TRANSFER LEARNING FOR MEDICATION ADHERENCE PREDICTION FROM SOCIAL FORUMS SELF-REPORTED DATA

A Thesis

Submitted to the Faculty

of

Purdue University

by

Kyle D. Haas

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Electrical and Computer Engineering

December 2018

Purdue University

Indianapolis, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Zina Ben-Miled, Chair

Department of Electrical and Computer Engineering

Dr. Brian King

Department of Electrical and Computer Engineering

Dr. Mohamed El-Sharkawy

Department of Electrical and Computer Engineering

Approved by:

Dr. Brian King

Head of the Graduate Program

ACKNOWLEDGMENTS

I would like to thank my thesis advisor Dr. Zina Ben-Miled for her guidance, support, dedication, and assistance not only on this thesis, but also toward the betterment of my professional development.

I would also like to give thanks to my committee members Dr. Brian King and Dr. Mohamed El-Sharkawy for their valuable feedback and insights, and I am also grateful for the technical and professional support of Dr. Malika Mahoui.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

ABBREVIATIONS

AED	Amount Taken Each Day
CND	Number of Documented Conditions
DVS	Number of Dental Visits through all Rounds
EVS	Number of Emergency Room Visits Through all Rounds
GEN	Gender
IMRF	Imputed Means Random Forest
MEPS	Medical Expenditure Panel Survey
MRA	Medication Refill Adherence
OPP	Out-of-Pocket Payments
PFT	Percentage each Feature was Traversed
REG	Region of living
RET	Has a Retirement Plan
RF	Random Forest
RFB	Random Forest Binning
RFT	Random Forest Ternary
SOM	Strength of Medication
SVM	Support Vector Machine
TOM	Type of Medication
YTT	Years Taking Treatment

ABSTRACT

Haas, Kyle D. M.S.E.C.E., Purdue University, December 2018. Transfer Learning for Medication Adherence Prediction From Social Forums Self-Reported Data. Major Professor: Zina Ben-Miled.

Medication non-adherence and non-compliance left unaddressed can compound into severe medical problems for patients. Identifying patients that are likely to become non-adherent can help reduce these problems. Despite these benefits, monitoring adherence at scale is cost-prohibitive. Social forums offer an easily accessible, affordable, and timely alternative to the traditional methods based on claims data. This study investigates the potential of medication adherence prediction based on social forum data for diabetes and fibromyalgia therapies by using transfer learning from the Medical Expenditure Panel Survey (MEPS).

Predictive adherence models are developed by using both survey and social forums data and different random forest (RF) techniques. The first of these implementations uses binned inputs from k-means clustering. The second technique is based on ternary trees instead of the widely used binary decision trees. These techniques are able to handle missing data, a prevalent characteristic of social forums data.

The results of this study show that transfer learning between survey models and social forum models is possible. Using MEPS survey data and the techniques listed above to derive RF models, less than 5% difference in accuracy was observed between the MEPS test dataset and the social forum test dataset. Along with these RF techniques, another RF implementation with imputed means for the missing values was developed and shown to predict adherence for social forum patients with an accuracy >70%.

This thesis shows that a model trained with verified survey data can be used to complement traditional medical adherence models by predicting adherence from unverified, self-reported data in a dynamic and timely manner. Furthermore, this model provides a method for discovering objective insights from subjective social reports. Additional investigation is needed to improve the prediction accuracy of the proposed model and to assess biases that may be inherent to self-reported adherence measures in social health networks.

1. INTRODUCTION

Non-adherence is one of the most costly medical expenditures. As of 2015, patients non-adherence cost in the United States reached $290 billion dollars [1]. In addition to being a substantial amount, this cost is a significant portion of healthcare spending where a large proportion of the cost of non-adherence arises from prescriptions that are never filled or not taken as prescribed [2]. While the financial losses are staggering, the most prominent motivation for better patient adherence is saving many of the patients whose conditions worsen due to poor compliance. Indeed, the US reported close to 125,000 deaths related to inadequate patient adherence [3].

Predicting why and when a patient becomes non-adherent has been a topic of research for decades [3, 4] and the costs and complications surrounding non-adherence makes it a focal issue in healthcare. Traditional prediction methods often use claims data aggregated from various health institutions. While claims data tend to be structured, they are often not accessible and may suffer from a time lag due to the needed pre-processing and de-identification. Social media data, on the other hand, are easily accessible, available in real-time, and cost-effective. Indeed, many studies have taken advantage of these data to study health trends and to gain insight into large-scale population health.

This thesis investigates multiple machine-learning models constructed from RF classifiers. Using transfer learning, the thesis also explores the possibility of predicting medication adherence using self-reported health information from publicly available profiles on a social forum. In particular, we are targeting medication adherence for two disease conditions: diabetes and fibromyalgia. These RF adherence models are trained using survey data and shown to predict self-reported social forum adherence measures. The contribution of this thesis consists of a model built upon accurate survey data that can:

- provide insights into patient adherence from unverified social forum data, and
- help define an objective measure for adherence from subjective social forum reports.

The remainder of this thesis is organized as follows: Chapter 2 of the thesis summarizes related work, Chapter 3 describes the methodology and the datasets used in this study to develop the proposed models, Chapter 4 discusses the prediction accuracy of these models, and Chapter 5 summarizes the main findings of the study and outlines direction for future work.

2. RELATED WORK

Identifying the patients at risk and the reasons for medication non-adherence can help guide the development of remedial and preventive plans. For many years, researchers have stipulated that many factors can influence non-adherence [3] including: poor patient-doctor interactions and lack of overall health understanding. The multitude of factors and their potential interdependence make profiling patients at risk for non-adherence difficult. Recently, the availability of large datasets from various sources allowed the development of successful non-adherence prediction models. For instance, Express Scripts developed a model that uses over 300 input parameters to predict adherence [5]. These parameters include patients demographic, clinical, and genomic features. A prediction accuracy of 98% was reported for this model with a lead-time of 6-12 months before patients stop taking their treatments [5]. Similarly, Allazo Health [6] and FICO [7] developed data-driven medication adherence prediction models. The latter assigns each patient a score representing the likelihood the patient will become non-adherent. These models are based on different input features, but they all seek to identify patients at risk of non-adherence. The focus of other research studies related to medication adherence is on the type of data used to study adherence. In [8], pharmaceutical retail data were compared to insurance claims data. While the findings of [8] were inconclusive, the source and validity of the data remains to be a topic of interest.

The use of advanced technology in predicting adherence or assisting non-adherent patients has significantly improved. Within the past year, studies have been conducted to better track and monitor non-adherent patients with the help of smartphones [9]. The result of these studies show that many patients better followed their prescribed treatments when they were required to take a selfie of themselves con-

suming their pills [9]. Improvements rates ranged from 50 to 67%. The smartphone application used in this case relied on a machine-learning engine that analyzes the image and validates whether or not the patient took the right treatment.

