AUTOMATED LEARNING OF HEALTH BEHAVIORS THROUGH CONSUMER AUTHORED NATURAL LANGUAGE TEXT

By

Zhijun Yin

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Computer Science

February 28, 2018

Nashville, Tennessee

Approved:
Bradley Malin, Ph.D.
Ching-Hua Chen, Ph.D.
Daniel Fabbri, Ph.D.
Jeremy Warner, M.D.
Yevgeniy Vorobeychik, Ph.D.
Yuan Xue, Ph.D.

To my lovely daughter, Nora Yin,

who has always been trying to be a paper helper since her one year old

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest appreciation and gratitude to my academic advisor and committee chair, Dr. Bradley Malin, for his continuous invaluable support of my study in computer science and biostatistics, and my research in this dissertation. Dr. Bradley Malin is not only a great mentor with patience, but also a brilliant researcher with enthusiasm and diligence. I am very proud of working with him.

I would also like to give sincere thanks to my committee members, Dr. Jeremy Warner, Dr. Daniel Fabbri, Dr. Yevgeniy Vorobeychik, Dr. Ching-Hua Chen, and Dr. Yuan Xue, for their very constructive advice to help improve the work in this dissertation. Particularly, I would like to thank Dr. Ching-Hua Chen for sharing her remarkable insights in shaping my research direction when I had summer internship in IBM T.J. Watson Research under her supervision. Meanwhile, I want to express many thanks to Dr. Yuan Xue who introduced me to Vanderbilt University so that I can work with so many talented people.

Further, I want to thank my friends in both the Health Information Privacy Laboratory and Advanced Network System Laboratory in Vanderbilt University, as well as other collaborators in IBM T.J. Watson Research. They are like families sharing frustration, joy and hope. I will never forget their help and encouragement.

Most importantly, I am particularly grateful to my daughter, my wife, and my parents for their consistent love and support, without which I can never make it.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Traditional methods for collecting data in support of clinical research include prospectively collected surveys (e.g., [1]), retrospective analyses of existing medical records (e.g., [2, 3]), and a combination of the two (e.g., [4]). Over the past decade, computerized methods for data collection have emerged, with traditional surveys for health research moving onto the Internet [5] and increasingly widespread electronic medical records (EMRs) able to be mined to investigate a wide range of acute and longitudinal phenotypes [6, 7, 8]. At the same time, these approaches tend to focus only on a medically centric worldview and are likely to provide only a partial view of a patient's life.

As distributed systems, cloud services and mobile grow in sophistication and market penetration, large amounts of personal data are generated every day, particularly online environments, where a range of aspects of people's lives are disclosed, including health related information. For instance, it has been shown that certain people use general online social media platforms (GOSMPs) to share information about their mental health, including symptoms, treatments received, and the influence such a problem has on their social life [9]. Additionally, there are online health communities (OHCs), where individuals patients can freely and actively share many things related to their conditions. These online platforms range from general, all-purpose systems (e.g., MedHelp, PatientsLikeMe, DailyStrength, Tudiabetes, CureTogether, and Asthmapolis [10]), to more specialized environments such as breastcancer.org, copdfoundation.org, depressionforums.org, and myptsd.com. It should be noted that currently a growing number of patients use messaging functionality in patient portals to seek (or provide) health related information directly from (or to) their health provider [11, 12]. All of these together provide great opportunities for health providers and biomedical researchers to learn about their patients from their own voice and beyond traditional data sources.

There are an increasing number of studies that demonstrate that the data disseminated via GOSMPs (e.g., Twitter and Facebook), discussion conducted in online forums, and messages communicated through patient portals, can inform health-related investigations. We review these studies later on. Note we highlight that studies have shown, for instance, that such data can be mined to model aggregate trends about health (e.g., detection of statistically significant adverse effects of pharmaceuticals [13, 14]). Recent investigations have also demonstrated that an individual's health status can be corroborated by the statements they publish over GOSMPs (e.g., confirmation of flu diagnoses [15]). For example, the posts or discussion published on GOSMPs or OHCs can be applied to build effective depression prediction models [16, 17]. Secure messages through patient portals have also been shown to be valuable to improve quality of the clinic visit [18].

Given the potential benefits, we must continue to investigate what types of health related behaviors through these patients or consumers emerge through patient or consumer generated information: In particular - When and why do individuals choose to disclose their personal health status on GOSMPs? How does discussion in OHCs relate to treatment behaviors? Finally, what kind of information are patients seeking through secure messages in patient portals? Collecting, processing, and acting upon self-authored natural language text imposes challenges on automatically extracting health-related information, including, but not limited to, ambiguity in communication, noisy data, long exposition that contains many different types of health information, and high-dimensionality in predictive model interoperability.

1.1 Research Goals

In this dissertation, we focus on applying a combination of text mining, machine learning and statistical inference to study how data generated by potential healthcare consumers can be utilized to learn their health related behaviors. Particularly, we investigate three research goals (RGs), as illustrated in Figure 1.1.

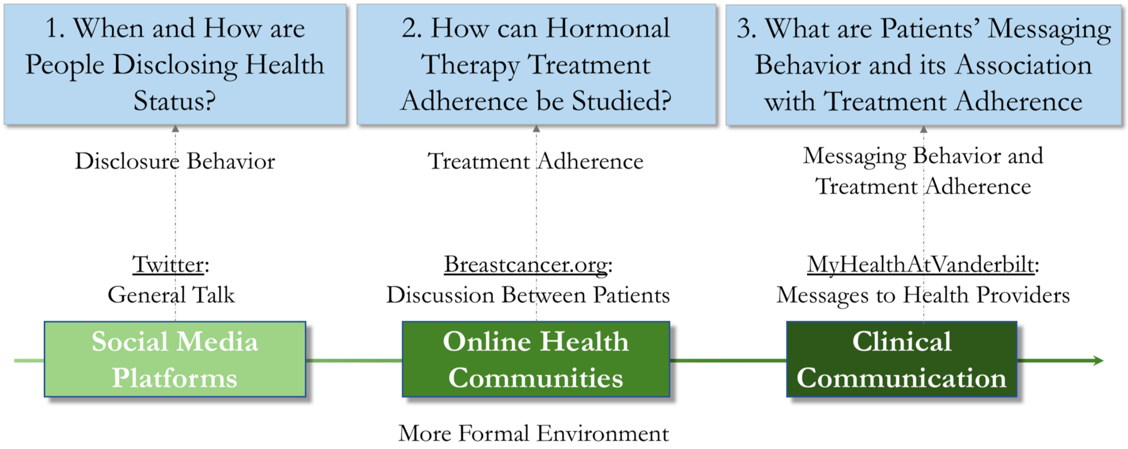

1. When and How are People Disclosing Health Status?
2. How can Hormonal Therapy Treatment Adherence be Studied?
3. What are Patients' Messaging Behavior and its Association with Treatment Adherence
Disclosure Behavior
Treatment Adherence
Messaging Behavior and Treatment Adherence
Twitter: General Talk
Breastcancer.org: Discussion Between Patients
MyHealthAtVanderbilt: Messages to Health Providers
Social Media Platforms
Online Health Communities
Clinical Communication
More Formal Environment

Figure 1.1: The three tasks that are investigated in this dissertation.

Figure 1.1 shows that the proposed automated learning process is conducted around three different types of consumer authored natural language text: 1) short posts in GOSMPs (for analysis of general health status); 2) discussion between patients in OHCs (for analysis of detailed treatment experiences in and out of the healthcare setting); and 3) messages sent through a patient portal from patients to their health providers. We defer the detailed computational steps, including data collecting, classification, inference and prediction to the description of each individual research goal.

RG1: Learning Semantics Behind Health Status Disclosure on Social Media Platform

The first research goal of this dissertation is to investigate the semantic factors that drive users to disclose personal health status about themselves and others on GOSMPs. While there has been research investigating the factors driving people to share personal health status [19], there has been little research into how collections of health issues, driven by communication semantics, relate to whom disclosures of health information on social media pertain to. In this dissertation, we introduce a scalable framework to automatically detect personal health status mentions a GOSMP, as well as propose a novel framework that relies on semantic analysis and non-negative matrix factorization (NMF), to computationally uncover similar health issues communicated through social media, as well as their association with an individual's propensity to disclose.

RG2: Learning Hormonal Therapy Adherence in An Online Breast Cancer Forum

However,GOSMPs such as Twitter are not suitable for patients to develop deep discussion on their health conditions, thus making it unlikely to accumulate sufficient information to learn their other health related behaviors. Hence, we move to more specialized OHCs to gain more insights on particular health issues. Specifically, we investigate breast cancer patients using data from breastcancer.org to fulfill this research goal. Particularly, we learn how medical concepts, emotions and personalities conveyed in posts are related to hormonal therapy treatment adherence. This includes building an efficient classifier to identity different adherence behaviors, extracting related factors and designing statistical models to learn associations between those extracted factors and adherence behaviors.

RG3: Learning Patients' Messaging Behavior and Its Association with Hormonal Therapy Medication Adherence

Finally, we go further to examine if secure messages generated in a clinical setting can be applied to learn patients' health-related behaviors as well. In this research goal, we examine breast cancer patients' messaging behaviors in *MyHealthAtVanderbilt* (MHAV), a patient portal in the Vanderbilt University Medical Center (VUMC). Specifically, we want to examine how breast cancer patients with hormonal therapy use the service, what topics are communicated in those messages, and how these messaging behaviors, in terms of messaging rates and topics, are associated with hormonal therapy medication discontinuation.

1.2 Dissertation Overview

The remainder of this dissertation is organized as follows: we describe our research regarding each research goal in Chapters II, III, and IV. In each chapter, we follow the convention of overview, background, related work, data preparation, main work and findings, and discussion. We will conclude the dissertation in Chapter V. We will user generated information (UGC) to refer to consumer generated natural language text when necessary, although UGC can contain other types of information such as images and videos. Our work has been published in Journal of Medical Internet Research (JMIR) and International AAAI Conference on Web and Social Media (ICWSM).

CHAPTER 2

LEARNING THE SEMANTICS BEHIND HEALTH STATUS DISCLOSURE ON A SOCIAL MEDIA PLATFORM

In this chapter, we present an investigation into behaviors of personal health status disclosure in GOSMPs. Discovering and characterizing disclosure behaviors in this domain can provide insight into what factors (e.g., which types of health issues) drive an individual to disclose. This is further notable because GOSMPs are often publicly accessible, such that when an individual discloses the nature of another individual's health status, it can induce privacy concerns. At the same time, the learned factors may provide healthcare researchers with a more insightful general picture into what information is available in GOSMPs and adapt their study designs accordingly.

2.1 Background

Traditional methods for collecting data in support of clinical research include prospectively collected surveys (e.g., [1]), retrospective analyses of existing medical records (e.g., [2, 3]), and a combination of the two (e.g., [4]). Over the past decade, computerized methods for data collection have emerged, with traditional surveys for health research moving onto the Internet [5] and increasingly widespread electronic medical records (EMRs) able to be mined to investigate a wide range of acute and longitudinal phenotypes [6, 7, 8]. At the same time, these approaches tend to focus only on a medically centric worldview, and may provide only a partial view of a patients life. Recognizing this limitation, investigators have suggested that that data contributed through non-traditional domains, such as mobile applications [20, 21, 22] and OHCs where patients self-report on their status [23, 24], will provide a more complete view of an individuals health and population-based health trends.

An increasing number of studies demonstrate that the data disseminated via GOSMPs, such as Twitter, can inform health-related investigations. We review such studies in the following section, but we highlight that studies have shown, for instance, that such data can be mined to model aggregate trends about health (e.g., detection of statistically significant adverse effects of pharmaceuticals [13, 14]). Recent investigations have also demonstrated that an individual's health status can be corroborated by the statements they publish over GOSMPs (e.g., confirmation of flu diagnoses [15]). Despite the power of such investigations, they are limited in that the associated approaches do not filter data from social media streams for any arbitrary health-related concept.

Further evidence suggests that disclosing information about the self online can be intrinsically rewarding [25], while sharing one's health status can assist in organizing networks and obtaining social support [26]. Yet, health information is considered one of the most sensitive aspects about an individual [27] and there is a perception that its disclosure has the potential to negatively impact personal privacy [28]. This begs the question of why (and when) individuals choose to disclose such information.

Several recent studies have investigated this issue by inquiring about which factors drive self-disclosure. In particular, one recent survey looked into the ways that youths disclose their personal health issues on GOSMPs [19], highlighting factors associated with trust and uncertainty. While it may be argued that concerns over personal privacy can be addressed by allowing an individual to choose when to disclose health information [29, 30], it must further be recognized that GOSMPs provide an opportunity for the disclosure of information about other individuals, often without consideration of their approval or consent. Specifically, it has been shown that individuals disclose information about a wide range of acquaintances, ranging from family members to friends to high profile persons in the media [31, 32].

Our *ad hoc* review of the social media posts suggests that the decision to disclose information may be context-dependent. As such, we anticipate that the semantics (e.g., language categories) an individual evokes when discussing a health problem could influence when they choose to disclose. Moreover, such semantics may correlate with whom the disclosure pertains to (e.g., the author of a post or a related individual). The potential for a semantic analysis with respect to posts about health information in GOSMPs is justified through evidence from prior investigations. Notably, semantic analysis has been applied to compare how the severity and social stigma of health issues drive people to seek via search engine or share in social media [33]. However, there has been little investigation into how collections of health issues, driven by communication semantics, relate to whom disclosures of health information on GOSMPs pertain to. Gaining an understanding of such factors could provide intuition into when an individual's privacy is put in jeopardy and if it is done so maliciously or simply to seek assistance or support. Moreover, by characterizing the semantics associated with such disclosure, it may be possible to develop programs to educate and, subsequently, mitigate the disclosure of other individual's information without their consent.

The objectives of our work in this chapter is to 1) develop a scalable framework for detecting mentions about personal health on a specific GOSMPs, namely Twitter, and 2) investigate the semantic factors driving people disclose whether their own or other people's health status. While the first objective is aimed to demonstrate whether it is possible to apply tweets regarding several health issues to detect health status mentions regarding a broad range of (other) health issues, the second objective is to learn the factors driving people health disclosure behaviors.

The detection system introduced in this research is composed of several core processes. First, the Twitter stream is filtered for tweets that are likely to contain health-related information. Next, a subset of the tweets are labeled with respect to the type of information that is communicated (e.g., health status of the author versus a metaphorical statement) and applied to train a classifier. While it is possible to label a large number of tweets given a substantial budget, it is unlikely that a classifier could be specialized for each specific health issue. For instance, imagine a researcher is interested in studying

10,000 distinct health issues, each of which will require at least 500 tweets to train a robust classier. If the cost to label each tweet is $0.10, it would cost $500,000 to build the necessary corpora! Our framework demonstrates that a scalable classifier, which discovers health mentions across a broad range of health issues, can be composed by leveraging a mixture of tweets from various health issues, which could make large-scale investigations much more cost-effective. In doing so, however, our system is oriented towards a high precision while maintaining a reasonable recall.

Further, in order to investigate people's health status disclosure behavior, we propose a novel framework that relies on semantic analysis and non-negative matrix factorization (NMF), to computationally uncover similar health issues communicated through social media, as well as their association with an individual's propensity to disclose. In this framework, we first apply a supervised classification model to distinguish tweets that communicate personal health issues from known confounding concepts (e.g., metaphorical statements that include a health-related keyword [34, 35]). Next, we annotate the tweets of each health issue with linguistic and psychological categories (e.g., social processes, affective processes and personal concerns). Then, we apply NMF over a space of health issue-by-language categories to obtain natural aggregations of the investigated health problems. Finally, we demonstrate that the semantics behind health issue mentions on Twitter are correlated with disclosure behavior.

There are several primary contributions of this research:

- **Labeled Health Mention Corpus.** We leverage Amazon Mechanical Turk (MT) to create a labeled corpus of tweets with health mentions for 34 health issues. These include certain high impact health issues investigated in the Medical Expenditure Panel Survey [36], such as arthritis, asthma, bronchitis, cancer, diabetes, hypertension, and stroke.

- **Health Mention Detection.** We introduce a system to automatically detect personal health mentions in tweet streams. We show that this system can be trained with a relatively small number of labeled tweets from several health issues. Moreover, it can effectively detect personal health mentions across a range of health issues on Twitter. For instance, training on 2000 tweets associated with four health issues (cancer, depression, hypertension, and leukemia) can yield a classifier that achieves a precision of 0.77 on the aforementioned corpus of tweets of 34 health issues.

- **Health Mention Attribution.** To demonstrate the potential for the data filtered from Twitter, we investigated how people reveal information about themselves and others. In doing so, we show that the likelihood an individual self-discloses is dependent on the health issues communicated. For example, for personal health status is revealed more than 50% for 11 of the 34 health issues. For certain health issues (e.g., allergies, bronchitis, insomnia, migraines, and ulcers), people are more likely to disclose their own heath status, while for other health issues (e.g., Alzheimer's, Down syndrome, leukemia, miscarriage, and Parkinson's), people are more likely to disclose another person's status.

- **Discovering Latent Factors.**We introduce a health issue-by-language category model (as opposed to the traditional document-by-term model) to study groups of health issues. By applying NMF on this model, we show the existence of four groups of health issues and their semantics, which correspond to: 1) common semantics, such as feeling and cognitive processes (e.g., insight and tentative), 2) biological processes (e.g., health - medicine, clinic, ingestion - eat, and taste), 3) social processes (e.g., family, friends, humans - girl, and women), and 4) negative emotions.

- **Interpreting Factors Driving Disclosure Behaviors.** Using over 200,000 tweets from a four-month period, we find that disclosure behavior is associated with semantically similar groups of health issues. Specifically, we show that major life-altering health issues related with family members, high medical costs and searching for social support (e.g., Alzheimers Disease, cancer, and Down syndrome) are more likely to have tweets disclosing other individual's health status. By contrast, we show that more benign health issues related with simple chronic biological processes and negative emotions (e.g., allergy, arthritis, asthma, and bronchitis) tend to have tweets with self-disclosed health status.

2.2 Related Work

Social Media and Health Research. As alluded to, various investigations have demonstrated that social media can be successfully leveraged to 1) enable individuals to discuss their health status, 2) influence an individual's health behavior and 3) support the analysis of aggregate trends around health activities.

First, a certain portion of studies have focused on the extent to which, as well as how, health information is self-reported via social media. It has been showed that users discuss their health conditions on public Facebook pages, but recognized that such pages tend to be overly general to attract users to contribute to a discussion [37]. However, another study found that individuals who use social media discuss certain ailments with high accuracy on Twitter [15]. Specifically, it was demonstrated that college students tend to talk about their influenza diagnosis and associated symptoms. More generally, latent topic model discovery was performed over self-reported health status in Twitter to detect complex and potentially novel phenotypes [38]. It has further been shown, that some Twitter users reveal genome sequencing results (in relation to ancestry information according to 23andme.com services) over Twitter [39].

Second, the previous investigations show that individuals publish information about themselves, but there is also a growing body of evidence to suggest that an individual's health behavior can be influenced by social media. In certain cases, social media may be exploited to bring about negative health behaviors. For instance, based on discussions about prescription abuse over Twitter, it was observed that social media may aggravate such problems [40, 41]. In a similar vein, a content analysis of tweets, in association with the demographics of the followers of marijuana Twitter accounts, showed that social media may allure young people to establish substance use patterns. However, it has also been shown that social media can encourage more positive changes in health behavior. Notably, it was shown that increasing communications with smokers on social media can promote free cessation services [42]. Moreover, Cobb and colleagues [43] developed a Facebook application that was able to track the significant elements of an intervention on smoking cessation. It was also found that the design and enaction of a community opinion leader model may mitigate the spread of HIV [44].

Third, social media can be mined to learn and characterize aggregate trends with respect to health activities. For instance, it was shown that flu trends can be effectively extracted from Twitter using standard machine learning strategies [45]. More specifically, the analysis of daily tweets across a major metropolitan region (e.g., New York) can enable the prediction of which health issues are currently influencing the health of the public [46]. Meanwhile, It was showed that both the keywords chosen to filter and create subgroups of tweets affected prediction accuracy [47]. Beyond health status, it has been illustrated that the rare or unknown side-effects of drugs can be discovered through sentiment analysis over Twitter [14].

Though social media can support a wide array of health-related investigations, there are a number of hurdles to making the associated methodologies scalable. As Curtis and colleagues [48] point out, for instance, insufficient procedures for protecting participants' privacy was one of the challenges to recruiting members from social media to conduct HIV research. In addition, it was recently revealed that the unreliability of big data and continuous changes of search algorithms contributed to failures in the Google Flu Trends program [49].

Our work differs from the aforementioned studies in that we focus on personal health status disclosure on Twitter. We note that a study [50] discussed the similar topic, but their work is limited in that 1) it relied on regular expressions for classification, 2) focused on a limited number of health issues, and 3) examined whether personal health status is disclosed, but did not differentiate whether heath status was disclosed for authors versus others.

Classification on Social Media. To mine health-related information from social media, it is critical to develop a classifier. However, tweets are constrained in size[1] and are, thus, composed of limited content. As a consequence, it is essential to define and select discriminative features to support automated health status detection. In certain studies, tweets were enriched with features by referencing external sources, such as Wikipedia [51, 52], to improve topic modeling, but their generality hamper them in the support of personal health mention detection.

As an alternative, it has been shown that punctuation, emoji characters, hashtags, and the *@username* designation, as well as text (including n-grams of words or characters [53]) from the webpage referenced by the URL in a tweet, can form meaningful features for classification purposes [54, 51, 55]. Features generated using natural language processing tools, such as part of speech tags and dependencies between terms have also been successfully incorporated as features in social media classifiers [56]. Building on previous studies, our work illustrates that nouns, verbs, pronouns, punctuation, emoji, hashtags, as well as dependencies, can serve as effective features for personal health mentions.

Social Media Corpus Construction. If we rely on a classifier to filter and analyze social media, then it is essential to obtain (or create) a labeled corpus to train the classifier. Crowdsourcing over online platforms, such as Amazon Mechanical Turk (MT), has been employed to generate labeled gold standard corpora [54]. Notably, AMT was leveraged to label when tweets were related to the health status of the author of a tweet in the latent topic modeling analysis discussed above [38]. However, it should be recognized that the survey utilized by [38] is limited in that it only related tweet content to the author and not another person's health status.

[1] The length of tweets was limited to 140 characters when this work was conducted, which has been changed to 280 and may change in future.

Seeking and Sharing Health Information. Various studies have provided intuition into what type of health information is communicated and/or sought over social media. For instance, it has been shown that, on Twitter, users with depression tend to publish content with a negative connotation and an expression of religious involvement [17]. On the other hand, as alluded to earlier, on Reddit[2], individuals with mental health problems (not limited to depression) provide information about challenges faced in daily life, as well as pose queries regarding certain treatments [9]. Similarly, cancer survivors in online forums on Reddit often shared information with personal narratives, while other online participants tend to ask for assistance immediately after diagnosis [57]. One study on loneliness on Twitter [58] showed that female users tend to be more likely to express more severe, enduring loneliness, but receive fewer responses (and presumably support) than male users. Several studies have also shown parents often turn to Reddit or Facebook to seek social support, but that their activities are often constrained by privacy concerns regarding the sharing of their children' health status [59, 60].

Detection of Personal Health Mentions. While the aforementioned studies discuss what behaviors people exhibit on social media, they do not necessarily address how to mine such information in an automated fashion. However, a growing number of approaches are being developed and applied to extract information from such settings. For instance, a character-based *n*-gram language model was shown to be effective for detecting tweets focused on post-traumatic stress disorder (PTSD) and depression [16, 61]. Additionally, language categories obtained from the *Linguistic Inquiry Word Count* (LIWC) framework, which we invoke in this study, have proven to be notable features for classifying specific health issues [17, 9, 62]. A word-based *n*-gram [38], as well as natural language processed outputs [32], have also been shown to be useful for building a universal health mention classifier, which cuts across a range of health issues. We note that the classification model we introduce in this work differs from previous investigations in

[2] a web content rating and discussion website.

that 1) we build a model to classify tweets associated with a broad range of health issues and 2) we combine the character *n*-gram model (which focuses on the level of a single tweet) and language categories (which focus on the broader level of a health issue and incorporate multiple tweets).

Factors Driving Health Disclosure. Studies that investigate the driving factors behind information disclosure have relied upon direct inquiry through surveys and inference via computational methods. Notably, one recent survey delved into the ways that youths disclose their personal health issues on social media [19]. The results suggested that these decisions were driven by 1) trust in social media platforms and 2) uncertainty about their physician's advice. Severity and social stigma of health issues have also been shown to be factors that motivate people to seek health information via web searches (e.g., via the Bing search engine) or share information in social media [33]. It has also been shown that some individuals with serious mental problems comment or upload videos to YouTube to seek peer support [63].

