQ-9 score of a new patient at the next time interval can be predicted by using a multivariate nearness approach. The accuracy of the model is further validated under distinct scenarios by using five-fold validation. A simulation-based monitoring system is further established, with which a visit schedule table can be designed for each patient according to the predicted depression level and a given criteria. The analysis offers important insights into depression prediction and management.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I feel very fortunate to have worked with and learned from so many incredible individuals at University of Washington during my M. Sc. study. First and foremost, I wish to express my gratitude to my supervisor, Dr. Shan Liu, who guided me into this interesting area of healthcare modeling and decision making, and walked me through those tough yet fascinating moments over the past year. I have benefited tremendously from her technical insights, vision and enthusiasm. I will always remember and miss those interesting and stimulating meetings we had, where she inspired me to overcome all the difficulties during my research. All I can say is that I could not ask for any more from a supervisor.

I also thank my reading committee members, Dr. Shuai Huang and Prof. W. Art Chaovalitwongse, for their careful reading of this thesis and valuable comments. I also thank Prof. Chaovalitwongse for his outstanding teaching and warm encouragement. Moreover, I also owe my fellow students a great debt for their companionship and friendship. I thank Ying Lin, Jiaqi Huang, Jingyi Lu, Yi Zhou, with whom I spent the most happy leisure time during the past two years in Seattle.

Last but not least, I thank my family, my parents, my little brother, Y. Gao, and all my friends for always supporting me, caring about me and for making my life as enjoyable as my work.

DEDICATION

This thesis is dedicated to all the people that are suffering from depression, and their family who have walked them through the bitterness with love and care.

CHAPTER 1. INTRODUCTION

1.1 Depression Diagnostics

Depression is a state of low mood and aversion to activity that can affect a person's thoughts, behavior, feelings and sense of well-being. It has been reported by the Centers for Disease Control and Prevention (CDC) that about 9% of Americans suffer from depression at least occasionally, which has become the most common mental illness seen in primary care [1]. People with a depressed mood would not only suffer from a number of negative feelings, including sadness, anxiousness, hopelessness, worthlessness, irritability and so forth, but also have difficulties in daily activities such as eating, concentrating, remembering details or making decisions. If depression symptoms are not promptly or effectively treated, the patients may develop psychiatric syndromes such as major depressive disorder. Depressive disorders are associated with functional impairment, decreased productivity, and increased risk for suicide [2-13]. In the United States, around 3.4% of people with major depression commit suicide, and up to 60% of people who commit suicide has depression or another mood disorder [13].

Due to the large population that suffering from depressive moods and the severity of this mental illness, significant clinical efforts have been made to effectively diagnose and monitor the depression symptoms of patients. Since depression ranges in severity from mild, temporary episodes of sadness to severe, persistent depression, it is of great importance for the doctors to accurately identify the level of depression of each patient and design the most appropriate treatment accordingly. At the current stage, it seems to the doctors that talking with the patient may be the most important diagnostic tool, during which they could learn about the patients' daily moods, behaviors, lifestyle habits and family history of depression or other illness. However, most diagnostic screening and monitoring interviews of depression depend on recall of past symptoms, which may be vulnerable to recall bias. Moreover, the growing

number of patients that suffer from depression and go to clinics to ask for advice makes it extremely difficult, if not possible, for the doctors to continue this kind of time-consuming diagnostic and monitoring methodology in the future.

To rule out the recall bias in the diagnostic screening and monitoring interviews as much as possible, a questionnaire-based diagnostic tool, known as Patient Health Questionnaire-9 (PHQ-9) was introduced [14]. The Patient Health Questionnaire (PHQ) is a self-report version of the Primary Care Evaluation of Mental Disorders (PRIME-MD) diagnostic tool for common mental disorders. The PHQ-9 is a brief, 9-item scale that includes only the depression-related items from the PHQ, which is shown in Table 1.1. The PHQ-9 has been validated for use in primary care settings and can be used to make a tentative diagnosis of depression and to monitor depression severity and response to treatment in the past 2 weeks. PHQ-9 has been field-tested in office practice, which is shown to be quick and user-friendly, improving the recognition rate of depression and anxiety and facilitating diagnosis and treatment [14-17].

Table 1.1 Patient Health Questionnaire-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been	0	1	2	3
9. Thoughts that you would be better off dead or hurting yourself in some way	0	1	2	3

PHQ-9 is less reliant on recall using an assessment scale and relevant both to understanding major depression epidemiology and to assessing a possible role for the PHQ-9 as a screening instrument [14-17]. It was proved in [14-16, 18] that PHQ-9 is a responsive and reliable measure of depression treatment outcomes. Suggested treatment actions in response to these various levels of PHQ-9 depression severity are shown in Table 1.2.

Table 1.2 PHQ-9 Scores and proposed Treatment Actions

PHQ-9 Score	Depression Severity	Proposed Treatment Actions
0 to 4	None	None
5 to 14	Mild to Moderate	· Support, community resources and education · May need Antidepressant therapy · Score>9 initiates treatment planning and follow up related to depression · Virtual Psychiatric Guidance
15 to 19	Moderate Severe	· Antidepressant and/or psychotherapy · Score>9 initiates treatment planning and follow up related to depression · Virtual Psychiatric Guidance · Referral to Psychiatry if warranted
20 to 27	Severe	· Antidepressant, Possible augmentation · Regular Follow up · Virtual Psychiatric Guidance · Referral to Psychiatry if warranted

The diagnostic validity of the 9-item PHQ-9 was established in studies involving 8 primary care and 7 obstetrical clinics. PHQ-9 scores > 10 had a sensitivity of 88% and a specificity of 88% for Major Depressive Disorder. Reliability and validity of the tool have indicated it has sound psychometric properties. Internal consistency of the PHQ-9 has been shown to be high. A study involving two different patient populations produced Cronbach alphas of .86 and .89. Criteria validity was established by conducting 580 structured interviews by a mental health professional. Results from these interviews showed that individuals who scored high (≥ 10) on the PHQ-9 were between 7 to 13.6 times more likely to be diagnosed with depression by the mental health professional. On the other hand, individuals scoring low (≤ 4) on the PHQ-9 had a less than a 1 in 25 chance of having depression [14]. In this thesis, PHQ-9 will be adopted as the key variable that accurately represents the depression level of a patient in the proposed model.

1.2 Depression Prediction and Monitoring

Prediction of disease progression can be of significant importance in clinical medicine, which, nevertheless, has always been a difficult challenge existing in the entire field of healthcare. With accurate and effective prediction on the development of a patient's disease, doctors could set up and update the treatment and monitoring schedules in a timely manner and make effective and cost-effective decisions about disease management.

To address this issue, there have been extensive studies on developing approaches to improve the accuracy of predictions in the field of health care, including machine learning, statistical data processing and so forth. Learning methods [19-22] were mostly adopted to investigate the predictive value of genetic variants in genetic disorders. In the statistical analysis of observational data such as propensity score matching (PSM) [16, 23-26], a key concern is that the dimensionality of the observable characteristics is very high in most applications, making it difficult to determine which dimensions should be taken into consideration. Nevertheless, those methodologies are found to be very difficult to implement in practical clinical systems due to the unaffordable complexity.

As for the prediction of depression, the difficulties mainly lie in the complex dynamics and large number of related variables of individual patients' depression conditions. It has been well known that biological, psychological and social factors all play a role in causing depression, which makes it more difficult for the clinician to predict how the disease will develop. A detailed literature review will be provided in Section 2.1. How to develop an accurate and applicable model with EHR data to estimate a patient’s depression progression at the individual level remains a significant challenge.

During the diagnostic assessment conducted by a psychiatrist or psychologist, a mental state examination, which is an assessment of the person’s current mood and thought content, plays a crucial role assisting the psychologist to determine the

severity of the patient. In the past decade, the PHQ-9 has been widely accepted as a standard instrument and proved to be a reliable and accurate assessment in the clinical diagnosis of depression, and has addressed positive contribution in detecting clinical change. A question naturally arises: With the current and past record of PHQ-9 scores of a patient, can the clinician predict what would be his/her score after certain time interval, say two weeks, a month, or two months? If yes, how?

At the current stage of clinical experiences for depression treatment, regular follow-up contacts and symptom monitoring are considered to be very important for each patient. Due to the limited resources, it is desirable to develop a cost-efficient method to assist physicians making an accurate monitoring and scheduling for each patient. A number of research efforts have been made to study this crucial issue, which is summarized in Section 2.2.

However, due to the lack of a statistical learning method using the EHR data to effectively predict the depression trajectory of an individual patient, the development of depression screening and monitoring strategies becomes even more difficult. It is thus of great importance to develop a monitoring system to design the scheduling table for each individual patient according to the predicted trajectory, i.e., severity of his/her future depressive symptoms.

1.3 Thesis Contributions and Outline

This thesis is devoted to establishing a prediction model for the individual patient's depression progression and designing a personalized monitoring scheduling system for the clinics.

First, PHQ-9 has been widely accepted as a standard instrument and proved to be a reliable and accurate assessment in the clinical diagnosis of depression, and has made positive contribution in detecting clinical change. To our best knowledge, little work has been done in the literature to study the depression progression by predicting development of PHQ-9 score overtime. A question naturally arises: With the current

and past record of PHQ-9 scores of a patient, can the clinician predict what would be his/her score after certain time interval, say two weeks, a month, or two months? If yes, how?

Moreover, it is desirable for the doctors to be able to set up an adaptive schedule for each patient seeking depression care. The scheduling should be dependent on the severity of the patients' predicted conditions in the near future. This furthers leads to the second question: With a depression prediction model, how to establish a monitoring scheduling system such that long-term scheduling decisions can be appropriately made?

To answer the first question, the modeling methodology in [2] is further extended to study the prediction of depression progression. In particular, the original EHR data is analyzed and processed to generate the data set to be used in this thesis in Chapter 3. A nature-history model is established in Chapter 4 to predict the PHQ-9 score development, i.e., the level of depression severity, for each individual patient. By using B-spline to fit the PHQ-9 observational data of each patient in the dataset, time triplets of PHQ-9 scores with a standard time interval are obtained. The model stratifies the time triplets by a number of variables, including patients' age, gender, visiting location, item 9 and the Carlson comorbidity score. To predict the PHQ-9 score development of a new patient, a multivariate nearness measure method is adopted. In particular, the model would search the entire database to find the most similar data points according to some given criteria, based on which the progression PHQ-9 score of a new-arrival patient in the following time interval can be estimated.

The accuracy of the predictive ability of the natural-history model is further examined in Chapter 5. In particular, five-fold cross-validation will be used to validate and assess the predictive ability of our natural-history model. The accuracy of the proposed model on long-term prediction and regular-timed observations prediction are studied, respectively.

To address the second question, the prediction of PHQ-9 score development of individual patients is further applied into the decision making process of monitoring and scheduling. In particular, a simulation-based patient monitoring system is introduced in Chapter 6. Upon a patient finished the PHQ-9 for two times, the proposed nature-history model can generate a predicted PHQ-9 score in the next regular time interval for him/her. According to the value of the predicted score, the monitoring system would suggest an appropriate scheduling time for the patient's next interview. This monitoring system can also be used for the assessment of treatment outcomes. The clinician could obtain useful information from the system to validate how effective the current treatment is and determine how to adjust it in the next course.

CHAPTER 2. LITERATURE REVIEW

Due to the high prevalence of depression and its profound impact on personal and family lives, a great deal of efforts from both clinical care practice and research have been made to better understand, control and cure this mental health problem. In this chapter, a detailed literature review will be presented, mainly in two aspects, depression prediction in Section 2.1 and depression screening and monitoring in Section 2.2.

2.1 Depression Prediction

It has been long observed that depression is a complex phenomenon due to the interaction of on a large number of factors including biological, psychological, social and so forth. For instance, epidemiological evidence suggests that the prevalence of major depressive disorder declines with age [27], yet may increase again at older ages [28-30]. Previous research also suggested that not everyone is progressing in an identical way. Mean levels of depressive symptoms were shown to differ by sex [31, 32], ethnicity [33], and educational level [34, 35]. Feng, et. al. [36] proposed an exploratory statistical approach to combine all the available information about an individual's depression, and represent it with one single value to simplify this complicated system.

The complex nature of depression makes it difficult, if not impossible, to predict how a patient's prediction would change over time. Most existing studies have been focused on predicting the effect of personal factors on patients' progression of depressive symptoms. Gunn, et. al. [37] studied how the personality measures affect the prediction of depression treatment outcomes. It was shown that baseline personality is crucial to predict the development of depression of individual patients. It was suggested in [38] that individuals who tend to both make negative inferences from negative life events and then constantly activate these negative interpretations

through rumination are at particularly high risk for developing episodes of major depression and for experiencing depressive episodes with longer duration.

The emergence of electronic health record (EHR) data in health system provides a possibility to develop statistical prediction models. For instance, The Patient Health Questionnaire (PHQ)-9, as introduced in Section I, is a self-administered questionnaire that includes 9 multiple-choice questions to assess depression levels. With the expanded use of EHR, quite a few health systems can now administer and store longitudinal PHQ-9 results for a large number of patients. It was revealed in [37, 38] that the dynamics in individual's depression trajectory became more sophisticated to estimate and interpret due to the widely reported heterogeneity of such dynamics in the population. Moreover, the irregular or sparse visits of most of the patients with depressive symptoms further results in statistical challenges to the PHQ-9 records in the current EHR dataset. Existing statistical methods [39, 40] to classify depression trajectories, nevertheless, may not be adequate to address these difficulties. To the best of our knowledge, how to establish a model that can effectively analyze the collected heterogeneous depression trajectories of a given dataset and further use the knowledge to predict the depression progression or trajectory of any new patients still remain largely unexploited.

2.2 Depression Screening and Monitoring

Depression screening and monitoring in primary care yields high numbers. The U.S. Preventive Services Task Force is updating their 2016 guidelines that recommend, “Screening should be implemented with adequate systems in place to ensure accurate diagnosis, effective treatment, and appropriate follow-up for both adult population and adolescents.” [41, 42].