In this thesis, two prediction medication adherence models are developed for each disease condition. Both models are based on RF [10] classifiers. RF has been successfully used for patient outcome related classifiers [11] primarily because of its ability to process high dimensional feature spaces and to handle categorical features. For instance, RF was used in [12] to predict the response of patients to various drugs. This study showed that RF outperformed several other classification techniques over a feature space ranging from as few as five features to as many as 1,000 features. In another study [13], RF was used to classify patients with liver disease. This study reports that RF outperformed other classifiers.

The methodology proposed in this thesis leverages publicly available data while previous medication adherence models [5–7] primarily used insurance claims data. The benefits of social media data for trend analysis in large population health have been shown in several previous applications. For instance, twitter data have been used to study allergen effects across the United States [14] and were shown to produce a high volume of information related to adverse event monitoring of pharmaceutical products [15]. With regards to patient adherence, social media was used to engage patients in order to improve their compliance [16, 17]. Predicting adherence using social data was not previously addressed, however social media data were used to build a machine-learning model that can correctly predict stress [18]. This study also compared different machine learning techniques including neural networks, RF, support vector machines, and Bayesian networks, all leading to an accuracy level higher than 70%.

Claims data are accurate and include detailed information for each patient, while social health data often lack accuracy due to the subjectivity of the unverifiable, self-reported information posted by social forum participants [19]. The growth and abundance of social data make them a valuable source of knowledge. However, using

this data is constrained by these accuracy and validity concerns. Indeed, models trained using social data may be learning inaccurate reports. Transfer learning can be used to address this issue. Transfer learning is able to transfer the learning acquired in one environment to another environment [20] thereby relaxing accuracy requirements which are often associated with traditional machine learning methods. This approach was successfully used in many studies, including social media applications [21, 22].

For instance, transfer learning was used to help predictive modelling of degenerate biological systems [23]. These systems are defined as degenerate if they are structurally different, but yield the same output or perform the same functionality. These output and functionality consistencies were able to translate through transfer learning and were found to help model new biological domains.

The objective of this thesis is to complement the above-mentioned efforts by using a novel transfer learning approach that can leverage social forum data. Social health forum data can enable adherence analysis at a very large scale and with a wide spectrum of coverage extending beyond single institutions. However, this self-reported data may suffer from overestimation [24] and missing values. We explore the impact of these characteristics as well as the predictive ability of various features in the proposed models.

3. METHODS

3.1 Data Sources

Two data sources are used to collect patient information for the proposed adherence prediction models: the Medical Expenditure Panel Survey (MEPS) [25] and PatientsLikeMe [26]. Patients from both data sources were filtered in order to extract only the patients that are taking the treatments associated with diabetes or fibromyalgia. The list of treatments for each of these disease conditions is given in Table 3.1. Treatments were selected if they were taken by at least a single patient on PatientsLikeMe for the given condition.

Table 3.1: List of targeted treatments for each condition

Diabetes	Insulin Glargine	Glimepirade	Insulin Aspart
	Liraglutide	Insulin Lispro	Metformin
		Glyburide	
Fibromyalgia	Zolpidem	Duloxetine	Pregabalin
	Gabapentin	Tramadol	

PatientsLikeMe [26] is a public forum where patients post, discuss, and review many of their current medications and conditions. The data collected from PatientsLikeMe originated from the most current, publicly available treatment evaluations and resulted in a total of 92 diabetic records and 357 fibromyalgia records. Patients provide evaluations for treatments in a structured format that includes their self-reported adherence to the treatment. While this aspect forgoes the need for an adherence metric, relying uniquely on the patients assessment makes the adherence

classification subjective. This form of self-reporting from public sources has been shown in previous studies to suffer from overestimation [24]. This is a limitation of the predictive models developed by using self-reported data and is the primary reason for using transfer learning.

The second data source is the MEPS database which is provided by the Agency for Healthcare Research and Quality (AHRQ) [25]. It is a collection of surveys assigned to a national representing population of individuals. Participants are questioned in a series of five rounds over a two-year interval. During each round, participants are asked to answer a survey questionnaire that focuses on pertinent health information consisting of health status, medical conditions, prescribed medications, insurance coverage and more. Due to the span of the study, multiple panels of participants overlap. Each year, a new panel of participants is enrolled in the study while previous year panel finishes the second year. This panel overlap provides an insight into nationwide dynamic changes.

The demographic breakdown mostly follows similar trends across the two disease conditions (Table 3.2) with few differences. For both diabetes and fibromyalgia, the average social forum patient was about 5 to 10 years younger. In regards to the region of residence, the largest variance occurs in the southern region population. The patient population in this region was approximately 40% for both conditions in MEPS, while no single population exceeded 30% in the social forum data set. Females represent the majority population in all four cases (two MEPS data sets and two forum data sets) with the fibromyalgia female population in the social forum data set exceeding 90%.

3.2 Features

The features extracted for each patient from both data sources include:

- TOM: Type of Medication (Table 3.1)
- YTT: Years Taking Treatment

Table 3.2: Demographic statistics for each condition

	Diabetes		**Fibromyalgia**	
	MEPS (3242 Samples)	Social (92)	MEPS (3044)	Social (357)
Age	59.9 ± 14.0	54.5 ± 11.6	58.0 ± 14.9	49.1 ± 10.7
Gender				
-Male	1450 (44.7%)	35 (38.0%)	995 (32.7%)	27 (7.6%)
-Female	1,703 (52.5%)	56 (60.9%)	1,944 (63.9%)	325 (91.0%)
-Not Listed	89 (2.7%)	1 (1.1%)	105 (3.4%)	5 (1.4%)
Region				
-Northeast	479 (14.8%)	3 (3.3%)	400 (13.1%)	55 (13.6%)
-Midwest	668 (20.6%)	24 (26.1%)	652 (21.4%)	84 (20.8%)
-South	1,292 (39.9%)	26 (28.3%)	1,308 (43.0%)	110 (27.3%)
-West	693 (21.4%)	16 (17.4%)	561 (18.4%)	92 (22.8%)
-Not Listed	110 (3.4%)	23 (25.0%)	123 (4.0%)	16 (4.5%)

- AED: Amount Taken Each Day
- SOM: Strength of Medication (e.g., 500, 1000 mg etc.)
- OPP: Out-of-Pocket Payments (Table 3.3)
- REG: Region of Living (Northeast, Midwest, West, or South)
- AGE: Age at the end of the study, or last known age
- GEN: Gender

One of the challenges of using multisource data is ensuring that each feature is present in the same format in both sources. In particular, the Out-of-Pocket Payments (OPP) feature is categorized according to Table 3.3. While MEPS gave the exact amount paid for each prescription, the social forum data only listed ranges for OPP. Similarly, patient records from the social forums were mapped in this study to the appropriate US census region based on their state of residence. All MEPS records directly list census regions.

