Talking About My Care: Detecting Mentions of Hormonal Therapy Adherence Behavior From an Online Breast Cancer Community

By

Zhijun Yin

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

December 16, 2017

Nashville, Tennessee

Approved By:

Qingxia Chen, Ph.D.

Bradley Malin, Ph.D.

ACKNOWLEDGMENTS

This research is sponsored by grant IIS1418504 of the National Science Foundation.

I would like to thank my thesis committee for their guidance in this work. Dr. Qingxia Chen, who served my committee chair, has been consistently supporting my thesis and giving me flexibility in building research topics. Dr. Bradley Malin, my primary advisor in PhD degree in Computer Science, has also been completely supportive for my study and research in Biostatistics.

I would also like to express my great thank to Dr. Jeffrey Blume, who is very kind to introduce, encourage and support me to pursue this wonderful MS degree in Biostatistics. I really appreciate him as well as other faculty and students for their dedication to providing such a great, unique learning experience, which makes me very proud to be one of them.

Finally, thank you, my lovely daughter Nora, my wife Tao, and my parents, for their endless supports, without which I can never make it.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Hormonal therapy adherence is challenging for many patients with hormone-receptor-positive breast cancer. Gaining intuition into their adherence behavior would assist in improving outcomes by pinpointing, and eventually addressing, why patients fail to adhere. While traditional adherence studies rely on survey-based methods or electronic medical records, online health communities provide a supplemental data source to learn about such behavior and often on a much larger scale. In this paper, we focus on an online breast cancer discussion forum and propose a framework to automatically extract hormonal therapy adherence behavior (HTAB) mentions. The framework compares medical term usage when describing when a patient is taking hormonal therapy medication and interrupting their treatment (e.g., stop/pause taking medication). We show that by using shallow neural networks, in the form of word2vec, the learned features can be applied to build efficient HTAB mention classifiers. Through medical term comparison, we find that patients who exhibit an interruption behavior are more likely to mention depression and their care providers, while patients with continuation behavior are more likely to mention common side effects (e.g., hot flashes, nausea and osteoporosis), vitamins and exercise.

Chapter 1

Introduction

Breast cancer is the most prevalent cancer among American women [1] and the second leading cause of death among women with cancer (just behind lung cancer) [2]. It is estimated that close to 12% of American women will eventually develop invasive breast cancer during their lifetime [3]. A common initial treatment for breast cancer is surgical intervention (e.g., lumpectomy or mastectoctomy), while adjuvant therapy (i.e., treatment after surgical intervention) is often invoked to reduce the risk of cancer recurrence [4]. In particular, hormonal adjuvant therapy is a popular treatment with a proven track record of significantly improving the long-term survival rate of patients with hormone-receptor-positive breast cancer [5]. This is notable because this disease subtype comprises 75% of all breast cancer cases [1]. To maximize this benefit of hormonal therapy, patients are prescribed a regimen of medication that is expected to continue for a minimum of five years [6]. For instance, taking tamoxifen (an oral hormonal therapy drug) for five years reduces breast cancer mortality by 33% in the decade after initial treatment [7]. Moreover, more recent evidence [8] suggests that maintaining a tamoxifen regimen for an additional five years can further reduce mortality by approximately 50%.

Despite the benefit of hormonal therapy for women, only around half complete a full five-year treatment [9]. There can be various reasons why breast cancer patients would fail to complete the regimen, ranging from adverse side effects [10] to progression of the disease into a terminal form [11, 12]. Still, there are many women who fail to stay on a recommended regimen for less obvious reasons [10]. As such, learning the factors associated with why women choose to stop (as well as stay on) hormonal therapy is critical to improving a patient's treatment experience. While there have been various investigations into regimen adherence [13, 14, 15, 16, 17, 11], most studies rely on *traditional* clinical re-

sources and methodology, such as formal survey-based studies [13, 15, 18, 16], electronic medical records (EMRs) and other clinical resources [19, 17, 11]. Though such traditional methods and data are valuable in healthcare research, there are certain drawbacks that should be recognized. In particular, survey-based methods are limited in that they typically incur high costs in time and money, often restricting a study to a smaller number of participants. Moreover, considering that breast cancer patients with hormonal therapy generally only have follow-up with their doctors every six months, this leads to a large information and time gap in traditional EMR systems about the patients' treatment (e.g., their feelings and experiences) between two visits.

The Internet, and social media in particular, has provided patients with the opportunity to seek and share treatment experiences in online environments. For instance, the *BreastCancer.org* website maintains an online discussion board for breast cancer patients to discuss any aspect of their daily lives they deem relevant. This includes, but is not limited to, their concerns, diagnoses, treatments, side effects and social support structure. This self-reported information provides a new opportunity to learn about breast cancer patients' treatment adherence - and on a much larger scale. With thousands of patients posting and interacting regularly and accumulating tens of thousands or greater (up to millions) of posts on discussion boards such as *BreastCancer.org*, one immediate research challenge that arises is how to efficiently leverage such rich text, ideally in an automated and less labor-intensive manner. More concretely, notable research challenges in this domain include: 1) mapping behavioral and health research questions to *ad hoc* self-reported information and 2) extracting useful knowledge from an environment that lacks formal mediation where true signals are often hidden in a vast amount of noise (e.g., patients can discuss anything they wish, including topics that are not directed related to cancer).