To tackle this issue, a number of screening and monitoring strategies were proposed in the literature. In [43], an online tool was introduced for patients to self-monitor depressive symptom severity and adverse effects, including suicidal

ideation and mania symptoms. Typically, a patient would be asked to complete a brief online set of questions periodically at multiple times, based on which the graph of mood scores over time could be generated with guideline recommendations for consultations. Noorden, et. al. [44] adopted multivariable Cox regression models to predict remission and response of patients, and adjusted for clinical and demographic characteristics. Sacks, et. al. [45] further investigated how patient activation, a measure of individuals' knowledge, skill, and confidence for managing their health, affects their depression remission and treatment response. It was found in [45] that more activated patients show greater improvement in depressive symptoms over one year of monitoring.

Research studies further considered the efficiency and effectiveness of the depression screening programs. Moore, et. al. [46] reported a secondary analysis from a randomized controlled trial comparing two approaches to the management of those with mild to moderate depression in primary care. Valenstein, et. al. [47] studied the cost of annual and periodic screening for depression and recommended that health care organization should improve the quality of depression treatment in their primary care clinics before implementing screening. O'Conner, et. al. [48] showed that depression screening programs without substantial staff-assisted depression care supports were unlikely to improve depression outcomes. Rush, et. al. [49] proposed a monitoring system to compare the acute and long-term outcomes of depressed outpatients where multiple treatment steps were included.

With the rapid expansion of EHR data, more and more attention has been paid to study how to develop statistical learning method to exploit the unknown characteristics of depression development and decision-making capabilities hidden in the collected big data. However, the difficulties in analyzing the collected heterogeneous depression data make it difficult to further propose depression screening and monitoring strategies by using EHR data for the intermediate timeframe (2-5 years). The 2-5 years' time window, nevertheless, is most clinically relevant for screening and treatment follow-up in real-world practice. In summary, how to develop

a cost-effective monitoring system that can be implemented in clinical practice is a operational challenge.

CHAPTER 3. PRELIMINARY ANALYSIS

In this chapter, the EHR data that used in this thesis will be summarized. This chapter is organized as follows. Section 3.1 introduces the dataset formulation and description. The original dataset is then processed for further analysis, which is presented in Section 3.2.

3.1 Data Description

The depression natural-history simulation model uses electronic health records of patients taken the depression questionnaire survey PHQ-9, which was proposed by the primary care evaluation of mental disorders (PRIME-MD) [3]. The data contains the following variables as listed in Table 3.1.

Table 3.1. Variables in the EHR data

ID	person-level ID
phqnbr	a sequence number for observations within that person (1 for that person's first observation, 2 for the second, and so on)
dayssincefirstphq	number of days for this observation in relation to the first observation for this person (takes value 1 for each person's first observation and so on)
daystonextphq	the number of days between this observation and the next one in the sequence for this person
txstatus	person's treatment status at the time of this observation (none = no treatment in past 5 years including on date of phq, past only = treatment in past 5 years but not in past 180 days including index date, new = no treatment in past 180 days but some treatment started on index date, ongoing = treatment in past 180 days)
phqscore	Phq-9 scores at time of the observation
qx9	item 9 of the PHQ9 regarding suicidal ideation.
vistype	P=primary care, M=mental health specialty, O=other
phqagecat	person's age at the time of observation (2=18 to 29, 3=30-44, 4=45-64, 5=65+)
female1	1 indicates female
charlson	person's the Carlson comorbidity score at time of the observation

3.2 Data Analysis

We first excluded patients with incomplete data, including those with no record of the PHQ-9 score at any observation time, and those were seen less than six times. As to the six-time observation constraint, we refer to the patients with at least six observations based on a standard time interval of two weeks, which means that if

there exists more than one observation in each time interval, i.e., two weeks, we average the PHQ-9 scores of these observations and regard them as one data point.

Moreover, it should be pointed out that the focus of our research is to make individual progression prediction for the patients with ongoing treatment in the past 180 days starting from baseline. Therefore, by considering this ongoing treatment group instead of the entire database, we aim to better examine the performance and effects of depression treatment.

Our final research sample consists of 3159 unique patients in total. For each patient, our data set contained his/her demographic and clinical data, including the PHQ-9 scores at time of the observation, the visit location type when the survey was taken, the person's age at the time of observation, the gender of the patients, the item9 of the PHQ-9 regarding suicidal ideation, and the person's Carlson comorbidity score at the time of observation.

Table 3.2. Data Summary

	Num_obs	Ave_PHQ9	Ave_item9	Ave_carlson
Mean	10.57	11.43	0.43	0.73
Min	6	0	0	0
Max	112	26.57	3	8.71
Median	9	11.33	0.19	0.11
Age	Num_patient	Percentage		
18-29	292	9.24%		
30-44	813	25.74%		
45-64	1532	48.50%		
65+	522	16.52%		
Gender	Num_patient	Percentage		
Female	2237	70.81%		
Male	922	29.19%		
Location	Num_patient	Percentage		
Primary care	70	2.22%		
Mental health	2987	94.56%		
Other	102	3.23%		

The data summary of these 3519 patients is shown in Table 3.2. Here Item9 refers to as the 9th item of the PHQ-9 survey, which indicates the suicidal ideation of a patient to some degree. As to the number of observations, the range is from 4 to 112, which is a relatively large gap. Age is aggregated into four categories of 18-29, 30-44, 45-64, and older than 64. A close look shows that the majority of our subjects in this

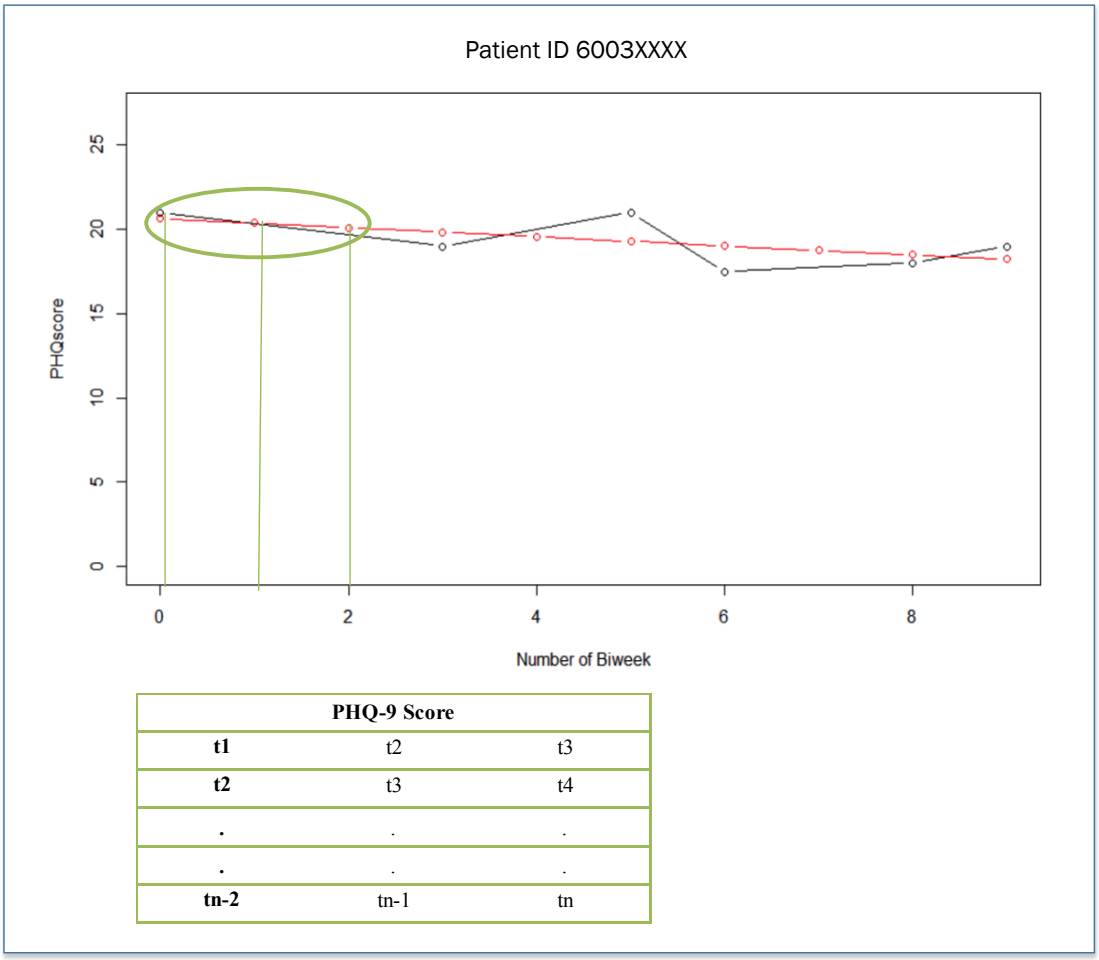
study are aged from 45 to 64 and our database contains more females than males. As for the variable of location type, it can be clearly seen that most of these PHQ-9 surveys were completed at the mental health institutions.

CHAPTER 4. THE NATURE-HISTORY MODEL

In this chapter, a stochastic depression natural-history model will be proposed to predict the PHQ-9 score of an individual patient. This chapter is organized as follows. Section 4.1 briefly introduces how to generate data over time for an new patient with PHQ-9 scores. Section 4.2 demonstrates the ideal prediction for a new patient by adopting the proposed model. A multivariate nearness measure approach is further studied in Section 4.3 to show the prediction process of the development of a patient's PHQ-9 score.

4.1 Disease Progression in a New Patient

We first take a look at how to model PHQ-9 scores trajectory of each patient. By using B-splines, interpolations of additional PHQ-9 scores are simulated between the existing PHQ-9 score in the EHR data. For each patient, regular-interval sampled, complete longitudinal history of PHQ-9 score can then be generated. Based on the spline of PHQ-9 scores, each set of 3 sequential PHQ-9 values is selected and converted into a time triplet (t_1-t_2-t_3 triplet) with a standard time interval of two weeks, as illustrated in Figure 4.1. Then, the model stratifies the time triplets by patients' age, gender, visiting location, item 9 and the Carlson comorbidity score. An important note about this matching strategy is that for visiting location, item 9 and the Carlson comorbidity score, we take the average value for each patient to establish our database, which may be biased since the value of this variable changes when the patients take this survey from one time to the next.



PHQ-9 Score		
t1	t2	t3
t2	t3	t4
.	.	.
.	.	.
tn-2	tn-1	tn

Figure 4.1. Interpolation of training data with B-splines

For training patients, the model creates a B-spline to fit their PHQ-9 scores. This figure shows one randomly selected patient. In the top panel, the red line represents the B-spline for that specific patient and the black line connects the real observations of that patient, where black circles represent the observed (actual) PHQ-9 score values. With a standard time interval of 2 weeks, these curves provide a number of points to sample from. Each spline is then decomposed into time triples of 3 consecutive PHQ-9 scores. Then the model merges these triplets with other variables from same times, for instance, gender, age, location, items 9 score, and Carlson comorbidity score. These triplets form the training database.

4.2 Ideal Prediction for a New Patient

Under ideal conditions, to predict the progression of disease in a patient with PHQ-9 score at current time t, the disease progression can be predicted by combining his/her PHQ-9 score at the current time (*t*) and the previous time (*t*-1). The model searches the entire list of patients within the same gender, age, visiting location, item 9 and the Carlson comorbidity score to find the 1st patient who is similar to the new patient in terms of PHQ-9 score. After this similar patient is found from the database, our model assigns this patient's t_3 PHQ-9 score to the new patient's t+1 PHQ-9 value. The detail of this searching process is described in Figure 4.2. Under real situations, this procedure should be achievable as long as we ask our patients to take the PHQ-9 survey every two weeks. However, using the current EHR dataset, we don't have data drawn at regularly specified intervals, therefore, couple of validation approaches are proposed in Section 5, which are used to measure how accurately the model performs in real practices.

ID	PHQ-9			Age	Gender	Location	Item9	Carlson
	t-1	t	t+1					
Index	7	11	?	2	1	MH	3	1

DATABASED

ID	PHQ-9			Age	Gender	Location	Item9	Carlson
	t-1	t	t+1					
6006XXXX	6	7	8	5	1	MH	3.5	1,6
6000XXXX	6	8	3	4	0	MH	0	0
6007XXXX	0	4	4	1	0	PC	1	0.5
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
6006XXXX	7.5	10	13	2	1	MH	3.5	1
.	.	.	.	.	.	.	.	.

Figure 4.2. Illustration of the search procedure for the similar patient to the new patient

A new patient with time *t*-1 and t PHQ-9 scores is shown in the above figure and the searching routine will estimate the value for time *t*+1. The model searches the whole database to find the patient who is in the same age group, same gender, same location, and who has the most similar PHQ-9 scores, item9 value and Carlson comorbidity score to the new patient. Once this most similar patient is found, the *t*+1 PHQ-9 score of this patient will be used as the new patient's *t*+1 score value.

4.3 Multivariate Nearness Measure

When searching for the most similar patient to the new patient from our database, a level-based search approach is proposed. First, for the input variables of age, gender, and visit location, our model only consider those patients who have exactly matched information with this new patient. As to the input variables of average item9 and average Carlson, our model consider them to be “near” only if they fall within predefined upper and lower limits. These thresholds of average item9 and average Carlson are 0.55 and 1.23, respectively, which are the standard deviations for each of them. Second, for the PHQ-9 value, the model defines that a patient is considered as "near" to the new patient if and only if the difference between its PHQ-9 score and the new patient's is lower than 2 at time t-1 and t, respectively. It can be written as

$$|P_{t-1}^{i} - P_{t-1}^{new}| \leq 2 \;\&\; |P_{t}^{i} - P_{t}^{new}| \leq 2$$

After all the patients satisfying the criteria of “similar” are found, the model then selects the one whose PHQ-9 scores of time t-1 and t are closest to the new patient. In particular, it chooses the one that has the minimum value S, the sum of the absolute differences between values of the PHQ-9 scores of the new patient and the matched patient at time t-1 and t, which is given by

$$S = |P_{t-1}^{i} - P_{t-1}^{new}| + |P_{t}^{i} - P_{t}^{new}|$$

In this case, it can be always guaranteed that the optimal patient which has the most similar PHQ 9 scores is chosen. Then, the t+1 PHQ-9 score of this most similar patient will be adopted as the new patient’s t+1 score value.