Table 3.3: Categorization of out-of-pocket payments (OPP)

Out-of-pocket payments (per month)	**Quantized Value**
<$25	0
$25-50	1
$50-100	2
$100	3
>$200	4

An additional data-preprocessing step was performed in the case of diabetic treatments. Some diabetic medicines are in the form of injectable liquids (i.e., ml) while others are in the form of pills (i.e., mg). Each of these types of dosage forms was assigned its own Strength of Medication feature (i.e., SOM-mg, SOM-ml).

3.3 Target Value Assignment

The target outcome for the model is medication adherence. Patients are assigned a value of 1 if they are adherent or 0 otherwise. As mentioned above, the social forum dataset includes a self-reported adherence metric. Patients select one of four adherence categories (i.e., always, usually, sometimes, or never taken as prescribed).

For this study, the always category is mapped to the adherence class (i.e., adherence = 1) while the other three categories are mapped to the non-adherent class (i.e., adherence = 0).

The MEPS survey data does not include a quantitative metric for medication adherence. Therefore, a medication adherence metric was developed based on previous studies by other researchers. In [27], eleven different medication adherence metrics were evaluated and the study recommended Medication Refill Adherence (MRA), which is defined as: The total days supply divided by the number of days of study participation, multiplied by 100. For example, a patient with a total of 200 days supplied over a period of 365 days will have an MRA of 55%. This metric can easily be derived from the MEPS dataset for all patients. However, one of its shortcomings is that it does not account for scenarios where patients are proactive in refilling their prescriptions or accidentally misplace medications. In these cases, the patients MRA value would become skewed due to the irregularity of pills supplied. Since the number of days supply was necessary to calculate MRA, MEPS records that did not include this parameter for a given medication were omitted. Moreover, four different threshold MRA values are considered since there is no predefined correspondence between MRA ranges and adherence. These thresholds are: 80%, 65%, 45%, and 35%.

All records containing a therapy listed in Table 3.1 for each target disease were retrieved from panels 17-19 (2012-2015) of the MEPS dataset. The above data extraction and scrubbing steps resulted in 3,242 MEPS diabetic and 3,044 MEPS fibromyalgia patient records.

3.4 Classifier

The classifier model proposed for medication adherence prediction is based on Random Forest (RF). RF consists of an ensemble of decision trees [10]. In this case, the number of decision trees is set to 100. Consensus among the decision trees is developed by using a technique called Bagging or Bootstrap Aggregation. That is,

for each tree a predefined number of records are selected randomly from the input dataset with replacement. Based on this selection, a given record may be selected more than once in a given tree while other records may not be selected entirely. During the construction of each tree, a random subset of the input features is considered at each node. In this thesis, the size of this subset was set to $\sqrt{n}$ where n is the total number of input features in the dataset. The randomized selection of both records and features from the input dataset help generate unique decision trees in the RF ensemble.

The best feature at each level of a given decision tree is determined by using the greatest reduction in impurity [28]. The parent node in the tree always has a higher impurity (less homogenous set of records) than all its children. A homogenous set of records corresponds to the case were all the records belong to the same class (i.e., adherent class or non-adherent class). The impurity of a node [28] is defined by:

$$I = 1 - (A_+)^2 - (A_-)^2 \tag{3.1}$$

where, A_+ and A_- are the percentage of adherent and non-adherent patients presented to the node in a given tree, respectively. The change in impurity between the parent node (p) and its left (l) and right (r) child nodes is given by:

$$\triangle I = I_p - P_l I_l - P_r I_r \tag{3.2}$$

where, P_{l} and P_{r} represent the percent of the total number of the parent records that are mapped to the left and right branches, respectively.

In order to split the records into the appropriate right or left branches at each node, the selected feature requires a reference value. All possible values for a given feature are iteratively evaluated until the best split is found (i.e., branches with the lowest impurity). In general, features can either be numeric or categorical. For instance, the feature AGE is numeric. When AGE is used as a feature for a given node in the tree, records that have an AGE value greater than the reference are assigned to the right branch of the node and the remaining records are assigned to the left branch. For the

categorical features, such as REG, patient records that have the same value as the reference are assigned to the right branch while the remaining records are assigned to the left branch.

One of the limitations of the above split classification is that it does not adequately handle patient records with missing feature values. Approximately 40% and 55% of the fibromyalgia and diabetes patients, respectively had at least one missing entry in the social forum dataset. A default split can be adopted in this case, where the record with missing value is arbitrarily assigned to the left branch. Alternatively, the record can be eliminated. Because of the prevalence of missing values in social data, the latter alternative was not viable. A default split also does not appropriately handle missing values. Two RF models are therefore proposed in this thesis in an attempt to address this shortcoming.

3.4.1 Random Forest - Binning (RFB)

In the first model, all numeric features are binned using k-means clustering [29]. An additional bin is then added to represent the case when the feature value is missing. Once the features are distributed across the bins, each node can split the records categorically.

K-means clustering is an iterative algorithm designed to find a given number of cluster centroids. Its heuristic approach provides a versatile and effective partitioning of a dataset. The algorithm begins by randomly selecting a defined number of means, and assigns each sample to the nearest mean. Following this assignment, the centroid is calculated and becomes the new mean for this cluster. All samples are then reassigned to the nearest cluster mean and the process continues until the algorithm converges to a local minimum.

In this study, there were four numeric features:

- YTT: Years Taking Treatment
- AED: Amount Taken Each Day

- SOM: Strength of Medication
- AGE: Age at the end of the study, or last known age

These four features were assigned five bins in order to stay consistent with the approximate number of bins for other categorical features such as OPP and REG. In the case of AED, no clustering algorithm was used as most patients administered their therapies 1-4 times daily. Any patient with a frequency above four was assigned to the 4-time daily cluster. For the other three features, the centroids determined by the k-means clustering for each disease condition are given in Table 3.4. Since this study covers multiple knowledge domains (social forums and MEPS survey), all data from each domain were aggregated together by feature, ensuring all feature entries would be represented before the partitioning began.

Table 3.4: Clustering centroids for three numeric features for patient records under a given condition

	Diabetes			**Fibromyalgia**		
Bin	**YTT**	**SOM**	**AGE**	**YTT**	**SOM**	**AGE**
0	Missing	Missing	Missing	Missing	Missing	Missing
1	1.82 yr	9922.3 mg	79.1 yr	22.1 yr	13.64 mg	62.2 yr
2	36.3 yr	500.0 mg	35.5 yr	11.64 yr	121.6 mg	77.1 yr
3	17.7 yr	5.5 mg	52.8 yr	5.41 yr	444.5 mg	49.6 yr
4	8.9 yr	39.0 mg	65.6 yr	1.04 yr	54.0 mg	33.7 yr

3.4.2 Random Forest - Ternary Tree (RFT)

In the second model, each node of the decision tree has three children: a left child, a middle child, and a right child. Unlike the RFB implementation, records with missing values are now assigned to the third child instead of being assigned to a specific bin signifying a missing record. The advantage of the RFT model is that it does not require additional data preprocessing as in the case of binning for the RFB model. However, the underlying RF ensemble-learning algorithm has to be modified in order to accommodate the additional child and to ensure that the missing value is never selected as a reference in the split at any node. In particular, we modified the change in impurity (Equation 3.2) to account for the addition of the third child node.

$$\triangle I = I_p - P_l I_l - P_r I_r - P_m I_m \tag{3.3}$$

3.4.3 Training/Testing using Random Forest

Classifier models are built using a training set, while their performance is evaluated using a testing set. The training set is comprised of samples the model uses to learn the factors that help assign these samples to their target classes. In this study, each sample represents a unique patient record for a given treatment (Table 3.1), while the target classes are adherent/non-adherent. Moreover, training classes are balanced (i.e., an equal number of sample records in both classes), which ensures any knowledge the model learns is not an artifact of a bias towards the larger represented class.