2.3 Data Preparation

In this study, we relied on the Twitter streaming API to collect tweets in English and published in the contiguous United States during a four-month window in 2014, which was approved by the institutional review board of Vanderbilt University (protocol 141150). The corpus of collected tweets (approximately 261 million) was filtered by a set of keywords associated with notable health issues. Specifically, we selected 34 health issues based on their high impact on healthcare as noted in the Medical Expenditure Panel Survey of the Agency for Healthcare Research and Quality (AHRQ) of the U.S. Department of Health and Human Services[3], as well as their popularity in Google Trends during the data collecting period. These health issues include chronic diseases (e.g., diabetes, hypertension, and arthritis), as well as more acute debilitating phenomena (e.g., stroke).

[3] http://meps.ahrq.gov/mepsweb/

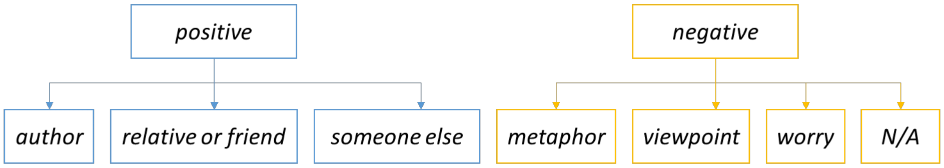


Figure 2.1: Label hierarchy.

Filtering the tweet stream resulted in a set of 281,357 tweets (i.e., a reduction of 99.89%).

To obtain the ground truth, we implemented a data annotation schema on Amazon Mechanical Turk (AMT) to investigate whether a given tweet containing the keywords discloses personal health status. Specifically, we randomly selected 100 tweets for each of the 34 health issues. Due to the diversity of the content and to help participants better understand the task, we provided seven options, defined as follows:

1. *The tweet discloses the health status of the author.*

2. *The tweet discloses the health status of the author's family members or friends.*

3. *The tweet discloses the health status of someone else, excluding the author, the author's family members and friends.*

4. *The tweet uses the health issue as a metaphor.*

5. *The tweet expresses a viewpoint on the health issue, or some kind of support to general patients with the health issue (excluding those specific persons mentioned in option 1, 2 and 3).*

6. *The tweet expresses a worry related with the health issue.*

7. *None of the above.*

Each participant, who was a certificated AMT master that continuously demonstrated high accuracy in the AMT marketplace, was required to select one, and only one, of these options to best describe the given tweet.

The answers used in data annotation schema were designed hierarchically, such that the seven options were compressed into several types of information for each investigated health issue as needed (see Figure 2.1): i) the number of self-disclosed tweets (option 1) and the number of tweets disclosing others (option 2 and 3), and ii) the number of tweets on health disclosure, including their own, and other people' health status (which we refer to as the *positive* class: options 1, 2 or 3), and the number of tweets not on health disclosure (which we refer to as the the *negative* class: options 4, 5, 6, or 7). We apply the first type of information to assess, for a certain health issue, if individuals are more likely to disclose their own health status or that of another person. We leverage the second type of information as the gold standard when building a binary classifier to automatically detect tweets with health mentions.

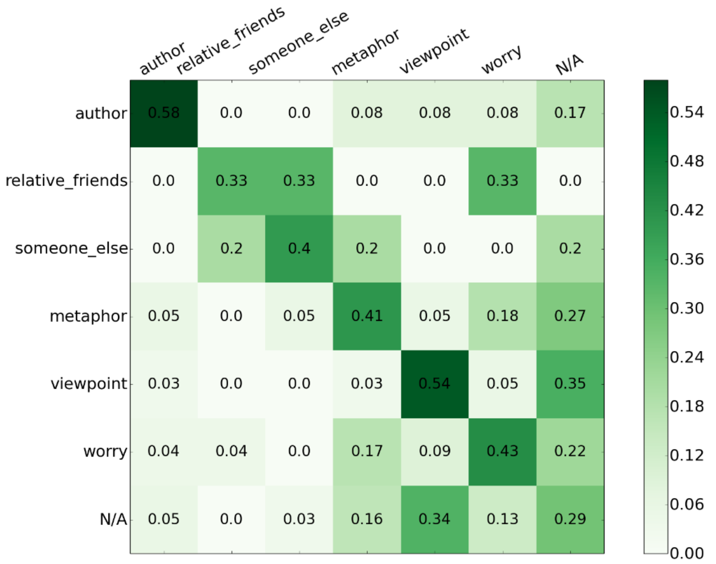


Figure 2.2: Confusion matrix of AMT labels for hepatitis.

Each tweet was labeled by two masters, while a third master was employed to break the tie when there is a disagreement on whether the tweet disclose personal health status or not[4]. There were 65 AMT masters who participated in the labeling task and 21 AMT masters were invoked to assist in the breaking of conflicting labels. The kappa score of

[4] In other words, we only handle labeling conflicts at the positive- and negative-class level. If one tweet, for instance, received option 2 and option 3 from two masters, respectively, then it is labeled as positive.

the agreement on the seven options level was 0.59 (an indicator of the complexity of the specific labeling task), but the kappa for the simpler positive- vs. negative-class level was 0.79. Figure 2.2 shows a confusion matrix of the labels received for hepatitis tweets. As the figure shows, the positive class exhibits mild confusion with the negative class. As such, the investigated tweets were categorized as two types: a positive class (41.5%) and a negative class (58.5%). Next, this dataset was applied to train a binary classifier to label additional personal health mention tweets in the larger corpus (i.e., the remaining 277,957 tweets) for further analysis.

For the purposes of this study, we created four types of datasets. The formalization of the design of these datasets is available in Appendix B. The first is referred to as the *gold standard* dataset and consists of all tweets with labels agreeing at the positive (negative) level. This dataset represents an ideal case where readers can determine when a tweet communicates personal health status. For example, tweets labeled as *author* by one MT master and *someone else* by a second MT master are treated as positive. By contrast, tweets labeled as *relative or friend* and *worry* are discarded.

Given the difficulty in labeling tweets in practice, three additional datasets are generated to resolve label conflicts. The first is the *conflict as positive* (CAP) dataset, which treats tweets with conflicting labels as positive. The second is the *conflict as negative* (CAN) dataset, which treats tweets with conflicting labels as negative. The third is the *TieBreak* dataset, which uses a third MT master to break the tie. These datasets represent the best case, the worst case, and the general case in the real world and will be relied upon to assess the system's scalability.

Additionally, in preparation for our analysis, we computed the ratio of the number of tweets disclosing the author's personal health status to the number of tweets disclosing another person's personal health status. We refer to this as the *Me vs. You*, or MvY ratio, for each health issue. The larger the MvY ratio for a health issue, the more likely it is that the corresponding tweets disclose their authors' personal health status. Figure 2.3

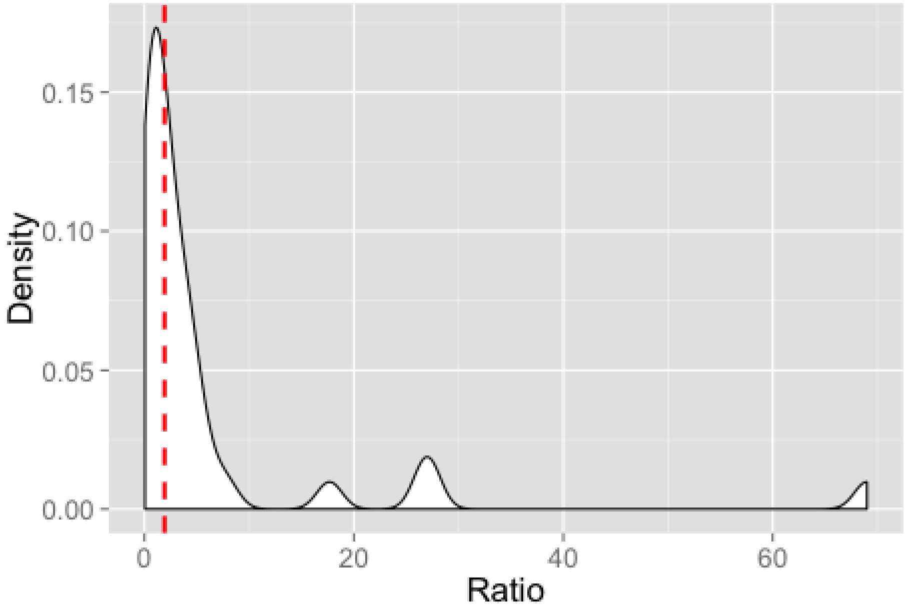


Figure 2.3: The density of the me vs. you (MvY) ratio. The red dashed line represents the median MvY ratio.

illustrates the density of the MvY ratio per health issue. It was observed that there was a strong positive skew suggesting that there are many health issues for which the author is more likely to self-disclose.

2.4 Observation

To demonstrate the opportunities for a personal health mention detection system, we conducted an investigation to test three hypotheses that are defined as follows:

H1: People discuss personal health status on Twitter.

H2: Personal health status disclosure rate is health issue dependent.

H3: The likelihood that people disclose their own versus other people's personal health status is health issue dependent.

We chose 100 tweets, at random, for each of the 34 health issues as shown along the x-axis of Figure 2.4, to generate the *TieBreak* dataset. These health issues are based on common and high impact health issues as defined by the Medical Expenditure Panel Survey [36].

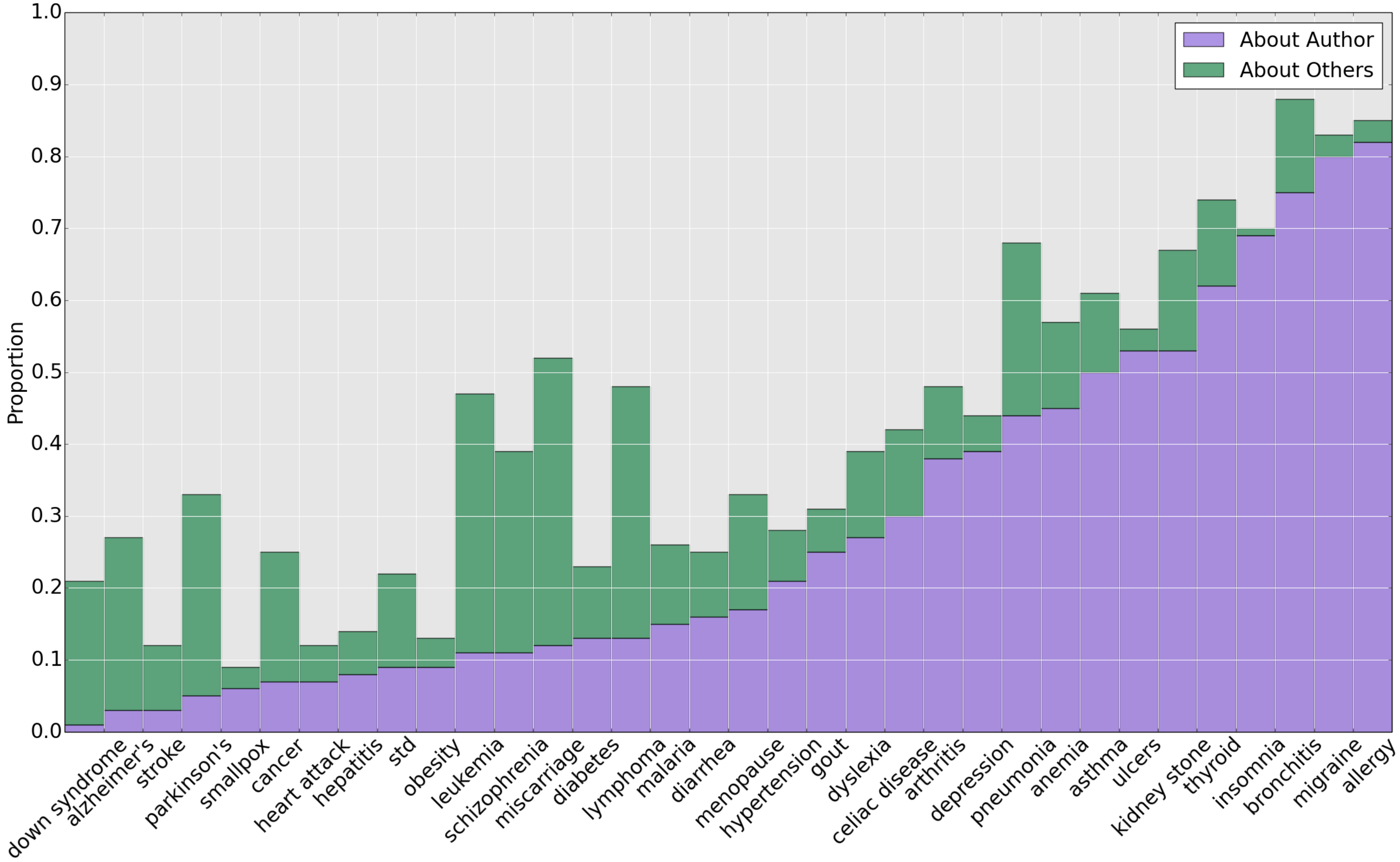


Figure 2.4: The extent to which people tweet about themselves versus others when disclosing personal health status. Note that this is a stacked bar chart, such that the sum of the *author* and *others* proportions corresponds to the overall proportion of positive instances.

Figure 2.4 illustrates how often people disclose their own health status as opposed to other individuals' status. The black bar, "About Author", represents the proportion of positive tweets with the *author* label. The gray bar, "About Others", represents the proportion of positive tweets with the label *relative or friends* and *someone else*. For a specific health issue, the sum of the two values is equal to the proportion of positive tweets for this health issue. For example, 40% of the tweets about miscarriages disclosed other people's status, while only 12% disclosed the author's status (such that 52% of the tweets were positive instances).

To test hypothesis H2 (personal health status disclosure rate) and H3 (who the disclosure is about), we define the following null hypotheses:

$H2_o$: The rate of positive and negative tweets is independent of the health issues.

$H3_o$: The rate of tweets disclosing the author's health status and others' health status is independent of the health issues.

To test these hypotheses, we used the *TieBreak* dataset, the 100 samples from each of the 34 distributions regarding how people disclose health status. To test H2, we applied a χ-square test on two variables: the number of positive tweets and the number of negative tweets in each health issue samples. To test hypothesis H3, we apply a Wilcoxon signed-rank test on two variables: the number of tweets disclosing the author's health status and the number of tweets disclosing the others' health status. We set the α level of significance to 0.05.

The results reveal several notable pieces of evidence, which are related to the the three hypotheses:

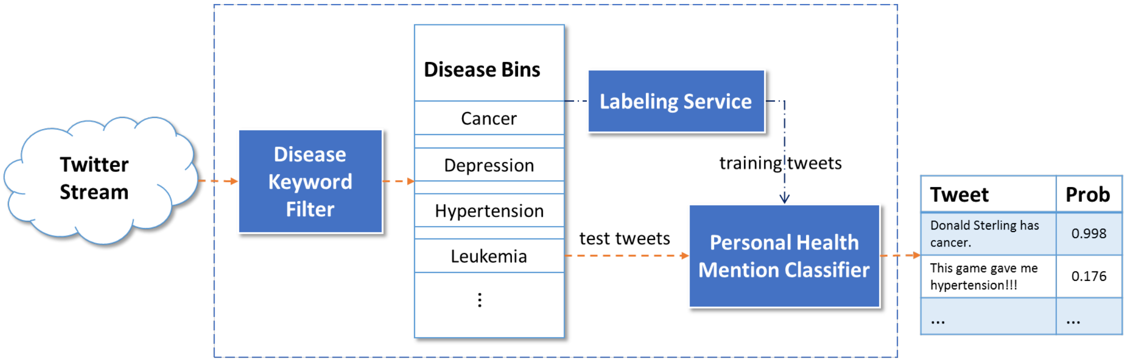
- **People disclose personal health status on Twitter for a range of health issues (H1)**: The disclosure rate for each of the 34 health issues is greater than 9%. There are 29 health issues with disclosure rates greater than 20% and 11 health issues with disclosure rates greater than 50%. The latter group includes: allergies, anemia, arthritis, asthma, bronchitis, insomnia, kidney stones, migraines, miscarriages, pneumonia, thyroid problems, and ulcers.

- **Health status disclosure rate is dependent on the health issue,** $\chi^2(33, N = 100) = 669$**,** $p < 0.001$. For instance, more than 80% of the tweets about migraines and allergies communicate personal health status. By contrast, only $\sim$ 10% of tweets about obesity and heart attacks communicate personal health status. Bronchitis exhibits the largest proportion of tweets ($\sim$88%) that disclose personal health status, while smallpox exhibits the smallest proportion ($\sim$9%).
- **The likelihood that people disclose their own versus other people's health status is dependent on the health issue,** $Z = 3.370$**,** $p < 0.001$. For instance, more than 70% of tweets about insomnia disclose the author's personal health statuses compared, while only 1% disclose another person's status. By contrast, $\sim$2% of the tweets for Down syndrome disclose the author's status, while $\sim$20% disclose another person's status.

2.5 Building A Scalable Classifier

2.5.1 System Pipeline

Figure 2.5 provides a high-level summary of the system engineered to detect personal health mentions on Twitter. The system is composed of three primary components: 1) a filtering service (e.g., a keyword filter based on health issues), 2) a labeling service, and 3) a health mention classification service. First, tweets collected via the Twitter streaming API are passed into a filter and stored in a bin indicative of a specific health issue. Next, a sample of the tweets associated with these health issues are sent to a labeling service (e.g., AMT). Once labeling is complete, a personal health mention classifier is trained and applied to report the probability that new incoming tweets correspond to such mentions.



Tweet	Prob
Donald Sterling has cancer.	0.998
This game gave me hypertension!!!	0.176
...	...

Figure 2.5: The framework for personal health mention detection over Twitter. First, tweets are filtered into bins according to their health issue topic. A portion of the tweets are supplied to a labeling service. The labeled data is then applied to train a classifier to detect personal health mentions.

2.5.2 Construction of a Health Mention Corpus

For the purposes of this study, we created four types of datasets. The formalization of the design of these datasets is available in Appendix B. The first is referred to as the *gold standard* dataset and consists of all tweets with labels agreeing at the positive (negative) level. This dataset represents an ideal case where readers can determine when a tweet communicates personal health status. For example, tweets labeled as *author* by one AMT master and *someone else* by a second AMT master are treated as positive. By contrast, tweets labeled as *relative or friend* and *worry* are discarded.

Given the difficulty in labeling tweets in practice, three additional datasets are generated to resolve label conflicts. The first is the *conflict as positive* (CAP) dataset, which treats tweets with conflicting labels as positive. The second is the *conflict as negative* (CAN) dataset, which treats tweets with conflicting labels as negative. The third is the *TieBreak* dataset, which uses a third AMT master to break the tie. These datasets represent the best case, the worst case, and the general case in the real world and will be relied upon to assess the system's scalability.

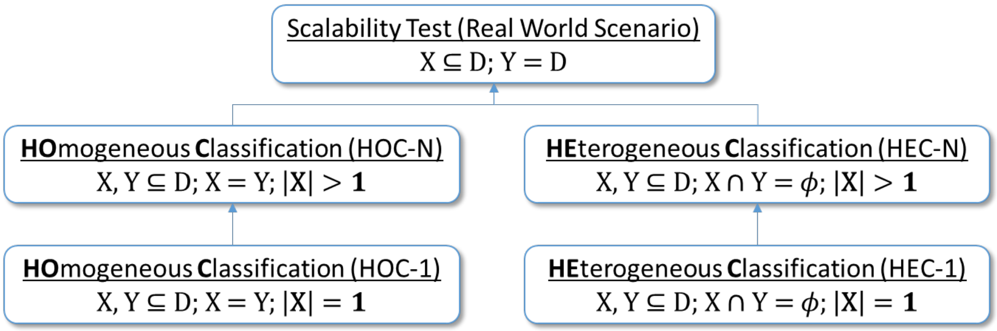

Scalability Test (Real World Scenario)
X ⊆ D; Y = D
HOmogeneous Classification (HOC-N)
X, Y ⊆ D; X = Y; |X| > 1
HEterogeneous Classification (HEC-N)
X, Y ⊆ D; X ∩ Y = ϕ; |X| > 1
HOmogeneous Classification (HOC-1)
X, Y ⊆ D; X = Y; |X| = 1
HEterogeneous Classification (HEC-1)
X, Y ⊆ D; X ∩ Y = ϕ; |X| = 1

Figure 2.6: An overview of the evaluation strategies for the personal health status mention classifier. Note, $D = \{d_1, d_2, \ldots, d_n\}$ is the set of health issues, X is the set of health issues selected to train the classifier, and Y is the set of health issues used to test the classifier.

System Classifier Evaluation Roadmap

System scalability emphasizes the ability to detect mentions for many, potentially unknown, health issues communicated via social media, using the labeled tweets from a limited number of health issues.

To formalize the scenario, let D be the set of health issues and X and Y be the set of health issues selected to train and test the classifier, respectively. By default, $X, Y \subseteq D$.

As depicted in Figure 2.6, there are two variations on classification that we assess. The first, which we refer to as *homogeneous classification*, corresponds to the traditional machine learning setting where a classifier is trained and tested on tweets from the same health issue. The second, which we refer to as *heterogeneous classification*, corresponds to when the classifier is trained and tested on tweets from disparate health issues. This type of scenario arises when a researcher attempts to reuse a classifier developed for one health issue on a different problem. Figure 2.6 further illustrates two training strategies to scale the system in a real world scenario: 1) train the classifier on tweets from one health issue, which results in homogeneous classification with $|X| = 1$ (HOC-1) and heterogeneous classification with $|X| = 1$ (HEC-1), and 2) many health issues, which results in homogeneous classification with $|X| > 1$ (HOC-N) and heterogeneous classification with $|X| > 1$ (HEC-N).

The ideal scalability test is to train an HOC-1 classifier for every health issue in *D* with a sufficient quantity of labeled tweets. However, it is difficult to realize this scenario in practice because of limited budgets (e.g., time and funds) for gathering and annotating such corpora. As such, we performed a series of experiments to compare the performance of the various models (i.e., HOC-1, HOC-N, HEC-1, and HEC-N) and leverage the best model to conduct scalability tests in a real world scenario.

2.5.3 Performance Measures

To assess the performance of the system, we rely upon the standard measures of precision and recall. In our setting, (*P*)recision corresponds to the proportion of tweets classified as positive that are in fact positive. (*R*)ecall corresponds to the fraction of real positive tweets that are classified as positive. Given the large volume of tweets and the often unbalanced positive/negative class ratio per health issue (see Section 2.6.1), we emphasize P while setting R to a reasonable level.

Henceforth, we report the area under the PR curve (AUPRC) to evaluate how a classifier performs in general. We consider the PR curve, which can be more indicative of a classifier's performance when the class ratio is highly imbalanced [64]. To characterize general performance, we report on AUPRC when testing the scalability of the system.

Health Status Classifier

We use a Naïve Bayes (MNB) classifier based on four types of features associated with tweets. Other learning algorithms, such as logistic regression, linearSVM, can also be plugged into the framework as the base classifier.

1. **Nouns, verbs and pronouns**. Each word is converted into its lemma form. Though pronouns are often defined as stop terms (which are discarded in traditional natural language processing), they are retained because they can disclose the personal

health status of a friend or family member (e.g., "*My mom makes having cancer look good*").

2. **Dependencies**. These are grammatical relations [65] between words in a tweet, such that one of the words is a health issue. Terms for health issues are replaced with the keyword *diagnosis* to compact the feature space. For example, the dependency (*"dobj", "have", "cancer"*) is converted into a feature that can be supplied to MNB, *dobj_have_diagnosis*.

3. **Punctuation** and **Emoji**. These can indicate an author's emotion and may improve classification (e.g., "*my uncle is cancer free !!!!!! lol*").

4. **HTTP_LINK**, **#hashtags**, and **@username**. These features represent the existence of link, hashtag, and @username in a tweet, respectively. **HTTP_LINK** represents the links that mentioned in the context, **#hashtag** is a keyword to describe a topic, and **@username** is a notification to inform the user with username about the posting of the context.

2.5.4 Experiment Design

In our experiments, we highlight the evaluation of two important factors that can affect the scalability of a classifier: 1) the diversity of health issues in the training data and 2) the quantity of training tweets. When we compare different classifiers, we focus on the former. When we test system scalability, we also evaluate the performance of the classifiers with different sizes of training dataset. The details of the experiment design are described as follows:

Dataset. We use the 34 health issues depicted in Figure 2.4 to represent *D* and define a synthetic health issue, or SYND, as the union of *cancer*, *depression*, *hypertension*, and *leukemia*. We select *cancer* and *leukemia*, for which tweets are skewed towards communicating about other people's heath status, and *depression* and *hypertension*, for which tweets are skewed towards communicating about the author's health status. We choose 1000 tweets, at random, for each of the four health issues to obtain the gold standard datasets. We also choose 100 tweets, at random, for each of health issue in D to generate gold standard, CAN, CAP and TieBreak datasets.