In this paper, we aim to develop a machine learning framework to distinguish mentions of hormonal therapy adherence behavior (HTAB) from other less relevant free-text contents in online health forums. In particular, we are interested in studying patient

behaviors (and their associated factors), such as taking a prescribed medication or interrupting treatment (e.g., stopping or pausing a regiment, or switching to a different medication). In our framework, the task of distinguishing mentions and non-mentions of HTAB is cast as a classification problem. To maximize the predictive performance of our framework, we extensively adapt and compose preprocessing and feature engineering techniques, as well as validate and interpret their effects. Our framework demonstrates that, through applying natural language processing and machine learning techniques, we can obtain an effective classifier to automatically detect mentions (and non-mentions) of hormonal therapy treatment adherence behaviors. Finally, we perform content analysis (through medical terms) to gain insight into the factors affecting how people communicate taking medication behavior and interrupting medication behavior.

Our work contributes to the field of user (or patient) generated online data (e.g., in social platforms and discussion communities), specifically where it is applied to supplement traditional data sources (e.g., EMRs) to study health related problems. In this research domain, we acknowledge that there is a growing collection of studies that cover a range of areas, including flu trends [20], mental health [11, 12], privacy issues about health mentions [21, 22] as well as how to build online communities to provide local cancer support [23]. Further, regarding this specific research topic, Freedman et al. [24] studied a large number of posts mentioning cancer treatments (including hormonal therapy) and identified treatment barriers that manifest from various aspects, including emotions, preferences and religious belief. Mao et al. [25] found that joint pain is the main reason patients stop taking Aromatase Inhibitors (AIs) treatment in online discussions of drug side effects. There have also been several studies that focus on *BreastCancer.org*, as discussed in a recent review [26], though the focus has been on different prediction problems.

Chapter 2

Methods

Our goal is to build an automatic framework to distinguish HTAB status (mentions and non-mentions) and learn their associated factors. Figure 2.1 shows the three main components of the proposed framework: 1) data preparation, 2) classifier building and 3) content analysis. Specifically, free-text data from users' postings are first collected from the hormonal therapy forum in *breastcancer.org* online discussion board. This yields a large amount of unlabeled text. Next, a subset of sentences containing at least one of seven common hormonal therapy medication keywords (e.g., Tamoxifen) are manually labeled based on their contents through a majority voting model. The labeled sentences are then applied to fit several candidate classifiers and the model with the best performance is applied to boost the number of labeled data. Finally, after extracting different HTAB, a regression analysis is applied to study associated factors.

2.1 Data sources

The online breast cancer community studied in this paper is maintained by Breastcancer.org, a non-profit organization that provides information about breast cancer. There are more than 170,000 registered members and 80 main forums in this online health community. Each forum is organized into different threads. Each thread has an initial post indicating a general topic (e.g., a question) that will be discussed (asked), with zero or more posts that follow the initial post. However, the posts that follow do not necessarily respond the initial post and could simply respond to each other. Additionally, certain threads may have a very small number of posts, while others could span years and consist of hundreds of posts. In this paper, we focus on one particular forum, *Hormonal Therapy*

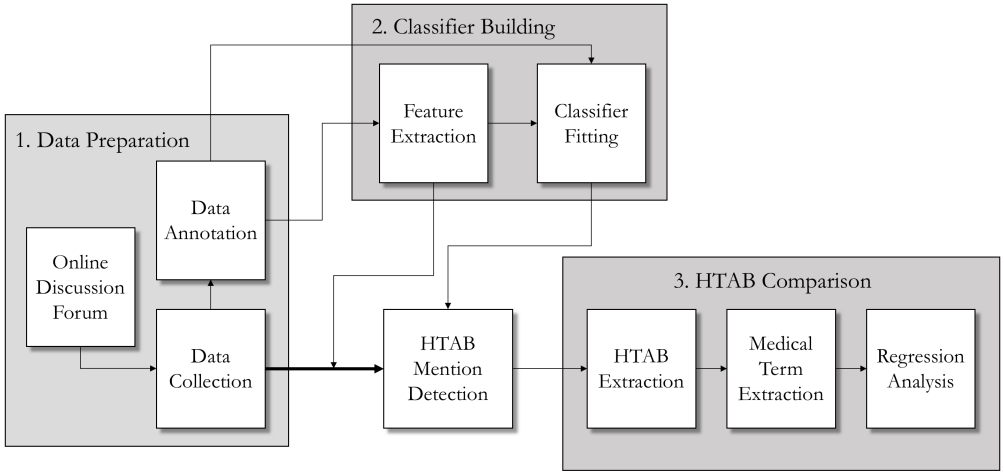

Figure 2.1: Framework for studying HTAB through an online breast cancer forum data. Three core components are highlighted in the figure: 1) data preparation, 2) HTAB mention classifier and 3) HTAB comparison.