Each tree in the RF algorithm is trained using a bag of samples from the training set. Additionally, the depth of the tree grows until the leaf nodes map to a single class (i.e., the only samples remaining in the node are a homogeneous set of adherent/non-adherent patients) or the leaf node can no longer split further. In most cases, the former is true. However, there are select cases where all the remaining samples may

have the same features characteristics in a leaf node. If this infrequent event happens, the tree no longer grows down this branch and is treated differently for the purposes of testing.

Once the model is trained, a testing set of patient records which are independent of the training set is used to evaluate the classifier model. This is a fundamental component of machine-learning as this can now assess whether the model actually learned patient adherence or whether the model may have overfit to the training data. A good prediction accuracy for the testing set is an indication that the model generalizes well.

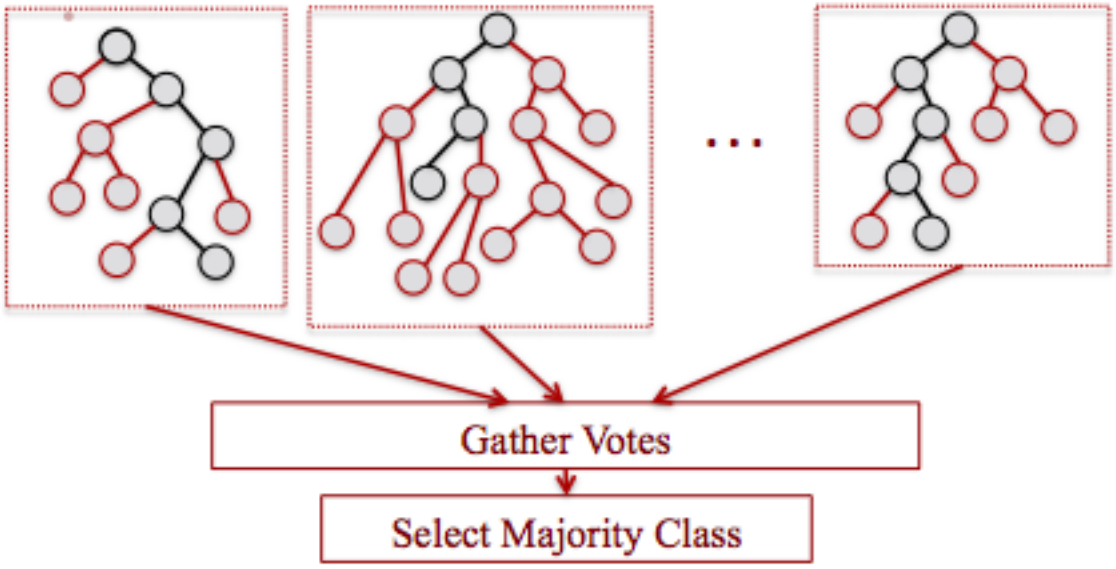

Fig. 3.1: Traditional RF implementation

Testing is performed in RF by taking each sample in the testing set and traversing each tree. Once the sample reaches to a leaf, the tree provides a single vote for the class represented by the leaf node. After all trees have voted for a given test sample, a collective democratic decision from the forest is made by selecting the class with the most votes. If the leaf node of a tree is not a homogeneous set, the tree

provides a partial vote for each class which is representative of the distribution of the heterogeneous node. For instance, if the leaf node has one non-adherent sample and three adherent samples, the trees provides a quarter vote for the non-adherent class and a three-quarter vote for the adherent class.

4. RESULTS

4.1 MRA Threshold and Class Breakdown

Transfer learning between the MEPS data and the social data is used towards adherence prediction. Using this approach, models are trained on MEPS survey data and then tested by using the social forum data. As previously mentioned, the MEPS dataset does not contain a quantitative adherence measure. Therefore, adherence was varied according to four different MRA thresholds (80%, 65%, 45%, and 35%). The breakdown of the number of adherent/non-adherent records in the MEPS dataset for different MRA thresholds is shown in Table 4.1. The last row of Table 4.1 also shows the breakdown of social forum patients.

Table 4.1: Breakdown of records according to adherence/non-adherence for varying MRA thresholds for the MEPS dataset

	Diabetes			**Fibromyalgia**		
MRA Threshold	Adherent	Non-Adherent	Percent Split	Adherent	Non-Adherent	Percent Split
80%	1,374	1,868	42/58	714	2,330	23/77
65%	1,698	1,544	52/48	906	2,138	30/70
45%	2,321	921	72/28	1,314	1,730	43/57
35%	2,577	665	79/21	1,571	1,473	52/48
PatientsLikeMe	70	22	76/24	281	76	79/21

This thesis relied upon publicly available patient self reports from a health social forum. From Table 4.1, the adherence splits for both disease conditions in the social forum data are highly skewed towards adherent patients compared to other high MRA threshold splits in the MEPS data. While each model tested in this study was trained on balanced datasets, it is of note that most patients on the social forum classify themselves as adherent, however most patient records from the MEPS survey indicate a much more balanced adherent/non-adherent distribution.

This difference may be attributed to two causes, either the social forum patients are overestimating their adherence, or these patients who are active participants in social health forums may be more engaged in their health which may be indicative of a greater potential for being adherent compared to the average patient. Both of these potential differences between social forum and MEPS survey patients are considered throughout this thesis.

4.2 Diabetes Adherence Prediction

Diabetes was selected as the first disease condition of interest. Two models were trained at each MRA threshold listed in Table 4.1 using the two RF algorithms described above (i.e., RFT and RFB). Samples were randomly removed from the higher represented class in each case until a 50/50 balance between adherent and non-adherent classes was obtained in the training dataset. For instance, 494 random non-adherent samples were removed for the models at the MRA 80% threshold. This process was repeated for both the 45% and 35% thresholds. The MRA 65% threshold was close enough to an even split that no samples were removed for this training dataset. Table 4.2 shows the number of samples used for training and testing.