CHAPTER 5. MODEL VALIDATION

In this chapter, the accuracy of the predictive ability of the natural-history model proposed in Chapter 4 will be examined. This chapter is organized as follows. Section 5.1 introduces the five-fold cross-validation methodology. Section 5.2 presents a validation approach for long-term prediction and the accuracy of the proposed model in the long-term prediction scenario. The validation is further extended to the regular-timed prediction case, which is studied in Section 5.3.

5.1 Five-fold Cross-validation

In this chapter, five-fold cross-validation will be used to validate and assess the predictive ability of the natural-history model proposed in Chapter 4. Cross-validation is a well-known model validation technique for assessing how the results of a statistical model will generalize to an independent data set, which is widely used when the goal of the model is predication and researchers want to estimate how accurately the model performs in practice. The fundamental idea is to define a validation dataset, which is selected from the known database such that the progression of the samples in it is already known, to "test" the model in order to shed more light on how the model will generalize to an unknown dataset.

In this chapter, the original data sample with 3159 subjects is randomly split into five parts with equal number of samples, among which one subsample is taken out from the database and used as validation data. This small proportion of validation data will be used to examine the model accuracy and the rest 80% samples will be served as training data to be used as the matching database. To reduce variability, this validation process repeats five rounds to make sure that each of these five subsamples is used exactly once as the validation data. The estimation result could be obtained by averaging the validation results over the five rounds. It is important to note that for the validation patients, the model creates perfectly fitted lines by connecting all the observations for each patient. By doing this, we assume that connecting dots is the

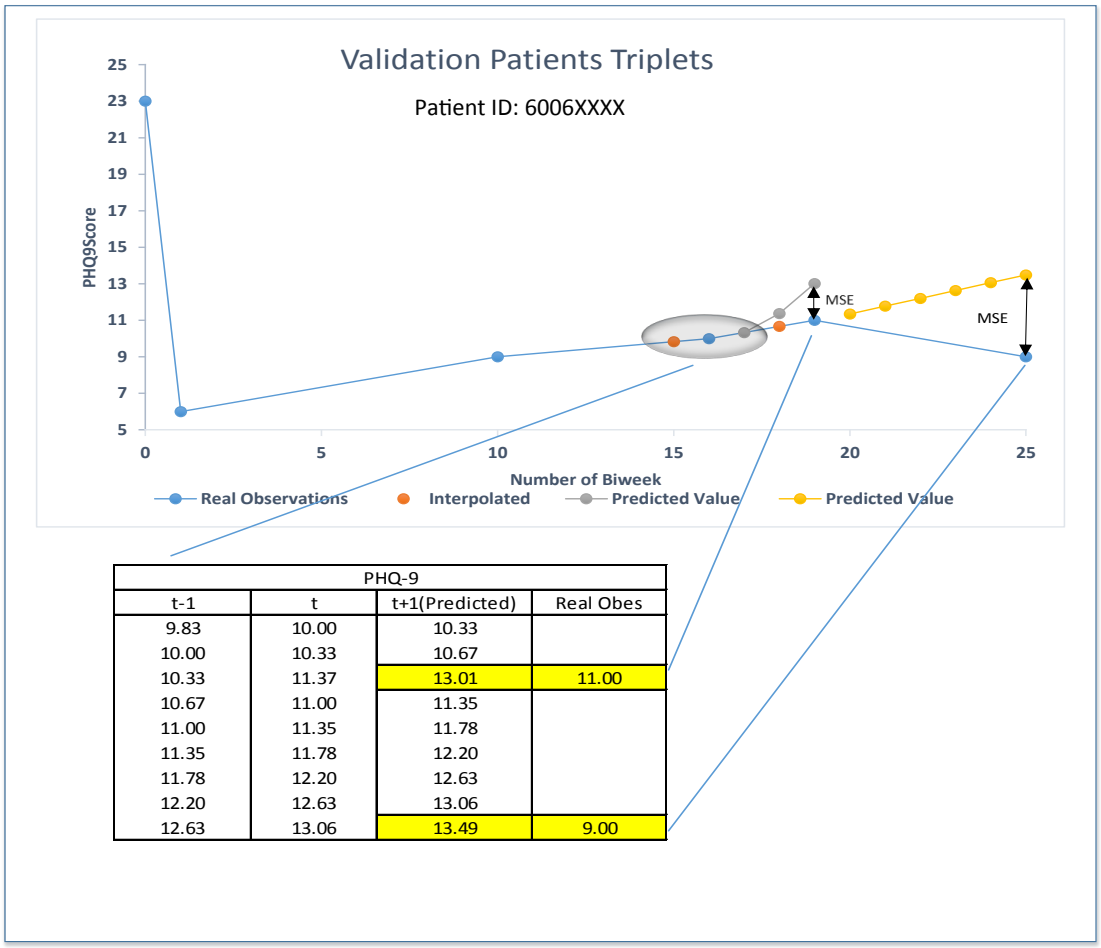
best way to find the "big picture" in a mass of data and reflects the true information patterns behind these data.

5.2 Long-term Prediction

In this section, a validation approach used for long-term prediction is proposed, which aims to predict the PHQ-9 score of a patient after a random time period based on his/her current observations. To verify the accuracy of the proposed model, in particular, we predict the PHQ-9 score of a patient's i^{th} observation basing on the generated triplet of his $(i-1)^{th}$ observation, and compare the predicted value to the actual observed one. In the following, we will show the procedure to make the predication and the accuracy of the proposed model in this scenario, respectively.

5.2.1 Procedure Description

Firstly, the model creates perfectly fitted lines by connecting all the observations for each patient in the validation dataset, shown as the blue line in Fig. 5.1. For each patient, starting from their second real observation, we go one previous period back to form a pair of points, and these two points are served as the PHQ-9 values at t-1 and t. With these two points, we use the proposed model to find the most similar patient from the training database, and updates the PHQ-9 value of t+1 as the corresponding value of the most similar patient, as shown in Fig. 5.2. Then we use the PHQ-9 values of t and t+1 as the validation data, and updates the PHQ-9 value of t+2 similarly, until the time point reaches that of the next real observation of this patient. Then, the mean square error (MSE) will be calculated as the square of the difference between the real observation and the predicted value. Detailed procedure is visualized in Figure 5.1, where the last two observations of patient 6006XXXX is adopted as an example to demonstrate the procedure. It will repeat multiple times for each real observation of each validation patient, which is determined by the time interval between two continuous observations.



PHQ-9			
t-1	t	t+1(Predicted)	Real Obes
9.83	10.00	10.33	
10.00	10.33	10.67	
10.33	11.37	13.01	11.00
10.67	11.00	11.35	
11.00	11.35	11.78	
11.35	11.78	12.20	
11.78	12.20	12.63	
12.20	12.63	13.06	
12.63	13.06	13.49	9.00

Figure 5.1. Interpolation of validation data with connecting observations

In the top panel, the solid blue line represents the perfectly fitted line for that specific patient. With a standard time interval of 2 weeks, these curves provide a number of points to sample from. The blue circles represent the observed (actual) PHQ-9 score values while the orange circles represent the interpolations for the two previous periods. Based on these pair of points, our model will search for the most similar patient from the training database, such that the model keeps updating the value of t3 until it meets a time point where a real observation exists.

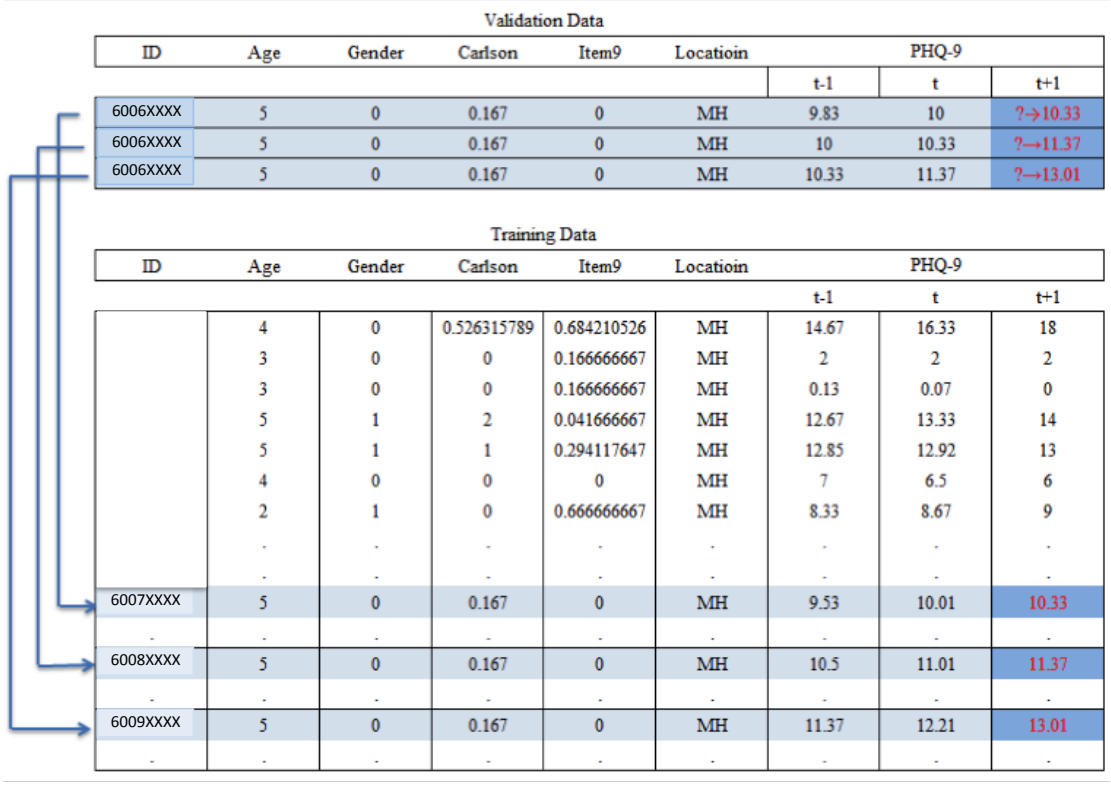


Figure 5.2. Searching for the match in long-term period

A new patient with time t-1 and t PHQ-9 scores is shown in above and the searching routine will provide the value for time t+1. The model searches the whole training database to find the most similar patient who is in the same age group, same gender, same location, and who has similar PHQ-9 scores, item9 value and Carlson comorbidity score to the new patient. Once this similar patient is found, the t+1 PHQ-9 score of this similar patient will be used as the new patient's t+1 score value. Our model will keep searching and updating the value of t+1 in the validation data set until we meet up with a time point, where a real observation exists.

5.2.2 Results and Discussions

To verify the accuracy of the proposed model in terms of long-term prediction, the predictions of last two real observations for each patient basing on their former observations are made for each validation group. The validation results of long-term prediction are summarized in Table 5.1. Overall, 9.34% of validations are shown to be unable to find any sample in the training data that satisfy the given criteria. For those successfully predicted ones, an average 6.38 root-mean-square deviation (rMSE) is

obtained, which is a comparatively large error since an interval of 5 points is considered as a different level in the PHQ-9 depression severity definition. It indicates that the model has its limitations in terms of long-term prediction.

Table 5.1. Prediction results of long-term prediction

Long-term period next real observation prediction					
ValidationGroup	MSE	rMSE	Find	NotFind	OverallNum
1	50.04	7.07	1155	109	1264
2	45.71	6.76	1150	114	1264
3	44.35	6.66	1145	119	1264
4	47.43	6.89	1140	124	1264
5	45.8	6.77	1138	124	1262
Percent of NotFind	9.34%				
Overall MSE	46.67				
Overall rMSE	6.38				

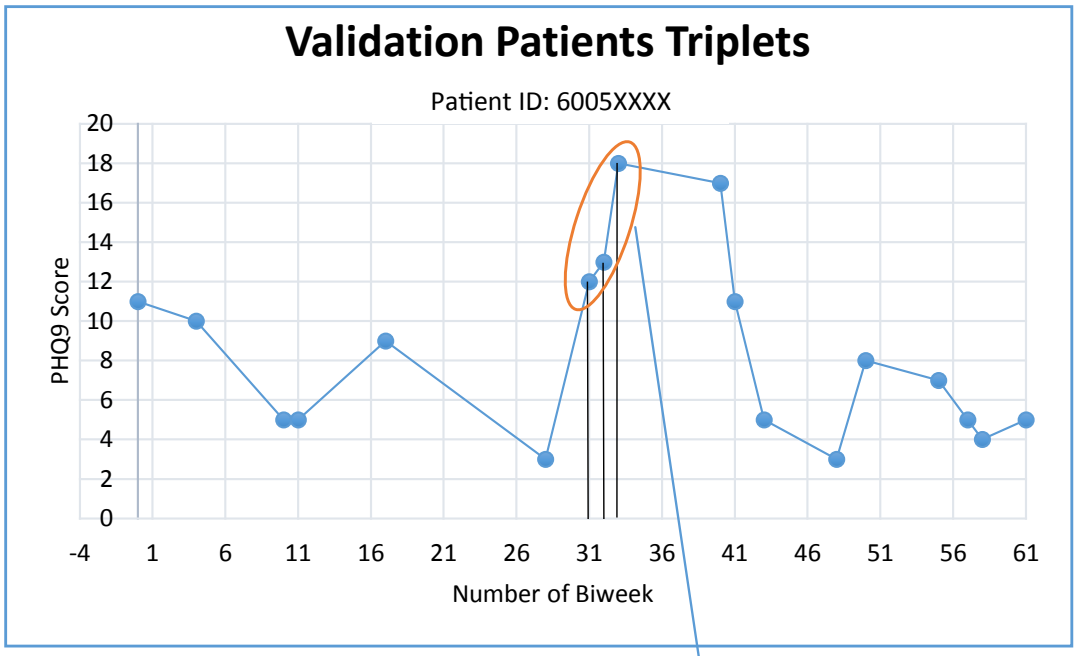
5.3 Regular-timed observations prediction

5.3.1 Procedure Description

In Section 5.1, it can be easily seen that the validation points may include PHQ-9 scores obtained by interpolation rather than real observations. In this case, it may introduce additional errors into the prediction since the interpolated values are already approximated values.