- *Before, During and After*[1]. We collected all the posts published in this hormonal therapy forum before June 22, 2016. There are 9,996 users who participated in 5,995 threads with more than 130,000 posts over a 9-year period.

2.2 Data Annotation

In this study, we rely on supervised machine learning to distinguish mentions from non-mentions of HTAB. To engineer efficient classifiers, we manually annotated a gold standard dataset. Considering the rich context of the posts (e.g., posts in this forum can consist of multiple paragraphs), we focus on features at the sentence-level for the classification task. This is because we observe that most sentences used in this forum follow general grammatical rules and are sufficiently verbose to convey information of interest. However, there still exists a substantial number of irrelevant sentences and building a classifier on such an unbalanced dataset would seriously affect its performance. As such, we further constrain our classification objects to those sentences containing at least one of seven common hormonal therapy medication keywords (*Arimidex*, *Aromasin*, *Fe-*

[1]https://community.breastcancer.org/forum/78/

mara, Tamoxifen, Evista, Fareston and Faslodex). We also extend the keywords to include their corresponding generic names (*Anastrozole, Exemestane, Letrozole, Tamoxifen, Raloxifene, Toremifene, and Fulvestrant, respectively*) to account for variation in communication.

The goal of annotation is to obtain binary labels that indicate whether a given sentence communicates an HTAB or not. However, directly providing such labels to annotators may not be sufficient for them to understand the task. Thus, we adopted the hierarchical labeling strategy that was applied in [21]. Specifically, based on our observation of how patients discuss their treatments in the forum and guidance in a decision making codebook introduced by Beryl and colleagues in a hormonal therapy survey [16], we provide seven options for annotators to choose from to best describe whether a given sentence, containing at least one of the seven common medication keywords as mentioned above, conveys information about: 1) taking medication action, 2) stopping taking medication, 3) switching medication action, 4) taking medication intent, 5) not taking medication intent, 6) not decided yet, and 7) none of the above. The first three options correspond to mentions of HTAB, while the other four options belong to non-mentions of HTAB. We randomly select 1000 sentences and assign them to each of two annotators who are familiar with the online discussion forum to obtain labels. A third annotator is called upon if the first two annotators do not agree in their labels. The final label of each sentence is based on a majority rule.

2.3 HTAB Classification

N-gram features. There are two important steps in a typical classifier fitting process to note: feature extraction and algorithm tuning. While both steps can influence the performance of a classifier, in this task we focus on feature extraction with several common off-the-shelf learners. In text mining and natural language processing, term frequency - inverse document frequency (TF-IDF) is a common technique to help extract features. As the name suggests, the intuition behind this concept is that if a term is frequent in a docu-

ment (high TF) but scarce in other documents (high IDF), then the term will be distinctive to this document. As a result, the higher the TF-IDF value that a term exhibits, the more important the term is to distinguish the document from others. A natural choice for a term is a single word in a document. However, using words as features to represent a document loses the power in the ordering of the words. An alternative solution is to apply an n-gram word combination (i.e., linking n adjacent words together) as a term (e.g., "hot flash" as a 2-gram). Further, to reduce the impact of sparsity in natural language, the term could also be an n-gram of characters. Compared to other traditional high dimensionality reduction techniques (e.g., PCA), the advantage of this technique is that it maximizes interpretability [27].

Low dimensional representation features. Recently, another popular technique aimed at high dimensionality reduction in text mining is based on learning a low-dimensional representation of a word. These techniques make use of large amounts of data to discover the semantic similarities between terms through an embedding. For instance, word2vec [28] is a technique where each word is represented as a vector of a predefined dimension (e.g., 100). By building a model to predict terms through their neighbors, the model has been shown to learn vector representations of terms quite well. For example, when applying our collected hormonal therapy forum data to fit a word2vec model for the word of *tamoxifen*, the ten most similar words provided by the model (based on cosine similarity) were: *taxmox* (0.97), *tam* (0.91), *arimidex* (0.89), *tamoxifin* (0.88), *tamo* (0.87), *femara* (0.85), *aromasin* (0.79), *anastrozole* (0.78), *letrozole* (0.77) and *armidex* (0.77). The model provides semantically similar words (e.g., other hormonal therapy medications such as femara, aromasin and anastrozole), as well as its common misspellings (e.g., taxmox, tam and tamo). This is notable from a natural language processing perspective because they serve as a different type of feature to fit a classifier, while accounting for misspellings and abbreviations, which is common in online environments.