Table 4.2 also shows the accuracy of the models (i.e., the number of correctly classified test samples against the total number of samples) along with the F_1 score which is a composite metric that represents a weighted balance between the recall and the precision of the models. Recall accounts for the number of correctly classified

adherent samples against the total number of actual adherent samples in the test set. Precision is the total number of correctly classified adherent samples against the total number of samples classified as adherent. These performance metrics are reported in Table 4.2 for the more accurate of the two RF implementations at each threshold.

Table 4.2: MEPS diabetes test results from MEPS trained models

MRA Threshold	Best Model	Training Size	Testing Size	Accuracy	F1 Score
80%	RFT	2,198	550	58.9%	60.6
65%	RFB	2,603	650	62.3%	66.9
45%	RFB	1,424	356	61.6%	62.7
35%	RFT	1,064	266	64.3%	67.6

We then selected the best performing model (RFT - 35% MRA) and tested the model against the 92 public diabetic patient records from PatientsLikeMe. The prediction accuracy for this test is 41.3% with an F_1 score of 40.6. These results show a large difference in prediction accuracy between the social forum records (41.3%) and the MEPS records (64.3%) and appear to imply that transfer learning is not a valid approach for social medication adherence prediction. However, this difference may be due to the small size of the social forum testing dataset.

4.3 Fibromyalgia Adherence Prediction

In order to investigate the root cause of this difference, we performed the same analysis for records with fibromyalgia therapies. As in the case of diabetes, MRA thresholds were also varied and random samples were removed from the larger class until the training datasets for MRA thresholds 80%, 65%, and 45% were balanced between adherent and non-adherent classes. For the 35% MRA threshold, the distri-

bution between the adherent and non-adherent classes was approximately balanced in the fibromyalgia dataset. Therefore, no samples were removed from this training dataset. The results for the fibromyalgia models are shown in Table 4.3.

Table 4.3: MEPS fibromyalgia test results from MEPS trained models

MRA Threshold	Best Model	Training Size	Testing Size	Accuracy	F1 Score
80%	RFT	1,143	285	70.2%	72.5
65%	RFT	1,450	362	69.9%	70.0
45%	RFT	2,103	525	73.0%	74.6
35%	RFB	2,436	609	77.3%	79.3

The best performing model (i.e., RFB - MRA threshold 35%) in Table 4.3 was selected and its aptitude for transfer learning was tested by using the 357 fibromyalgia records retrieved from PatientsLikeMe. The prediction accuracy for this test is 54.9% with an F_1 score of 65.5. As in the case of diabetes (Table 4.2), there is a large difference in prediction accuracy between the social forum test set (54.9%) and the survey test set (77.3%).

These results for both diabetes and fibromyalgia seem to indicate that transfer learning between survey data and social forum data is not possible. In order to understand the reasons behind the failure of the above transfer learning approach, we conducted an experiment to analyze the differences between the models for fibromyalgia. Diabetes was not chosen because of the limited number of PatientsLikeMe records.

4.4 Feature Investigation

In this second experiment, a model was trained on the fibromyalgia social forum dataset. As in the case of the previous experiment, we randomly removed adherent samples until there was a balanced distribution of 152 records in the training dataset.

Next, we tested the model using the MEPS survey data at the 35% MRA threshold as this threshold had the best accuracy (Table 4.3). The predictive accuracy in this case was 55.9% with an F_1 score of 51.2. Based on the previous experiment, these low accuracy levels were expected. However, the resulting RF model, which was trained by using the social forum dataset, can be compared to the RF model derived from the MEPS dataset.

The comparison between the two models was based on the number of times a feature was traversed in the underlying RF. A percentage for each feature (PFT) was obtained by dividing this number by the total number of nodes traversed in the RF for all features. As described in Chapter 3, RFs are built by selecting features that provide the greatest reduction in impurity. While there are measures that ensure that the same feature is not selected repeatedly at each branch, the total times a feature is selected is indicative of its relative entropy compared to other features. In an ideal scenario, for transfer learning to be effective between these domains, we expect to see similar PFT values for each feature in the social forum model and in the survey model. For some features this was true. However for others, the models had different PFT values. The result of this analysis is shown in Table 4.4.

In Table 4.4, four features had less than 5% difference in PFT across all four models. These features are TOM, YTT, REG and GEN. The type of medication (TOM) being the most important feature in predicting adherence, while gender (GEN) is the least predictive feature. The small difference in PFT between these four features indicates each model weighted these features consistently. Two additional features (i.e., SOM and AGE) had a larger PFT difference between the RFT and RFB models. This difference may be due to the binning applied to these numeric features in the case of the RFB model. Indeed, each numeric entry is generalized during the binning. However, the difference does not explain the low performance of transfer learning across the social forum and the survey datasets since the RFT (i.e., models 1 and 3) models, on the one hand, and the RFB (i.e., models 2 and 4) models, on the other hand, differ in accuracy by less than 5% in Table 4.4.

Table 4.4: PFT for each feature in the fibromyalgia model. Model 1: RFT with MRA threshold 35% using MEPS. Model 2: RFB with MRA threshold 35% using MEPS. Model 3: RFT using social forum data. Model 4: RFB using social forum data

	Model 1	**Model 2**	**Model 3**	**Model 4**
TOM	24.6	26.8	24.6	26.9
YTT	12.6	11.7	12.7	10.0
AED	11.0	12.5	4.6	9.1
SOM	14.9	11.6	18.0	12.0
OPP	10.0	11.8	13.6	13.9
REG	8.2	10.2	10.2	12.3
AGE	14.5	9.6	13.8	10.5
GEN	4.5	5.8	2.4	5.4

The remaining two features (i.e., OPP and AED) in Table 4.4 were the only two features where either both the social models had less PFT values than their survey counterparts or vice-a-versa. This indicated a potential difference between the social forum and survey datasets. Based on this finding, we retrained the RFB and RFT models at the 35% MRA thresholds and tested against both the survey data and the social forum data without the OPP and AED features. The results are shown in Table 4.5.

In the case of the new RFT model, the prediction accuracy for the social forum test data is 68.6% compared to the 54.9% accuracy that was obtained earlier with the RFT model trained by using all the features. Moreover, the difference in prediction accuracy between the survey test set and the social forum test set is less than 5%. This result shows that transfer learning between adherence models developed by using survey data and social forum is possible. However, careful consideration must be given

Table 4.5: MEPS fibromyalgia trained models without OPP and AED

Model	Testing Set	Training Size	Testing Size	Accuracy	F1 Score
RFT MRA 35	MEPS	2,436	609	73.5%	74.4
RFT MRA 35	Social Forum	2,436	357	68.6%	80.4
RFB MRA 35	MEPS	2,436	609	73.5%	76.0
RFB MRA 35	Social Forum	2,436	357	63.9%	75.3

to the features that are used to construct the model. Indeed, removing OPP and AED improved the prediction accuracy of the transfer learning for the social forum models. However, as evident by Table 4.3 and Table 4.5, the prediction accuracy decreased for the MEPS test set (77.3% to 73.5%). This result illustrates the benefit of a larger feature space, but also demonstrates that some features may not share a one-to-one relationship between the social and survey domains.