To improve the validity of the validation data and rule out as much interfering variables as possible, in this subsection, the validation dataset is only constituted of regular-timed triplets, which represents that we are only interested in the patients who have one or more three-consecutive real observations drawn at regularly specified intervals. The training dataset, on the other hand, consists of all the triplets of spline estimated PHQ-9 scores. Figure 5.3 illustrates the observation curve of patient no. 6005XXXX as an example, where a continuous observation at time slots 31-32-33 can be found as a set of validation data. In this case, we can guarantee that all the PHQ-9 scores of the validation dataset are real observations. For all of these

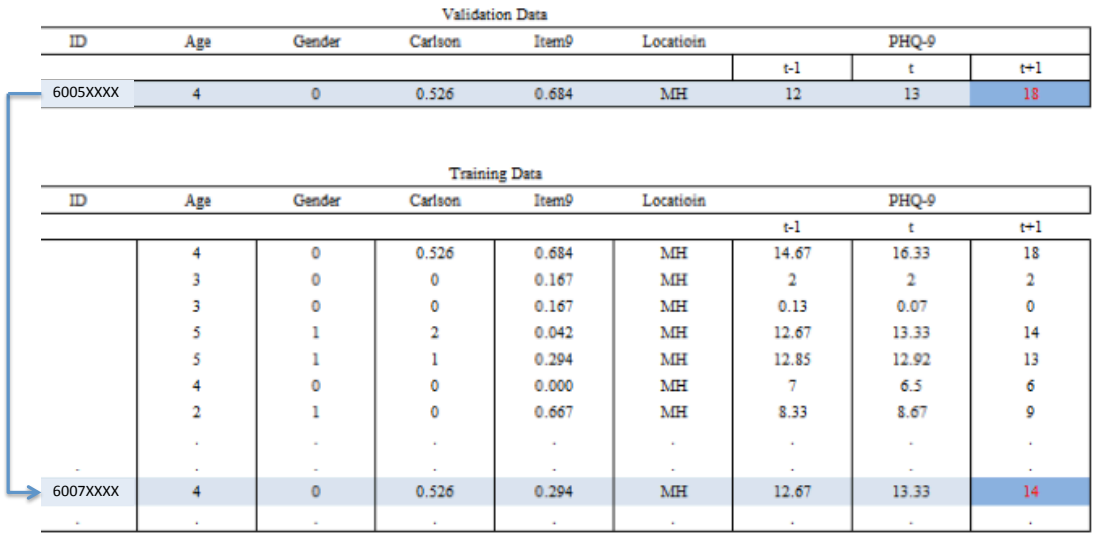
validation data triplets, we use the first two points to predict the third one using the proposed model as illustrated in Figure 5.4, and obtain the average MSE and rMSE.



PHQ9 Score		
t-1(31)	t(32)	t+1(33)
12	13	18

Figure 5.3. Illustration of validation patients: triplets of Patient ID 6005XXXX

For validation patients, the model connects dots to fit their PHQ-9 scores. In the top panel, the blue line represents the connecting dots line for that specific patient and the blue circles represent the observed (actual) PHQ-9 score values. With a standard time interval of 2 weeks, these curves provide a number of points to sample from. Each spline is then decomposed into time triples of 3 consecutive PHQ-9 scores and our model will search for the triplets where three points are all real observations instead of estimated values. And these regular-timed triplets will be used as validation patients to estimate the accuracy of this model. Additionally, the model merges these triplets with other variables from same times, for instance, gender, age, location, items 9 score, and Carlson comorbidity score.



Validation Data

ID	Age	Gender	Carlson	Item9	Locatioin	PHQ-9		
						t-1	t	t+1
6005XXXX	4	0	0.526	0.684	MH	12	13	18

Training Data

ID	Age	Gender	Carlson	Item9	Locatioin	PHQ-9		
						t-1	t	t+1
	4	0	0.526	0.684	MH	14.67	16.33	18
	3	0	0	0.167	MH	2	2	2
	3	0	0	0.167	MH	0.13	0.07	0
	5	1	2	0.042	MH	12.67	13.33	14
	5	1	1	0.294	MH	12.85	12.92	13
	4	0	0	0.000	MH	7	6.5	6
	2	1	0	0.667	MH	8.33	8.67	9
	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
6007XXXX	4	0	0.526	0.294	MH	12.67	13.33	14
.	.	.	.	.	.	.	.	.

Figure 5.4. Searching for the match for the regular-timed observations.

The triplet with three real observations is shown in above and the searching routine will provide the value for time t+1. The model searches the whole database to find the most similar patient who is in the same age group, same gender, same location, and who has similar PHQ-9 scores, item9 value and Carlson comorbidity score to this new patient. Once this optimal similar patient is found, the t+1 PHQ-9 score of this similar patient will be used as the new patient's t+1 score value.

5.3.2 Results and Discussions

Similar to Section 5.1.2, experiments are repeated for each validation group and the corresponding prediction results are summarized in Table 5.2. Since the number of valid three continuous observations varies from group to group, the total number of validation data for each group is drastically different compared to that for long-term prediction given in Table 5.1. Moreover, we can clearly see from Table 5.2 that the percentage of data that cannot find a similar patient in the training data is 27.22%, which is much higher than the long-term prediction case. By taking a closer look at the PHQ- 9 scores of three continuous observations, we find that a fierce fluctuation can be found between two continuous observations with high probability. In the training data, however, since all the triplets are generated using B- splines, which smooth out the possible fluctuations as shown in Fig. 4.1. For those largely fluctuating observations, therefore, it is of increasing probability that the model

cannot find a similar patient for it, leading to a much higher percentage of NotFind. It indicates the spline method may significantly affect the accuracy of the proposed model, which requires further investigations in the future. For those successfully predicted ones, the average rMSE is 4.68, which is much improved comparing to that of the long-term prediction. It implies that the proposed model provides more accurate predictions in short-term scenarios.

Table 5.2. Results of regular-time prediction

Regular-time real observations prediction					
ValidationGroup	MSE	rMSE	Find	NotFind	OverallNum
1	12.85	3.58	330	108	438
2	24.29	4.93	208	94	302
3	24.87	4.99	244	92	336
4	24.30	4.93	204	84	288
5	28.02	5.29	209	69	278
Percent of NotFind	27.22%				
Overall MSE	21.90				
Overall rMSE	4.68				

By comparing the prediction errors in Table 5.1 and 5.2, the proposed model may be more suitable for short-term predictions. Based on the analysis, plenty of insights can be provided to the prediction of depression progression in clinical treatments. First of all, we could predict the onset of major depression in the immediate future, which is of great importance to prevent depression-related suicides; secondly, we could provide decision aid on treatment to a patient according to the predicted results. In addition, the inaccuracy of other prediction methods may help us to find undiscovered variables that affect the patients' depression conditions, which would help in further improving the model.

CHAPTER 6. SIMULATION-BASED MONITORING

In this chapter, we will establish a simulation-based monitoring scheduling system based on the prediction model proposed in Chapter 4. This chapter is organized as follows. Section 6.1 presents the system description and the detailed steps of the monitoring and scheduling process.

6.1 System Description

In this section, the detailed description of the monitoring scheduling system will be provided. The basic idea is to design the visiting interval rules according to the value of the predicted PHQ-9 score of the patient at the next time slot, i.e., two weeks. Note that in the simulation system, the PHQ-9 scores of new patient and matching dataset are all based on spline estimates as used in the prediction model in Chapter 4. Fig. 6.1 illustrates the flow chart of the system.

The scheduling of an individual patient for simulation monitoring includes the following steps:

a. Since the patients in the dataset did not visit their doctors at regular interval, the original PHQ-9 scores are interpolated using B-spline, where the time slot is set to be two weeks. Therefore, each patient would be stored with interpolated PHQ-9 scores at each bi-week time slot during their entire follow-up period in the simulation monitoring system.

b. Recall that in the natural-history model, the prediction of PHQ-9 score at the next time slot, i.e., two weeks later, needs to be provided with two PHQ-9 scores of the patient according to the time-triplets methodology. In the practical scenarios, therefore, each patient will be asked to visit the clinic at two times to complete the PHQ-9 test, denoted as time slot t_0 and t_0+1. In the simulation scheduling system, correspondingly, we will select the first two interpolated values of PHQ-9 score of the patient as the first two observational data of the real PHQ-9 scores in clinics.

c. With these two observed PHQ-9 scores and all the other variables of the patient, the natural-history model could be applied to predict the patient's PHQ-9 score at the next time interval, i.e., two weeks later.

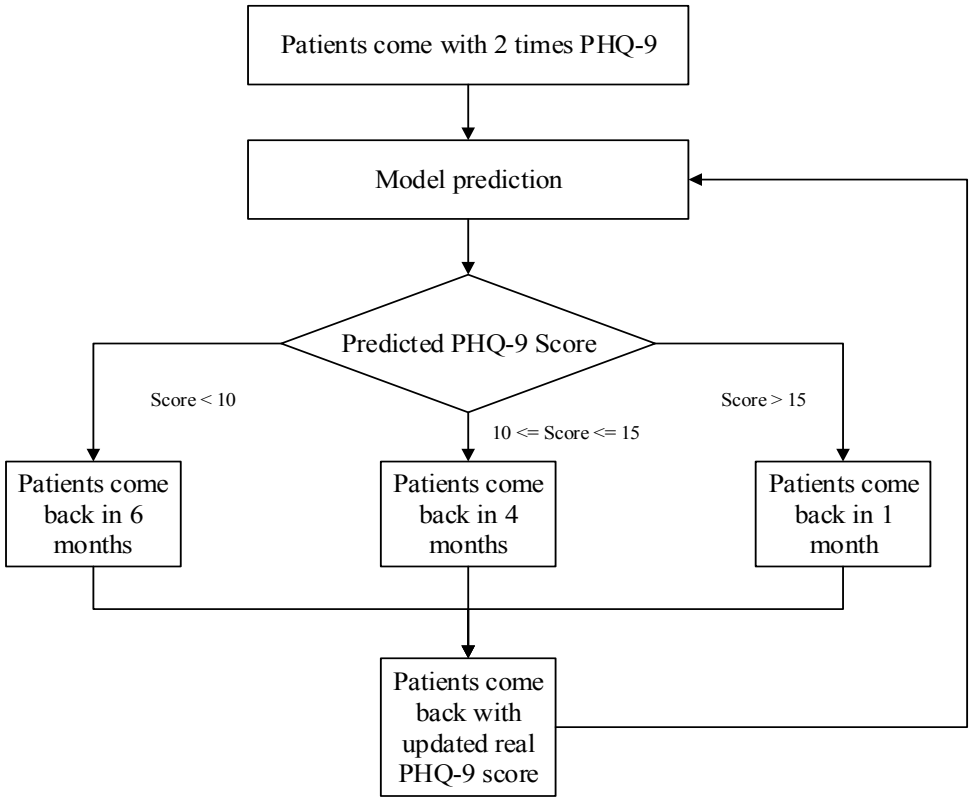


Figure 6.1. Flow chart of the monitoring scheduling system based on the depression prediction model

d. With the predicted PHQ-9 score, the system schedules the time of next interview according to a given criteria. The criteria should be set according to clinical availability and doctors' experiences, and can be revised according to the specific economic or working load conditions of the clinics. As shown in Fig. 6.1, for instance, if the score is lower than 10, which means the depression symptoms of the patient is predicted to be mild, the patient may be asked to come back in six months; if the score is between 10 to 15, which means the depression symptoms of the patient is predicted to be moderate, the patient may be asked to come back in four months; if the score is higher than 15, which means the depression symptoms of the patient is predicted to be severe, the patient may be asked to come back in one month. Note that in the EJR dataset, each patient visited the clinics for a certain period of time. In the simulation

system, therefore, if the scheduled time for the next time interval is larger than the maximum time slot t_{max} of the patient, i.e., the time that the patient had his/her last interview, the simulated scheduling cannot go on and will stop.

e. In practice, the patient would come back to the doctor at the schedule time, denoted as time slot t_1+1, and will do the PHQ-9 test again to check his latest depression level. At this time, a new PHQ-9 score at time slot t_1+1 is obtained, which is the corresponding splined PHQ-9 score of the patient at time slot t_1+1 in the dataset.

f. The system would then generate the time schedule for the next interval. In the proposed simulation-based monitoring scheduling system, an approximation is made to obtain the PHQ-9 score at time slot t_1. In particular, we connect the two PHQ-9 scores at former time slot of t_0+2 (the previous predicted score) and the current time slot of t_1+1 (current testing score), such that we gain a straight line L. We assume that the PHQ-9 score at time slot t_1 is equal to the value of L at time slot t_1, i.e., the depression symptoms of the patient is developing in a linear manner during the time period. Then we have the patient's PHQ-9 score at time slot t_1 and t_1+1, and the natural-history model can generate the predicted PHQ-9 score at the next time slot.

g. Go to Step 3.

To further demonstrate the prediction and scheduling procedure of the proposed simulation monitoring system, Fig. 6.2 illustrates the detailed procedure of a patient in the dataset. We can clearly see from Figure 6.2 that after the first and second visit at Bi-week 1 and 2, the proposed system predict the patient's score at Bi-week 3 by using the prediction model. The predicted score is equal to 3.01, which is lower than 10, such that the doctor asked the patient to come back 12 Bi-weeks later, i.e., Bi-week 14. When the patient visited again on Bi-week 14, he did another test and got a score. The system then predicted his score on Bi-week 15, which is 9.09 by using

the test score at Bi-week 14, and the expected score at Bi-week 13, which is obtained by drawing a line through the score points at Bi-week 3 and 14, and used the value of the straight line at Bi-week 13. The procedure then continues repeatedly until the next scheduled time exceeds the maximum observed time slot for the patient.