Comparison between HOC-1 and HOC-N. We use the cancer, depression, hypertension, and leukemia gold standard datasets to train each homogeneous classifier. There are two situations where we can evaluate how homogeneous classifiers are impacted by the diversity of health issues in the training data. First, suppose that we aim to detect multiple health issues. Given a fixed number of training tweets, how does an HOC-N classifier (e.g., trained with SYND) differ from a group of HOC-1 classifiers (e.g, four HOC-1 classifiers)? Second, now imagine we wish to perform detection for only one single health issue (e.g., cancer). Given a fixed number of training tweets, how does a HOC-N classifier (e.g., trained with SYND and test on cancer) differ from the associated HOC-1 classifier (e.g., cancer HOC-1 classifier)?

Comparison between HEC-1 and HEC-N. To evaluate the diversity of health issues in training dataset, we compare HEC-1 with HEC-N ($2 \leq |X| \leq 4$). In particular, we use the cancer, depression, hypertension and leukemia gold standard datasets for training and the gold standard dataset of $D \setminus SYND$ to test all of the heterogeneous classifiers.

System scalability test. When assessing system scalability, we test the classifier on the CAN, CAP, and TieBreak datasets of D. This enables the evaluation of the performance of the system in a real word scenario. We also test the classifier trained with different number of tweets.

Experimental Methodology. For each experiment, the related labeled tweets are stratified to generate 30 train-test sets. Each set preserves the proportion of samples for each positive (negative) class. The data is partitioned, such that model training is performed on 80% of the tweets, while testing is performed on the remaining 20%. To control the comparison, the size of the training set for each compared classifier is equivalent.

Tweet	**Cancer**	**Depression**	**Hypertension**	**Leukemia**	**SYND**
Positive	166 (19.2%)	261 (36.1%)	211 (27.7%)	436 (50.8%)	1074 (33.5%)
Negative	697 (80.8%)	461 (63.9%)	551 (72.3%)	423 (49.2%)	2132 (66.5%)

Table 2.1: The number of positive and negative tweets in the *gold standard* datasets.

2.6 Classification Evaluation

2.6.1 Classification Data Set

We extracted the gold standard datasets for each of the four health issues mentioned in Section 2.5.4. Table 2.1 summarizes the number of tweets in each class. Except leukemia, which has a balanced positive and negative instance space, there were substantially more negative than positive tweets. Due to the definition of SYND, the number of positive and negative tweets of the synthetic health issue is the sum of the four health issues.

2.6.2 Most Informative Features

Rank	**Cancer**	**Depression**	**Hypertension**	**Leukemia**	**SYND**
1	I	I	I	I	I
2	my	my	my	my	my
3	!	@username	have	@username	@username
4	@username	you	@username	HTTP_LINK	!
5	you	it	dobj_have_diagnosis	!	have
6	have	go	!	she	HTTP_LINK
7	she	poss_diagnosis_my	get	have	she
8	he	!	she	he	you
9	HTTP_LINK	get	it	battle	obj_have_diagnosis
10	obj_have_diagnosis	have	blood	help	he

Table 2.2: The most informative features for homogeneous health mention classification.

Before conducting an in-depth empirical investigation, we inspected the classifiers and their corresponding features to determine if they are intuitive. Here, we report on the top 10 informative features by training in a homogeneous classification setting with tweets of each of the five health issues (cancer, depression, hypertension, leukemia and SYND). The features are selected based on their average coefficient values of classifiers in cross validation. Table 2.2 reports these features for each classifier.

The results show the effectiveness of feature selection in several ways. First, more than five features are pronouns, such as *I*, *my* and *she* (which was also confirmed in [50]). These are stop words that are typically removed in the context of general text classification. However, in our scenario, they appear to signify users who disclose health information about themselves and others (e.g., "my mom makes having cancer look easy"). Second, certain words, such as *get*, *have* and *battle*, when applied in conjunction with a health issue, can disclose personal health status (e.g., "*my friend lost his battle to leukemia*"). Third, dependencies, such as "*obj_have_diagnosis*", are strong positive indicators (e.g., "*I have seasonal allergy*").

This table also provides several notable results about other behaviors when people disclose personal health status. For instance, people often include *@username* in health mentions. They use links to provide additional information such as pictures, locations or texts, or use exclamation mark to express strong feelings about personal health status.

The hypertension classifier was notable because it had specific health-related terminology ranked highly. Specifically, the term *blood* is highly informative for this classifier. We suspect this is because hypertension is commonly referred as high blood pressure.

2.6.2.1 Homogeneous and Heterogeneous Classification

In this experiment, we compared the effectiveness of homogeneous and heterogeneous classifiers then testing on tweets from each of the five health issues. Table 2.3 provides the AUPRCs for each homogeneous (along the diagonal) and heterogeneous

	Cancer	**Depression**	**Hypertension**	**Leukemia**	**SYND**
Cancer	0.732 ± 0.058	0.528 ± 0.018	0.552 ± 0.014	0.869 ± 0.009	0.728 ± 0.009
Depression	0.441 ± 0.007	0.663 ± 0.054	0.611 ± 0.014	0.821 ± 0.006	0.666 ± 0.006
Hypertension	0.451 ± 0.009	0.646 ± 0.011	0.664 ± 0.062	0.726 ± 0.008	0.616 ± 0.006
Leukemia	0.638 ± 0.011	0.603 ± 0.011	0.559 ± 0.019	0.936 ± 0.019	0.579 ± 0.007
SYND	0.625 ± 0.022	0.618 ± 0.026	0.626 ± 0.019	0.831 ± 0.023	0.820 ± 0.018

Table 2.3: AUPRC for homogeneous and heterogeneous classifiers. Classifiers were trained with row health issue tweets and tested on column health issue tweets. All the results in each column are statistically significant with Wilcoxon signed-rank test, $p < 0.003$.

(off diagonal cells) health mention classifier. Each row corresponds to the health issue relied upon for training the classifier, while each column corresponds to the health issue the classifier was applied to.

First, it should be noted that each homogeneous classifier outperforms the heterogeneous classifiers when testing the corresponding health issue tweets, but such classifiers do not generalize. It can be seen that the leukemia HOC-1 classifier achieved the highest AUPRC. This may be due to the balance in the positive and negative classes for this health issue. However, it was observed that the homogeneous classifiers exhibited much higher variance compared to the heterogeneous classifiers. The suggests that heterogeneous classifiers may yield stable results.

Second, the HEC-1 classifier may tend to obtain a better AUPRC when testing on health issues with a similar author-to-others disclosure rate. For instance, cancer achieved the best AUPRC when testing on leukemia tweets. Meanwhile, leukemia achieved the best AUPRC when testing on cancer tweets. Depression and hypertension also achieved the best AUPRC when testing on each other. We suspect that tweets with cancer keywords have much more noises than leukemia, thus making it more difficult to detect health status mentions on cancer. This can be verified by that the AUPRC of cancer classifier is 0.732 when classifying cancer tweets, while the AUPRC of leukemia classifier is 0.936 when classifying leukemia tweets. Because leukemia is ”blood cancer”, a cancer classifier can result in AUPRC of 0.869 when classifying leukemia tweets (which is better than 0.732, but still lower than 0.936, which makes sense).

Third, it also shows that the SYND heterogeneous classifier (HEC-N) was the second best heterogeneous classifier when testing on cancer, depression and leukemia tweets, and the best heterogeneous classifier when testing on hypertension. Considering that the HEC-1 classifier is specialized to a certain health issue, the HEC-N classifier may provide a more scalable alternative when filtering for personal health mentions on other health issues.

2.6.2.2 Comparison of Homogeneous Classifiers

Classifier	**Cancer**	**Depression**	**Hypertension**	**Leukemia**
HOC-1	0.732 ± 0.058	0.663 ± 0.054	0.664 ± 0.063	0.936 ± 0.019
HOC-N	0.723 ± 0.061	0.645 ± 0.053	0.672 ± 0.070	$0.927 \pm 0.022^{*}$
HOC-N‡	$0.756 \pm 0.050^{**}$	$0.681 \pm 0.050^{**}$	$0.702 \pm 0.059^{**}$	$0.940 \pm 0.021^{**}$

Table 2.4: AUPRC of homogeneous health mention classifiers, given the same number of training tweets. $^{*}P < 0.001$, $^{**}P < 0.05$, comparing to the test results for HOC-1.

In this experiment, we evaluated how homogeneous classifiers are influenced by (1) the number of health issues in the training set and (2) the number of tweets used for training classifiers. Table 2.4 shows the results for the HOC-1 and HOC-N classifiers when testing on the tweets of each health issue. In each column, homogeneous classifier HOC-1 and HOC-N were trained with the same number of training tweets. The number of training tweets for HOC-N‡ classifier equaled to the number of all the tweets training for each HOC-1 classifier.

The Wilcoxon signed-rank tests showed that HOC-1 and HOC-N classifiers are statistically significant only when testing on leukemia tweets ($P < 0.001$). This suggests that HOC-N classifiers are expected to have similar performance with HOC-1 classifiers.

The Wilcoxon signed-rank tests for the classification results in each column show that HOC-1 and HOC-N‡ are statistically significant ($P < 0.05$). This indicates that if the total number of training tweets is fixed, HOC-N classifier outperforms the combination of HOC-1 classifiers.

These results suggest show that the HOC-N classifier can serve as a substitute for HOC-1 classifiers.

2.6.2.3 Comparison between Heterogeneous Classifiers

In this experiment, we evaluated how heterogeneous classifiers are influenced by the number of health issues in the training set. Figure 2.7 shows the results of HEC-1 and HEC-N ($N \in \{2,3,4\}$) when testing on the other 30 health issues. For HEC-1, it should

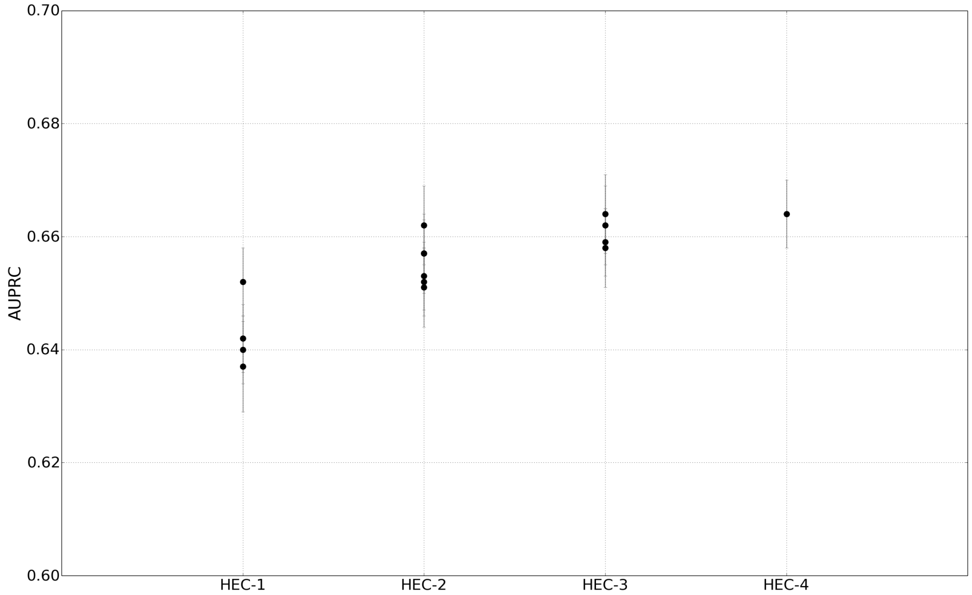


Figure 2.7: Comparison between HEC-1 and HEC-N trained on cancer, depression, hypertension and leukemia, and tested on the remaining 30 health issues. The tweets of each test health issue stratified with respect to their rate of observation.

be noted that the best AUPRC was achieved by the cancer HEC-1. This may be due to the fact that cancer can be invoked to communicate a wide variety of concepts beyond an individual's health status, such as the Zodiac, the name of a physical building or a metaphor. The results also indicate that HEC-N tends to outperform HEC-1.

This suggests hypothesis H4 may be true, provided the classifier is based on an appropriate mixture of health issues. However, determining an optimized group of health issues to achieve an HEC-N classifier with performance comparable to the HEC-1 classifier is left to future investigation. Based on these findings, we use HOC-N and HEC-N to conduct the system scalability tests.

2.6.3 System Scalability

After breaking ties, 43.7% of the *TieBreak* dataset are positive instances. As such, there are approximately 120,260 positive instances out of 281,357 tweets in the health issue bins (or 0.046% of all the collected tweets). The distribution of positive and negative tweets in

Tweets	Gold	CAN	CAP	TieBreak
Positives	1082 (41.3%)	1082 (33.2%)	1718 (52.7%)	1366 (41.9%)
Negatives	1539 (58.7%)	2175 (66.8%)	1539 (47.3%)	1891 (58.1%)

Table 2.5: Class distribution of tweets in the datasets.

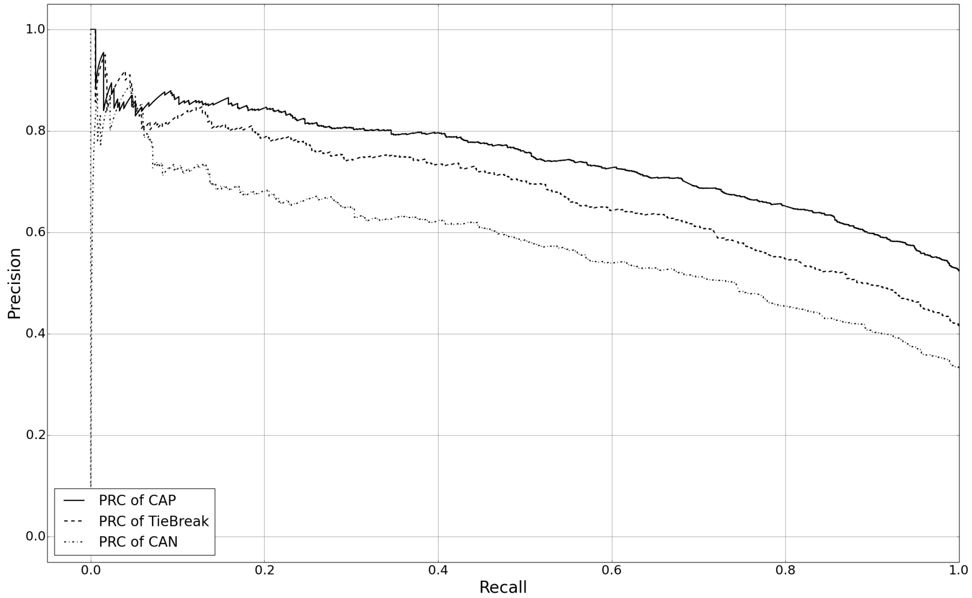


Figure 2.8: PR curves for testing on the gold, CAN, and CAP datasets.

each dataset is reported in Table 2.5.

We trained the SYND classifier with the gold standard datasets for cancer, depression, hypertension and leukemia, and tested it on the other three types of datasets. Figure 2.8 depicts the PR curves for each dataset and shows the average and standard deviation of AUPRC. The upper line corresponds to testing on the CAP dataset (AUPRC = 0.753, st.Dev. = 0.005), the middle line corresponds to testing on the TieBreak dataset (AUPRC = 0.685, st.Dev. = 0.005) and the lower line corresponds to testing on the CAP dataset (AUPRC = 0.594, st.Dev. = 0.007).For example, when fixing the recall to 0.4, it was observed that the CAP, TieBreak, and CAN scenarios yield a precision of 0.8, 0.77, and 0.62, respectively. These results demonstrate the scalability of the system classifiers to obtain a high precision with a reasonable recall when testing many other health issues in the Twitter environment.

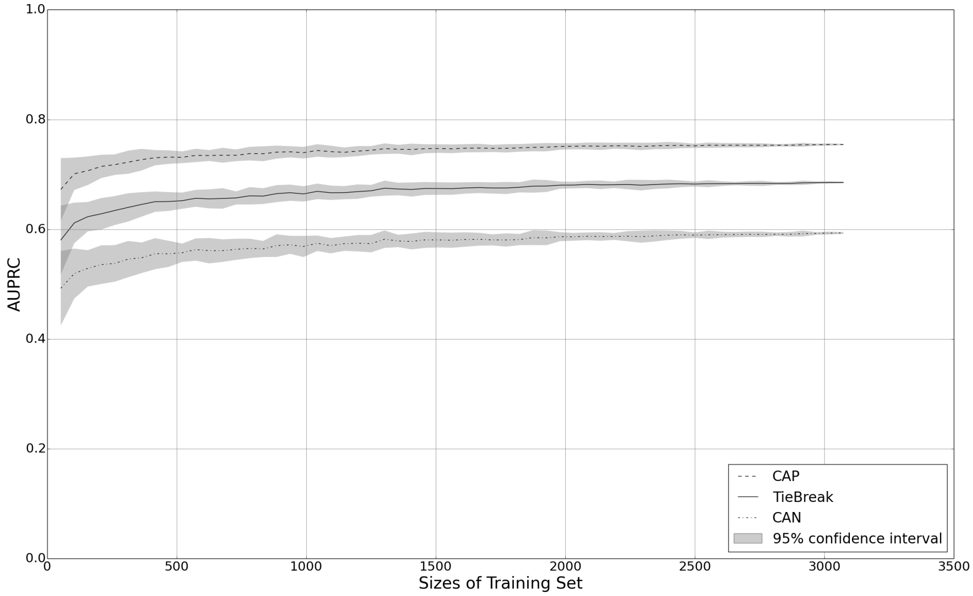


Figure 2.9: Performance of the SYND classifier with a varying amount of training data.

Figure 2.9 shows how the size of the training set influences the AUPRC of the classifiers. For each training set, the mean AUPRC and a 95% confidence interval is illustrated in the gray area. For each dataset, the results suggest that AUPRC achieves stability when the training set consists of approximately 2000 tweets.

2.7 Health Mention Classification

Manual discovery and annotation of tweets that disclose personal health status is a timely, as well as costly, process. Thus, as alluded to in the previous section, we engineered a classification strategy, based on the labels provided by the AMT masters, to automatically detect tweets communicating the mentions of health status and augment the dataset for investigation[5].

[5] It should be recognized that we aimed to build a classifier that is sufficient for detecting a large number of tweets with health mentions, so that we may investigate the extent to which language categories influence disclosure. It is impossible to engineer a perfectly accurate classifier by incorporating many other features (e.g., part of speech, grammar features and word2vec) and, thus, we acknowledge that there is a certain degree of error in the labels of the tweets we investigate.

2.7.1 Building Classification Models

We observed that the tweets can be naturally clustered by health issue keywords (e.g., all tweets associated with asthma), and each cluster may have different properties (e.g., one issue may be associated with a larger proportion of tweets that disclose health information than others). To incorporate these inter-cluster differences into classifiers, one natural candidate solution is to build a hierarchical model, where the parameters are governed by hyperparameters for each health issue. However, this would result in a very expensive computational model, due to the high-dimensionality of the features (e.g., word unigrams) involved in text classification.

Rather, in this research, we introduce an approach based on language categories to reflect the differences between health issues. We construct features at both 1) the tweet-level and 2) the health-level. The tweet-level features consisted of 2000 character n-grams ($2 \leq n \leq 5$) from all of the labeled tweets, according to the ranking of their TF-IDF (Term Frequency - Inverse Document Frequency) values. To obtain health-level features we extracted language categories using LIWC from all of the unlabeled tweets. LIWC has been invoked by many social data analysis studies with some successes [17, 9, 33, 61]. Basically, LIWC counts the number of words (e.g., from all the tweets related to a health issue) that match each of the language categories[6] and converts them as the percentages of total words. In total, we use 64 language categories as for health-level features. Note that all the tweets related to the same health issue are defined over the same language categories.

The baseline model trains classifiers with character n-gram features at the tweet-level. Our proposed model is an augmentation that includes the language categories as features at the health-level.

[6] http://www.kovcomp.co.uk/wordstat/LIWC.html

2.7.2 Predicting Health Mentions

We considered three common learning models[7] for each model: i) a logistic regression, ii) a linear support vector machine (SVM), and iii) a random forest. All of the parameters were set to their defaults. The 3400 labeled tweets were applied as a gold standard. We applied 10-fold shuffled and stratified cross-validation and reported the mean and standard deviation of the area under the receiver operating characteristic curve (AUC) for each classifier in Table 2.6. A t-test was applied to assess if there is a statistically significant difference (at the 0.05 significance level) between the classifiers in their capability.

Model	**Baseline**	**Proposed**
Linear Regression	0.825 ± 0.007	0.837 ± 0.007
Linear SVM	0.811 ± 0.010	0.839 ± 0.007
Random Forest	0.823 ± 0.012	0.833 ± 0.012

Table 2.6: A comparison of the AUC for the baseline model and proposed model.

The t-test confirmed ($p < 0.05$) that introducing language categories as features (at health-level) can improve the performance of logistic regression and linear SVM. The results further indicate that the linear SVM with language categories as features (at health-level) significantly outperforms the classifiers that are devoid of such features ($p < 0.05$).

To understand the importance of these features, we used the random forest classifier in the proposed model to select the 20 most informative features for health mention detection, as shown in Table 2.7. It should be noted that 13 out of the 20, corresponding to features at health-level, were obtained via the application of LIWC. Considering that only 64 out of the 2064 features are health-level features ($P(group\ feature) = 0.031$), a sign test implied there was a strong significant difference between these two types of features (where 13 successes out of 20 trials with $p < 0.001$). Table 2.7 also shows that biological processes, and the health language categories in particular, are critical for health mention detection. We further recognized notable language features pertained to time (notably

[7] As implemented in the Scikit-Learn package (version 0.15.0). http://scikit-learn.org/

Rank	Feature	Rank	Feature
1	***health***	11	**funct**
2	nee	12	***negate***
3	***pronoun***	13	***conj***
4	***auxverb***	14	***we***
5	my	15	i′
6	***i***	16	***verb***
7	n	17	***bio***
8	has	18	***present***
9	***time***	19	i′m
10	ne	20	***humans***

Table 2.7: The 20 most informative features selected by the random forest classifier. The group features (i.e., the aggregated language categories at the level of a health issue) are depicted in bold italicized font.

the present time) and those associated with humans. Interestingly, pronouns are also important in both types of features. We suspect this stems from the fact that many tweets disclose health status about the authors' family members and friends. The following tweet is a clear example of this observation:

> *"Just found out that* ***my grandmother*** *has cancer. Thyroid cancer to be exact."*

Finally, we applied the logistic regression classifier and obtained 54,247 health mention tweets (with an expected precision of 81.7%) to conduct the NMF analysis.

2.8 Discovery of Similar Health Issues

We aim to investigate if semantically similar health issues associate with the MvY disclosure rate. In this section, we show how the health issues were grouped according to their semantics.

2.8.1 Grouping Health Issues with NMF

We applied NMF to the set of tweets to learn similar health issues. We applied NMF, as opposed to another matrix factorization strategy like singular value decomposition, because it has been shown to have better interpretability when the original matrix values are all positive [66]. However, applying the document-term model (as is traditional in matrix factorization) for the short texts encountered on Twitter will suffer from data sparsity. Many strategies have been proposed to overcome this problem, ranging from aggregation of documents [67] or words (e.g., the document by bi-terms model [68]) to a document-by-word embedding model [69]). In this research, we propose a health issue-by-language category model.

To build this model, we apply LIWC to extract language categories from the tweets with mentions for each health issue. This results in a matrix of 34 health issues by 64 language categories, which is subject to NMF[8]. We set the rank (i.e., the number of basis components in NMF) to 4 because this exhibited the best cophenetic correlation coefficient and dispersion coefficient. Note that this decision was based on the correlation of consensus matrix obtained from 100 NMF runs.

2.8.2 Health Issue Groups and Semantics

Figure 2.10 depicts the heatmap of the four basis components (denoted by B_1, B_2, B_3 and B_4, respectively). Figure 2.11 illustrates the heatmap of the mixture coefficients for each basis component. The health issues are grouped by assigning them to their most associated basis component. The associated semantics are thus explained via the corresponding coefficients in the basis component. Note that certain health issues are affiliated with more than one group. Examples of such issues include *Alzheimer's Disease*, *Down Syndrome* and *Parkinson's Disease*, all of which can be characterized by the first (B_1) and the third (B_3) basis components.