Additional investigation is needed in order to determine why removing these features improved the performance of the social forum models. We speculate that this may be due to over/under estimation by the patients [18] and to differences in the socio-demographic distribution of the patients. For instance, 80% of the MEPS patients had an OPP <\$25 each month, whereas less than 40% of the social forum patients had an OPP <\$25. Now that the AED and OPP features were removed, we updated the PFT values for the new models in in Table 4.6. All models that shared the same domain illustrated consistent PFT values (<5%).

As discussed earlier, the difference between the adherent distributions (Table 4.1) of the social forum and MEPS data are due to social forum patients either providing false adherence reports or being naturally more engaged in their healthcare. From the results of this study, it appears the latter is more indicative of the truth than the former. If patients were providing false self-reports there would be a stark difference between the results from the survey and social forums testing sets. For the RF model,

Table 4.6: PFT for each feature in the fibromyalgia model following the removal of AED and OPP. Model 1: RFT with MRA threshold 35% using MEPS. Model 2: RFB with MRA threshold 35% using MEPS. Model 3: RFT using social forum data. Model 4: RFB using social forum data

	Model 1	**Model 2**	**Model 3**	**Model 4**
TOM	29.3	33.2	25.2	29.5
YTT	16.2	16.0	17.4	14.4
SOM	18.5	15.0	20.2	17.2
REG	10.7	14.1	13.2	16.7
AGE	20.1	13.6	20.4	14.2
GEN	5.2	7.9	3.4	7.9

the difference between the survey and social forum test sets was 5%. This small error may be attributed to minor over-estimations and in general, the model identified most of these self-identified adherent patients on the social platform as adherent.

4.5 Additional Features

While refining the feature space improved the social adherence prediction from transfer learning, there was a reduction in performance in the survey test sets (77.3% to 73.5%) merely by removing two features. This result prompted the question of whether other features exist that can aid in improving adherence prediction. Currently, social forums do not typically ask questions beyond the feature space presented in this work, but if other features have a substantial impact on adherence prediction, social forums may be more inclined to request this information from participants.

The MEPS survey data provides a greater depth of patient information compared to the social forum data. Altogether, five more patient features were added to the model which were not present in any public patient treatment evaluation on the social forum. For this reason, only the MEPS test sets were evaluated in this experiment. The features are:

- RACE: Patient Race/Ethnicity
- DVS: Number of Dental Visits through all Rounds
- CND: Number of Documented Conditions
- EVS: Number of Emergency Room Visits Through all Rounds
- RET: Has a Retirement Plan

Furthermore, only the diabetic records were considered as these models did not perform as well as the fibromyalgia models and the hope was for these new features to provide additional insight into adherence prediction for diabetic patients. All the features were binned using the same k-means clustering procedure detailed in Section 3.4.1. These bins are shown in Table 4.7.

Table 4.7: Clustering centroids for six numeric features in MEPS Diabetes Model with additional features

Bin	**YTT**	**SOM**	**AGE**	**DVS**	**EVS**	**CND**
0	Missing	Missing	Missing	Missing	Missing	Missing
1	23.69 yr	37.13 mg	78.3 yr	20.2	4.8	6.4
2	11.34 yr	500.0 mg	64.9 yr	5.95	20	13.8
3	4.65 yr	975.8 mg	52.1 yr	10.6	1.5	35.8
4	0.9 yr	5.52 mg	34.7 yr	2.1	9.7	22.7

The experiment produced mixed results. Indeed, as seen from Table 4.8, while the accuracy increased at the 35% and 45% thresholds, in the case of the 80% threshold the performance is nearly identical and for the 65% threshold the performance actually decreased. These results appear to indicate that the five new features (RACE, DVS, EVS, CND, and RET) have little impact on adherence prediction, or that these features provide no new insight beyond what the model is able to learn from the original features.

Table 4.8: Comparison between the performance of the original MEPS model against the new MEPS model with additional features

	Original Diabetes Model			**Diabetes Model with Five New Features**		
MRA	**Best Model**	**Accuracy**	**F1 Score**	**Best Model**	**Accuracy**	**F1 Score**
80	RFT	58.9%	60.6	RFT	59.3%	60.4
65	RFB	62.3%	66.9	RFB	57.3%	62.3
45	RFB	61.6%	62.7	RFT	67.1%	69.4
35	RFT	64.3%	67.6	RFT	69.2%	70.3

4.6 Multi-Class Analysis

While a binary prediction for adherence provides substantial insight into patient compliance, many patients are not strictly adherent/non-adherent but represent varying degrees of compliance. For this reason, the model responsible for illustrating the potential for transfer learning (Table 4.5) was altered to accommodate additional classes. Indeed, these new classes provide a more refined prediction into patient medication adherence compliance.

As discussed in Chapter 2, the social forums allowed participants to self report which of the four varying degrees of adherence they believed described their adherence level (always, usually, sometimes, never taken as prescribed). Having used four varying MRA thresholds (80%, 65%, 45%, 35) for the MEPS data, each MEPS patient was assigned a level of adherence similar to that of the social patient. The mapping between the threshold and the new adherence target is given below:

- MRA thresholds >80: Always
- MRA thresholds <80 and >65: Usually
- MRA thresholds <65 and >45: Sometimes
- MRA thresholds <45: Never Taken as Prescribed

The class breakdown for each new adherence level is shown in Table 4.9.

Table 4.9: MEPS fibromyalgia breakdown for each of the four classes

	Class Breakdown for Fibromyalgia Records			
Threshold	MRA 80%	MRA 65%	MRA 45%	MRA 35%
Level	Always	Usually	Sometimes	Never Taken as Prescribed
Patients	714	192	408	1730

After class balancing, new models were trained using multi-class targets. Since a better performance for MEPS data was found with features AED and OPP (Table 4.5), two different models were created, one with these features and one without. The results for the MEPS test sets are given in Table 4.10. Even against a randomly guessing model (expected accuracy is 25%) the best of the four-class models only learned approximately 8% more information regarding adherence. This low performance may be attributed to the availability of a limited number of training samples as a result of class balancing and the reduced number of samples per class.

Table 4.10: MEPS fibromyalgia trained models with four class targets

	Best Model	Training Size	Testing Size	Accuracy
Without AED & OPP	RFT	616	152	28.9%
With AED & OPP	RFB	616	152	32.9%

In order to further investigate if poor performance is due to the limited number of samples in the training set, both the middle classes (usually, and sometimes) were combined resulting in a model with three classes. These two classes had the lowest number of samples (Table 4.9). The same training and testing procedure was then applied and the corresponding results are shown in Table 4.11.