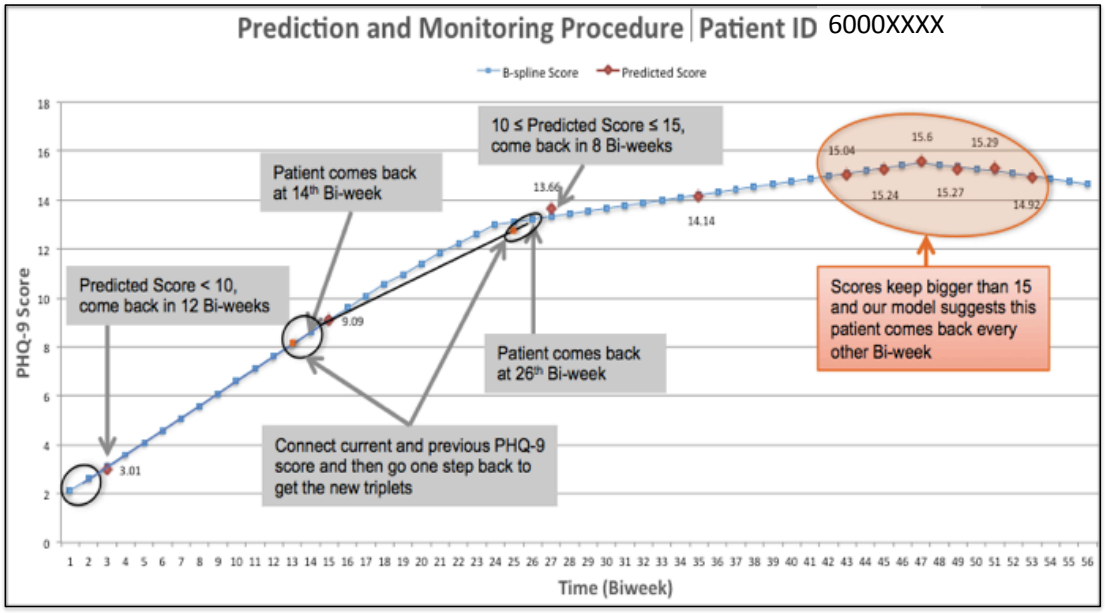


Figure 6.2. Prediction and monitoring procedure of according to the criteria set in Fig. 6.1.

6.2 Monitoring of 3159 Patients

In this section, we will show the scheduling table of the proposed monitoring scheduling system for individual patients in the dataset. Recall that we have 3159 patients in total, who have distinct time lengths of visit history and trajectories of depression symptoms. Due to limited space, here we omit the full version of the scheduling table. Table 6.1 presents 20 randomly selected patients' visiting schedule for illustration by using the criteria shown in Fig. 6.1.

Table 6.1. The visiting schedule table of the simulated scheduling system (V#: visit number. unit: bi-week)

Patient ID	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	2	14	26												
2	2	14	26	34	42	44	46	48	50	52	60				
3	2	10	18	26	34	42	50								
4	2	4	6	8	10	12	14								
5	2	4	6	14	22	30	38	46	58	70	78	86	94		
6	2	4	6	8	10	12	14	16	24	32	40				
7	2	4	6	14	26	38									
8	2	14	26	38	50	62	74	86	98						
9	2	14	26	38	50	62									
10	2	4	6	8	10	12	14	22	30						
11	2	4	6	14	22	30	38	46	54	62	70				
12	2	4	12	20	28	36	48								
13	2	10	18	26	34	42	54	66	78	90	102				
14	2	4	6	8	10	12	14	16							
15	2	14	22	30	38	40	42								
16	2	10	18	26	34	42	50	58	66	74	82	90			
17	2	4	6	8	10	18	26								
18	2	14	26	38	50	58	66	74	86	98	110	122	134	146	158
19	2	14	26	38	46	54	62	70	78	86	94	102	110		
20	2	14	26	38	50	62	74	86	98	110	122	134			

Let us take a closer look at the scheduling table. For patient No. 5 in Table 6.1, he/she begins with interviews scheduled with one-month apart, indicating that the depression symptoms are severe at first. Yet after the third meeting, the visiting interval becomes two months. It implies that the treatment the doctor has given to the patient is working, i.e., his/her symptoms are relieved. Similar patterns can be found on patient No. 6, 7, 11 and 17. On the contrary, patient No. 2's schedule reveals another story. In the beginning, his visiting interval was six months, indicating light depression levels. Yet after the third meeting, i.e., Bi-week 26, his conditions became worse and worse, where the visiting frequency became more frequent, from 4 months apart to one month apart. In this case, the doctor should design new treatment plans for him. After the tenth interview at Bi-week 52, the treatment seems working and his schedule becomes 4 months apart again.

6.2.1 Simulation Results with the Monitoring Scheduling System

Figures. 6.3-6.5 shows the histograms of average visiting interval of the 3159 patients under three different scheduling criteria, respectively. We can clearly see from Figs. 6.3-6.5 that no matter which scheduling criteria is chosen for the visiting interval, the percentage of patients that have minimum interval and maximum interval are much higher than the others. It implies that patients paying visit to depression clinics are more likely to have either severe depression symptoms and require frequent monitoring and visits, or have mild symptoms and do not need to come back in the immediate future.

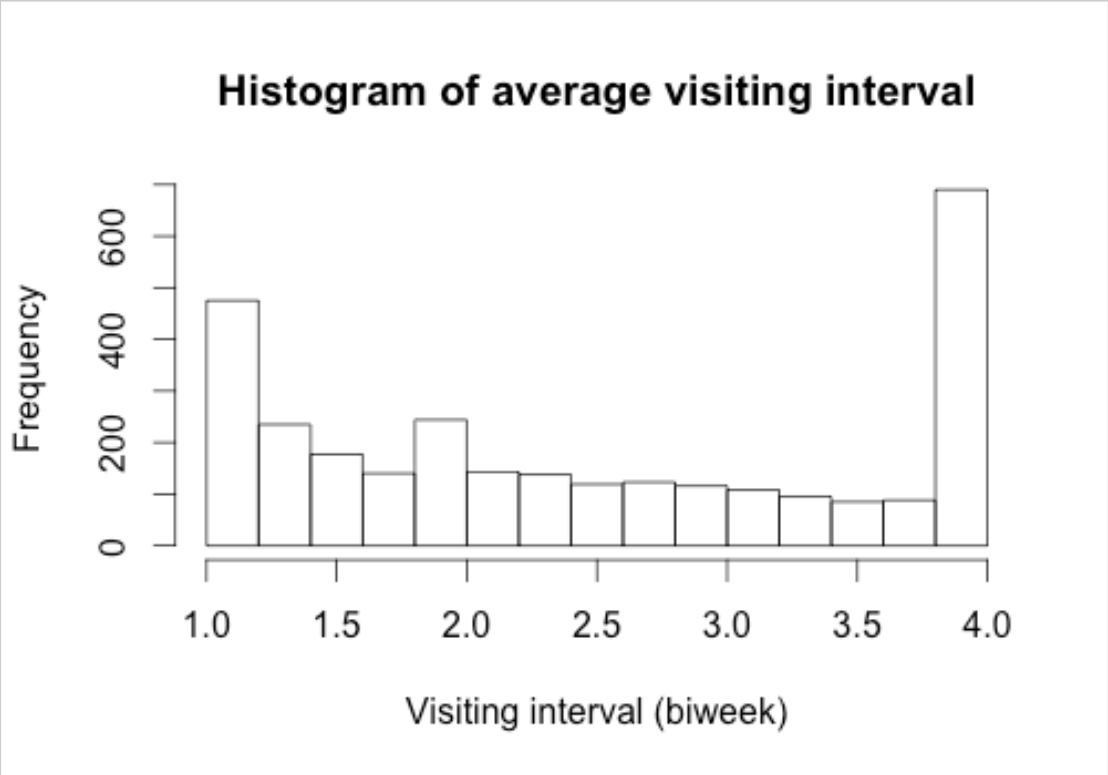


Figure 6.3. Histogram of average visiting interval with the monitoring scheduling system. Scheduling criteria: 2 months if score<10, 1 month if 10<score<15, 2 weeks if score>15

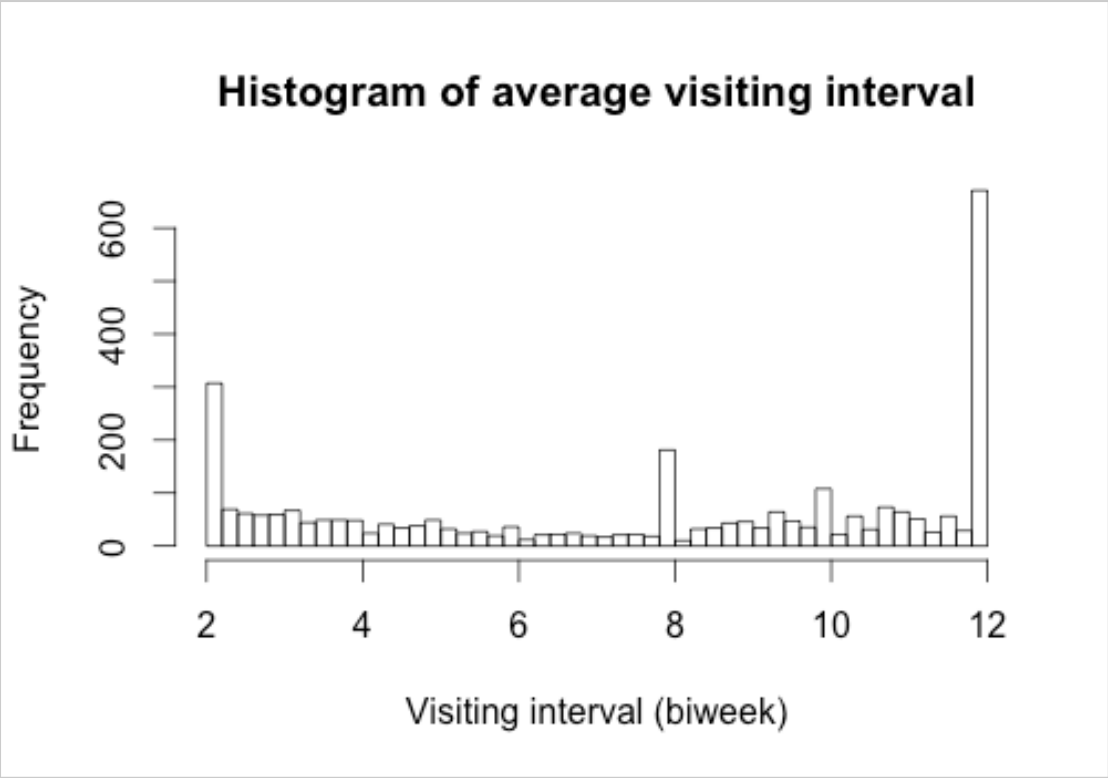


Figure 6.4. Histogram of average visiting interval with the monitoring scheduling system. Scheduling criteria: 6 months if score<10, 4 months if 10<score<15, 4 weeks if score>15

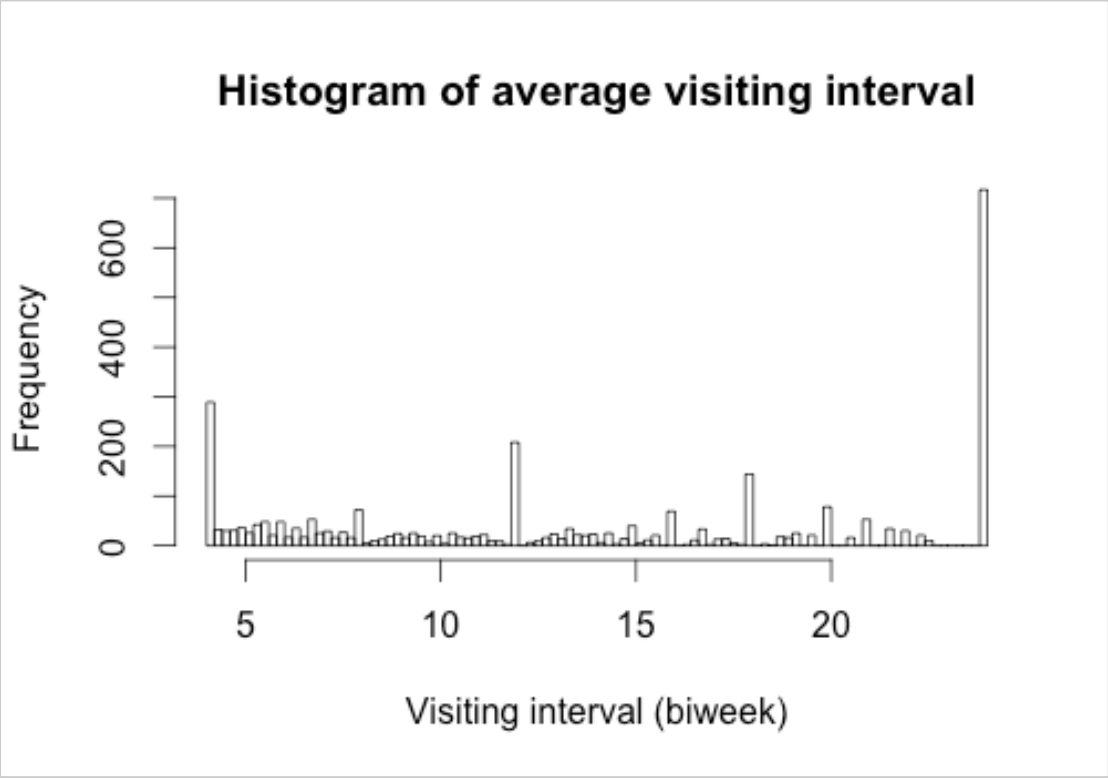


Figure 6.5. Histogram of average visiting interval with the monitoring scheduling system.. Scheduling criteria: 12 months if score<10, 6 months if 10<score<15, 2 months if score>15

Let us further take a look at the number of visiting times of the patients in the dataset. Figures. 6.6 -6.8 present the histograms of average visit times under three different scheduling criteria, respectively. Similar to the average visiting interval case, most patients in the simulation system have either very few times of meetings or a large number of meetings.