Classification Design. In this paper, we extract both n-gram words ($1 \leq n \leq 3$) and

n-gram characters ($2 \leq n \leq 5$) from sentences as features . We also introduce the mean of the vector representation (word2vec) of words in a sentence as another type of feature. For classification, we rely on logistic regression (LR), a SVM classifier (SVC) with RBF kernel and a random forest classifier (RFC). To compare the performance of different classifiers, we rely on stratified 10-fold cross-validation (CV), whereby for each 10 times of randomly data shuffling, we use 900 labeled sentences (out of a total of 1000) for training and the remaining 100 for testing. Note that CV can be applied for either 1) model parameter tuning to avoid overfitting or 2) performance evaluation to mitigate the effect of randomness. While it is rigorous to create another test data set from all of the collected corpus to evaluate the models performance, we did not apply CV to tune the hyperparameters in the model. Instead, we directly applied it to test the model. We report the average accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (AUC), along with their standard deviations. We also applied a t-test to ascertain if there is a significant difference in the performance of the classifiers. We trained the word2vec model by feeding all of the posts we collected (after removing the labeled sentences) into an implementation of gensim (version 0.31.0). We empirically set the dimensionality of the model to 100 and removed all of the words with frequency smaller than 5. The LR, SVC and RFC models were trained using scikit-learn [29] (version 0.18), where the hyperparameters were left as the default in the package.

2.4 HTAB Comparison

In this task, we aim to boost the amount of the labeled data using a machine learning technique. As such, we can continue to extract HTAB and conduct further behavior analysis. We will subsequently apply the classifier with the best precision to detect HTAB mentions from the remaining unlabeled sentences. In this paper, we mainly focus on two types of HTAB: 1) taking: where a user claimed that she is under treatment with some hormonal therapy medication, and 2) interruption: where a user claimed that she stops

(or pauses) a regimen, or switches to another different medication. We use patterns that we observed during labeling process to filter these two types of HTAB. For instance, we apply patterns that include (but are not limited to) *started*, *been on* and *stay on* to extract taking HTAB, and apply patterns such as *switched/switch*, *took me off*, *back on* and *stopped taking* to extract interruption HTAB.

After filtering the sentences, we retained the posts mentioning either a taking behavior or an interruption behavior (but not both). While people may have different reasons for interrupting a treatment, and they may even talk about the same behavior multiple times in the forum, in this study we do not aggregate posts published by the same user on the same type of behavior. This is because, when discussing a behavior such as taking, a person's feelings and health conditions may change with time. Moreover, a person may discuss their taking behavior with respect to different medications. Thus, to obtain meaningful concepts, we rely on ADEPT [30], a conditional random field (CRF) based classifier aimed to identify medical terms from patient-authored text. Due to space limitations, we refer the reader to the original publication for details on model performance. As such, we removed non-medical related terms from further analysis.

However, even if we focus on medical concepts only, more than 20,000 medical terms would remain. Thus, we applied two strategies to mitigate the high dimensionality problem: 1) we empirically retain the top 2,000 medical terms based on their TF-IDF values and 2) we use LR with lasso regularization, where posts mentioning an interruption behavior form the positive class and posts mentioning a taking behavior form the negative class. While lasso can serve as a feature selection approach by forcing the coefficients of many terms to zero, it has been found to be unstable [31]. This is because, given a different sample and random state, and when there are correlations between terms, lasso regression may result in different features selections. To resolve this problem, we adopt stable selection [32]. The basic idea behind this technique is to subsample the training data and fit a lasso regularized LR model, whereby the features that are repeatedly se-

lected across multiple runs of randomization receive higher scores (with a range from 0 to 1, the higher, the better). However, the scores cannot communicate which behavior the corresponding terms are informative for. Thus, to obtain directionality, we rely on the Pearson biserial correlation between the terms and the two behaviors [33]. We apply the implementation of stable selection sklearn (version 0.18) and the implementation of Pearson biserial correlation in scipy (version 0.81), with all possible hyperparameters set to their default.[2]

[2]Recently, open-vocabulary approaches [34] has become popular in natural language processing. In particular, while topic modeling is appealing and has the potential to complete a similar task, its results are often difficult to interpret. Further, the number of topics has to be determined ahead. Thus, we focused on medical terms because the they have domain-specific meaning and their interpretation is relatively straightforward.

Chapter 3

Results

3.1 Data Annotation

Table 3.1 shows the distribution of labeling results. There is an approximately even number of HTAB mentions and non-mentions. The two primary annotators exhibited a *very good* agreement level (Cohen's $\kappa = 0.80$) at the mentions vs. non-mentions level and a *good* agreement (Cohen's $\kappa = 0.72$) at the seven detailed options level. For the sentences designated as an action-related option, approximately 80% are about taking hormonal therapy medication, while only about 8% are about discontinuing hormonal therapy medication and 12% are about switching hormonal therapy medications.

3.2 Model Performance

Table 3.2 shows the performance of models fitted with various features and algorithm combinations. For each measure, the mean values and the standard deviations are reported. The maximal mean value of each measure is highlighted in blue. There are several notable findings to recognize.