Table 4.11: MEPS fibromyalgia trained models with three class targets

	Best Model	Training Size	Testing Size	Accuracy
Without AED & OPP	RFT	1440	360	49.2%
With AED & OPP	RFT	1440	360	50.0%

The three-class models (Table 4.11) have a better performance than their four-class model (Table 4.10) counterparts. Compared against a randomly guessing model (33%), the best model learned approximately 17% more information related to patient adherence. While this result is not significantly larger, it illustrates the potential for a more refined patient adherence multi-class model. The accuracy of each class in the multi-class model is shown in Table 4.12. The results indicate the middle class has the worst predictive accuracy of the three classes. The model is better at identifying patients on different ends of the adherence spectrum, namely very compliant patients and patients who never take their medications as prescribed.

Table 4.12: Accuracy distribution across each of the three classes

	Model Accuracy	"Always" Class Accuracy	"Sometimes/Usually" Class Accuracy	"Never Taken as Prescribed" Class Accuracy
Without AED & OPP	49.2%	70.0%	29.7%	51.9%
With AED & OPP	50.0%	66.0%	32.0%	55.0%

Table 4.13: PFT value comparison between the three class model and the binary target class model counterpart

	With AED & OPP		**Without AED & OPP**	
	RFT	**3 Class**	**RFT**	**3 Class**
TOM	24.6	25.1	29.3	28.8
YTT	12.6	13.9	16.2	17.1
SOM	14.9	13.9	18.5	18.1
REG	8.2	8.8	10.7	10.7
AGE	14.5	14.7	20.1	20.5
GEN	4.5	4.5	5.2	4.8
AED	11.0	10.4		
OPP	10.0	8.7		

With respect to the predictive features, the three class model PFT values compared to the RFT binary class model is shown in Table 4.13. The largest difference between the PFT of the two models was for the YTT feature for the models with AED and OPP, and the difference was only 1.3%.

4.7 RF with Imputed Means

As stated in Chapter 3, at least 40% of all the patients in the social forum for both disease conditions had at least one missing value. In order to properly predict patient adherence from a data source with a high prevalence of missing data, two modified approaches(RFB and RFT) using an RF classifier were presented. RFB required a pre-processing step where all features were binned using k-means clustering, and the missing values were placed in a separate bin. RFT did not require any pre-processing but did require a modification to the core algorithm where all missing values split into a separate third child. Indeed, both of these methods provided a process for handling missing values, without making an assumption about their values, while still differentiating them from known entries values.

An alternative to the above two approaches, considers imputing the missing values with the mean of the given feature. The results of the imputed means RF (IMRF) against the results presented in Tables 4.2 and 4.3 for RFT and RFB are shown in Table 4.14.

Table 4.14: The performance of IMRF compared to the previous RFT/RFB models. All models are trained using MEPS diabetes and fibromyalgia patients

	Diabetes				**Fibromyalgia**			
MRA	**IMRF**		**RFT/RFB**		**IMRF**		**RFT/RFB**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
80%	57.4%	61.9	58.9%	60.6	74.0%	75.5	70.2%	72.5
65%	56.4%	57.4	62.3%	66.9	76.8%	77.4	69.9%	70.0
45%	59.8%	59.8	61.6%	62.7	75.0%	75.1	73.0%	74.6
35%	65.8%	69.2	64.3%	67.6	75.0%	75.4	77.3%	79.3

Table 4.14, shows that IMRF is another viable RF implementation for adherence prediction. Moreover, the accuracy of IMRF is comparable to that of the RFT and RFB models. As in the previous procedure, both the AED and OPP features were removed from the best performing fibromyalgia model (MRA-65%) and the model was retrained and tested against a MEPS test and a social forum test set. The results are shown in Table 4.15.

Table 4.15: Comparison between MEPS IMRF and RFT fibromyalgia trained models without OPP and AED

Model	Testing Set	Training Size	Testing Size	Accuracy	F1 Score
RFT MRA 35	MEPS	2,436	609	73.5%	74.4
RFT MRA 35	Social Forum	2,436	357	68.6%	80.4
IMRF MRA 65	MEPS	1,450	352	77.9%	77.8
IMRF MRA 65	Social Forum	1,450	357	73.7%	79.4

The IMRF implementation is a better model for transfer learning than the RFT models. However, compared to RFT and RFB models, the IMRF is more sensitive to missing values as imputation cannot be used for a large number of records.

The PFT values of the IMRF models remained consistent between survey and social trained models as well as consistent with the RFT and RFB models. Table 4.16 shows the PFT values for all three models (RFT, RFB, and IMRF) in each domain for fibromyalgia.

The benefit of the RFB and RFT models is that they do not change the distribution of the data. Therefore, a mixed-mode model should be considered for features with a limited number of missing values. For features with a larger number of missing features either RFT or RFB should be used. This mixed-mode model will be investigated as part of future work.

Table 4.16: PFT for each feature in the fibromyalgia model following the removal of AED and OPP. Model 1: RFT with MRA threshold 35% using MEPS. Model 2: RFB with MRA threshold 35% using MEPS. Model 3: RFT using social forum data. Model 4: RFB using social forum data. Model 5: IMRF with MRA threshold 65%. Model 6: IMRF using social forum data.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
TOM	29.3	33.2	25.2	29.5	30.8	26.0
YTT	16.2	16.0	17.4	14.4	15.7	16.9
SOM	18.5	15.0	20.2	17.2	19.0	19.7
REG	10.7	14.1	13.2	16.7	10.4	11.0
AGE	20.1	13.6	20.4	14.2	19.3	22.5
GEN	5.2	7.9	3.4	7.9	4.8	3.8

5. CONCLUSION

A medication adherence prediction model was developed using MEPS survey data. This model was then tested on patients from the social forum PatientsLikeMe. While there were no clear demographic differences, we did not have access to other features such as race, education, or social status in the social forum dataset. Moreover, all the social forum records came from public profiles originating from only one source, PatientsLikeMe. A more comprehensive investigation will cover data aggregated from multiple public institutions. Moreover, while modified implementations and numerical methods were used to account for missing entries, many patients from the social form site contained at least one unknown entry.

Despite possible inconsistency in the social and demographic distribution of the two patient populations, the results show that using transfer learning between these two environments is possible. That is, a model trained using accurate but limited survey data can then be used to predict adherence from social forum data, which is available at scale and in a timely manner. Also, due to the large amount of adherent responses on the social platform, and the model identifying them as adherent, these results infer that many of these patients are not providing false reports and are actually more engaged in their healthcare than the average patient.

One of the reasons for successful transfer learning between these two domains can be attributed to the feature space used to predict adherence. Indeed, many of the patient features provided to the models were features that many social forum patients responded to accurately. For example, with features such as: AGE, GEN, REG, TOM, etc..., unless the patient is blatantly providing false information, the responses to these features are unambiguous. Whereas with OPP, one of the features that did not transfer between the two domains, many patients are taking multiple therapies and unless each patient is carefully documenting the average cost of each

specific therapy, false responses towards OPP is highly likely. This result also indicates that transfer learning may not be possible for all the features. In addition to OPP for the above reasons, AED also did not transfer between the two domains. Omitting these features improved the adherence prediction accuracy for social forum patients with models that are trained by using survey data.