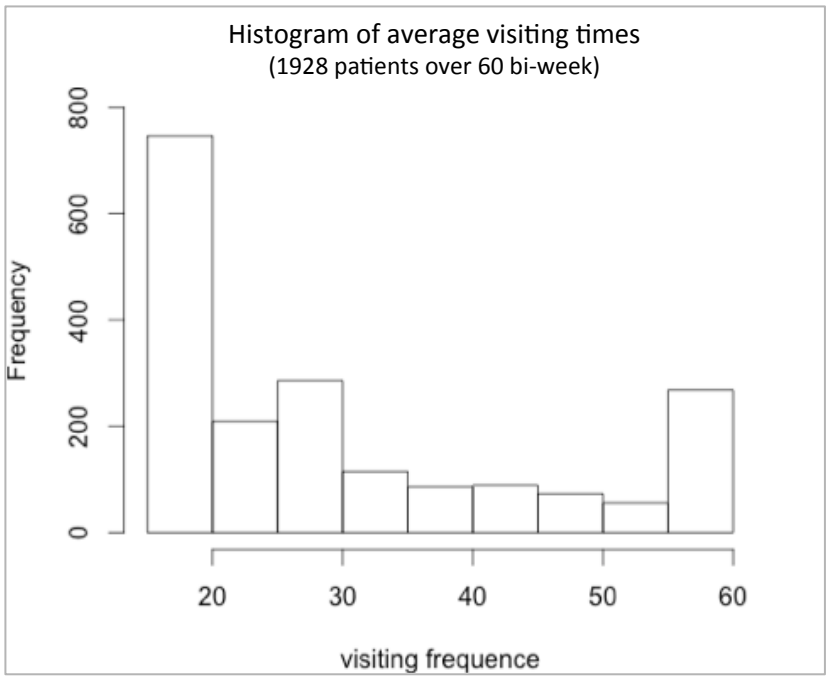


Figure 6.6. Histogram of average visiting times with the monitoring scheduling system. Scheduling criteria: 2 months if score<10, 1 month if 10<score<15, 2 weeks if score>15

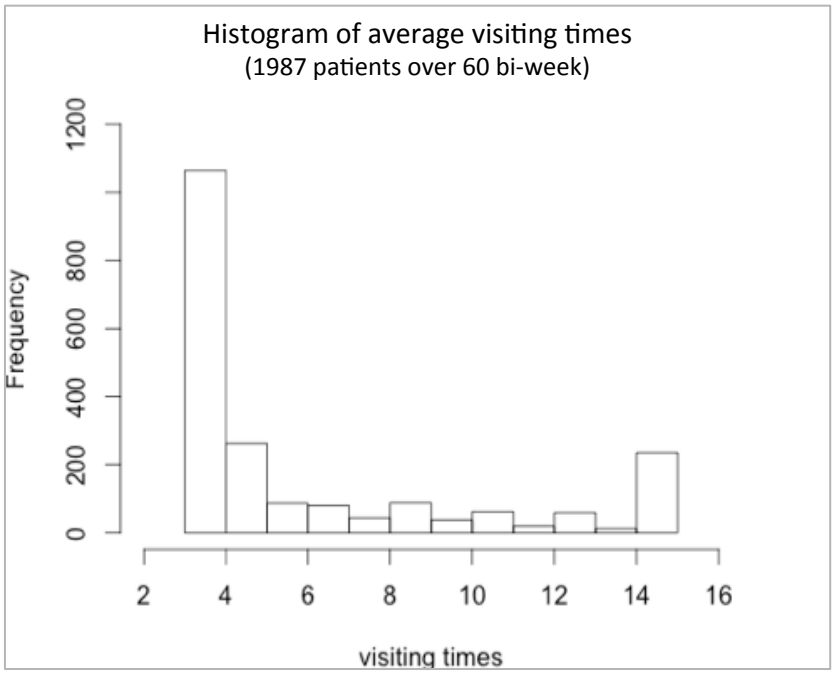


Figure 6.7. Histogram of average visiting times with the monitoring scheduling system. Scheduling criteria: 6 months if score<10, 4 months if 10<score<15, 4 weeks if score>15

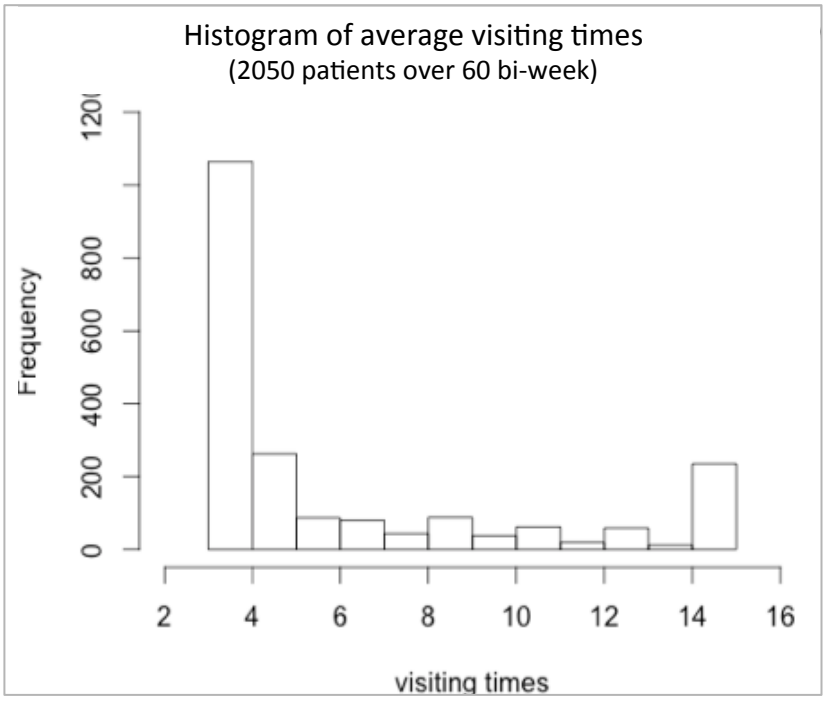


Figure 6.8. Histogram of average visiting times with the monitoring scheduling system. Scheduling criteria: 12 months if score<10, 6 months if 10<score<15, 2 months if score>15

6.2.2 Simulation Results with Perfect Model Prediction

Let us assume that the prediction model proposed in Chapter 4 provides perfect predicted PHQ-9 scores of each individual patient's trajectory. In this case, when a patient comes back to the clinics at time slot t1, we can directly use the corresponding splined PHQ-9 score of the patient at time slot t1+1 in the dataset as the predicted score at the next bi-week, and decides the next visiting time according to the score. With this assumption, Figures. 6.9-6.11 shows the histograms of average visiting interval of the 3159 patients under three different scheduling criteria, respectively.

By comparing Figures. 6.9-6.11 to Figures. 6.3-6.5, it can be clearly observed that the average visiting intervals under with the perfect model prediction have shown highly similar patterns to those with the scheduling monitoring system. It indicates that the proposed prediction model works accurately, at least with these three scheduling criteria in this thesis. The reason of the almost identical patterns mainly lies in the highly smooth B-splined PHQ-9 trajectories. In this case, it is highly possible that the predicted PHQ-9 scores and the B-splined PHQ-9 scores lie in the same range of scores, i.e., smaller than 10, 10 to 15 and higher than 15, leading to the same decision of visiting intervals. In the following subsection, we will further study a subset of patients with less smooth B-spline fit PHQ-9 scores.

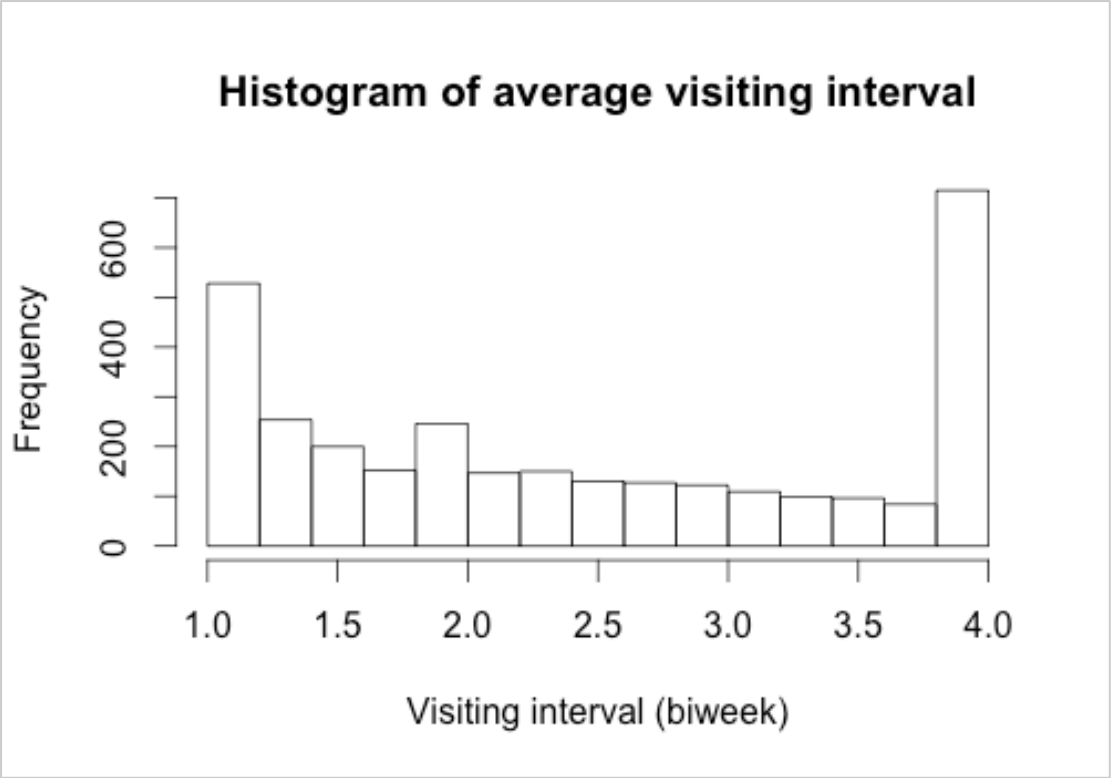


Figure 6.9. Histogram of average visiting interval with perfect model prediction. Scheduling criteria: 2 months if score<10, 1 month if 10<score<15, 2 weeks if score>15

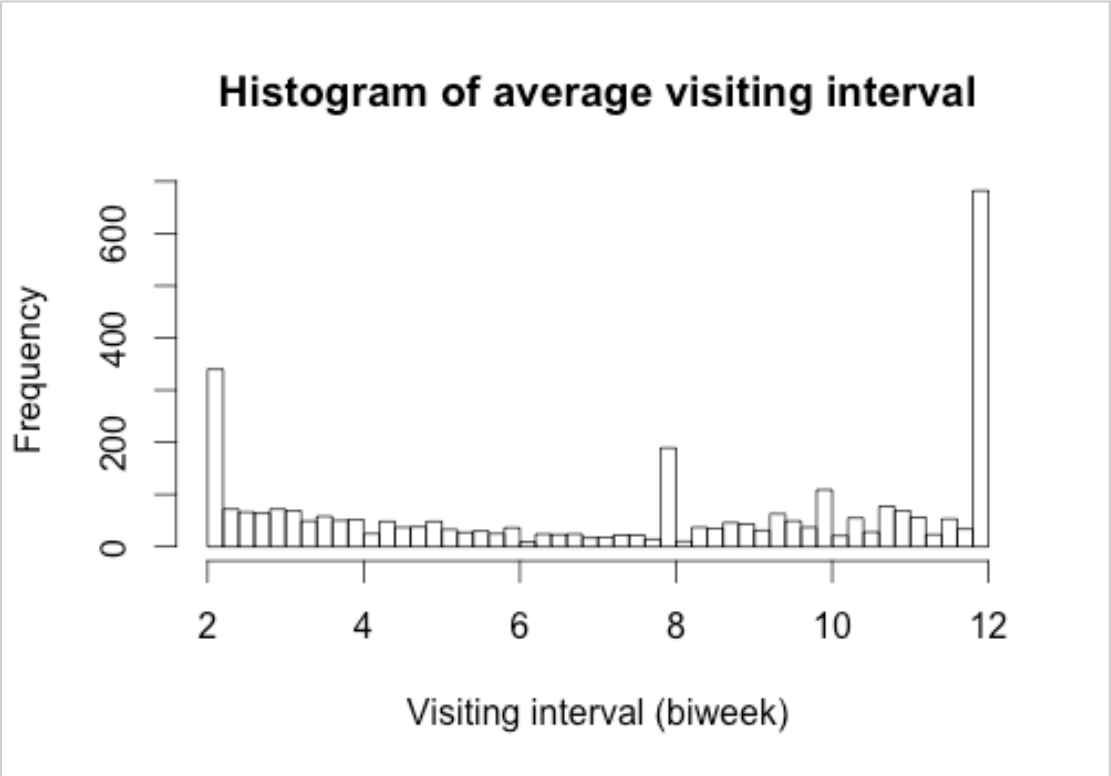


Figure 6.10. Histogram of average visiting interval with perfect model prediction. Scheduling criteria: 6 months if score<10, 4 months if 10<score<15, 4 weeks if score>15