First, models fit with word2vec features have a significant improvement over models fit with either single word or word n-gram features for almost all of the five metrics. For instance, in comparison to RFC fit with word features, RFC fit with word2vec had a 13.9%

	HTAB Mention			No Mention of HTAB			
	Action			Intent			
Option	Taking	Stop	Switch	Take	Not Taking	Undecided	None-of-Above
#Sent.	403	41	62	40	25	33	396

Table 3.1: The distribution of options in the 1000 labeled sentences (after the third annotator broke ties).

.

improvement on AUC, 11.9% improvement on precision, 21.5% improvement on recall, 16.2% improvement on F1 score, and 12.8% improvement on accuracy. This suggests word2vec features have a positive influence on HTAB mention detection.

Feature	Classifier	AUC	Precision	Recall	F1 Score	Accuracy
Single Word	LR	0.833 ± 0.030	0.757 ± 0.041	0.715 ± 0.032	0.735 ± 0.029	0.739 ± 0.031
	RFC	0.796 ± 0.029	0.755 ± 0.033	0.678 ± 0.049	0.708 ± 0.037	0.728 ± 0.035
	SVC	0.799 ± 0.022	0.506 ± 0.002	1.000 ± 0.000	0.672 ± 0.002	0.506 ± 0.002
Word n-gram(1, 3)	LR	0.829 ± 0.023	0.759 ± 0.033	0.735 ± 0.033	0.746 ± 0.026	0.747 ± 0.027
	RFC	0.803 ± 0.027	0.752 ± 0.037	0.686 ± 0.055	0.707 ± 0.032	0.721 ± 0.035
	SVC	0.829 ± 0.023	0.759 ± 0.033	0.735 ± 0.034	0.746 ± 0.025	0.747 ± 0.027
Character n-gram(2, 5)	LR	0.890 ± 0.025	0.806 ± 0.031	0.802 ± 0.034	0.804 ± 0.027	0.802 ± 0.027
	RFC	0.867 ± 0.031	0.753 ± 0.037	0.686 ± 0.037	0.706 ± 0.032	0.721 ± 0.036
	SVC	0.861 ± 0.022	0.506 ± 0.002	1.000 ± 0.000	0.672 ± 0.002	0.506 ± 0.002
Word2vec	LR	0.904 ± 0.025	0.817 ± 0.042	0.814 ± 0.027	0.815 ± 0.031	0.813 ± 0.034
	RFC	0.907 ± 0.026	0.845 ± 0.036	0.824 ± 0.039	0.823 ± 0.042	0.821 ± 0.033
	SVC	0.922 ± 0.022	0.830 ± 0.052	0.842 ± 0.036	0.836 ± 0.041	0.832 ± 0.044

Table 3.2: Performance of models fitted with different features and algorithms combinations. Stratified 10-fold cross validation was applied on the 1000 labeled data sentences. The maximal mean value of each metric is highlighted with blue color.

Second, models fit with character n-gram features exhibited significant improvement in AUC over models fitted with either single word features or word n-gram features ($p < 0.01$). As an example, the improvement for RFC was $8 \sim 9\%$. Further, the LR model fitted with character n-gram features significantly improved on the LR model fitted with single word features on all measures ($p < 0.01$).

Third, while the LR model fitted with character n-gram features have a similar performance to the LR model fitted with word2vec features, the latter improved both RFC and SVC on AUC, precision, F1 score and accuracy.

However, given the same type of features (e.g., either n-gran or word2vec features), there is no significant difference among the algorithms on the majority of the measures. It should also be noted that SVC with either single word features or character n-gram features obtained a perfect recall but with a severe cost of precision (almost equivalent to a random guess).

Based on Table 3.2, we can conclude that, for HTAB detection, word2vec features result in the best performance, followed by character n-gram features, and then word related features. Based on this finding, we chose to apply RFC fit with word2vec (with the highest average precision) to classify all of the remaining unlabeled sentences for further analysis. Moreover, the smaller standard deviation suggests that RFC is not overfitting.

3.3 Comparing Medical Terms Between HTAB

After filtering with the taking and interruption behavior patterns and extracting the medical terms, we obtained 19,174 posts published by 5,251 users. Among those posts, 13,461 (70.2%) discuss a taking behavior (the negative class), while 5,713 (29.8%) discuss an interruption behavior (the positive class). Among these users, 4,548 (86.7%) mention a taking behavior, 1,961 (37.3%) mention an interruption behavior, and 1,258 (24.0%) mention both taking and interruption behaviors.