[8] https://cran.r-project.org/web/packages/NMF/

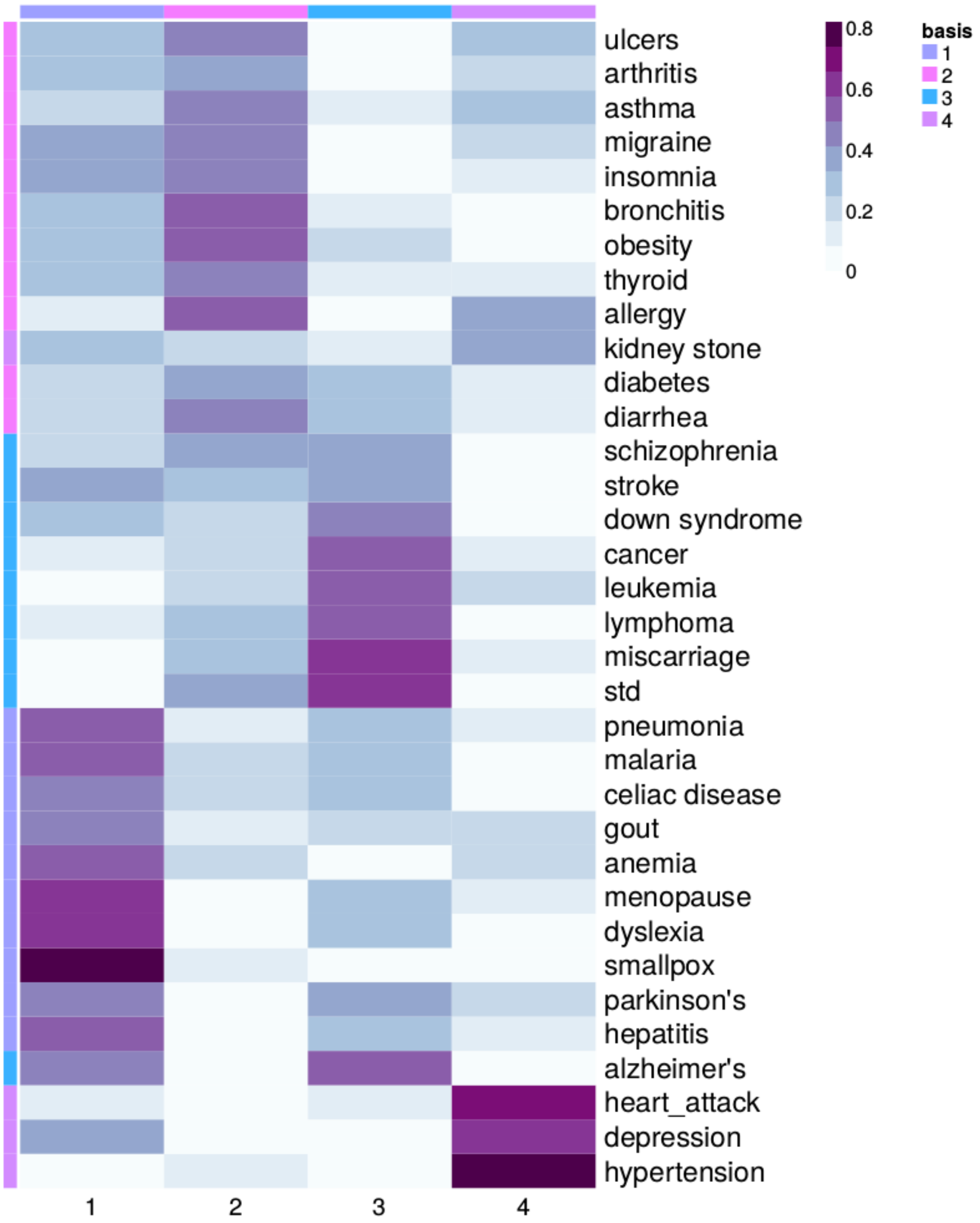


Figure 2.10: A heatmap of the basis components derived from NMF. Each cell is the probability that a health issue (along the row) belongs to the basis component (along the column).

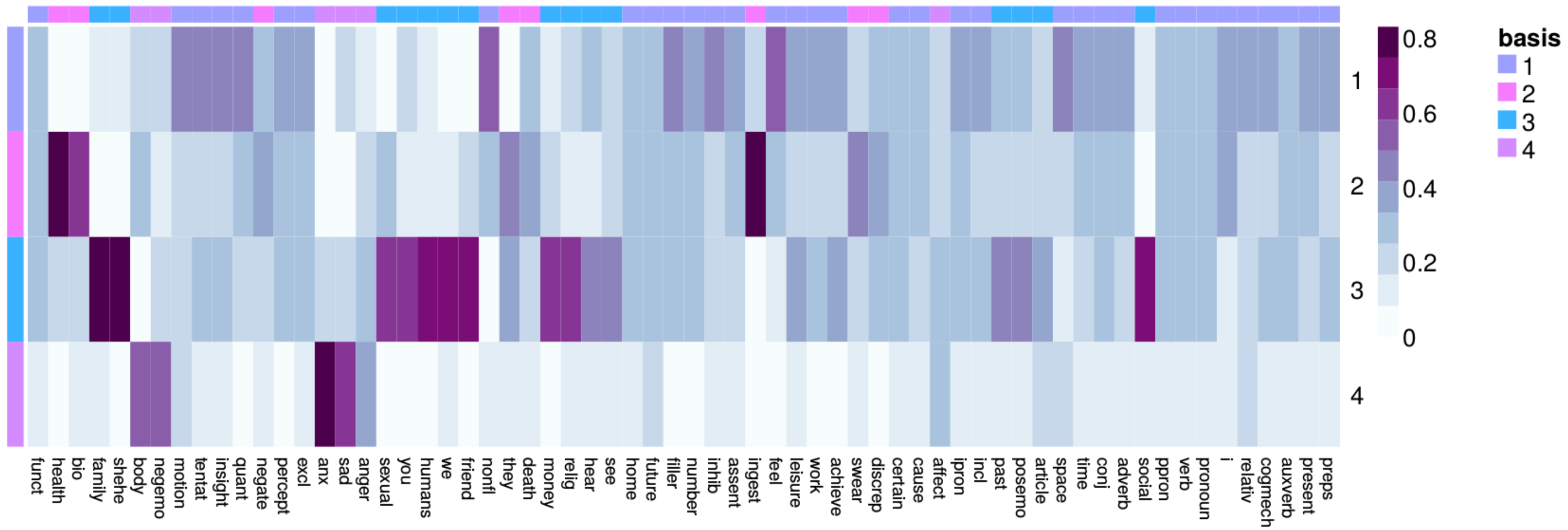


Figure 2.11: A heatmap of the mixture coefficients derived from NMF. Each cell is the probability that the language category (along by the column) is associated with the basis component (along the row).

Group I (corresponding to B_1). This basis component, in comparison to the others (see Figure 2.11), exhibits a set of similar probabilities for a broad range of language categories. These include cognitive processes (e.g., think, know, guess, and stop), quantifiers (e.g., much and lot), non-fluencies and perceptual process, and feeling. This basis component covers common semantics shared by a wide range health issues, such as *dyslexia*, *gout*, *hepatitis*, *malaria*, *menopause*, *Parkinson's Disease*, *pneumonia*, and *smallpox*. The following tweets are clear examples:

*"Not this year. She is **feeling** better from the pneumonia but still weak"*

*"I get picked on a **lot** for my dyslexia so I act like I **read** something faster than I really did."*

*"I **know** menopause only happens once, but my mother's an exception to that."*

*"PLease help me to **move** back to Oklahoma to get on a clinical trial for my hepatitis c."*

Group II (corresponding to B_2). This basis component exhibits a strong semantic of biological processes, especially health (e.g., clinic and pill) and ingestion (e.g., eat and taste). Note there are also less strongly semantic terms, such as swear words (e.g., damn) and the third person plural. Further note that swear words may be used to express negative experiences. The group of health issues corresponding to this basis component include more common disorders: *allergy*, *anemia*, *arthritis*, *asthma*, *bronchitis*, *celiac disease*, *diabetes*, *diarrhea*, *insomnia*, *migraine*, *obesity*, *thyroid* and *ulcers*. The following tweets serve as examples of this group:

*"Seriously in need of some allergy **medicine**."*

*"Well went to the **doc**. Gave me some shit for stomach Ulcers. Hopefully this works and I can **eat** in the next couple of days :)"*

*"I'm just glad my migraine went away, but I'm still **sick** to my stomach"*

"I just need one of my friends to have insomnia like me so ***they*** *can stay up and text me all night long"*

Group III (corresponding to B$_3$). This basis component has strong semantics associated with social processes (e.g., friends, family and humans - girl and woman), third person singular, first person plural, second person plural, money, religion (e.g., church and pray), and sexuality (e.g., love and incest). There are also less strong semantics associated with past tense and positive emotions. The group of health issues most associated with this basis component are more severe, and often debilitating, including: *Alzheimer's Disease*, *cancer*, *Down syndrome*, *miscarriage*, *leukemia*, *lymphoma*, *schizophrenia*, *sexually transmitted disease*, and *stroke*. The following tweets serve as examples of this group:

*"****#Ineedtoraisemoney*** *to help my* ***husband*** *who recently had a stroke need to raise $5000.00 any help out there!???"*

"He has leukemia. His parents go to my ***church****."*

"use ***love*** *as means to solve it. My dad has awful violent schizophrenia he has tried to kill me and mom."*

"Dad is officially ***cancer free****! They* ***caught*** *it in time and no chemo treatments are needed!!!* ***:)****"*

Group IV (corresponding to B$_4$). This component is mainly about affective processes, such as negative emotions, anxiety, anger and sadness. The group of health issues most associated with this basis component are associated with chronic and/or painful problems, including: *depression*, *hypertension*, *heart attack* and *kidney stone*. Note that the semantic of *body* in this basis component may be due to the last two health issues. The following tweets serve as examples of this group:

"I'm so ***afraid*** *that my depression is coming back. :("*

"I was so ***nervous*** *! I almost died of a heart attack"*

"I ***hate*** *medicine that's y I don't take it but right now my depression hittin hard af"*

"reminds me of ***grief*** *being diagnosed as depression and meds being prescribed"*

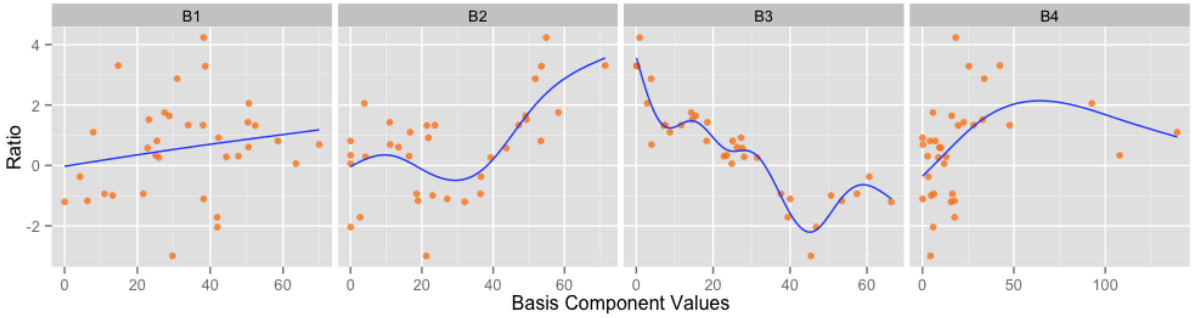


Figure 2.12: Correlation between the NMF basis components and the MvY ratio. The blue lines were smoothed via a thin plate regression spline. Note the positive effect of B_2 and the negative effect of B_3 with respect to the MvY ratio.

2.9 Linking to MvY Disclosure

In this section, we investigate how the learned health issue groups associate with the rate at which information is disclosed about the author or other individuals. We first regress MvY ratio on the predictors extracted from the four NMF basis components, in order to examine how these basis components contribute to MvY disclosure when considered independently and when combined. Then, by connecting to the associated health issues in each basis component, we explore how the semantics (as factors) drive MvY disclosure.

2.9.1 Factors Driving MvY Disclosure

We use the four NMF basis components to predict MvY. Specifically, we adopt generalized additive models (GAMs) with a thin plate regression spline smoother [70]. In the process, we apply a log function to the response to account for the positive skewness of

MvY. We apply an ANOVA to perform a Chi-square test on the deviance and compare the different GAMs at the 0.05 significance level. **Single Predictor Models.** To examine the effect of each individual basis component, we begin by predicting MvY with a single predictor. This corresponds to models M_1, M_2, M_3 and M_4. (for each of the corresponding enumerated components).

Figure 2.12 shows the relationship between the basis components and the MvY ratio. It can be seen that there is a direct correlation between the chance a health issue associates with Group II and the MvY ratio. By contrast, there is a negative correlation between a health issue associating with Group III and the MvY ratio. Neither Group I nor Group IV exhibits a strong association with the MvY ratio (and neither has a significant coefficient). The Chi-square tests indicate the best single predictor model is M_3, followed by M_2.

Model	**Predictor**	**EDF**	**Ref.df**	**F**	**R-sq. (adj)**	**Dev.**	**GCV**
M_2	s(B_2)	3.96 ***	4.83	5.85	0.45	51.5%	1.65
M_3	s(B_3)	7.91 ***	8.67	36.75	0.91	92.8%	0.33
M_{2+3}	s(B_2, B_3)	16.84 ***	21.40	17.73	0.92	96.1%	0.43
M_{1+2+4}	s(B_1)	3.65 ***	4.54	15.02	0.81	84.0%	0.62
	s(B_2)	1.00 ***	1.00	102.41			
	s(B_4)	1.21 ***	1.38	37.96			

Table 2.8: Smoothed terms for predicting the MvY ratio under different models. Note the linear effect of s(B_2) in M_{1+2+4}. *** $p < 0.001$.

Model	**Resid. Df**	**Resid. Dev**	**Df**	**Deviance**	**Pr(>Chi)**
M_2	29.042	40.958	-	-	-
M_{1+2+4}	27.143	13.509	1.899	27.449	$***$
M_3	25.088	6.083	2.055	7.426	$***$
M_{2+3}	16.162	3.330	8.926	2.753	0.143

Table 2.9: Comparison between models with Chi-square tests on deviance with ANOVA function. *** $p < 0.001$.

Multiple Predictors Models. Based on the results of M_2 and M_3, we investigated two additional GAMs: i) M_{2+3} by smoothing the marginal smooths of B_2 and B_3, and ii) M_{1+2+4} by applying a linear combination of the smoothed B_1, B_2 and B_4.

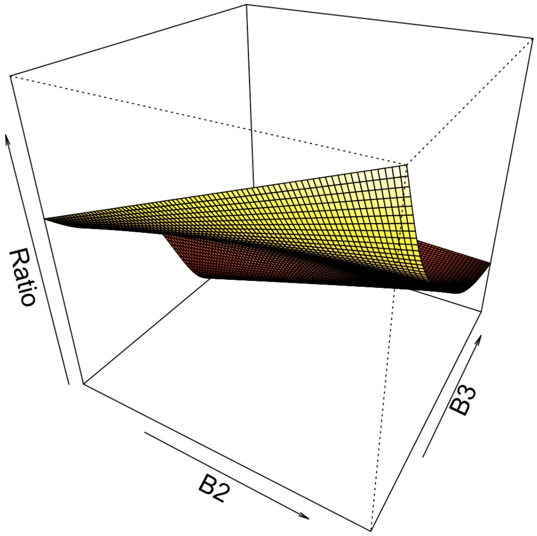


Figure 2.13: An illustration of the combined effect of B_2 and B_3 on predicting MvY. A larger MvY is positively correlated with Group II and negatively correlated with Group III.

Table 2.8 summarizes the MvY predictive capability for the models. There are statistically significant effects for each of the predictors in each model. The effects of the predictors in M_{2+3} and M_{1+2+4} are shown in Figures 2.13 and 2.14, respectively. Notably, Figure 2.13 shows that a health issue tends to exhibit a higher MvY (i.e., self-disclosure rate) when it positively correlates with Group II and negatively correlates with Group III. Figure 2.14 shows that, when combined together, B_1, B_2 and B_4 enhance the prediction of a higher MvY.

Table 2.9 summarizes the results of the comparison on these models with a Chi-square test under an ANOVA function. Although M_{2+3} has a larger adjusted R^2 and deviance, there is not a statistically significant difference[9] with respect to M_3. It was, however, observed that M_3 outperforms M_{1+2+4} in a statistically significant manner. Furthermore, it was found that M_{1+2+4} outperforms M_2 in a significant manner as well. This suggests that Group III more strongly associates with an author disclosing another individuals'

[9] Interestingly, the best model we obtained is M_{1+2+3}; however, there is no statistically significant effect on B_1 in M_{1+2+3}.

health status, whereas the combination of Group I, II and IV associate with self-disclosure.

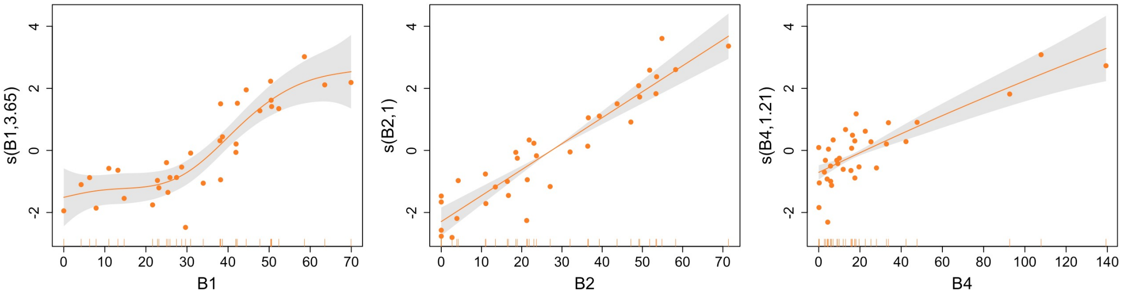


Figure 2.14: Effects of B_1, B_2 and B_4 when combined to predict the MvY ratio. Each of these basis components positively influence a higher MvY prediction.

2.9.2 Semantics Behind MvY Disclosure

To confirm these results, we ran a Spearman rank correlation test between the language categories and the MvY ratio. The results of the test, with correlation coefficient greater than or equal to 0.5, are shown in Table 2.10. The results suggest that, as expected, there is a strong correlation between the use of first person singular and self-disclosure of health issues. In addition, tweets communicating self-disclosure tend to apply adverbs, time, quantifiers and present tense. For instance, it was observed that the top four health issues with the largest proportion of tweets using the words of “morning”, “afternoon”, “tonight”, “tomorrow”, “now” and “soon” (more than 8%) were *insomnia*, *migraine*, *kidney stone* and *asthma*. Additionally, the top four health issues with the largest proportion of tweets containing words of “pill” and “med(icine)” were *malaria*, *thyroid*, *anemia* and *hypertension* (more than 8%).

Next, we turn our attention to tweets in which the author discloses the health status of another person. As expected, we find a strong correlation with the 3rd person singular. We also find that authors tend to disclose information about family members. Moreover, it appears that the religion category indicates social support. For instance, we observed that *Alzheimer’s Disease*, *leukemia*, *Parkinson’s Diseases* and *cancer* are the top four health issues with the largest proportion of tweets containing the words “mother

Category	CC.	Statistic
1st person singular	0.82 ***	1170.45
Adverbs	0.69 ***	2051.63
Time	0.58 ***	2768.42
Quantifiers	0.56 ***	2847.81
Present tense	0.55 ***	2977.65
Body	0.52 **	3169.48
Relativity	0.51 **	3196.49
2nd person	−0.50 **	9833.52
Religion	−0.52 **	9934.07
1st person plural	−0.59 ***	10412.72
Humans	−0.67 ***	10934.02
Family	−0.79 ***	11746.19
Social processes	−0.92 ***	12561.19
3rd person singular	−0.93 ***	12642.93

Table 2.10: Language categories with a Spearman correlation ≥ 0.50 for the MvY ratio. Note that "CC" represents the correlation coefficient. $^{**}p < 0.01$, $^{***}p < 0.001$.

(mom)" and "father (dad)" (more than 20%). Also, it should be noted that more than 15% of *Alzheimer's Disease* tweets contain words relating to "grandmother (grandmom)" and "grandfather (grandpa)". Note that *cancer*, *lymphoma*, *leukemia* and *pneumonia* are the top four health issues with the largest proportion of tweets containing words of "bless", "pray" and "support" (more than 15%).

Looking back at the NMF results in Figures 2.10 and 2.11, it can be seen that, for tweets disclosing another person's health status, the related health issues[10] and language categories are consistent with Group III. For tweets with a self-disclosed health mention, the related health issues[11] and language categories distribute in Group I, II and IV. This is consistent with M_{1+2+4}, which indicates that all three basis components, in combination, have a positive effect on predicting a higher MvY ratio.

[10] Note that pneumonia has a weak signal in Group III.

[11] Note that half of the health issues belong to Group II: *asthma*, *insomnia*, *migraine* and *thyroid*.

2.10 Discussion

2.10.1 Building Scalable Classifiers

There are several notable findings from this investigation. First, it appears that Twitter users disclose the health status of themselves and others. Second, the health status disclosure rate may depend on the health issue. Third, how people disclose their own and other people's health status may also be health issue dependent. Fourth, tweets related to a small group of health issues can train a scalable classifier to detect health mentions on Twitter streams. Another interesting phenomenon illustrated in the PR curves (Figure 2.8) is that the system classifier, trained with the tweets for which AMT masters exhibited high concordance in their labels, is more likely than AMT masters to classify tweets with conflict labels as positive. One possible explanation is that the classifier makes its decision based on thousands of examples, while most AMT masters made decisions only with the description of the annotation schema, which indicates that the classifier may be more effective with the labeling task. This suggests there may be a difference between using a expert and crowdsourcing to generate the labeled corpus. However, determining how best to leverage the crowd to mimic an expert is beyond the scope of this investigation.

2.10.2 Effective Language Categories

It is important to recognize that the language categories extracted by LIWC are essential to our framework in several ways. First, the language categories play a more significant role for health mention detection than traditional features based on character *n*-grams. Second, the language categories enable an avoidance of data sparsity when applying NMF. Third, the groups of health issues, driven by language categories and their semantics (as expressed by the associated language categories), act as factors for learning the motivation behind MvY disclosure.

2.10.3 Groups of Health issues

By applying NMF on the health issue-by-language category model, our investigation suggests there are (at least) four groups of health issues. It is interesting to note that, although B_3 (Group III) is associated with the high cost of medicine and social support, there is a relatively strong signal for the semantic of a positive emotion. This may be due to the fact that some tweets celebrate reversals in diagnosis (e.g., a family member is now cancer free) or an expression of support, such as the following tweet:

> *"Love 2 my wife, my hero, whose done w radiation treatment today ..."*

Basis component B_4 (Group IV) is also notable because it mainly focuses on negative emotions. These emotions appear to cut across various health issues, including depression, heart attack, hypertension, and kidney stones. Such health issues appear to align with literature on these topics, particularly [17, 9] where users with mental health problems (on Twitter and Reddit) have been shown to express negative emotions.

2.10.4 MvY Disclosure

Our findings show that basis component B_3 has a stronger impact on predicting the MvY ratio than the combination of the other three components (i.e., B_1, B_2 and B_4). At the same time, B_3 has a negative effect, while the other three groups tend to have positive effects. This suggests that the health issues that occur for family members, are associated with the high cost of medicine, and require social support, tend to have a lower MvY ratio. By contrast, for other health issues, where the semantics are associated with biological processes (e.g., health and ingestion) and negative emotions, the authors tend to disclose their own health status.

2.10.5 Impact on Health Related Research

According to our investigation, roughly 44% of the tweets containing health issue keywords disclose personal health status. We believe there is a potential for information to assist healthcare professionals in learning about their patients or their patients' family medical history, information often missing in the EMRs. This indicates that social media platforms, such as Twitter, contain huge amounts of personal health care related information that may complement traditional EMRs in research and practice. We recognize that the veracity of such data must still be verified, but an opportunity exists nonetheless.

2.10.6 Limitation and Future Work

There are several limitations of this investigation that we wish to highlight regarding the data annotation and building scalable classification.

First, two parameters to extract tweets from Twitter streams require configuration: 1) the set of keywords invoked in the filter and 2) the geolocation applied to discover tweets. Compared to keywords, geolocation can filter tweets disseminated by authoritative organizations (due to the absence of "coordinates" and "place" information in these tweets), such as the American Cancer Society, and thus greatly reduce noise, but at a cost of excluding tweets without geolocation.

A second limitation exists in the annotation provided to the AMT masters for labeling the corpus. Specifically, the *N/A* option was assumed to be the negative class in this research, but this assumption could, in certain instances, be incorrect. Third, this investigation was restricted to only 34 health-related phenomena, which is clearly only a sample of all possible health issues. The health issue keywords based filter service can be enhanced by applying a layman health vocabulary [71]. Given that this study shows there is 1) high variability in the rate at which people tweet about a certain health issue and 2) to whom the statement of health issue corresponds, it will be critical to investigate how these methods fare in the context of other health issues.

Finally, we presented a framework, relying on language categories, to demonstrate that groups of health issues and their semantics are associated with the rate of disclosure for one's self vs. another individual on Twitter. While it is not necessarily the case that disclosure of another individual's health information has transpired without their consent, it is likely that many such disclosures have not been approved. As such, we believe this investigation shows there are opportunities to develop support programs for individuals to utilize (e.g., via private discussion or counseling) before unveiling the health status of their relative or friends. At the same time, we believe that our ability to automatically detect such revelations, suggesting that interventions can be invoked after an initial disclosure to mitigate further revelations.