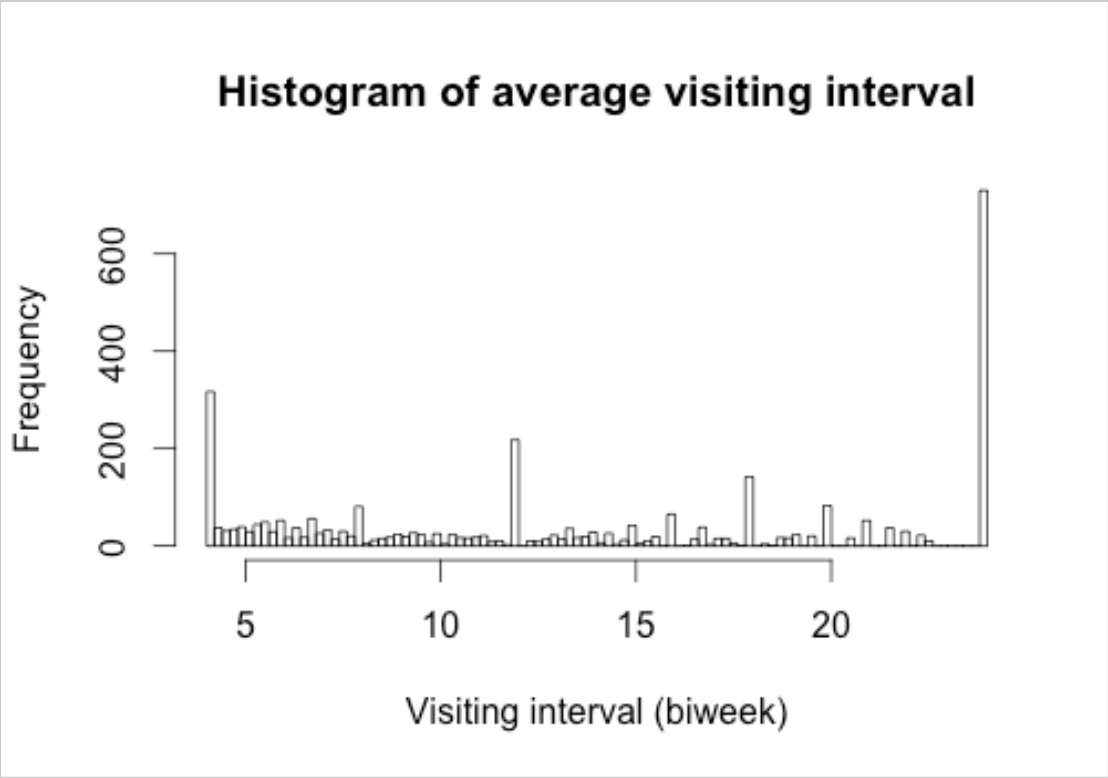


Figure 6.11. Histogram of average visiting interval with perfect model prediction. Scheduling criteria: 12 months if score<10, 6 months if 10<score<15, 2 months if score>15

6.3 Monitoring of 610 Patients

It should be pointed out that in practical scenarios, the actual visiting frequencies by the adaptive monitoring rules proposed in the simulation system are closely dependent on the EHR dataset, for instance, how smooth are the B-spline fit. So far, the b-spline adopted in this thesis provides a smooth fit for each patient's PHQ-9 trajectory. In this subsection, we will take a step forward to investigate how the proposed simulation monitoring system works for patients with less smooth PHQ-9 trajectories.

To put together an appropriate dataset for such analysis, we select the subjects in the original MHRN dataset that have more than six observations within a 20 Bi-week time window. In this case, we find 610 patients in total. Then, the individual trajectory of each patient is fit by using B-spline model. Nevertheless, we further add a random

error with normal distribution, N(0,3), on the B-spline trajectories. Figure 6.15 illustrates the simulated PHQ-9 trajectories of 9 randomly selected subjects. We can clearly see from Figure 6.15 that the simulated curves is much less smooth compared to the previous results. The PHQ-9 scores of each patient at each time slot are then interpolated from these trajectories.

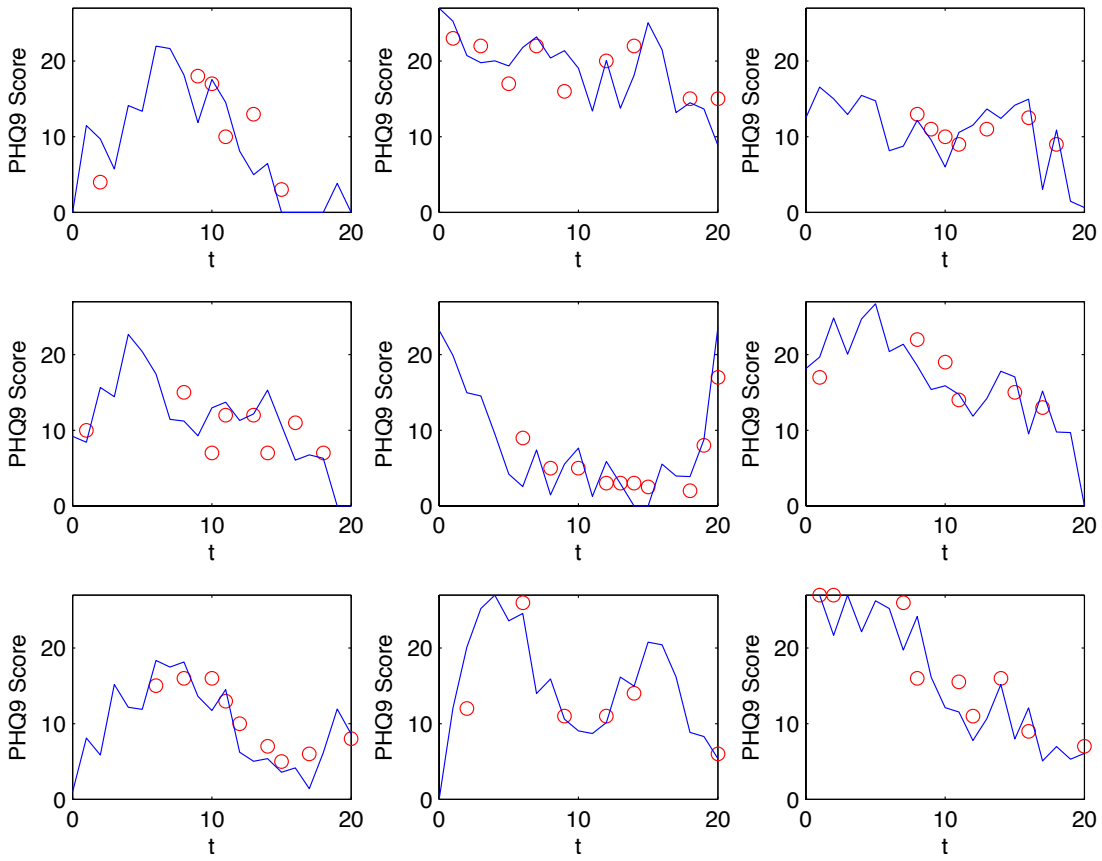


Figure 6.12. The real observations (red dots) and the simulated trajectories (blue line) on 9 randomly selected subjects from the 610 patients

Let us now apply the simulation monitoring system for such a 610 patient dataset. By following the steps of the natural history model in Chapter 4, we first generate the triplets to build the training database. For each patient's prediction in the model, we use all the data of the other 609 patients as the training database. Furthermore, the scheduling procedure provided in Figure 6.1 can be then adopted to generate the scheduling table of each patient in the dataset. Note that in the new dataset, all the patients are observed in a 20-Biweek time window only. In this case, we use the

tightest scheduling criteria of 2 months if score<10, 1 month if 10<score<15, 2 weeks if score>15 for illustration.

Table 6.1 presents the results of the number of patients that can be successfully scheduled by using the proposed simulation monitoring system for the new dataset of 610 patients and the original dataset of 3159 patients. Note that during the scheduling procedure of a patient, if the natural history model cannot find a match for the patient's PHQ-9 scores in the training database, it cannot make a schedule for him/her. We can clearly see from Table 6.2 that a much larger proportion of unsuccessfully scheduling can be found when the new dataset is used. The reason lies in the less smooth B-spline that was used. Due to the more fluctuating PHQ-9 scores, it becomes more likely that the model could not find a similar record in the database. It indicates that the proposed model may have some limitations when it faces less smooth B-spline or more fluctuating trajectories.

Table 6.2. Comparison of successfully scheduling of the simulation monitoring system between the new dataset and the original dataset

Dataset	No. of patients	No. of patients that can be scheduled	Percentage of successfully scheduling
New dataset	610	449	73.60%
Original dataset	3159	3024	95.70%

Similar to Section 6.2, let us take a further look at the histograms of the average visiting interval for the new dataset, which are shown in Figure 6.13. By comparing Figure 6.13 to Figure 6.3, we can clearly see that in contrast to the extremes in Figure 6.3, where most of the patients in the 3159 database either visit the clinics with very small intervals or large intervals, with the new 610 database, the average visit interval of the patients tend to be more uniformly distributed.

Let us further consider the perfect model prediction case. As demonstrated in Section 6.2.2, in this case, we directly use the splined value of PHQ-9 scores at the next bi-week of the visiting time as the predicted value, and decide the time schedule for the next meeting. Figure 6.14 illustrates the corresponding histogram of the

average visiting interval for the 610 patients. For this new subset of 610 patients with less smooth PHQ-9 trajectories, we can clearly see from Figure 6.14 that the histogram of average visiting interval with the perfect model prediction still show similar patterns with that with the scheduling monitoring system, more uniformly distributed for instance. Nevertheless, they have obvious deviations. It is different from the almost identical patterns for the 3159 patients with very smooth B-splined PHQ-9 scores. It implies that the proposed prediction model works less effectively when the PHQ-9 trajectories have less smooth patterns and more strong fluctuations.

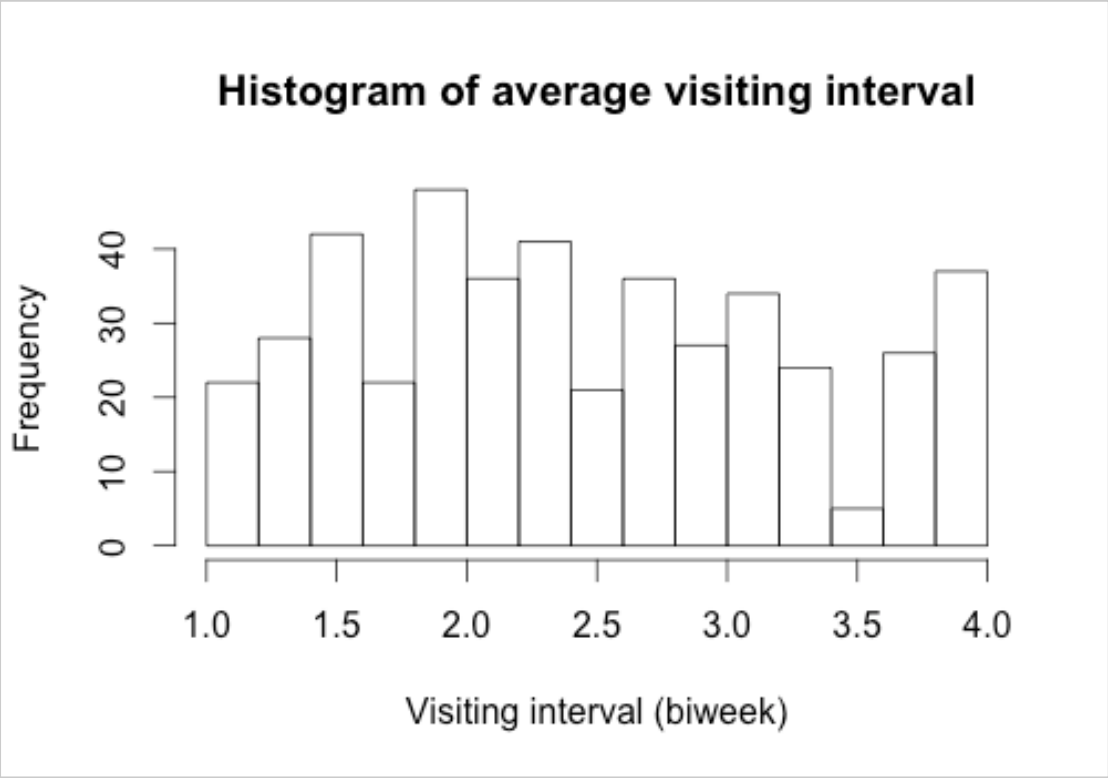


Figure 6.13. Histogram of average visiting interval with the monitoring scheduling system. Scheduling criteria: 2 months if score<10, 1 month if 10<score<15, 2 weeks if score>15

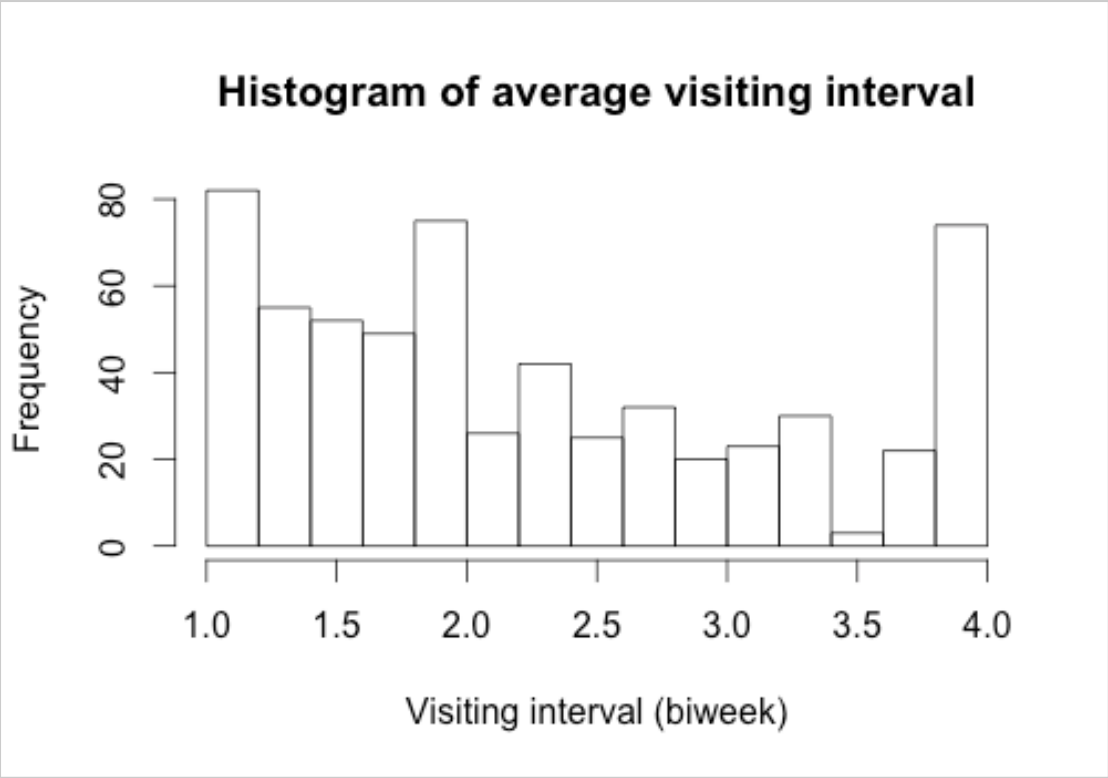


Figure 6.14. Histogram of average visiting interval with perfect model prediction. Scheduling criteria: 2 months if score<10, 1 month if 10<score<15, 2 weeks if score>15

CHAPTER 7. CONCLUSION

Depression is a major mental health problem that affects the lives and works of millions of people in the U.S. Given the pervasiveness of depression and its associated morbidity and mortality, it is important and imperative to develop accurate approaches to predict depression progression to help these patients be treated and cured on a timely manner. A number of interesting and crucial problems have arisen up. Is it possible to predict how the patient's depression level will develop in the following several weeks or months? How should the doctors make appropriate schedules to monitor each patient? Due to the complex nature of the natural history of depression, how to predict each individual patient's depression progression, or effectively design a monitoring scheduling system for the clinics according to the predicted results, still remain largely unexploited.