After applying stable selection through 200 randomly generated subsamples, Table 3.3

Rank	Score	Term	Behavior	Rank	Score	Term	Behavior
1	1.0	femara	interruption	26	0.515	patient	interruption
2	1.0	exemestan	interruption	27	0.5	recurr	interruption
3	1.0	drug	interruption	28	0.495	carpel	interruption
4	1.0	aromasin	interruption	29	0.49	sweat	taking
5	1.0	arimidex	interruption	30	0.48	menopaus	interruption
6	1.0	ai	interruption	31	0.48	induc	interruption
7	0.985	pain	interruption	32	0.47	vacat	interruption
8	0.95	chemo	taking	33	0.47	tylenol	taking
9	0.945	radiat	taking	34	0.465	med	interruption
10	0.81	calcium	taking	35	0.46	exercis	taking
11	0.76	flash	taking	36	0.455	lumpectomi	taking
12	0.745	hot	taking	37	0.45	trigger	interruption
13	0.735	hormon	interruption	38	0.445	improv	interruption
14	0.68	sever	interruption	39	0.435	wellbutrin	interruption
15	0.655	depress	interruption	40	0.43	reaction	interruption
16	0.62	anastrozol	interruption	41	0.43	osteopenia	taking
17	0.565	hair	taking	42	0.43	estrogen	interruption
18	0.56	onc	interruption	43	0.43	anti	interruption
19	0.55	fog	interruption	44	0.425	bilat	interruption
20	0.545	therapi	interruption	45	0.42	vit	taking
21	0.54	oncologist	interruption	46	0.42	period	taking
22	0.53	surgeri	interruption	47	0.415	ekg	interruption
23	0.53	inhibitor	interruption	48	0.41	ultrasound	taking
24	0.53	effect	taking	49	0.41	cortison	interruption
25	0.525	nausea	taking	50	0.405	methotrex	interruption

Table 3.3: 50 medical terms that are the most useful in distinguishing posts mentioning taking behavior from posts mentioning interruption behaviors. The Pearson biserial correlation between all of these terms is significant at a level of 0.01 ($p < 0.01$).

shows the 50 medical terms that were most important for distinguishing between posts that mention a taking behavior and an interruption behavior. The Pearson biserial correlation between all of the terms was significant at the 0.01 level ($p < 0.01$). The goodness of fit measured with AUC is 0.765 ± 0.004, suggesting that the lasso models used in stable selection fit the observations well. Here, there are several notable results to highlight.

First, hormonal therapy medications, such as Aromatase Inhibitors (AI) (e.g., femara, exemestane (aromasin), and arimidex (anastrozole)), are more likely to be mentioned in posts related to an interruption behavior. It should also be noted that some users may

switch between different AIs, as stated by one user:

Example 1 *"... I had this problem with Femara and was taken off and switched to Aromasin ..."*

Second, common side effects of hormonal therapy medications are more likely to be mentioned with a taking behavior, such as hot flashes, hair (loss), nausea, sweat(ing) and osteoporosis. Common drugs or supplements mentioned with a taking behavior include Tylenol, vit(amin) and calcium. By contrast, depress(ion), pain, and fog are often likely to be mentioned with an interruption behavior. As a user complained:

Example 2 *"This spring I switched to Exemestane due to terrible depression, sleep issues, back-/shoulder/neck pain."*

It is not surprising to see that Wellbutrin, an antidepressant, is also more likely be mentioned with an interruption behavior. Note that methotrex(ate), a medication to treat cancer and "*usually given after other medications have been tried without successful treatment of symptoms*" [35] is also more likely to be mentioned with an interruption behavior. Heart-related terms such as carpel, cortison and ekg are also more likely to be mentioned with an interruption behavior.

Third, there are some types of terms that are mentioned with only one of the two types of behaviors. For instance, professionals (e.g., oncologist or onc) are more likely to be mentioned with an interruption behavior. This suggests that these users' interruption behavior may be associated with their physicians and possibly with permission (or by the suggestion of) other care professionals. As one user stated:

Example 3 *"In the 3 years I've been on Femara, I have negotiated 2 one-month "vacations" from Femara with my onc when the S/Es got too bad ..."*

People are more likely to mention exercise when discussing their taking behavior, but may mention recurr(ence) when talking about an interruption behavior. As one user said:

Example 4 *"I actually felt so bad on tamoxifen that I went off of it last year, then had recurrence 4 months later ..."*

Finally, people with an interruption behavior tend to mention surgery, which may be a possible reason why they pause a medication. People with a taking behavior are more likely to mention lumpectomy, ultrasound, chemo, radiat(ion), and period.

Chapter 4

Discussion

Findings. In this study, we built a framework to learn about hormonal therapy adherence behaviors, with a focus on classifier design and content comparison between taking and interruption behaviors. There are two notable contributions in this work: 1) we find that features based on embeddings (e.g., word2vec) can assist in establishing efficient detectors of HTAB mentions. For instance, compared to an RFC model fit with single word features, an RFC model fit with word2vec features resulted in a $\sim$21% improvement of recall and a $\sim$12% improvement on precision. The performance of the model ensures high quality for the extracted content; 2) by focusing on medical terms, we gain insights into the different factors associated with taking and interrupting hormonal therapy treatment behavior. For example, we find that people with an interruption behavior tend to mention depression and related antidepressant. This is in alignment with other studies [36, 37] where depression was found to be significantly associated with non-adherence to hormonal therapy treatment. Our findings further suggest that certain common sides effects (e.g., hot flashes, nausea and osteoporosis) may not be as likely to induce an interruption behavior severe as depression.