There are also several limitations regarding MvY investigation, which we believe can serve as the basis for further investigation in the topic of health information disclosure in social media. The first limitation is in the fidelity of the health mention prediction model. We tuned the precision of the logistic regression classifier to 81.7%, thus mixing tweets without health mentions into the NMF analysis. As this research evolves, it will be useful to build more robust health mention classifiers, which may be possible by incorporating more variety in the training space (e.g., via additional health issues) and features (e.g. extracted from social connections). The second limitation is in the factorization and resulting groups. Specifically, it should be noted that the number of basis components is determined by optimization for several coefficients associated with NMF decomposition. Regularization, constrained by external factors (e.g., the severity and social stigma of health issues), are worth considering to derive a more interpretable matrix factorization. In particular, it would be worthwhile to investigate if there exist certain clinical factors that drive the formation of groups of health issues and MvY disclosure.

CHAPTER 3

LEARNING HORMONAL THERAPY ADHERENCE IN AN ONLINE BREAST CANCER FORUM

The previous chapter demonstrated that data from GOSMPs can be leveraged to learn personal health status disclosure behaviors. However, these online environments were not designed for people to discuss their health conditions, which makes it difficult for patients to have targeted discussion about long-term treatments. Thus, in this chapter, we present an investigation with data from an OHC that was established over a decade ago. Specially, we illustrate that the posts of breast cancer patients who underwent hormonal therapy can be used to learn about their treatment adherence.

3.1 Background

Social media platforms have received substantial attention from individuals who are seeking, or looking to share information about their treatment experiences. There are many OHCs that have been established, many of which have been in existence for over a decade, such as depressionforums.org and breastcancer.org. These environments are notable because it allows individuals to speak candidly and at length with others who have been diagnosed with similar problems. At same time, these online communities open up novel opportunities to learn about long-term adherence to treatment.

Ensuring adherence to long-term treatment protocols is crucial to improving survival rates for many health issues (e.g., cancer [72], diabetes [73], and sexually-transmitted diseases like HIV [74]). For example, hormonal therapy is a proven treatment known to boost the survival rate of breast cancer patients [75], but the benefits are maximized when patients are on the protocol for at least five years [76]. However, adherence to long-term treatment is challenging for many patients. For example, it is reported that only around half of patients who start a hormonal therapy regimen actually complete the five-year treatment [77]. There are numerous obvious reasons for lack of adherence, such as the cost of medication, adverse side effects, and the recurrence of disease during therapy. Still, other reasons are more subtle. For example, [78] found that 14% of studied patients stopped treatment “for no particular reason”. A deeper understanding of the factors associated with adherence is necessary if society aims to improve current adherence rates.

There have been various investigations into regimen adherence. However, a large portion of these investigations have relied on primary data collected through formal survey methodology [79, 80, 81]. Surveys are notable because they standardize and control the questions asked, but they are time-consuming and are often restricted to a limited number of participants. At the same time, investigations have focused on secondary data derived from electronic medical records (EMRs) and other traditional clinical resources [82, 83, 84]. EMRs enable observational studies on large cohorts with a wide range of factors documented in the clinical setting [85], but they often lack the voice of the patients themselves. Patient-reported outcomes, if collected, are usually structured [86] and unlikely to reveal nuances such as patients’ emotions, feelings and experiences.

Thus, in this research, we aim to demonstrate that treatment adherence can be studied in OHCs. Specifically, we focus on learning hormonal therapy adherence (HTA) from patients’ self-reported information on the breastcancer.org online discussion board. Specifically, for the purposes of this investigation, we label HTA behaviors as three types of events: 1) *taking* - where a prescribed medication is consumed according to an oncologist’s recommendation, 2) *interruption* - where the patient stops (or pauses) a regimen, or switches to a different medication (with or without clinician advice), and 3) *completion* - where a patient achieves the endpoint of a five-year treatment protocol. Given these types of events, we address three research questions:

R1: To what extent are breast cancer patients’ emotions correlated with different treatment decisions?

R2: Are personality types associated with treatment adherence?

R3: Can we predict future interruptions based on the information posted by patients in OHCs?

To investigate these three research questions, we begin by extracting statements related to adherence events via a combination of rule-based filtering and statistically-informed classification models. Next, we apply a one-way ANOVA test on the emotion scores of sentences that mention adherence events. Then, we study the association between personality traits and two HTA groups (with completion events vs with interruption events) using logistic regression (LR) analysis. Finally, we build a LR model to examine the extent to which posts predict interruption events in future. The main contributions of this research are summarized as follows:

- **Emotions**. We find that patients in OHCs tend to exhibit fear with taking events, anger with interruption events, and joy (with a tinge of sadness and disgust) with completion events.

- **Personalities**. We demonstrate that personality types, extracted from patient self-reported online information, confirm the majority findings in more traditional treatment adherence studies. At the same time, we show there is a discrepancy with a particular personality type, suggesting further opportunities for investigation.

- **Predictability**. Based on features derived from discussion posts, a classification model can obtain an area under receiver operating characteristic (ROC) curve (AUC) of ~ 0.8. The most informative features suggest that patients who are at the beginning of therapy and mention side effects (e.g., depression) are more likely to experience an interruption than those further into treatment.

3.2 Related Work

Understanding the factors associated with HTA can be useful for identifying strategies to improve medication adherence. However, as noted earlier, most of these studies either rely on formal survey methods or the secondary use of EMRs (or other similar resources), which have limitations. Our goal is to investigate the potential value of using social media data to learn some of these factors.

3.2.1 Factors Associated With HTA

In this section we review some of the factors that others have found to be significantly associated with adherence during hormone therapy.

Like many treatment regimens, side effects are known to be important factors leading to low HTA [78]. At the same time, there are many other factors that associate with low HTA. For instance, by examining medical cost and low-income patients undergoing hormonal therapy, it was [87] showed that high healthcare costs are associated with suboptimal adherence. It was [88] observed that patients with non-adherence experiences for chronic diseases are less likely to adhere to hormonal therapy. It is also suggested that patients with stage IV cancer, as opposed to earlier stages, are more likely to have lower HTA [84, 89]. In a study of Swedish breast cancer patients, a study [82] found that larger tumor sizes tend to have higher HTA. They also found that younger patients are more likely to continue treatments. By contrast, it was [89] found that patients in the 40 to 60 age group are at higher risk of discontinuing treatment than older patients.

Various associations between negative habits and HTA have also been discovered. For instance, it was [84] showed that patients who drink alcohol tend to have low HTA. There are some studies that have investigated the association between emotions and HTA. Generally, these studies have found that negative emotions are related with low HTA [90, 91]. These studies are limited in that they focus solely on interruption types of behaviors.

Our observation is that significant factors will vary based on cohort characteristics, and it is difficult to generalize across cohorts. Social media data allow us to observe a potentially more diverse set of patients than may be included in studies where the cohort is carefully selected.

3.2.2 HTA Study on Social Media

Social media has increasingly been relied upon to conduct health-related studies. These studies have a broad range, investigating flu trends [92, 93], mental health [84, 89], extracting medical related languages [94], how to build online communities to provide local cancer support [95], and privacy issues associated with health mentions [32, 96]. Pertinent to our investigation, there is a growing body of work that focuses on breast cancer treatment and social media and we refer the reader to the excellent review [97]. We highlight that [98] illustrated the breast cancer symptoms reported on MedHelp.org exhibit consistency with symptoms reported in the clinical setting. It has been demonstrated that breast cancer patients' have reductions in anxiety when attending patient-support groups via Twitter [99] . Similarly, it has been found that breast cancer patients tend to report more positive emotions as they engage in online discussion [100].

While Internet-based interventions have been applied to improve patients' adherence with mental health [101] or anti-retroviral medications [102], there are few studies that focus on the factors related to HTA through self-reported information on social media. For example,an investigation [103] 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. It was [104] found that joint pain is the main reason patients stop taking aromatase inhibitors (AIs, a subclass of hormonal therapy, in online discussions of drug side effects and HTA discontinuation. In this research, we characterize HTA with three types of events: taking, interruption and completion. These are notable because taking events may provide insight into a patients' current state (i.e., when they are in the midst of treatment), while interruption and completion events allow for characterizing the difference between low and high adherence patients. To the best of our knowledge, this is the first work to investigate the association between personality traits and HTA through patients' self-reported information in an OHC.

3.3 Data Preparation

Breastcancer.org is a non-profit organization that disseminates information about breast cancer to healthcare consumers. Additionally, it operates an online discussion board where breast cancer patients can seek, share, and respond to information about their experiences. The discussion board is organized into 80 forums, with more than 135,000 annotated topics. In this chapter, we focus on the *Hormonal Therapy - Before, During and After*[1] forum. We collected all topics and posts within this forum published before June 22, 2016. In total, there are 9,996 patients who participated in 5,995 threads with more than 130,000 published posts.

3.3.1 Data Annotation

For the purposes of this study, we investigate discussions related to seven hormonal therapy drugs[2]: *Arimidex* (chemical name: *anastrozole*), *Aromasin* (chemical name: *exemestane*), *Femara* (chemical name: *letrozole*), *tamoxifen*, *Evista* (chemical name: *raloxifene*), *Fareston* (chemical name: *toremifene*) and *Faslodex* (chemical name: *fulvestrant*). Since the same drug may be referred to in a variety of ways, we standardized the data by replacing the aliases of each medication (e.g., brand name) with their corresponding chemical names.

There were 913,493 sentences voiced in the forum. We found 123,633 sentences (13.5%) contained at least one of the chemical names of interest. These sentences were communicated in 66,617 posts, published by 8,563 patients. We selected 1,000 sentences, at random, for annotation by human reviewers. The annotators were asked to assign each sentence to one of seven options: 1) *Action: Taking medication*, 2) *Action: Stopped taking medication*, 3) *Action: Switched medications*, 4)*Plan: Taking medication in future*, 5) *Plan: Not taking medication in future*, 6) *Plan: Not yet decided* and 7) *None of the Above*. These options were based on our observation of how patients discuss treatments in this forum and guid-

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

[2] http://www.breastcancer.org/treatment/hormonal/for_you

	Relevant						Non-relevant
Option	action:taking	action:stop	action:switch	plan:take	plan:not taking	plan: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). The relevant versus non-relevant classes are approximately 3:2 in size.

ance in a decision making codebook introduced by Beryl and colleagues [81]. While not every option is utilized in the following analysis, we believe that this activity enabled the human reviewers to gain a better understand of the labeling task. For the purposes of our investigation (which focuses on two-class prediction), we labeled all of the first six options as *relevant* sentences and the final *None of the Above* option as *non-relevant* sentences.

We employed a majority rule annotation strategy with three reviewers. The first two reviewers annotated every sentence, while the third annotator was employed to break ties[3]. The primary two annotators exhibited a *very good* agreement level (Cohen's $\kappa = 0.82$) at relevant vs. non-relevant level; *good* agreement (Cohen's $\kappa = 0.72$) at the level of the seven options. After the third annotator broke ties, we obtained 604 relevant sentences and 396 non-relevant sentences. The distribution of different options after annotation is shown in Table 3.1.

3.4 Adherence Event Classification

3.4.1 Adherence Event Extraction

To document adherence events with high precision, we adopted a hierarchical methodology similar to that invoked by others [105, 106], which works as follows. First, we built an LR model to distinguish relevant from non-relevant sentences. Second, we applied a rule-based method to search relevant sentences for each adherence event.[4]

3.4.1.1 Relevant Sentence Classification

To distinguish between relevant and non-relevant sentences, we translated each sentence into a low-dimensional representation. This representation serves as the features for a LR model (as described below). We used the mean of the low-dimensional representation vectors of words, namely, word2vec [107], in a sentence to represent the feature set

[3] All three annotators spent at least one month in this forum and were familiar with this topic.

[4] Note that adherence events may not align with labeled actions.

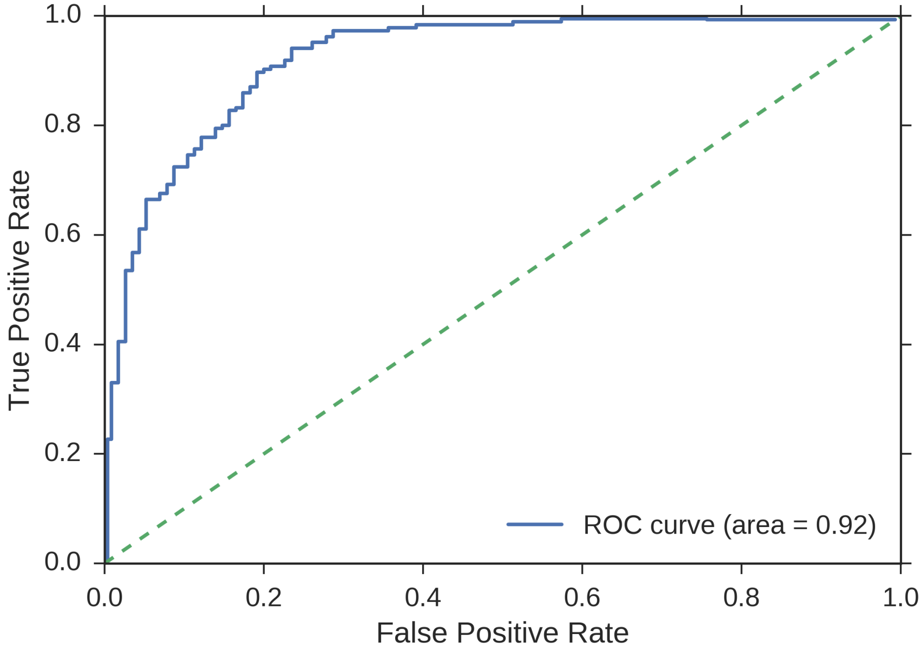


Figure 3.1: The ROC curve of the logistic regression model (with the mean of word2vec as features) that classifies sentences as relevant or non-relevant.

for the classification model. We restricted our word2vec representation to words with a frequency of at least 5 instances in the hormonal therapy forum. We set the dimensionality of the word vectors to 100. We use the skip-gram model with negative sampling implemented in gensim (version 0.13.0) [108] to fit the word2vec model.

We used the LR model implemented in sklearn (version 0.18.0) [109] and applied a stratified shuffle/split method to create five cross-validation iterations. In each iteration, 80% of the instances were used to fit LR model and the remaining 20% were used for testing purposes. All parameters of the LR model were set to their default values in the software package. The LR model achieved an area under the ROC curve (AUC) of 0.932 ± 0.010. For illustration purposes, Figure 3.1 depicts the ROC of the fitted LR for one arbitrary split of the train-test data (i.e., 20% test data).

By adjusting the class weights, we tuned the LR model to achieve a precision of 0.882 ± 0.023 and recall of 0.882 ± 0.022. We then refit the model with all of the 1000 labeled

	Pattern	κ.	Prec.
Completion	5 years, finished, ended, completed, done, ...	0.86	0.83
Interruption	back on, vacation, switch, took a break, took me off, gave up, stopped taking, ...	0.82	0.85
Taking	started, been on, stay on, ...	0.76	0.89

Table 3.2: A sample of the patterns applied for extracting adherence events. We account for variations in the spelling of a discovered word by applying word2vec (e.g., *yrs* for *years*, *vacations* and *vaca* for *vacation*).

sentences before applying it to extract the relevant sentences from the entire forum. Upon doing so, we obtained 80,510 relevant sentences that were distributed across 51,826 posts and authored by 8,023 patients.

3.4.1.2 Rule-based Event Extraction

To extract additional sentences for each adherence event, we empirically created patterns that were based upon annotating experience. For example, when patients mentioned that stopping a medication, the possible patterns could be 1) *Took me off*, 2) *Stop taking*, and 3) *Being off* a medication. Similarly, when patients mentioned that they were taking a vacation[5], the possible patterns could be 1) *Vacation*, 2) *Holiday*, and 3) *Took a break* from a medication. When patients mentioned taking a medication, the possible patterns could be 1) *Started* and 2) *Been on* a medication. We refer the reader to Table 3.2 for additional examples of the patterns applied in our model.

To ensure high precision, we iteratively labeled events as follows: First, we extracted and labeled the completion events and removed these from further consideration. Second, from the remaining set, we extracted and labeled interruption events, which again were removed from further consideration. The remaining sentences are used to extract

[5] It should be noted that vacation events for certain medications were not captured by any label in the initial annotation task. However, upon re-examination, we determined that this group of sentences was not labeled as non-relevant. This is notable because it means that we can still extract such instances from the set of relevant sentences through a deterministic rule-based method.

taking events. We followed the same process to extract patient groups with different adherence events. To assess the performance of this methodology, we directed two of the annotators to assess 100 randomly selected sentences from each classified event category. The agreements, in terms of the Cohen's kappa, between these two annotators and the precision for each type of events are summarized in Table 3.2. Finally, we obtained 1,172 posts published by 513 patients for completion events, 8,681 posts published by 2,525 patients for interruption events, and 15,116 posts published by 4,826 patients for taking events.

3.5 Emotion Analysis (R1)

To investigate if there exist significant differences in emotions between adherence event types when patients mentioned them, we randomly selected 500 sentences from each of the three adherence event categories. We chose sentences instead of entire posts because, in this forum, sentences are sufficiently verbose to convey information of interest (see examples below). By contrast, posts are too long to obtain precise emotion scores. These sentences were fed into the IBM Watson Tone Analyzer Service[6] to obtain emotion scores for each sentence, which together with IBM Watson Personality Insights service (see below) have been recently adopted for many emotion and personality related studies (e.g., [110, 111, 112]).

The service returns scores with a range of 0 (the weakest) to 1 (the strongest) for five different emotion categories: *anger*, *disgust*, *fear*, *joy* and *sadness*. After obtaining emotion scores, we apply a one-way ANOVA test, with a significance level of 0.05, for each category. In this hypothesis test, the null hypothesis is that there is no significant difference in emotion when different adherence events are mentioned. Figure 3.2 depicts barplots of the emotion scores for each adherence $\langle event\ type, emotion \rangle$ pair. Table 3.3 reports on the one-way ANOVA test results for each of the five tests. Each of the p-values are smaller

[6] https://www.ibm.com/watson/developercloud/tone-analyzer.html

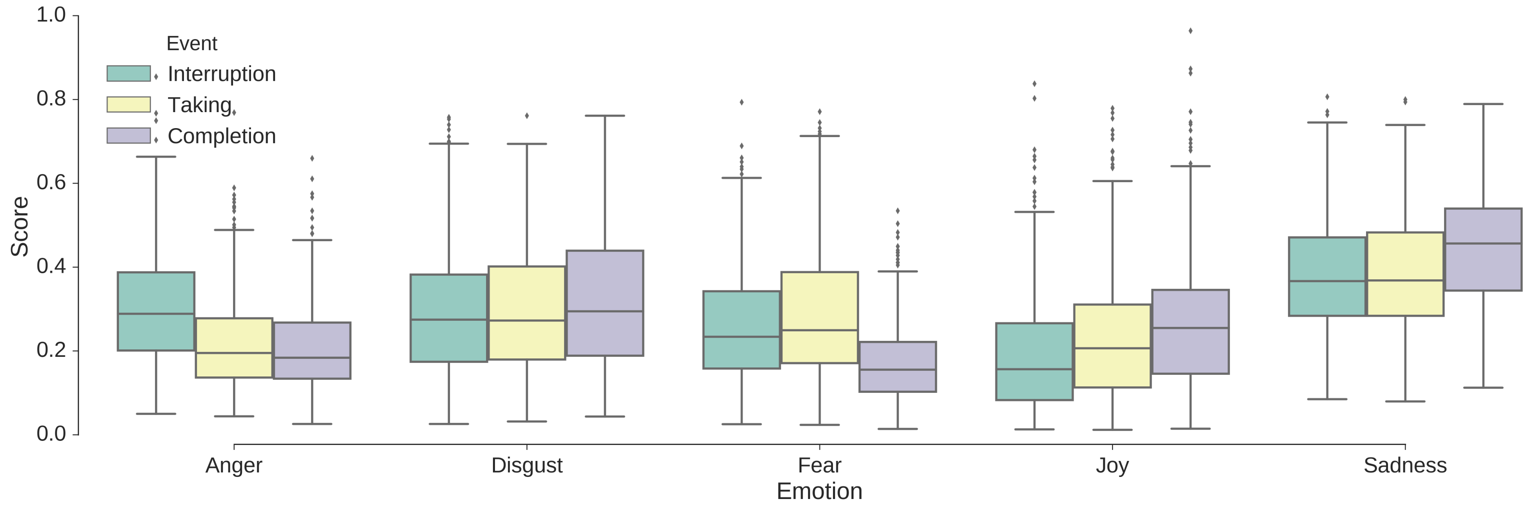


Figure 3.2: A boxplot of the emotion scores for adherence ⟨*event type*,*emotion*⟩ pairs. Mentions with interruption events tend to exhibit greater levels of anger. By contrast, mentions with completion events tend to exhibit greater levels of disgust, joy and sadness and lower levels of fear.

	Anger	**Disgust**	**Fear**	**Joy**	**Sadness**
F	100.449	6.866	107.977	25.327	40.592
p	< 0.001	0.001	< 0.001	< 0.001	< 0.001

Table 3.3: The results of the one-way ANOVA test on five emotions for three types of adherence events.

than the predefined significance level of 0.05. This implies that there exists a significant difference between the emotions across the adherence event type.

From Figure 3.2, we found that patients tend to exhibit a relatively higher degree of anger when mentioning interruption events. This may be due to multiple reasons, such as frustration with the side effects of medications. A clear example of this phenomenon is in the following patient post:

> *"letrozole was hell and i'll be starting exemestane tomorrow."*

We also note that mentions with completion events tend to have a slightly higher level of disgust in comparison to the other two events. This may arise because, after five years of treatment, some patients may refuse to continue further treatment after re-balancing their quality of life and cancer recurrence. As one patient noted:

> *"I finished 5 years of Tamoxifen and declined the Letrozole because my chance of recurrence was very very low and i wanted to feel more alive than the Tamoxifen allowed."*

Yet it appears that completing a five-year treatment makes patients relatively less fearful and more joyful. This is not unexpected because, in spite of various side effects, approximately half of the women on hormonal therapy medications achieve this goal. For example:

> *"I am happy to be done with the Anastrozole – but I am so glad I made the whole 5 years!"*

At the same time, completing a five-year treatment does not necessarily imply the end of hormonal therapy. Instead, it may just be the beginning of a second five-year treatment period. Moreover, the cancer may reoccur after the initial five year period. As one patient noted:

> *"On a side note I was on Tamoxifen for five years and still got a recurrence so I'm not married to the idea of taking pills anyway."*

Interruption and taking events did not exhibit a significant difference on disgust or sadness. However, there was a relatively higher joy score in taking event mentions, in comparison to interruption events. This may be because patients who continue taking a medication may experience side effects that are quite different to the degree that patients who stop the medication do. As was voiced in one post:

> *"I have been on Fulvestrant since January of 2014, very little side effects."*

Still, not everyone voices a lower degree side effects when taking hormonal therapy medications. It should be noted that some patients who start taking medication often fear the side effects. As one patient noted:

> *"I just started tamoxifen 3 days ago and i am sitting here in fear of getting fat ..."*

3.6 Association Between Personalities and Adherence (R2)

Next, we investigate if there exists an association between specific personality traits and HTA. Specifically, we focus on which personality traits are associated with the patient group with interruption events and the patient group with completion events. We apply a classical LR analysis to study this problem. To obtain personality scores, we leverage the IBM Watson Personality Insights service[7].

[7] https://www.ibm.com/watson/developercloud/personality-insights.html

Personality	Coef	Std Err	Z	$P > \|z\|$
Excitement seeking	11.973	4.402	2.720	0.007
Self-discipline	-6.105	1.967	-2.104	0.002
Outgoing	5.767	2.495	2.312	0.021
Melancholy	-5.680	2.405	-2.361	0.018
Achievement striving	4.268	1.963	2.174	0.030
Imagination	-2.657	1.209	-2.198	0.028
Log-Likelihood				-377.72
LL-Null				-422.83
LLR p-value				0.0006

Table 3.4: The significant predictors in the logistic regression (LR) model, ranked by their absolute coefficients.

This service inspects documents (e.g., emails, tweets or posts) and returns personality characteristics along three dimensions: the "Big Five" traits [113], values, and needs. In this research, we only apply the "Big Five" traits, the facets of which are shown in Figure 3.3. Note that the scores for categories are reported as percentiles instead of absolute measures. For instance, a 90% on *Extraversion* suggests that the writer is more extroverted than 90% of the people in the population. After applying a threshold of 3,000 published words or greater (reaching the Personality Insights service's maximum precision), there were 1,402 patients who mentioned interruption events and 348 patients who mentioned completion events. We conduct a LR analysis using the 35 personality trait scores for the 1,750 patients.