The work in this thesis provides a step towards filling that void at both the method and practical levels. From the method viewpoint, a nature-history model is proposed to address the depression prediction problem by using EHR data. By using the depression dataset provided by the Mental Health Research Network (MHRN), a dataset with 3159 patients are selected, including information on age, gender, location, items 9 score, Carlson comorbidity score, and PHQ-9 scores overtime. In the proposed model, each patient's trajectory of PHQ-9 scores are translated into a set of time triplets by using interpolation and B-spline. For any new patient, a multivariate nearness approach is adopted to predict its next time point PHQ-9 score. In particular, the model always finds the most similar patient in the current dataset and uses his/her PHQ-9 score at the next time slot as the predicted value for the new patient.

To demonstrate the prediction accuracy of our model, extensive experiments are conducted by using five-fold cross-validation under distinct circumstances including long-term predictions and regular-time predictions, and the MSEs and rMSEs are obtained. The prediction errors indicate that the proposed model may be more suitable for short-term predictions and are very sensitive to fierce fluctuations in PHQ-9 scores. Based on this analysis, insights can be provided to the prediction of depression progression in clinical treatments. First of all, we could predict the

onset of major depression in the immediate future, which is of great importance to prevent depression-related suicides; secondly, we could provide decision aid on treatment to a patient according to the predicted results.

By adopting the proposed depression prediction model, a simulation-based monitoring and scheduling system is further developed, which aims to provide effective and efficient guidance for the depression clinics to design a visit scheduling table for each individual patient, and help the doctors better understand the effect of an ongoing treatment plan on the depression trajectory of a patient. The scheduling procedure of the proposed system is further introduced, and simulation results are presented to demonstrate the practicability of the proposed monitoring system. The applicability of the system is also examined when a less smooth B-spline is used for the PHQ-9 trajectories.

There are several limitations of our study, which may affect the generalizability of our method and require in-depth analysis in the future. Firstly, it is difficult to justify the accuracy of using interpolated PHQ-9 scores as validation data, since there is only a small number of available validation dataset containing real PHQ-9 scores observations drawn at regular time intervals. Moreover, it is widely observed in the data source that the observed PHQ-9 scores are likely to have significant fluctuations among continuous time points. The triplets in the training dataset, however, are obtained using splined curves and more likely to be flat. This contradiction leads to comparatively large prediction errors using our model. It is therefore of great interest to investigate the effect of parameters in the B-spline fits on the accuracy of the proposed model. Furthermore, the monitoring scheduling system proposed in this thesis can only be evaluated by using simulation tools. Nevertheless, the practicability of such systems should be examined in real-world clinics.

APPENDIX A

B-splines [4]

A function f is called a basis spline (B-spline) on $[x_1, x_N]$ if the following criteria are met:

- The function f is defined on $[x_1, x_N]$.
- The function f is a piecewise polynomial function of degree $<N$, and each piece of the function is a polynomial of degree $<N$ between and including adjacent knots.
- The function f is continuous at the knots. When all internal knots are distinct its derivatives are also continuous up to the derivative of degree k-1. If internal knots are coincident at a given value, the continuity of derivative order is reduced by 1 for each additional knot.
- The knots must be in ascending order, and the number of internal knots is equal to the degree of the polynomial if there are no knot multiplicities.

BIBLIOGRAPHY

[1] Chris Iliades, *Stats and Facts about Depression in America*, online resources, http://m.everydayhealth.com/health-report/major-depression/depression-statistics.

[2] Johnson J, Weissman MM, Klerman GL. *Service utilization and social morbidity associated with depressive symptoms in the community*, JAMA. 1992;267: 1478-83.

[3] Wells KB, Stewart A, Hays RD, Burnam MA, Rogers W, Daniels M, et al. *The functioning and well-being of depressed patients, Results from the Medical Outcomes Study*. JAMA. 1989;262:914-9.

[4] Murray CJ, Lopez AD. *Evidence-based health policy—lessons from the Global Burden of Disease Study*, Science. 1996;274:740-3. [PMID: 0008966556]

[5] Ormel J, VonKorff M, Ustun TB, Pini S, Korten A, Oldehinkel T. *Common mental disorders and disability across cultures. Results from the WHO Collaborative Study on Psychological Problems in General Health Care*, JAMA. 1994;272:1741-8. [PMID: 0007966922]

[6] Hays RD, Wells KB, Sherbourne CD, Rogers W, Spritzer K. *Functioning and well-being outcomes of patients with depression compared with chronic general medical illnesses*, Arch Gen Psychiatry. 1995;52:11-9. [PMID: 0007811158]

[7] Broadhead WE, Blazer DG, George LK, Tse CK. *Depression, disability days, and days lost from work in a prospective epidemiologic survey*, JAMA. 1990;264: 2524-8. [PMID: 0002146410]

[8] Olfson M, Fireman B, Weissman MM, Leon AC, Sheehan DV, Kathol RG, et al. *Mental disorders and disability among patients in a primary care group practice,* Am J Psychiatry. 1997;154:1734-40. [PMID: 0009396954]

[9] Simon GE, VonKorff M. *Recognition, management, and outcomes of depression in primary care*, Arch Fam Med. 1995;4:99-105. [PMID: 0007842160]

[10] Simon G, Ormel J, VonKorff M, Barlow W. *Health care costs associated with depressive and anxiety disorders in primary care*, Am J Psychiatry. 1995;152: 352-7.

[11] Simon GE, VonKorff M, Barlow W. *Health care costs of primary care patients with recognized depression, Arch Gen Psychiatry*. 1995;52:850-856.

[12] Rice DP, Miller LS. *The economic burden of affective disorders*, Br J Psychiatry Suppl. 1995:34-42.

[13] Barlow DH; Durand VM. Abnormal psychology: *An integrative approach (5th ed.*), Belmont, CA, USA: Thomson Wadsworth. 2005.

[14] Kroenke, Kurt, and Robert L. Spitzer. *The PHQ-9: a new depression diagnostic and severity measure*, Psychiatr Ann 32.9 (2002): 1-7.

[15] Piegl, Les, Tiller, Wayne. *The NURBS Book*, Springer-Verlag Berlin Heidelberg (1997).

[16] R. H. Dehejia and S. Wahba, *Propensity score matching methods for non-experimental causal studies*, Discussion Paper Series, Columbia University, 2002.

[17] Patten, Scott B., and Don Schopflocher. *Longitudinal epidemiology of major depression as assessed by the Brief Patient Health Questionnaire (PHQ-9),* Comprehensive psychiatry 50.1 (2009): 26-33.

[18] Löwe, Bernd, et al. *Monitoring depression treatment outcomes with the patient health questionnaire-9,* Medical care 42.12 (2004): 1194-1201

[19] Nicholas A. Furlotte1, Babak Alipanahi1 and David A. Hinds, *Deep learning and the prediction of human disease risk*, 23andMe Inc..

[20] T. Nguyen and T. Ho, *A Semi-Supervised Learning Approach to Disease Gene Prediction.*

[21] Dasgupta A, Sun Y, Konig I, Bailey-Wilson J, Malley J. *Brief review of regression-based and machine learning methods in genetic epidemiology: the GAW17 experience*, Genet Epidemiol, 2011

[22] C. Wu, K. Walsh, A. DeWan, J. Hoh, and Z. Wang, *Disease risk prediction with rare and common variants*, BMC Proc, 2011

[23] Paulsen. JS, Long JD, et al, *Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study,* Lancet Neurol, 2014, 13: 1193-1201.

[24] M. Caliendo and S. Kopeinig, *Some practical guidance for the implementation of propensity score matching,* IZA Discussion Paper, May 2005.

[25] P. C. Austin, *A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003,* Statistics in Medicine, vol. 27, pp. 2037-2049, 2008.

[26] Alagoz, Oguzhan, et al. *Incorporating biological natural history in simulation models: empirical estimates of the progression of end-stage liver disease,* Medical Decision Making 25.6 (2005): 620-632.

[27] Kessler RC, Birnbaum H, Bromet E, Hwang I, Sampson N, Shahly V. *Age differences in major depression: results from the National Comorbidity Survey Replication (NCS-R)*, Psychol Med. 2010;40(2):225-237.

[28] Blazer DG, Landerman LR, Hays JC, Simonsick EM, Saunders WB. *Symptoms of depression among community-dwelling elderly African-American and white older adults,* Psychol Med. 1998;28(6): 1311-1320.

[29] Kessler RC, Foster C,Webster PS, House JS. *The relationship between age and depressive symptoms in two national surveys*, Psychol Aging. 1992;7(1):119-126.

[30] Fiske A, GatzM, Pedersen NL. *Depressive symptoms and aging: the effects of illness and non-health-related events,* J Gerontol B Psychol Sci Soc Sci. 2003;58(6):320-328.

[31] Kessler RC. *Epidemiology of women and depression*, J Affect Disord. 2003;74(1):5-13.

[32] Van de Velde S, Bracke P, Levecque K, Meuleman B. *Gender differences in depression in 25 European countries after eliminating measurement bias in the CES-D 8*, Soc Sci Res. 2010;39:396-404.

[33] Bromberger JT, Harlow S, Avis N, Kravitz HM, Cordal *A. Racial/ethnic differences in the prevalence of depressive symptoms among middle-aged women: the Study of Women's Health Across the Nation (SWAN),* Am J Public Health. 2004;94(8):1378-1385.

[34] Kim J, Durden E. *Socioeconomic status and age trajectories of health,* Soc Sci Med. 2007;65(12): 2489-2502.

[35] Miech RA, Shanahan MJ. *Socioeconomic status and depression over the life course,* J Health Soc Behav. 2000. 41:162-176.

[36] Feng QY, Griffiths F, Parson N, Gunn J, *An exploratory statistical approach to depression pattern identification*, Physica A. 2013;392: 889-901.

[37] Gunn, J., et al., *A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care,* J Affect Disord, 2013. 148(2-3): p. 338-46.

[38] Sutin, A.R., et al., *The trajectory of depressive symptoms across the adult life span,* JAMA Psychiatry, 2013. 70(8): p. 803-11.

[39] Twisk, J. and T. Hoekstra, *Classifying developmental trajectories over time should be done with great caution: a comparison between methods,* J Clin Epidemiol, 2012. 65(10): p. 1078-87.

[40] Bartolucci, F., A. Farcomeni, and F. Pennoni, *Latent Markov Models for Longitudinal Data,* 2013: Chapman and Hall/CRC press.

[41] U.S. Preventative Services Task Force, *Depression in Adults: Screening*, Accessed at http://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/depression-in-adults-screening1, 2016.

[42] USPSTF, *Depression in Children and Adolescents: Screening,* http://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/depression-in-children-and-adolescents-screening1. 2016.

[43] Hetrick, Sarah, et al. *Primary care monitoring of depressive symptoms in young people,* Australian family physician 43.3 (2014): 147.

[44] van Noorden, Martijn S., et al. *Predicting outcome of depression using the depressive symptom profile: the Leiden Routine Outcome Monitoring Study,* Depression and anxiety, 2012, 29(6): 523-530.

[45] Sacks, Rebecca M., et al. *How well do patient activation scores predict depression outcomes one year later?* Journal of affective disorders , 2014, 169: 1-6.

[46] Moore M., Byng R., et al, *'Watchful waiting' or 'active monitoring' in depression management in primary care: Exploring the recalled content of general practitioner consultations,* Journal of Affective Disorders, 2013, 145, 120-125.

[47] Valenstein M., Vijan S., et al, *The Cost-Utility of Screening for Depression in Primary Care*, Annals of Internal Medicine, 2001, 345-360.

[48] O'Conner, EA, Whitlock, EP, et al, *Screening for Depression in Adult Patients in Primary Care Settings: A Systematic Evidence Review*, Annals of Internal Medicine, 2009, 151(11), 793-812.

[49] Rush AJ, Trivedi MH, et al, Acute and Longer-Term Outcomes in Depressed Outpatients Requiring One or Several Treatment Steps: A STAR*D Report, Am J Psychiatry, 2006, 163(11), 1905-1917.

[50] Lin Y., Qian X., el al, *A Rule-Based Prognostic Model for Type 1 Diabetes by Identifying and Synthesizing Baseline Profile Patterns,* PloS one, 2014. 9(6): e91095.