It is further interesting to note that people exhibiting an interruption behavior are likely to mention their care professionals. This suggests that their interruption behavior may be an artifact of, or even suggested by, their care professionals (e.g., take a break due to the following surgery). This is noteworthy because this interruption behavior might still be under their care professionals' control.

Implication. By relying on natural language processing techniques, machine learning models and statistical inference tools, we built an automated framework to study treatment adherence through an online breast cancer community. Given that patients with

long-term hormonal therapy tend to see an oncologist only twice per year, our framework provides a supplemental perspective (beyond routine clinical information) to learn about these patients' treatment experiences. Our findings, which are based on a comparison of medical term predictive ability, demonstrate the potential power of this type of online data to conduct treatment adherence research. Moreover, this framework may be extendable to assist in the study of adherence for other chronic diseases (e.g., depression or diabetes) through patient-authored online data. However, we acknowledge that domain-specific knowledge would be needed to customize behavior patterns for each disease. Still, once such information has been generated, the framework proposed in this paper could easily be adopted.

Limitation and further work. There are certain limitations that we highlight, which can serve as guidance for future research. First, we did not tune the hyperparameters of the models. It will be useful to determine if a combination of the proposed features can boost model performance. Second, we applied our observed patterns to filter the taking and interruption behaviors. While this strategy leads to high precision, it will miss adherence behavior mentions that fail to follow these patterns. Thus, as a next step, we anticipate including posts that communicate both taking and interruption behaviors with algorithms that achieve higher fidelity. Finally, we extracted medical terms without consideration for the grammar and context in which they are situated. We believe that more efficient models may be develop to detect different types of behaviors and interpretable medical concepts.

Chapter 5

Conclusion

In this paper, we proposed a framework to learn about hormonal therapy treatment adherence behaviors (HTABs) through an online breast cancer discussion forum. The framework consists of three core components: 1) data preparation, 2) classifier engineering and 3) comparison of HTABs. We analyzed a dataset consisting of over 130,000 posts across a 9-year period and demonstrated that features based on embeddings (e.g., word2vec) can help build a more efficient and effective classifier for HTAB mentions in online generated data. Furthermore, by comparing the predictive capability of medical terms used in describing taking and interruption behaviors, we discovered that people with an interruption behavior are more likely to mention depression and their care professionals, while people with a taking behavior are more likely to mention common side effects (e.g., hot flashes, nausea and osteoporosis), vitamins and exercise. This study demonstrates that an individual's discussion of hormonal therapy in an online environment may provide insight into treatment adherence behaviors. We further believe that this framework has the potential for extension to learn adherence for other chronic diseases treatment (e.g., diabetes and depression), provided domain-specific guidance is available.

REFERENCES

[1] Daniel F Hayes Kathleen I Pritchard and Sadhna R Vora. Adjuvant endocrine therapy for non-metastatic, hormone receptor-positive breast cancer. Webpage, 2016. Retrieved Sep 1, 2016 from http://www.uptodate.com/contents/adjuvant-endocrine-therapy-for-non-metastatic-hormone-receptor-positive-breast-cancer.

[2] Cancer among women. http://www.cdc.gov/cancer/dcpc/data/women.htm, 2016.

[3] U.s. breast cancer statistics. Webpage, 2016. Retrieved Sep 1, 2016 from http://www.breastcancer.org/symptoms/understand_bc/statistics.

[4] Adjuvant therapy for breast cancer. Webpage, 2016. Retrieved Sep 1, 2016 from https://www.mskcc.org/cancer-care/patient-education/adjuvant-therapy-breast.

[5] Caitlin C Murphy, L Kay Bartholomew, Melissa Y Carpentier, Shirley M Bluethmann, et al. Adherence to adjuvant hormonal therapy among breast cancer survivors in clinical practice: a systematic review. *Breast Cancer Res Treat*, 134(2):459–478, 2012.

[6] Carolyn Gotay and Julia Dunn. Adherence to long-term adjuvant hormonal therapy for breast cancer. *Expert Review of Pharmacoeconomics & Outcomes Research*, 11(6):709–715, 2011.

[7] Early Breast Cancer Trialists' Collaborative Group and others. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: Patient-level meta-analysis of randomised trials. *Lancet*, 378(9793):771–784, 2011.

[8] Christina Davies, Hongchao Pan, Jon Godwin, Richard Gray, et al. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: Atlas, a randomised trial. *Lancet*, 381(9869):805–816, 2013.

[9] Rowan T Chlebowski, Jisang Kim, and Reina Haque. Adherence to endocrine therapy in breast cancer adjuvant and prevention settings. *Cancer Prevention Research*, 7(4):378–387, 2014.

[10] Sayaka Kuba, Mayumi Ishida, Yoshiaki Nakamura, Kenichi Taguchi, et al. Persistence and discontinuation of adjuvant endocrine therapy in women with breast cancer. *Breast Cancer*, 23(1):128–133, 2016.

[11] Claudia Brito, Margareth Crisóstomo Portela, and Mauricio Teixeira Leite de Vasconcellos. Adherence to hormone therapy among women with breast cancer. *BMC Cancer*, 14(1):1, 2014.