Figure 3.3 shows the coefficients and their 95% confidence intervals (sans intercept). Note that the coefficients are basically indicative of the log-odds of the probability that a user is in the patient group with completion events. The coefficients with intervals not crossing the $x = 0$ vertical line are the most notable. The positive significant coefficients (shown to the right in a dark green color) suggest the personality traits are more related with the patients with completion events, while the negative significant coefficients (shown to the left in a dark red color) suggest the personality traits are more related with the patients with interruption events. A higher absolute coefficient value indicates a

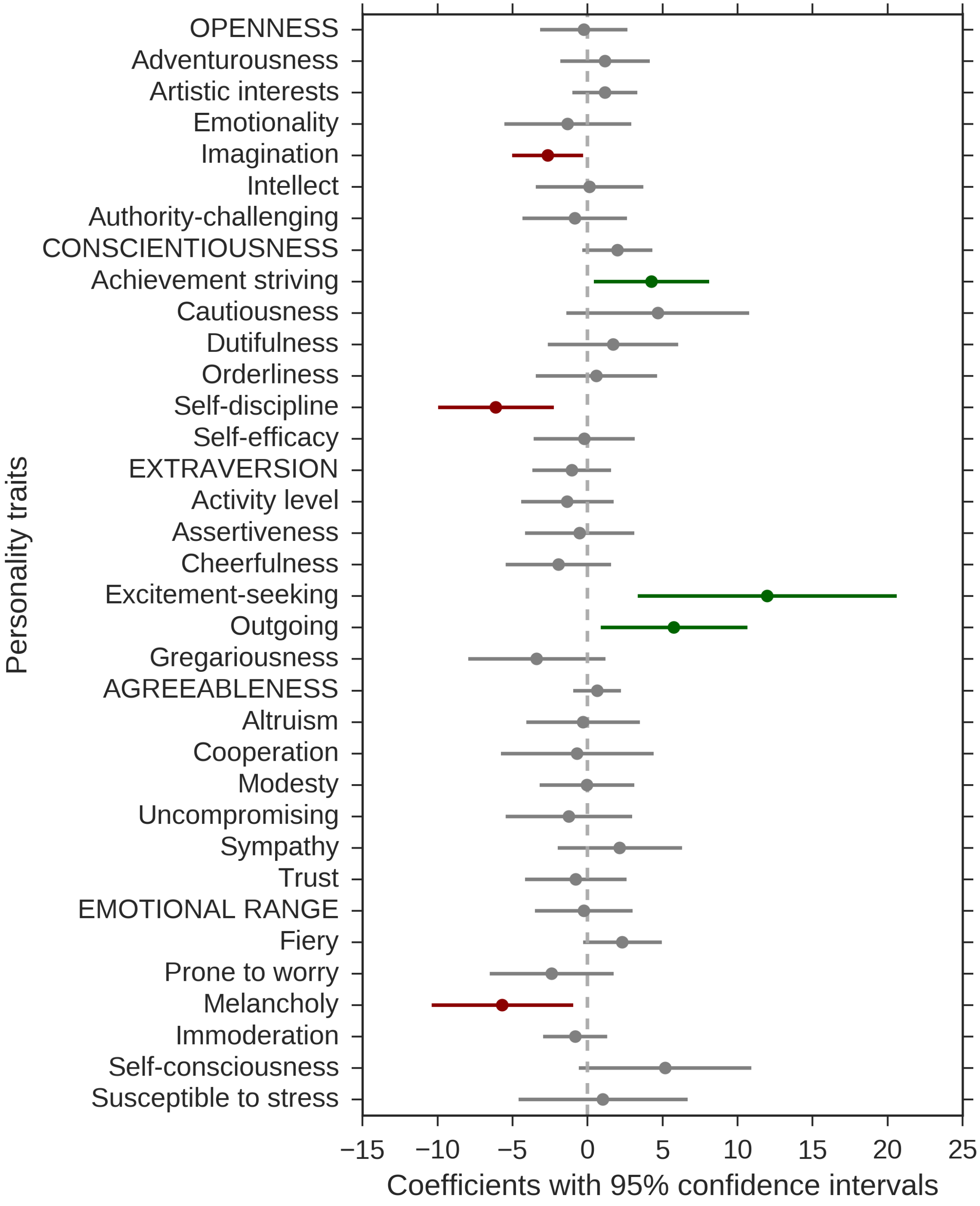


Figure 3.3: The coefficients and their 95% confidence intervals for the 35 personality traits. The coefficients are grouped by their categories: the labels in uppercase are the Big Five and the lowercase are their facets.

stronger association. Table 3.4 shows the statistics for six personality traits that are associated with HTA in a statistically significant manner. The model is significant in comparison against a baseline null model according to a likelihood ratio test (p = 0.0006).

Definitions of the personality traits presented in Table 3.4 are drawn from the service and listed below:

- **Excitement seeking:** *A need for environmental stimulation.*
- **Self-discipline:** *The capacity to begin tasks and follow through to completion in spite of boredom.*
- **Outgoing:** *Interest in and friendliness towards others; socially confident.*
- **Melancholy:** *Normal tendency to experience feelings of guilt, sadness, hopelessness, or loneliness.*
- **Achievement striving**: *The need for personal achievement and sense of direction.*
- **Imagination:** *Openness to creating an inner world of fantasy.*

Among the Big Five, Agreeableness is the only personality that does not have significant facets. The facets of *excitement seeking* (Extroversion), *outgoing* (Extroversion) and *achievement striving* (Conscientiousness) have significant positive association with patients with completion events. In contrast, the facets of *self-discipline* (Conscientiousness), *melancholy* (Emotional range), and *imagination* (openness) have significant positive association with patients with interruption events. Note that Conscientiousness has two facets with effects in opposite directions. We discuss our findings further in the Discussion section.

3.7 Interruption Events Prediction (R3)

In this section, we investigate the extent to which the existence of interruption events in the future can be predicted by earlier published posts. To do so, we focused on two

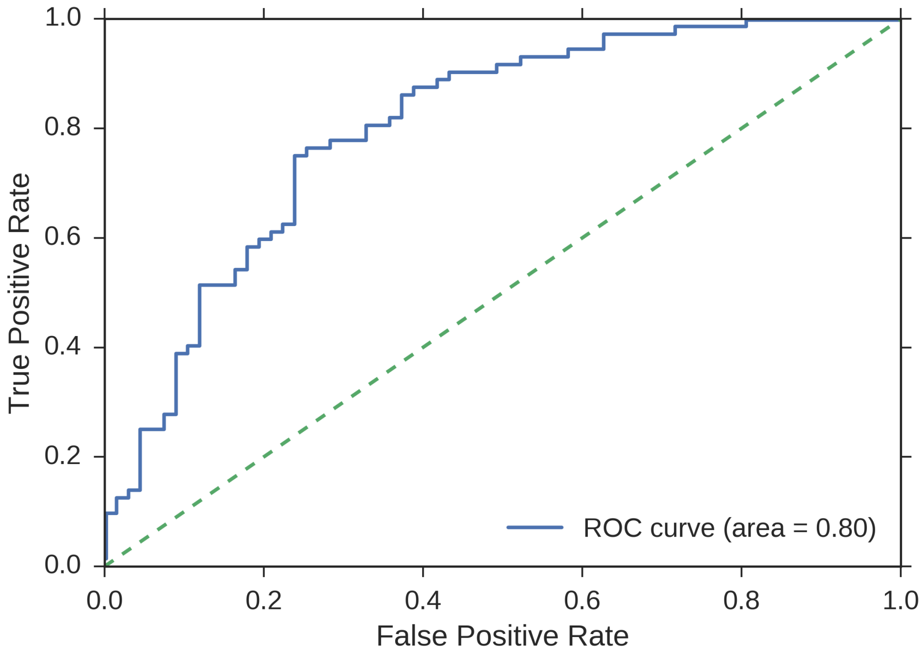


Figure 3.4: The ROC curve of the LR model for predicting Interruption events, with unigrams and bigrams of stemmed words as features.

classes of forum patients. The first corresponds to patients who mentioned interruption events, but never mentioned completion events. The second corresponds to patients who mentioned completion events. For each user in these classes, we extracted all posts (in the hormonal therapy forum) that were published before their first mention of an interruption or completion event. We sampled from the two patient groups to study an equal number of individuals in each class, as per [114].

To perform binary classification, we selected the 2000 unigrams and bigrams of *stemmed* words with the highest term frequency - inverse document frequency (TF-IDF) values as features. Due to the sparsity of the feature space, we apply LR with ridge regularization. We set all of the parameters of the model as default. To evaluate the performance of the model, we apply a stratified shuffe/split strategy to perform five-fold cross-validation. We measure performance using accuracy, precision, recall, F-score, and AUC.

There were 1,347 patients who mentioned interruption events and 347 patients who

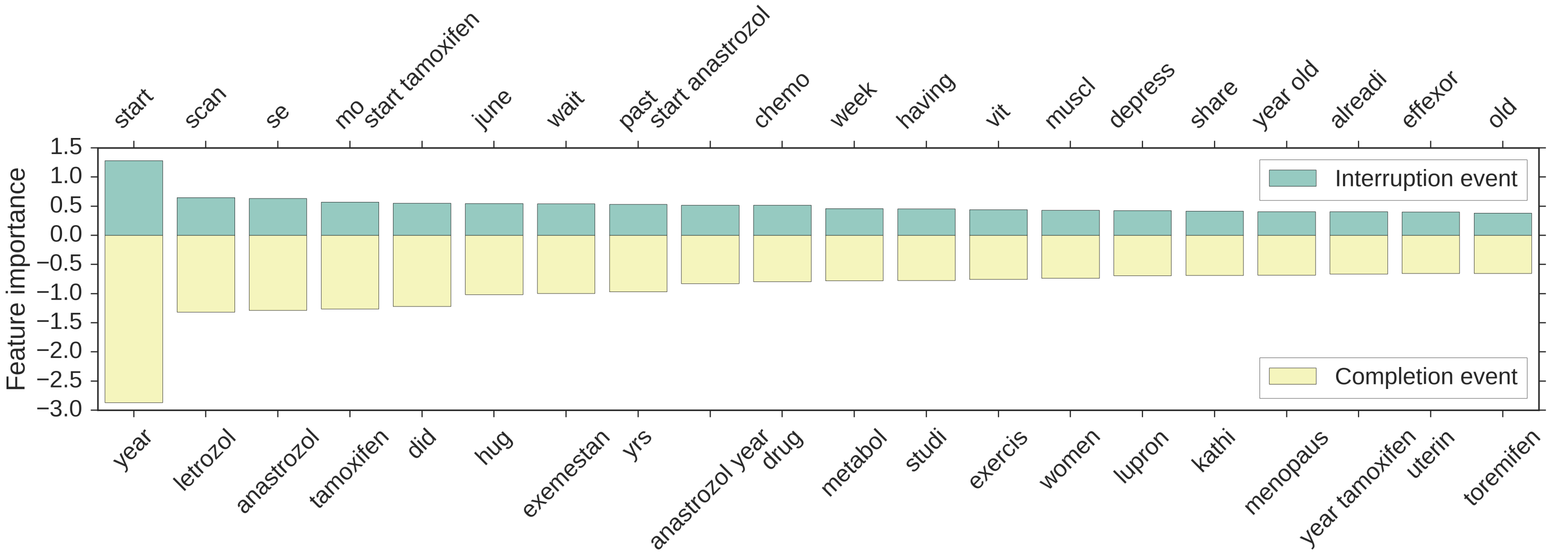


Figure 3.5: The 20 most informative features for each class. Features with positive weights correspond to the patient group with Interruption events, while features with negative weights correspond to the patient group with Completion events. The larger the absolute score for a feature, the greater its importance in the model.

Precision	0.709 ± 0.018	F1	0.723 ± 0.024
Recall	0.739 ± 0.046	Accuracy	0.717 ± 0.018

Table 3.5: Performance of the LR model for predicting Interruption events.

mentioned completion events. To balance the classes, we randomly sampled 347 patients from patients with interruption events (for a total of 694). Table 3.5 shows the performance of the model. The AUC of the model evaluated with five-fold stratified cross-validation is 0.801 ± 0.020. To illustrate the ROC curve, we report the AUC (with mean value) in Figure 3.4.

To obtain insight into the association between features and classification, we report the 20 most informative features for each class in Figure 3.5. These features are selected based on the rank of their mean coefficients in the classifiers that are obtained through the five-fold cross-validation. Features with a positive and negative weight correspond to the patient groups communicating interruption and not communicating interruption events, respectively. The features show that patients beginning a medication appear to be more likely to experience an interruption event in comparison to patients who have taken it for many years. Mentions about side effects, such as depression, are also strong signs that an interruption event will occur in the future.

3.8 Discussion

3.8.1 Insights on Factors Related to HTA

Our research is based on self-reported patient information in an online health forum. Self-reported information has the potential to provide a candid view of patients' daily experiences, thus allowing for more non-clinical insights into understanding HTA. For instance, our emotion analysis shows that patients who mentioned interruption events often exhibit a strong emotion of anger. If care providers could continuously monitor patients' posts (or be provided with interpretation services in the event patients do not wish doctors to listen in on everything they have to say), they may be provided with signs of potential interruption events before they occur (e.g., through rising rates of an anger emotion).

Our interruption event prediction model also suggests that interruption events are more likely to transpire for patients who are at the beginning of a hormonal therapy drug regimen and/or manifesting sides effect (e.g., depression). Care providers may consider paying special attention to this type of patient. However, even for those patients who have completed a five-year protocol, they may have strong sadness when mentioning completion events. Cancer survivorship is complex and it has been shown that completion of a treatment course can be accompanied by symptoms similar to post-traumatic stress disorder (PTSD [115, 116]. It is quite possible that the transition from the known setting of adjuvant treatment to the "unknown" setting of routine surveillance could trigger sadness, which is manifest in the forum writings.

Our findings indicate that long-term consistent support may be needed to correct patients' perspectives and improve their overall treatment experience. Given that the common practice in breast oncology clinics is to see patients twice per year while they are receiving long-term hormonal therapy, ancillary triggers for impending HTA problems, such as prompting patients for reflective comments through consumer health informatics interfaces, should be pursued.

3.8.2 Personality Traits and HTA

Our findings in personality analysis show that patients who completed treatment successfully are significantly more likely to display the traits of Extraversion (*excitement seeking* and *outgoing*) compared to patients who started but did not complete treatment. This is aligned with traditional offline adherence studies where Extraversion has been found to be positively associated with adherence to treatment with antidepressants [117]. It has been [118] showed association between Extraversion and exercise adherence. Interestingly, our interruption event prediction model also shows that *exercise* is one of top informative predictors for completion events prediction (see Figure 3.5).

Facets of Openness to Experience (*Imagination*) and Neuroticism (*Melancholy*) are found to be negatively associated with patients with completion events in our model. This is also aligned with related literature in which both of them were found to be associated with low adherence [119, 120, 121, 122]. While Conscientiousness has been found to be positively correlated with medication adherence in patients with chronic disease [119], we found two of its facets (*achievement striving*, *self-discipline*) have opposite effects on HTA. The finding of a negative effect of self-discipline on HTA is contrary to studies on adherence to diet and exercise [123]. We do not know the reason behind this discovery at this time, but believe it provides an opportunity for future investigation.

3.8.3 Impact on Social Health Studies

There are three main messages from our investigation. First, we demonstrate that it is feasible to study the adherence behavior of breast cancer patients undergoing hormonal therapy by analyzing their posts in an online community. Second, we show that emotions and personalities, which are scarce in traditional medical resources, but can be automatically extracted from self-reported information, may provide further insights into the nature of HTA. Finally, our interruption event predicting model built upon patients' history posts suggests the possibility for proactive interventions to improve overall adherence.

3.8.4 Limitation and Future Work

There are several limitations in our work that we wish to highlight. Although breastcancer.org is large, it may not contain representative samples from the full spectrum of breast cancer patients. Methodologically, we rely upon a rule-based approach to obtain sentences related to different adherence events. While this method promises a high precision, it neglects related sentences that do not quite follow the predefined rules. As such, there are clearly opportunities for enhancing the recall of this model. Another limitation of this work is that we only examine how word (in the form of unigrams and bigrams) features can be leveraged to predict interruption events. More semantically-meaningful features could be potentially applied to improve the model. For instances, such features could be derived from a patient's posting statistics, self-reported diagnosis and treatment history, and language categories. In this investigation we only considered emotions when patients mention different adherence events. This leads to an incomplete picture of the population and it is necessary to investigate how emotion changes before interruption events actually occur. Similarly, the extent to which the previous posts are predictive to interruption events is also deserving of further study. It will be interesting to investigate why self-discipline is negatively associated with HTA.

CHAPTER 4

LEARNING PATIENTS' MESSAGING BEHAVIOR AND ITS ASSOCIATION WITH HORMONAL THERAPY TREATMENT ADHERENCE

So far, this dissertation has shown that posts in publicly-accessible online environments can be relied upon to infer personal communications about one's health. In this chapter, we refine our investigation to the online environment associated with the clinical setting. Specifically, we consider the communications between breast cancer patients prescribed hormonal therapy and their healthcare providers over an online patient portal. We examine patient messaging behaviors and their association with hormonal therapy treatment adherence (HTA). This investigation enhances the evidence that user generated content in an online environment can be relied upon to effectively learn patients' health behaviors.

4.1 Background

User generated content (UGC) in online environments is increasingly being relied upon to supplement traditional electronic medical records (EMRs) in research about health and medicine. While there are numerous criteria that can be invoked to categorize such data sources, the two primary are: *GOSMPs* (e.g., Twitter and Facebook), and *OHCs* (e.g., breastcancer.org and depressionforums.org). GOSMPs are typically affiliated with a larger population that discusses an extensive range of health-related topics, but with relatively loose connections between the contributors. By contrast, OHCs are generally utilized by patients, relatives and friends proving social support, and healthcare professionals. These individuals typically engage in discussions about long-term treatment experiences, seeking and sharing information. As a result, the OHC tends to support much closer relationships.

Both types of environments have received attention from the healthcare research community. For instance, studies based on GOSMPs usually begin with building efficient classifiers to detect health status mentions [32]. This is because GOSMPs often contain many topics beyond health per se. After identifying health status mentions, further analysis can be accomplished into a variety of issues, such as public health surveillance [124] or personal privacy (e.g., determining when an author discloses health information about someone else [96]). Since OHCs provide the opportunity for individuals to discuss anything they want, investigations based on OHCs can often obtain insights into patients' long-term treatments, health-related behaviors, and their social connections in manners that are not typically encountered in traditional health research settings. These studies include, but are not limited to, discovering shifts in suicidal ideation [114], studying medication adherence [125], and investigating the impact of social support and influence between patients [126, 127]. Moreover, there is evidence that combining data from clinical and social media environments, new types of investigations can be established, such as early detection of adverse drug events [128].

At the same time, another type of UGC that has gained in popularity corresponds to the messages communicated in patient portals. For instance, *MyHealthAtVanderbilt (MHAV)* is such a portal where patients can check lab results, order prescriptions, and communicate with their healthcare providers through direct messaging. In comparison to the two aforementioned data sources where people establish discussions with all members of a community, *MHAV* only allows patients to communicate with their healthcare teams so that confidentiality can be maintained. While they are yet another form of UGC generated by patients (or their caregivers, e.g., spouse), messages in patient portals can be linked to a patient's EMR, thus making it feasible to directly investigate associations with a patient's health outcomes or behaviors.

The objective of our study is to investigate patients' messaging behaviors and their association with medication adherence. Specifically, we focus on a cohort of breast cancer patients prescribed hormonal therapy in Vanderbilt University Medical Center (VUMC).

Hormonal therapy is an adjuvant treatment for patients with hormone-receptor-positive breast cancer used to prevent cancer recurrence and death. This phenomena is significant and worthy of investigation because it comprises 75% of all breast cancer cases [129]. In this long-term treatment, breast cancer patients are recommended to take a particular medication for at least five years to achieve maximal benefit [76]. However, there are many potential factors that may trigger a patient to stop using a hormonal therapy medication, such as a high cost or an undesirable side effect that is induced by the treatment. While UGC in OHCs has been relied upon to gain some intuition into this problem [125], to the best of our knowledge, this is the first investigation to use UGC communicated directly to care providers in the clinical setting.

This investigation specifically addresses three research questions:

- "R1: How do breast cancer patients prescribed hormonal therapy medications use a patient portal messaging service to communicate with healthcare providers? "

- "R2: What are the topics mentioned in these messages during the journey of hormonal therapy?"

- "R3: To what extent are these messaging behaviors associated with hormonal therapy medication adherence? "

To investigate these research questions, we first explore the patterns of messaging volume because it has been shown that communication volume is associated with negative events (e.g., readmission [130]). Next, we examine the messaging rate along hormonal therapy timeline. Building on findings from these investigations, we then inspect the topics by looking into messaging content through an unsupervised learning approach. Finally, we apply a survival analysis to investigate whether the messaging behaviors, in terms of messaging rates and topics, are associated with hormonal therapy adherence. The main contributions of this research are summarized as follows:

- **Messaging patterns**. The results suggest that breast cancer patients who submit a relatively large number of messages are at greater risk of discontinuing hormonal therapy medications. In addition, these patients tend to send more messages before hormonal therapy and less as the treatment unfolds.

- **Messaging content**. An unsupervised analysis indicates that there is a broad range of health related topics communicated by the patients. These topics include common requests that have been shown to exist in studies that rely on manual review, such as appointment scheduling and requests for prescription refills [131], but we also find that some topics have notable temporal patterns that relate to breast cancer treatment workflow.

- **Survival analysis**. The findings indicate that messaging rate, mentions of side effects and surgery-related topics are associated with medication discontinuation. By contrast, seeking professional suggestions, expressing gratitude to healthcare providers, and mentions of drugs used to treat side effects are associated with medication adherence.

4.2 Related Work

The previous section provides evidence that UGC from GOSMPs and OHCs can be applied to effectively learn about an individual's health. In this section, we review prior research related to 1) UGC generated in patient portals and 2) HTA.

4.2.1 Investigation on Patient Portal Messages

As patient portals grow in their deployment and adoption, they are increasingly invoked in health-related research. However, most research on data from this domain focuses on the association between the number of messages sent and health outcomes. For instance, it has been shown that frequent usage of this service is associated with better glycemic control, increased outpatient utilization, higher rates of hemoglobin A1c testing adherence [132, 133], and reductions in both urgent care and primary care utilization [134, 135]. More recently, there have been investigations into the content of patient portal messages [131, 136, 137]. However, these studies are limited in several ways. First, some of these studies rely on humans to manually review and annotate message instances, which is both time-consuming and lacks scalability. Alternative, some investigations applied supervised learning to classify topics in the messages, which addresses the issue of scalability, but is unable to detect topics that fail to be defined *a priori* or are not found in the training data. Moreover, to the best of our knowledge, there are no studies that focus on linking the content of patient portal messages to health outcomes and behaviors. Yet this provides us with an opportunity because UGC generated in a patient portal may contain factors that are not captured in structured EMRs. As such, in this chapter we apply an unsupervised learning model to discover patient-communicated topics and subsequently associate the learned topics with breast cancer patients' medication discontinuation.

4.2.2 Hormonal Therapy Treatment Adherence

Traditional investigations in HTA tend to rely on patient information gathered through surveys or data in the official EMR. However, as noted in a recent study [125], these types of data sources are confounded by the fact that they are either time-consuming to gather (e.g., surveys) or lack information about a patient's treatment experience beyond what is seen (or documented) by a licensed healthcare provider. There is increasing evidence showing that UGC generated in online environments can be utilized to learn factors related to hormonal therapy medication discontinuation. For instance, one study based on a breast cancer forum suggested that patients who mentioned side effects such as depression are more likely to discontinue hormonal therapy [125]. Additionally, many potential barriers to breast cancer treatments (e.g., cost and trust) can be detected in UGC from the online environment [103]. However, there are no studies that examine the factors inferred from patient portal messages and their association with HTA. Our work strengthens the evidence that UGC generated in online environments can supplement traditional EMRs in healthcare research.

4.3 Data Preparation

In this section, we describe and summarize the study cohort, as well as the strategy applied to define the hormone therapy discontinuation events under investigation.

4.3.1 Extracting Study Cohort

This study relies on de-identified data from the patient portal and EMR of VUMC and was approved by the institutional review board of Vanderbilt University (protocol 162016). The data includes, but is not limited to diagnosis codes, procedural codes, medications, test reports, and messages communicated between patients and healthcare providers via the online portal. All patient identities were replaced with persistent pseudonyms by a third party honest broker and all dates within a record were consistently offset by a number of days uniformly sampled from a (-365,-1) range.

To create the study cohort, we rely on the VUMC Tumor Registry and extract patients who were diagnosed with stage I to III breast cancer. We further focus on breast cancer patients who were prescribed at least one of the following hormonal therapy medications [138]: *Anastrozole*, *Exemestane*, *Letrozole*, which are categorized as Aromatase Inhibitors (AIs), and *Raloxifene*, *Tamoxifen*, which are known as Selective Estrogen Receptor Modulators (SERMs). For each patient, we treat the first medication entry date of these medications as the start of the patient's hormonal therapy. This procedure yields a cohort of 1,473 patients with hormonal therapy medication entry dates ranging from late 1994 to early 2017.