This study did not find a single MRA threshold that proved to universally provide the best adherence prediction. For both diabetes and fibromyalgia models under the RFB and RFT implementations, the best performance in accuracy was observed at the 35% MRA threshold, while the IMRF implementation produced the best model for fibromyalgia at the 65% MRA threshold. A higher predictive performance at lower thresholds indicate models are better at differentiating between patients who rarely take their treatments, against all other patients. The inverse is true for all the models that are more accurate at higher MRA threshold predictions.

Finally, of all the RF implementations, the IMRF model produced the best results for social forum adherence prediction using transfer learning. However, the disadvantage for using imputed means is that it changes the entropy of the data. While it may have been effective in this case, this implementation may not be as scalable as the other RF implementations since social data is prone to missing data. One of the core objectives for this thesis is to develop a scalable and cost-effective approach to adherence prediction. The IMRF models may not comply with the very reason social data was selected. Based on the results of this thesis, RFT and RFB are recommended for social adherence prediction applications. A hybrid approach that combines imputed means and either RFT or RFB will be considered in future work.

Future work will also consider aggregating public data from other social sites and investigating temporal adherence prediction. This thesis appropriately identifies "if" patients will become adherent. An extension to this work could identify time to non-adherence. This can help support preventative measures for patients that are at risk for becoming non-adherent.

REFERENCES

REFERENCES

[1] T. Philipson, "Non-adherence in health care: Are patients or policy makers ill-informed?" May 2015, [Accessed: 2018-9-20]. [Online]. Available: https://www.forbes.com/sites/tomasphilipson/2015/05/08/non-adherence-in-health-care-are-patients-or-policy-makers-ill-informed/#1a0db3f44c4a

[2] A. O. Iuga and M. J. McGuire, "Adherence and health care costs," *Risk Management and Healthcare Policy*, vol. 7, pp. 35–44, 2014.

[3] L. R. Martin, S. L. Williams, K. B. Haskard, and M. R. Dimatteo, "The challenge of patient adherence," *Therapeutics and Clinical Risk Management*, vol. 1, pp. 189–199, 2005, 3.

[4] R. H. Friedman, L. E. Kazis, A. Jette, M. B. Smith, J. Stollerman, J. Torgerson, and K. Carey, "A telecommunications system for monitoring and counseling patients with hypertension: impact on medication adherence and blood pressure control," *American Journal of Hypertension*, vol. 4, pp. 285–292, 1996, 9.

[5] Express Scripts, "Predicting adherence," http://lab.express-scripts.com/lab/insights/adherence/infographic-predicting-rx-nonadherence, [Accessed: 2018-9-20].

[6] Allazo Health, https://allazohealth.com, [Accessed: 2018-9-20].

[7] FICO, "Fico medication adherence score," https://www.fico.com/en/products/fico-medication-adherence-score, [Accessed: 2018-9-20].

[8] A. A. Krumme, G. Sanfelix-Gimeno, J. M. Franklin, D. L. Isaman, M. Mahesri, O. S. Matlin, W. H. Shrank, T. A. Brennan, G. Brill, and N. K. Choudhry, "Can purchasing information be used to predict adherence to cardiovascular medications? an analysis of linked retail pharmacy and insurance claims data," *BMJ Open*, 2016.

[9] D. L. Labovitz, L. Shafner, M. Reyes Gil, D. Virmani, and A. Hanina, "Using artificial intelligence to reduce the risk of nonadherence in patients on anticoagulation therapy," *Stroke*, vol. 48, pp. 1416–1419, 2017, 5.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, 1.

[11] A. L. Boulesteix, S. Janitza, J. Kruppa, and I. R. Knig, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *WIREs Data Mining and Knowledge Discovery*, vol. 2, pp. 493–507, 2001, 6.

[12] D. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Random forest: a reliable tool for patient response prediction," in *The 2011 IEEE Conference on Bioinformatics and Biomedicine*, 2011.

[13] A. Gulia, R. Vohra, and P. Rani, "Liver patient classification using intelligent techniques," *Journal of Computer Science and Information Technologies*, vol. 5, pp. 5110–5115, 2014, 4.

[14] M. J. Paul and M. Dredze, "You are what you tweet: analyzing twitter for public health," in *The Fifth AAAI Conference on Weblogs and Social Media*, 2011.

[15] C. C. Freifeld, J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout, and N. Dasgupta, "Digital drug safety surveillance: monitoring pharmaceutical products in twitter," *Drug Safety*, vol. 37, pp. 343–350, 2014.

[16] C. L. Ventola, "Social media and health care professionals: benefits, risks, and best practices," *Pharmacy and Therapeutics*, vol. 39, pp. 491–499, 520, 2014, 7.

[17] E. Knight, R. J. Werstine, D. M. Rasmussen-Pennington, D. Fitzsimmons, and R. J. Petrella, "Physical therapy 2.0: leveraging social media to engage patients in rehabilitation and health promotion," *Physical Therapy*, vol. 3, pp. 389–396, 2015, 95.

[18] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *22nd ACM International Conference on Multimedia*, 2014.

[19] M. W. Newman, D. Lauterbach, S. A. Munson, P. Resnick, and M. E. Morris, "Its not that i dont have problems, im just not putting them on facebook: challenges and opportunities in using online social networks for health," in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 2011.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2010, 10.

[21] J. Sun, S. Staab, and J. Kunegis, "Understanding social networks using transfer learning," *Computer*, pp. 52–60, 2018.

[22] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Socialtransfer: cross-domain transfer learning from social streams for media applications," in *The 2012 ACM Conference on Multimedia*, 2012, pp. 649–658.

[23] R. Zou, M. Baydogan, Y. Zhu, W. Wang, and J. Li, "A transfer learning approach for predictive modeling of degenerate biological systems," *Technometrics*, vol. 57, pp. 362–373, 2015, 3.

[24] K. Pursey, T. L. Burrows, P. Stanwell, and C. E. Collins, "How accurate is web-based self-reported height, weight, and body mass index in young adults," *Journal of Medical Internet Research*, vol. 16, 2014, 1.

[25] Agency for Healthcare Research and Quality, "Medical expenditure panel survey," https://meps.ahrq.gov/mepsweb/, [Accessed: 2018-9-20].

[26] PatientsLikeMe, https://www.patientslikeme.com, [Accessed: 2018-9-20].

[27] L. M. Hess, M. A. Raebel, D. A. Conner, and D. C. Malone, "Measurement of adherence of pharmacy administrative databases: a proposal for standard definitions and preferred measures," *Annals of Pharmacotherapy*, vol. 40, pp. 1280–1288, 2006.

[28] L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," *Annals of Mathematics and Artificial Intelligence*, vol. 41, pp. 77–93, 2006.

[29] M. J. B., "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability*, 1967.