[12] Nina Schmidt, Karel Kostev, Achim Jockwig, Iannis Kyvernitakis, et al. Treatment persistence evaluation of tamoxifen and aromatase inhibitors in breast cancer patients in early and late stage disease. *Int J Clin Pharmacol Ther*, 52(11):933–939, 2014.

[13] V Ziller, M Kalder, U-S Albert, W Holzhauer, et al. Adherence to adjuvant endocrine therapy in postmenopausal women with breast cancer. *Ann Oncol*, 20(3):431–436, 2009.

[14] Anne S Oberguggenberger, Monika Sztankay, Beate Beer, Birthe Schubert, Verena Meraner, et al. Adherence evaluation of endocrine treatment in breast cancer: methodological aspects. *BMC Cancer*, 12(1):1, 2012.

[15] Sumita S Bhatta, Ningqi Hou, Zakiya N Moton, Blase N Polite, et al. Factors associated with compliance to adjuvant hormone therapy in black and white women with breast cancer. *SpringerPlus*, 2(1):1, 2013.

[16] Louise L Beryl, Katharine AS Rendle, Meghan C Halley, Katherine A Gillespie, et al. Mapping the decision-making process for adjuvant endocrine therapy for breast cancer the role of decisional resolve. *Med Decis Making*, pages 79–90, 2016.

[17] B Makubate, PT Donnan, JA Dewar, AM Thompson, et al. Cohort study of adherence to adjuvant endocrine therapy, breast cancer recurrence and mortality. *Br J Cancer*, 108(7):1515–1524, 2013.

[18] P Wuensch, A Hahne, R Haidinger, K Meißler, et al. Discontinuation and non-adherence to endocrine therapy in breast cancer patients: is lack of communication the decisive factor? *J Cancer Res Clin Oncol*, 141(1):55–60, 2015.

[19] Annette Wigertz, Johan Ahlgren, Marit Holmqvist, Tommy Fornander, et al. Adherence and discontinuation of adjuvant hormonal therapy in breast cancer patients: a population-based study. *Breast Cancer Res Treat*, 133(1):367–373, 2012.

[20] Sarah C Vos and Marjorie M Buckner. Social media messages in an emerging health crisis: tweeting bird flu. *J Health Commun*, 21(3):301–308, 2016.

[21] Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. A scalable framework to detect personal health mentions on twitter. *J Med Internet Res*, 17(6):e138, 2015.

[22] Zhijun Yin, You Chen, Daniel Fabbri, Jimeng Sun, and Bradley Malin. # prayfordad: Learning the semantics behind why social media users disclose health information. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[23] Jacob Berner Weiss. *Building an online community to support local cancer survivorship: combining informatics and participatory action research for collaborative design*. PhD thesis, Vanderbilt University, 2009.

[24] Rachel A Freedman, Kasisomayajula Viswanath, Ines Vaz-Luis, and Nancy L Keating. Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment. *Breast Cancer Res Treat*, 158(2):395–405, 2016.

[25] Jun J Mao, Annie Chung, Adrian Benton, Shawndra Hill, et al. Online discussion of drug side effects and discontinuation among breast cancer survivors. *Pharmacoepidemiol Drug Saf*, 22(3):256–262, 2013.

[26] Shaodian Zhang, Erin OCarroll Bantum, Jason Owen, Suzanne Bakken, et al. Online cancer communities as informatics intervention for social support: conceptualization, characterization, and impact. *J Am Med Inform Assoc*, 24(2):451, 2017.

[27] Vandana Korde and C Namrata Mahender. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85, 2012.

[28] T Mikolov and J Dean. Distributed representations of words and phrases and their compositionality. *Adv Neural Inf Process Syst*, 2013.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*, 12:2825–2830, 2011.

[30] Diana Lynn MacLean and Jeffrey Heer. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *J Am Med Inform Assoc*, 20(6):1120–1127, 2013.

[31] Edouard Grave, Guillaume R Obozinski, and Francis R Bach. Trace lasso: a trace norm regularization for correlated designs. In *Adv Neural Inf Process Syst*, pages 2187–2195, 2011.

[32] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[33] John Paparrizos, Ryen W White, and Eric Horvitz. Detecting devastating diseases in search logs. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 559–568. ACM, 2016.

[34] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[35] Methotrexate. https://www.drugs.com/methotrexate.html, 2016.

[36] Catherine M Bender, Amanda L Gentry, Adam M Brufsky, Frances E Casillo, et al. Influence of patient and treatment factors on adherence to adjuvant endocrine therapy in breast cancer. In *Oncology Nursing Forum*, volume 41(3), page 274. NIH Public Access, 2014.

[37] Brent T Mausbach, Richard B Schwab, and Scott A Irwin. Depression as a predictor of adherence to adjuvant endocrine therapy (aet) in women with breast cancer: a systematic review and meta-analysis. *Breast Cancer Res Treat*, 152(2):239–246, 2015.