The patient portal was established much later than the EMR system. As a result, we

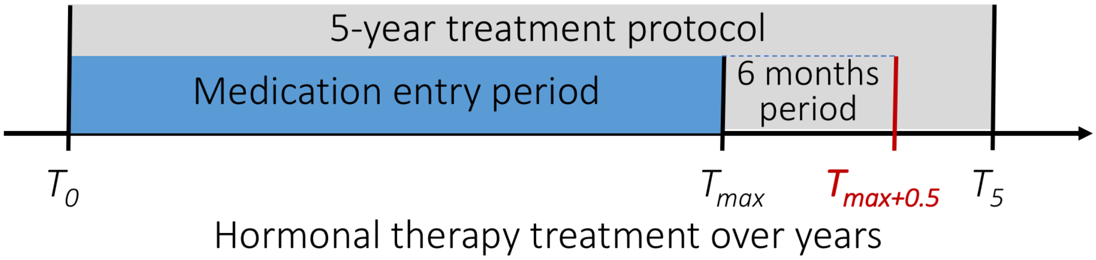


Figure 4.1: A schematic illustrating the critical time points used to define medication discontinuation event within 5 years' treatment. T_0 refers to the first hormonal medication entry date, and T_{max} is the maximum medication entry date. $T_{max+0.5}$ represents the 6 months after the T_{max} and T_5 is the end of the 5-year treatment protocol. A discontinuation event is experienced at $T_{max+0.5}$when $T_{max+0.5} < T_5$ and $T_{max+0.5}$ is smaller than the end of data collection window.

constrain the study cohort to breast cancer patients who started their treatment after the date of the first message sent by breast cancer patients, which took place in late 2005. After removing patients who were reported as deceased or have no messaging records, we end up with a cohort of 1,106 patients. It should be noted that the patients with deceased status are excluded due to their unsolved treatment adherence status.

4.3.2 Determining Medication Discontinuation Events

Adherence to hormonal therapy is important to prevent cancer recurrence. It is clinically recommended to take prescribed medications for at least 5 years as a treatment protocol, while there is some evidence to suggest that 10 years of treatment achieve greater benefits [139]. In this study, we focus on medication discontinuation realized within the initial 5 years of hormonal therapy. Although some patients in the cohort are affiliated with more than 5 years' medication entry history, there were not a sufficient number to merit investigation at the time of this study.

Figure 4.1 illustrates the critical time points that define a medication discontinuation event. Specifically, we estimate the event at 6 months after the maximum medication entry date (denoted as $T_{max+0.5}$) within a 5-year period. This is due to the fact that the breast cancer patients prescribed hormonal therapy are expected to have a follow-up every 6 months in the first 5 years after diagnosis [138]. However, it should be noted that for some patients, we do not observe a full 5 years worth of records (e.g., patients who started treatment after 2012). As such, their $T_{max+0.5}$ values are smaller than 5 years, but exceed the end of data collection window. We are unsure if these patients will discontinue medications in the future. In this circumstance, we treat their event status as right-censored in the survival analysis described below.

4.3.3 Study Cohort Summary Statistics

Using a 12-year data collection window, we obtain 245 (22.2%) right censored patients, 478 (53.2%) patients finishing a 5-year protocol, and 383 (34.6%) patients dropping off their medications. This latter observation is in alignment with a finding by a systematic evidence review, based on 29 studies, that indicated discontinuation rates of hormonal therapy ranges from 31-73% [75].

The patients in the cohort had an average age of 53.9 (±11.1) at breast cancer diagnosis. 91.3% of these patients are White, 5.9% are African American, 1.9% are Asian, and 0.9% are some other race. 12.9% of the patients were in an advanced cancer stage (i.e., stage IIIa, IIIb, or IIIc), while 87.1% were in early cancer stages (i.e., stage Ia, Ib, IIa, IIb) when their diagnosis was documented in the EMR. Among these patients, 52.4% were prescribed only AI medications, 8.6% were prescribed only SERM medications, and 39.0% were prescribed both AI and SERM medications during the course of their treatment.

4.4 Messaging Patterns

In this section, we investigate how patients use the portal messaging service, in terms of messaging volume and messaging rates along the treatment timeline.

4.4.1 Messaging Volume

We first examine the number of messages sent by the breast cancer patients. Figure 4.2 shows the log-log plot of the messaging frequency distribution, from which a clear heavy tailed phenomenon can be observed. Specifically, 10% of the patients sent only 1 message during the data collection window. By contrast, 0.2% of the patients sent more than 500 messages. This suggests that patients' messaging behavior in the private clinical environment might be consistent with that found in online health communities [140].

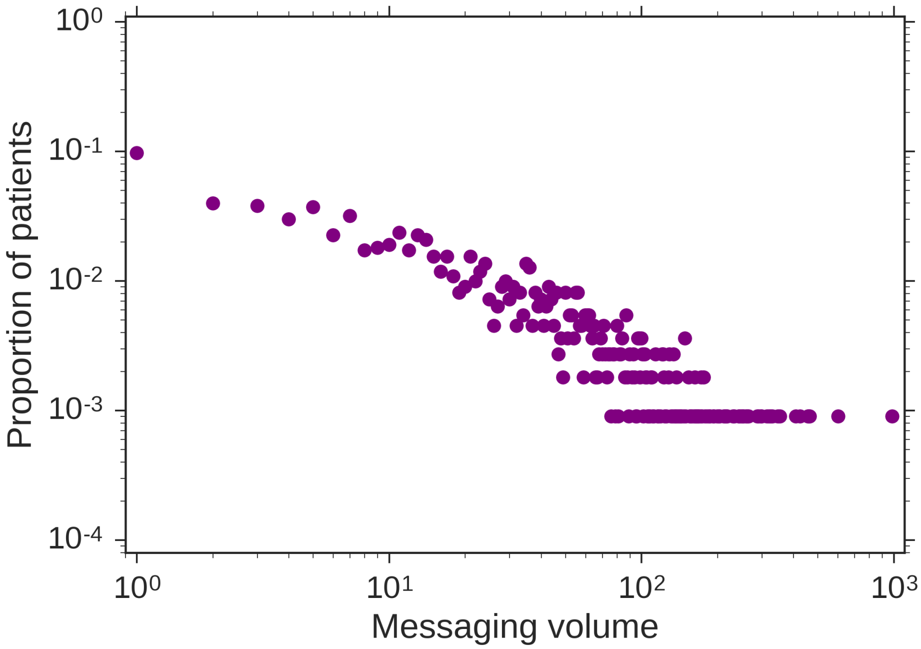


Figure 4.2: Log-log plot of the number of messages sent by patients through the patient portal.

We further explore the connection between messaging volume and medication discontinuation. To do so, we hold out the right censored patients and, thus, focus only on those who either completed 5-year treatment or experienced medication discontinuation events. We order these patients according to their messaging volume and apply a moving average to estimate the probability that a patient discontinues medication in each sliding window. Each window includes 12.5% of the selected patients for the risk estimation. We also apply the same moving average strategy to obtain the corresponding average log transformed messaging volumes. Figure 4.3 illustrates the relationship between messaging volume and medication discontinuation probabilities. The green line shows a smoothing cubic spline fit.

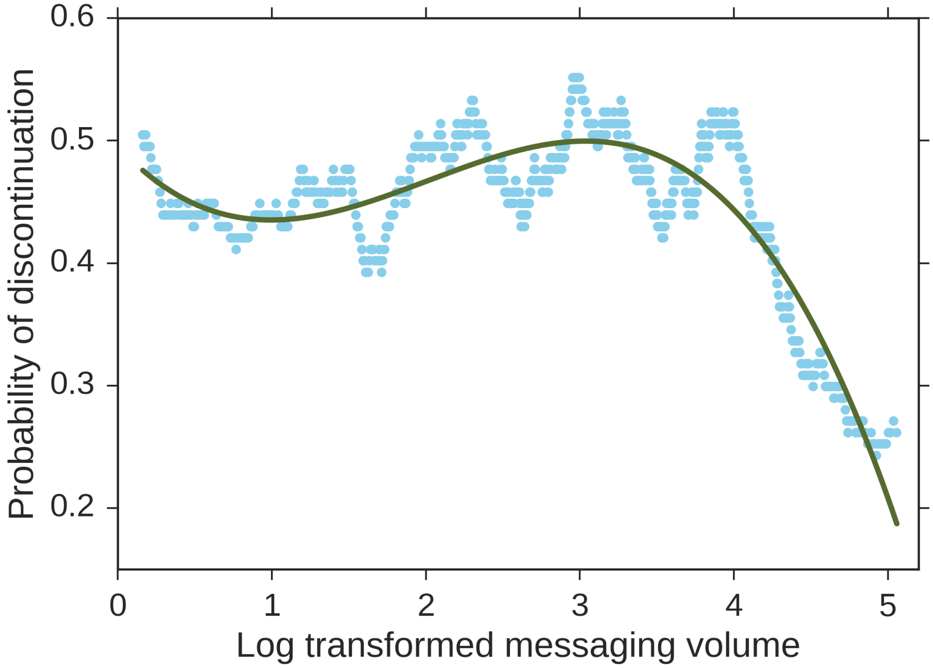


Figure 4.3: Log transformed messaging volume with respect to the probability of medication discontinuation (after smoothing with a moving average).

Figure 4.3 shows that patients are less likely to discontinue medication as the number of messages they send increases. However, the trend of decreasing likelihood falters when the messaging volume grows to around 3 messages on average. At this point, patients begin to exhibit an increasing risk of discontinuing medication until the messaging volume reaches another inflection point at around 20 messages. After this point the probability of discontinuing medication rapidly decreases. Since a large proportion of patients characterized by this decreasing trend completed the 5-year treatment protocol, and considering our 12-year data collection window, we suspect that these patients accumulated a high messaging volume by using the service over years. This suggests that messaging rates may be an important indicator when predicting treatment adherence.

4.4.2 Longitudinal Messaging Pattern

Given the observations based on messaging volume, we initiated an investigation into how messaging rates change across the treatment timeline. To obtain the overall trend, we set the first hormonal therapy medication entry date as the starting point (also known as the start of hormonal therapy treatment) and partition the timeline of messaging dates into a series of 6-month periods. As such, the time period with an index of 0 corresponds to the first 6-month hormonal therapy treatment, while time periods with negative (non-negative) index values correspond to time before (after) hormonal therapy treatment. We calculate messaging rates for each patient by counting the number of messages sent in each time period.

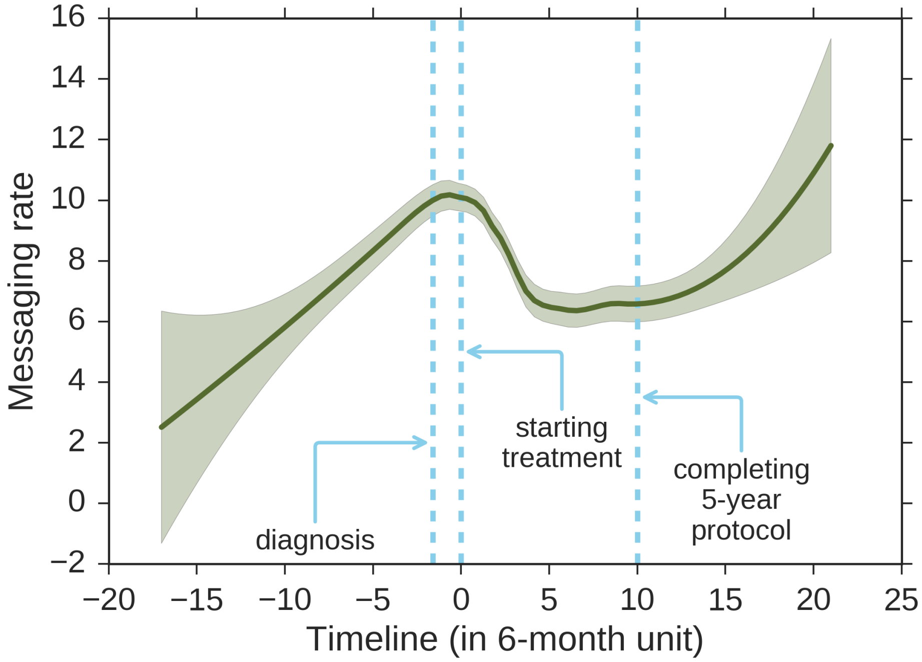


Figure 4.4: The LOWESS smoothed curve and 95% confidence interval of the number of messages sent per 6-month period. 25% of the data is applied to estimate each data point. The three dotted lines from left to right indicate: left) the 90% percentile of diagnosis dates (index of -1.6), middle) the therapy start date (index of 0), and right) the end of the 5-year protocol (index of 10).

Figure 4.4 depicts the LOWESS smooth curve with its 95% confidence interval for messaging rates along the treatment timeline. There are three clinically important dated events highlighted in the figure. These are indicated with vertical dotted lines: left) the diagnosis of breast cancer (index of -1.6, 90% percentile), middle) the start of hormonal therapy (beginning of the period with index equal to 0), and right) completing a five-year treatment protocol (beginning of the period with index equal to 10). These three events segment the system into four different trends:

1) Breast cancer patients send messages with an increasing rate before the disease is diagnosed.

2) The messaging rate reaches the maximum just before the start of the hormonal therapy.

3) The messaging rate drops quickly in the initial 2.5 years of treatment, and then remains unchanged until the end of the 5-year protocol.

4) The rate increases up in the second 5-year protocol.

Note that the wider estimated confidence intervals in the two ends of the timeline (beyond indices of ±10) are caused by the limited number of observations. The third trend supports our conjecture that messaging rate might associate with a medication discontinuation event. Meanwhile, a comparison between the second and third trends suggests that breast cancer patients tend to have less communications with healthcare providers as hormonal therapy progresses.

4.5 Messaging Content

Next, we investigate the content of the messages communicated through the patient portal. Specifically, we focus on inferred topics and their trends along the treatment timeline.

Topic	Word Samples	Rank
58	*checkup, annual, exam, appointment, appt, apt, yearly, appts, appointments, visit*	6
60	*bloodwork, labs, cbc, mammogram, ultrasound, tests, testing, images, films, biopsy, mri, reports, colonoscopy, test, report, results, records, labwork, mamogram, scope*	12
9	*cvs, publix, walmart, wal, kroger, costco, walgreens, krogers, pharm, mart, walid, rite, caremark, medco, kmart, target, tvc, mce, rightsource, riteaid*	18
38	*prednisone, exemestane, wellbutrin, cymbalta, metformin, gabapentin, celexa, paxil, prozac, zoloft, levaquin, lexapro, zetia, warfarin, lyrica, methotrexate, arimidex, aromasin, effexor, lovenox*	23
107	*lips, tingling, thighs, muscles, arms, shoulders, hips, joints, fingertips, tongue, knees, sensation, wrists, soreness, lip, ankles, forehead, toes, glands, tenderness*	29
69	*documentation, statement, certification, unum, fers, signature, receipt, disability, employer, necessity, document, verification, authorization, letter, account, fmla, approval, clearance, letterhead, files*	30
124	*herceptin, taxol, infusions, treatment, chemotherapy, chemo, zometa, treatments, infusion, radiation*	44
13	*diarrhea, headache, chills, vomiting, spotting, coughing, nausea, headaches, bleeding, appetite, spells, eating, cough, migraines, breath, bleed, periods, energy, shortness, urinate*	51

Table 4.1: Topics indicating the most common usages of the messaging service. It should be noted that many other topics such as verbs (e.g., related to needs, refills, actions, and schedules), dates, locations and healthcare providers are very popular as well.

4.5.1 Obtaining Messaging Topics

First, we extract messaging topics in an unsupervised fashion. We anticipate that an unsupervised approach will yield more categories than a traditional supervised approach. While topic modeling has been applied to personally contributed information (e.g., [38]), in this work, we apply hierarchical clustering directly on the words in the messages. By employing a word embedding technique for natural language processing (in the form of word2vec), we map words into a vector space with much lower dimensionality and calculate their semantic similarity in terms of their cosine distance [107].

We use a skip-gram model with negative sampling implemented in *gensim* package (version 0.13.1) to fit a word2vec model with all messages in our cohort. We keep the words that appeared at least five times, choose a sliding window with size of 5, and set the word vector length as 200. While the data volume is modest (58,569 messages with 3,164,848 words), the health related topics enable us to fit a promising model. For instance, the ten most related words to *depression* obtained from our model are: *anxiety*, *mood*, *panic*, *attacks*, *fatigue*, *neuropathy*, *constipation*, *vertigo*, *swings*, and *sadness*.

However, the fitted word2vec model still contained 13,171 distinct words, many of which are unlikely to assist in understanding messaging content. As a result, we chose the 2000 most frequent words, and enlarge the selected word collection by incorporating words that exhibit a cosine distance greater than 0.6 with any of the selected words. This is expected to include more words due to misspelling and semantic similarity. This process recovers an additional 2,010 words. We further remove several types of words that we suspect would contribute little towards constructing meaningful topics: 1) stop words (e.g., “the” or “of”), 2) years, and 3) words with a cosine distance greater than 0.6 from the words noted above. This process yields 3664 words.

We applied an agglomerative hierarchical clustering with complete linkage implemented in *sklearn* package (version 0.18.1) to extract topics. In practice, we do not want to generate word clusters that are extreme in size. When a cluster is too large, it tends to be composed of a mixture of topics. When a cluster is too small, it contributes little to dimensionality reduction. As such, we adopted a simple metric to help decide the number of clusters: the standard deviation of cluster size. When the number of clusters is greater than two, the standard deviation of the cluster size decreases towards zero as the number of clusters grows towards the vocabulary size with a proper step (e.g., 25 in our case). Based on this observation, we follow the elbow principle to locate the angle where marginal gain in cluster size begins dropping. We find the optimal number of clusters at 200.

4.5.2 Popular Topics

After detecting topics, we examine what are most popular messaging service usages. In this investigation, we confine our analysis to ±10 six-month periods. To measure topic popularity, we count the number of times that a topic was mentioned by any patient in any 6-month period. Table 4.1 shows the eight common topics regarding messaging usages, after removing those verbs, location, and time related word clusters. For each topic in the table, we rank the words based on their average similarity with other words in the same topic. If the size of the topic is greater than 20 (10), we display the top 20 (10) words; otherwise, we show all words in the topic[1].

It can be recognized that these topics are about appointments (#58), lab testing reports (#60), pharmacies (#9), hormonal therapy medications (#38, e.g., *exemestane*) and other drugs used for treating side effects (#38, e.g., *wellbutrin* used for treating depression), body and symptoms (#107), statements (#69), other breast cancer treatments (#124), and common side effects (#13). These topics characterize the most frequent concepts communicated when breast cancer patients message care providers through the online portal.

[1] We follow the same rule when depicting samples of words in all following tables that communicate topic-related information.

Topic	Word Samples	Coef	p
25	*upper, pelvis, fracture, abdomen, thoracic, neck, fractured, lumbar, pelvic, wall, spine, injury, inner, rt, cervical, nerve, stimulator, injured, lower, brace*	-0.681	0.010
168	*healthcare, services, united, human, resources, department, provider, social, worker, employee, group, billing, occupational, requirements, benefits, security, financial, dept, medical, portal*	-0.632	0.021
166	*culture, cd, urinalysis, rays, ray, cta, disc, dvd, sonogram, sample*	-0.615	0.025
161	*msg, message, notification, confirmation, voicemail, reminder, notice, mistake*	-0.571	0.041
170	*vaccine, shingles, flu, vaccination, h1n1, shot, pneumonia, shots*	0.626	0.022
147	*weak, fatigued, nauseous, dizzy, bothersome, constipated, nauseated, depressed, tired, congested, overwhelmed, discouraged, crying, hungry, annoying, upset, angry, badly, slowly, awful*	0.632	0.021
87	*drive, travel, live, trip, stay, closer, close*	0.720	0.006
3	*thick, phlegm, mucous, yellow, yellowish, mucus, brown, raised, darker, dark, colored, blisters, nose, bright, dry, formed, red, cloudy, greenish, wet*	0.747	0.003

Table 4.2: Topics with statistically significant temporal trends (at the 0.05 level). The topics are sorted according to the coefficients of Spearman's rank-order correlation. Positive (Negative) coefficients suggest increasing (decreasing) temporal topic trends.

4.5.3 Temporal Topic Trend

We further inspect how topics change along treatment timeline. To obtain the temporal topic trends, we first infer the topic distribution for each patient by calculating the percentage of topic words mentioned in a 6-month period. Then, we average the topic distribution across patients for each 6-month period. We measure the significance of temporal topic trends through calculating their Spearman's rank-order correlation with timeline at significance level of 0.05. Figure 4.2 shows topics with statistically significant temporal trend.

From Figure 4.2, it can be seen that the topics with an increasing trend in popularity are mainly about symptoms (topic #3), side effects and negative emotions (#147), vaccination (#170), and traveling (#87). By contrast, the topics with a decreasing trend in popularity are mainly about injure related (#25), notification (#161), examination (#166), and hospital visiting related (#168). We suspect that these trends are related to breast cancer treatments. For example, being *depressed*, *nauseous*, *dizzy* and *constipated* (#147) are common side effects caused by hormonal therapy medications [141]. However, before consuming hormonal therapy medications at home, breast cancer patients are generally undergoing examinations and receiving other treatments that can only be conducted within healthcare facilities (e.g., surgery, chemotherapy and radiotherapy), which may recur multiple topics (e.g., #168, #161) are more popular at that moment than in later hormonal therapy treatment.

4.6 Linking to Medication Adherence

We now investigate how the messaging behaviors, in terms of messaging rates and topics, associate with hormonal therapy medication discontinuation events.

4.6.1 Statistical Model

We apply a Cox proportional hazards regression model to learn associations between messaging behaviors and medication discontinuation. The Cox model is a method for investigating the effect of several independent variables with respect to the time when an event of interest happens. There are two primary benefits in applying a Cox model, instead of say logistic regression, in this study. First, the Cox model is a semi-parametric model that does not assume any particular survival distribution. Second, the Cox model can make use of right censored patients (whose medication discontinuation events are not observed at the end of the data collection window) by incorporating both time (to when an event happened) and adherence status (i.e., patients realized the events or right censored) into the model.

4.6.2 Control Predictors

We introduce four adjusted predictors into the model: *age at diagnosis*, *race*, *cancer stage*, and *hormonal therapy medication*. We impute missing values for age (1.1%) with the average age and scale the variable into the (0,1) range. We partition race into white and non-white and use 1 to encode an advanced cancer stage (and 0 for early cancer stage). The variable of hormonal therapy medication (denoted as *taking AI*) is represented as a proportion of the number of periods on AI divides to the number of periods on either AI or SERM medications.

4.6.3 Message Related Predictors

We construct two types of message related predictors. First, we build topic predictors as follows: 1) for each patient, we aggregate all the messages sent after the breast cancer diagnosis date and before either: i) when the medication discontinuation event occurred or ii) the patient is right censored. As such, we model each patient as a diary of messages; 2) we replace the words in each patient diary with the corresponding topic numbers (if present); 3) we calculate the TF-IDF values for each topic in each patient diary, which we use as topic predictor values. Second, we include messaging rate (in terms of the average number of messages sent per six-month period) as an additional variable. Since the messaging rate distribution is right-skewed, we applied a log transform and, subsequently, scaled the data into a (0,1) range before applying the Cox model.

4.6.4 Cox Model Results

We fit a Cox model with a concordance of 0.753 through applying *lifelines* package (version 0.9.4) . Note that concordance measures the proportion of patient pairs in which patients with a higher-risk predictor discontinue medications before patients with the lower-risk predictor.

4.6.4.1 Significant Control Predictors

The risk of a predictor is represented by the hazard ratio (*HR*) in the form of the expected exponential of its estimated coefficient in the Cox model. If the *HR* is significantly greater than 1, then the predictor is associated with an increased risk of discontinuing a prescribed medication. By contrast, if a *HR* is significantly less than 1, then the predictor is associated with a decreased risk of discontinuation. Among the four control predictors, *age at diagnosis* is significantly associated with an increased risk of discontinuation ($HR = 1.173$, $p = 0.026$), while *taking AI* is significantly associated with a decreased risk of discontinuation ($HR = 0.715$, $p < 0.001$).

4.6.4.2 Messaging Rate

After controlling for the four proposed predictors, average messaging rate is found to be positively associated with discontinuation ($HR = 1.373$, $p = 0.002$). This further lends evidence to our conjecture that the messaging rate might be associated with medication adherence.

4.6.4.3 Topic With Increased Risk

Tables 4.3 summarizes the topics with an increased likelihood (with $HR > 1$) of discontinuing a medication. From Table 4.3, it can be seen that patients who mention tests for assessing heart damage (#126, $HR = 1.216$) or describe common side effects caused by hormonal therapy (#13, $HR = 1.214$; #53, $HR = 1.164$) have an increased risk of discontinuing medications. As one patient communicated when taking Tamoxifen,

> *"I am having really bad* ***mood*** *sID*** and depression is becoming a problem. My appt is not until middle of **DATE and I don't think I can wait that long. I think its the tamoxifen that I'm taking."*

Patients who mentioned breast- or ovary-related surgeries (#44, $HR = 1.170$) are more likely to discontinue medications. As one patient once wrote:

> *"I just had a full **hysterectomy** last week and I had a **bilateral mast(ectomy)** at the end of **DATE. So...I won't feel comfortable starting any drugs, i.e., Tamoxifen unless we check my blood."*

Patients who discontinue medications might mention their blood test results (#0, $HR = 1.194$). As another patient noted:

> *"Ok, been off the tamoxifen for a month. I feel much better! ... Are there better options? Not really wanting to go back.... So, looking at my lab results, there appears to be a significant change in the **hemoglobin**, PCV, **platelet** and red blood cell counts... Does tamoxifen cause that?"*

Patients who discontinue medications also mention verbs (#189, $HR = 1.159$) or adjectives (#79, $HR = 1.130$) that might be related to health conditions or tests. As one patient voiced:

> *" Is it **normal** to feel a little dizzy and nauteous? It has been almost a week since my surgery."*

Another topic that was associated with an increased risk of medication discontinuation is mentions of a website (#56, $HR = 1.171$), through which patients may conduct research or gather information. As one patient said:

> *"Can I take Tamoxifen with Factor Five Leiden? I've read lots of things **online** in my research ... if I cant as it will greatly increase my DVT and Blood Clot risks."*

4.6.4.4 Topic With Decreased Risk

Tables 4.4 summarizes the topics with a decreased likelihood (with $HR < 1$) of medication discontinuation. These topics include seeking suggestions and good relationships with healthcare providers (#136, $HR = 0.766$; #42, $HR = 0.838$; #105, $HR = 0.872$):

> *"Do I need to modify any of medications (excluding RA medications) prior or post surgery? - Tamoxifen -Zoladex **Thank** you for your **expertise**."*

> *"**Thank** you so very much for our good appointment today. I really did need to see you, and I am grateful for the **expertise**, **knowledge**, and history you bring to my case and overall health issues."*

These patients may also take drugs to cope with side effects or symptoms (#38, $HR = 0.807$; #2, $HR = 0.822$; #25, $HR = 0.831$; #85 $HR = 0.829$):

> *"... The **Wellbutrin continues** to help with concentration and neuropathic pain. I am still taking it twice a day. At this time I do not need a new prescription ..."*

> *"The pain is joint-related and generalized (feet, knees, hips, **spine**, **neck**, shoulders, hands) ... In fact it worsened so much that I took myself off the Femara on Saturday and **restarted** the Arimidex instead."*

> *"He saw my eye issues as **indicative** of paraeoplastic syndrome. "*

Patients may also mention terms related to decision making (#187, $HR = 0.833$), expressing preference (#106, $HR = 0.846$), or others (#157, HR =0.849). As one patient wrote:

> *"Hi, I have **decided** that I will continue in the clinical trail for the next five years."*

While there are two statistically significant topics (#109, #127) that are difficult to explain their associations with medication discontinue, most of our findings are confirmed by the literature, as discussed in the following section.

4.7 Discussion

The findings of this study have several major implications that we wish to highlight. First, the topics communicated by breast cancer patients via patient portal messages appear to be effective indicators of their potential for treatment adherence. In particular, we discovered several topics that statistically significantly associate with medication discontinuation events and are clearly supported by evidence in the literature. For example, gastrointestinal reactions such as *nausea* and *vomitting* (topic #13) have been shown to be risk factors of hormone therapy discontinuation [142]. Additionally, while an *echocardiogram* (*echo*) or *EKG* (*electrocardiogram*) (topic #122) is often used to detect potential long-term side effects induced by radiation therapy or chemotherapy, cardiac complications are also recognized as a severe side effects of AI [143]. Another interesting finding is that our clustering algorithm puts *lumpectomy* and *mastectomy* in the same surgery-related topic (#44), which was found to be associated with an increased risk of discontinuing medication. This is notable because, in the literature, *lumpectomy* was found to be associated with early hormonal therapy discontinuation, while *mastectomy* was not [144].

We further observe that patients who request professional suggestions (#136) or express gratitude to healthcare providers (#105) are associated with HTA. This finding is in alignment with evidence that respect for the advice of caring physicians and family members is one of the factors driving breast cancer patients to adhere to prescribed treatments and follow-up appointments [145]. Further, there are also studies that have shown that a good relationship with one's physician and self-efficacy in taking medication are associated with better hormonal therapy adherence [146]. Meanwhile, it has been shown that managing side effects can help medication adherence [147] as well. We believe that this may explain why mentions of drugs for treating side effects are found to associate with treatment adherence in this study. These factors are rarely recorded in structured EMR systems, suggesting that patient portal messages can supplement traditional data resources to support healthcare investigations.

Second, we find that the average messaging rate has an increased risk of medication discontinuation. This suggests that healthcare providers who communicate with patients through patient portal could be encouraged to pay more attention to patients with abnormal messaging rates. While sending messages at a reasonable rate suggests a normal treatment routine, sending messages at a faster rate may indicate that patient have more concerns or questions about their health conditions or treatment, which, in turn, could lead to future medication discontinuation. Still, the messaging rate is only an indicator of a potential problem and the semantics communicated in the messages will need to be considered before taking any action.

Third, our data-driven approach leverages minimal human efforts to analyze patient portal messages. For example, we can identify common usages of the messaging service, which are aligned with the findings in previous studies based on manually explored message content [131]. With unsupervised methods, it may be possible to design and deploy an automatic alert system to monitor patients' messages. Such a system would help alleviate the burden of healthcare providers, as well as assist them in providing more effective (e.g., more timely and targeted) interventions to help patients achieve completion of long-term treatment regimens.

Despite the merits of this investigation, there are several limitations that we believe can serve as the basis of future work. First, the study cohort comes from one institution, which may limit the generalizability of our findings. Second, while the combination of word embedding and clustering promises quality in most topics, there are some topics with words that lack a clear implication. For example, we find some topics that exhibit statistical significance (e.g., #109, #127) lack a clear relationship with medication discontinuation. This is one limitation by adopting a data-driven approach. We believe it would be a fruitful endeavor to supplement the unsupervised methods with domain knowledge, so that topic extraction and selection could be guided before fitting a statistical model. Finally, it will be useful to compare the results of this investigation with patients who achieve 10 years of hormonal therapy treatment.

Topic	Word Samples	HR	95% CI	p
122	*xray, marrow, ekg, echo, abd, density, mets, emg, tail, echocardiogram*	1.216	(1.013, 1.460)	0.036
109	*came, arrived, got, went, ran, returned, moved, found, called*	1.214	(1.065, 1.384)	0.004
13	*diarrhea, headache, chills, vomiting, spotting, coughing, nausea, headaches, bleeding, appetite, spells, eating, cough, migraines, breath, bleed, periods, energy, shortness, urinate*	1.214	(1.057, 1.394)	0.006
0	*triglycerides, hdl, hemoglobin, creatinine, ferritin, glucose, bun, ldl, t3, crp, ast, protein, t4, platelet, fsh, phosphatase, wbc, serum, enzymes, sodium*	1.194	(1.046, 1.363)	0.008
56	*myhealthatvanderbilt, myhealth, line, online, computer, internet, sheet, summary, site, listed*	1.171	(1.044, 1.313)	0.007
44	*reconstruction, lumpectomy, mastectomy, diep, flap, bilateral, hysterectomy, surgery, procedure, ovaries*	1.170	(1.012, 1.353)	0.034
53	*irritability, irritable, worsening, sadness, intestinal, mood, swings, frequent, significantly, attacks, cognitive, functioning, appearance, swallowing, panic, vision, osteoarthritis, peripheral, inflammation, balance*	1.164	(1.024, 1.322)	0.020
189	*eliminate, prevent, improve, affect, reduce, minimize, impact, resolve*	1.159	(1.045, 1.285)	0.005
79	*negative, positive, expected, normal, tested, result, compared, spread, lead*	1.130	(1.008, 1.267)	0.036
192	*suppression, blocking, hormonal, induced, bkm120, osteoporosis, estrogen, nutrition, diabetes, non*	1.129	(1.003, 1.272)	0.044

Table 4.3: Topics that are positively associated with a medication discontinuation event (statistically significant at the 0.05 level). The topics are sorted based on their *HRs*. A larger *HR* suggests more increased risk of medication discontinuation.

Topic	Word Samples	HR	95% CI	p
105	*thx, thankyou, ty, thks, wishes, regards, promptly, thank, thanks, ed*	0.872	(0.763, 0.997)	0.044
102	*ovarian, oid, dcis, ductal, uterine, invasive, colon, stage, diagnosis, policy*	0.868	(0.756, 0.996)	0.043
24	*wondered, wondering, wandering, wonder, correctly, wrong*	0.861	(0.757, 0.979)	0.023
157	*informed, advised, notified, assured, offered, told, treated, given, diagnosed, treating*	0.849	(0.741, 0.973)	0.018
106	*prefer, like, mind*	0.846	(0.733, 0.977)	0.023
127	*happens, happening, exactly, means*	0.844	(0.731, 0.975)	0.021
42	*cardiologist, urologist, dermatologist, neurologist, gyn, gynecologist, psychiatrist, doc, oncologist, physician, rheumatologist, doctor, specialist, pcp, gp, friend, ob, onc, obgyn, woman*	0.838	(0.729, 0.964)	0.013
187	*decided, plan*	0.833	(0.717, 0.968)	0.017
25	*upper, pelvis, fracture, abdomen, thoracic, neck, fractured, lumbar, pelvic, wall, spine, injury, inner, rt, cervical, nerve, stimulator, injured, lower, brace*	0.831	(0.705, 0.980)	0.028
85	*variety, indicative, bouts, lack, importance, lots, tons, ahold, alot, signs, proof, none, course, lieu, episodes, instances, events, expense, plenty, rid*	0.829	(0.715, 0.962)	0.013
2	*start, begin, stop, discontinue, resume, finish, continue, restart, skip, wait*	0.822	(0.704, 0.959)	0.013
38	*prednisone, exemestane, wellbutrin, cymbalta, metformin, gabapentin, celexa, paxil, prozac, zoloft, levaquin, lexapro, zetia, warfarin, lyrica, methotrexate, arimidex, aromasin, effexor, lovenox*	0.807	(0.682, 0.955)	0.013
136	*knowledge, expertise, guidance, efforts, counsel, input, responses, feedback, kindness, recommendations, compassion, attentiveness, consideration, judgment, patience, advice, support, assistance, encouragement, suggestion*	0.766	(0.645, 0.909)	0.002

Table 4.4: Topics that are negatively associated with a medication discontinuation event (statistically significant at the 0.05 level). The topics are sorted based on their *HRs*. A smaller *HR* suggests more decreased risk of medication discontinuation.

CHAPTER 5

CONCLUSION

In this dissertation, we investigated how consumer generated information can be applied to learn about people or patients' health related behaviors. The dissertation can be summarized as follows:

5.1 Learning Health Status Disclosure on General Social Media Platform

We showed that a health mention detection system can be designed and deployed for microblogging systems, such as Twitter. At the same time, we illustrated that the information communicated through such mentions can disclose the health status of the authors and other individuals at a wide range of rates. Our investigation further showed that the combination of tweets from several health issues can yield a classifier that dominates a classifier based on the tweets of a single health issue. This may enable the system to use a smaller amount of training data to build a classifier that detects health status mentions across a range of health issues. Meanwhile, our findings highlight that the authors of tweets tend to disclose information about another persons health status when talking about the high cost of medicine or treatment and when searching for social support, but disclose information about their own health status when talking more benign health issues related with simple chronic biological processes and negative emotions. We anticipate extending this work to include more robust health mention classifiers and regularizing NMF to obtain more interpretable basis components from the factorization process.

5.2 Learning Hormonal Therapy Medication Adherence

We investigated hormone therapy adherence (HTA) based on patient self-reported information in a large, longitudinal online breast cancer forum. We focused on a dataset collected from breastcancer.org and characterized adherence behavior with three types of events: taking (medication), interruption (of the treatment regimen) and completion (of five-years of treatment). From an emotional perspective, we found that when patients mention taking (medication) events, they have a relatively higher rate of fear (for potential side effects); when patients mention interruption events, they have a relatively higher rate of anger; and when patients mention completion events, they exude more joy and less fear, but also experience relatively higher sadness. Most of our personality analysis confirmed results from treatment adherence studies based on data collected from offline settings (e.g., surveys), but we also found self-discipline is negatively associated with completion events, which should be interesting to be investigated in future. Our interruption event prediction model suggested that patients at the beginning of their treatments are more likely to realize interruption events in the future than patients who successfully make it through several years of treatment. We have demonstrated that patient-provided information in an online breast cancer community can be potentially applied to study HTA. We believe our methodology can be adopted to study adherence to treatment for other health issues through patient self-reported online information.

5.3 Learning Patients' Messaging Behavior and Its Association with Treatment Adherence

We investigated a cohort of breast cancer patients who underwent hormonal therapy at VUMC, and studied their messaging patterns, inferred topics from the messages they communicated to care providers through an online portal, and uncovered associations between these factors and hormonal therapy treatment adherence. Our analysis on patients' messaging volume and rates suggested that 1) patients tend to send more messages before the start of treatment and 2) patients with high messaging rates exhibited a greater risk of discontinuing a prescribed five-year medication regimen. We further applied a word embedding model (i.e., word2vec) and a hierarchical clustering strategy to extract a broad range of health related topics, and characterized these topics for their popularity and temporal trends. In doing so, we verified that patients use the messaging service to accomplish goals (e.g., schedule an appointment and request a prescription refill) observed in prior studies that relied on manual review of such messages. We conducted a survival analysis, which (after controlling for age at diagnosis, cancer stage, race and hormonal therapy medications) indicated that the average messaging rate, mentions of side effects and surgery-related topics are associated with medication discontinuation for breast cancer patients. By contrast, our model showed that seeking professional suggestions, expressing gratitude to healthcare providers, and mentions of drugs used to treat side effects are associated with treatment adherence. This research strengthens the evidence that UGC in online environments can supplement traditional medical data to study health-related behaviors. We also believe that there are opportunities for extending this research by replicating the study with data from additional institutions and refining the unsupervised learning model to incorporate expert guidance, so that the topics learned are clearly interpretable and meaningful for healthcare professionals.

We believe the research in this dissertation strengthens the evidence that UGC in online environment can be utilized to supplement traditional data sources (e.g., EMRs) to conduct biomedical informatics research. Further research directions include 1) building infrastructure to collect and link people's generated information in multiple disparate platforms, 2) extracting knowledge map from unstructured, longitudinal data to better represent patients' profile, and 3) applying statistical inference (e.g., causal inference, Bayesian analysis), machine learning and natural language processing technique to learn more health related behaviors and decision making. We anticipate our research strengthen the fields where UGC can be applied to improve health and help clinical decision making.

However, it should be noted that research regarding improving personal health using their UGC in online environments should fully consider and honor the privacy of the individuals. From this perspective, there are several open ethnical challenges that should be considered in future. First, are users informed that their data is applied in healthcare research after they sign up user agreement when using online service? While it is appealing for research, a large dataset brings great challenges to obtain informed consent from thousands of their generators. Second, will the research promise the anonymity of those passive “participants? The timely accumulated data is making it more challenging to decide when and how to present the research results (or even the data) without disclosing identities. For instance, while current presentation of research results may protect privacy, is it still true when more and more UGC becomes available in long run? Finally, once some patients are identified with high risk of discontinuing medications by our algorithm, how should healthcare providers provide effective interventions without offending these patients, especially in a situation that they may not be aware that their data is being studied? We refer the reader to a recent tutorial [148] and guidance [149] for further discussion on ethics issues and additional challenges associated with using UGC in academic research.

BIBLIOGRAPHY

[1] Andrew M Garratt, Danny A Ruta, Mona I Abdalla, J Kenneth Buckingham, and Ian T Russell. The sf36 health survey questionnaire: an outcome measure suitable for routine use within the nhs? *BMJ*, 306(6890):1440–1444, 1993.

[2] Gregory P Samsa, David B Matchar, Larry B Goldstein, Arthur J Bonito, Linda J Lux, David M Witter, and John Bian. Quality of anticoagulation management among patients with atrial fibrillation: results of a review of medical records from 2 communities. *Archives of Internal Medicine*, 160(7):967–973, 2000.

[3] Linda S Williams, Engin Y Yilmaz, and Alfredo M Lopez-Yunez. Retrospective assessment of initial stroke severity with the nih stroke scale. *Stroke*, 31(4):858–862, 2000.

[4] Lois Quam, Lynda BM Ellis, Pat Venus, Jon Clouse, Cynthia G Taylor, and Sheila Leatherman. Using claims data for epidemiologic research: the concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population. *Medical Care*, pages 498–507, 1993.

[5] Eysenbach Gunther and Wyatt Jeremy. Using the internet for surveys and health research. *Journal of Medical Internet Research*, 4(2):E13, 2002.

[6] Pascal Coorevits, M Sundgren, Gunnar O Klein, A Bahr, B Claerhout, C Daniel, Martin Dugas, D Dupont, A Schmidt, P Singleton, et al. Electronic health records: new opportunities for clinical research. *Journal of Internal Medicine*, 274(6):547–560, 2013.

[7] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395, 2012.

[8] Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A Oniki, Les Westberg, Calvin E Beebe, Cui Tao, Craig G Parker, Peter J Haug, Stanley M Huff, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of ehr data: the sharpn project. *Journal of Biomedical Informatics*, 45(4):763–771, 2012.

[9] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *The International AAAI Conference on Web and Social Media*, 2014.

[10] Melanie Swan. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. *Journal of Medical Internet Research*, 14(2):e46, 2012.

[11] Nancy P Gordon and Mark C Hornbrook. Differences in access to and preferences for using patient portals and other ehealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan. *Journal of Medical Internet Research*, 18(3), 2016.

[12] Bengisu Tulu, John Trudel, Diane M Strong, Sharon A Johnson, Devi Sundaresan, and Lawrence Garber. Patient portals: An underused resource for improving patient engagement. *CHEST Journal*, 149(1):272–277, 2016.

[13] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale Twitter mining for drug-related adverse events. In *Proceedings of the International Workshop on Smart Health and Wellbeing*, pages 25–32, 2012.

[14] Subhabrata Mukherjee, Gerhard Weikum, and Cristian Danescu-Niculescu-Mizil. People on drugs: credibility of user statements in health communities. In *Proceedings of the 20^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74, 2014.

[15] Todd Bodnar, Victoria C Barclay, Nilam Ram, Conrad S Tucker, and Marcel Salathé. On the ground validation of online diagnosis with Twitter and medical records. In *Proceedings of the* 23*rd* *International Conference on World Wide Web (Companion Volume)*, pages 651–656, 2014.

[16] Dredze Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in Twitter. In *The International AAAI Conference on Web and Social Media*, pages 579–582, 2014.

[17] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *The International AAAI Conference on Web and Social Media*, pages 128–137, 2013.

[18] Ashley E Wade-Vuturo, Lindsay Satterwhite Mayberry, and Chandra Y Osborn. Secure messaging and diabetes management: experiences and perspectives of patient portal users. *Journal of the American Medical Informatics Association*, 20(3):519–525, 2013.

[19] Wan-Ying Lin, Xinzhi Zhang, Hayeon Song, and Kikuko Omori. Health information seeking in the web 2.0 age: Trust in social media, uncertainty reduction, and self-disclosure. *Computers in Human Behavior*, 56:289–294, 2016.

[20] Deborah Estrin. Small data, where n = me. *Commun. ACM*, 57(4):32–34, 2014.

[21] Santosh Kumar, Wendy J Nilsen, Amy Abernethy, Audie Atienza, Kevin Patrick, Misha Pavel, William T Riley, Albert Shar, Bonnie Spring, Donna Spruijt-Metz, et al. Mobile health technology evaluation: the mhealth evidence workshop. *American Journal of Preventive Medicine*, 45(2):228–236, 2013.

[22] Mark Tomlinson, Mary Jane Rotheram-Borus, Leslie Swartz, and Alexander C Tsai. Scaling up mhealth: where is the evidence? *PLoS Medicine*, 10(2):e1001382, 2013.

[23] John Riedl and Eric Riedl. Crowdsourcing medical research. *Computer*, 46(1):89–92, 2013.

[24] Paul Wicks, Timothy Vaughan, and James Heywood. Subjects no more: what happens when trial participants realize they hold the power? *BMJ*, 348:g368, 2014.

[25] Diana Tamir and Jason Mitchell. Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences USA*, 109(21):8038–8043, 2012.

[26] Theodore F Claypoole. Privacy and social media. *Business Law Today*, 2014.

[27] Pew Research Center. Public perceptions of privacy and security in the post-Snowden era, 2014.

[28] Maja van der Velden and Khaled El Emam. 'Not all my friends need to know': a qualitative study of teenage patients, privacy, and social media. *Journal of the American Medical Informatics Association*, 20(1):16–24, 2013.

[29] Kelly Caine and Rima Hanania. Patients want granular privacy controls over health information in electronic medical records. *Journal of the American Medical Informatics Association*, 20(1):7–15, 2013.

[30] Eric Meslin, Sheri Alpert, Aaron Carroll, Jere Odell, William Tierney, and Peter Schwartz. Giving patients granular control of personal health information: using an ethics "Points to Consider" to inform informatics system designers. *International Journal of Medical Informatics*, 82(12):1136–1143, 2013.

[31] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: an analysis of privacy leaks on Twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12, 2012.

[32] Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. A scalable framework to detect personal health mentions on twitter. *Journal of Medical Internet Research*, 17(6), 2015.

[33] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W White. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1365–1376. ACM, 2014.

[34] Michael Hayes, Ian Ross, Mike Gasher, Donald Gutstein, James Dunn, and Robert Hackett. Telling stories: News media, health literacy and public policy in Canada. *Social Science and Medicine*, 64(9):1842–1852, 2007.

[35] K. McNeil, P.M. Brna, and K.E. Gordon. Epilepsy in the Twitter era: a need to re-tweet the way we think about seizures. *Epilepsy and Behavior*, 23(2):127–130, 2012.

[36] Agency for Healthcare Research and Quality. Medical expenditure panel survey.

[37] Timothy M Hale, Akhilesh S Pathipati, Shiyi Zan, and Kamal Jethwani. Representation of health conditions on Facebook: content analysis and evaluation of user engagement. *Journal of Medical Internet Research*, 16(8):e182, 2014.

[38] Michael J Paul and Mark Dredze. Discovering health topics in social media using topic models. *PLoS One*, 9(8):e103408, 2014.

[39] Lukasz Olejnik, Agnieszka Kutrowska, and Claude Castelluccia. The beginning of genetic exhibitionism? In *Proceedings of the* 1^{st} *Workshop on Genome Privacy*, 2014.

[40] Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. Tweaking and tweeting: exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *Journal of Medical Internet Research*, 15(4):e62, 2013.

[41] Carl Lee Hanson, Ben Cannon, Scott Burton, and Christophe Giraud-Carrier. An exploration of social circles and prescription drug abuse through Twitter. *Journal of Medical Internet Research*, 15(9):e189, 2013.

[42] Jennifer C Duke, Heather Hansen, Annice E Kim, Laurel Curry, and Jane Allen. The use of social media by state tobacco control programs to promote smoking cessation: a cross-sectional study. *Journal of Medical Internet Research*, 16(7):e169, 2014.

[43] Nathan K Cobb, Megan A Jacobs, Jessie Saul, E Paul Wileyto, and Amanda L Graham. Diffusion of an evidence-based smoking cessation intervention through facebook: a randomised controlled trial study protocol. *BMJ Open*, 4(1):e004089, 2014.

[44] Devan Jaganath, Harkiran K Gill, Adam Carl Cohen, and Sean D Young. Harnessing Online Peer Education (HOPE): integrating C-POL and social media to train peer leaders in HIV prevention. *AIDS Care*, 24(5):593–600, 2012.

[45] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, 2011.

[46] Ruchit Nagar, Qingyu Yuan, Clark C Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and John S Brownstein. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research*, 16(10):e236, 2014.

[47] Anna C Nagel, Ming-Hsiang Tsou, Brian H Spitzberg, Li An, J Mark Gawron, Dipak K Gupta, Jiue-An Yang, Su Han, K Michael Peddecord, Suzanne Lindsay, et al. The complex relationship of realspace events and messages in cyberspace: Case study of influenza and pertussis using tweets. *Journal of Medical Internet Research*, 15(10):e237, 2013.

[48] Brenda L Curtis. Social networking and online recruiting for HIV research: ethical challenges. *Journal of Empirical Research on Human Research Ethics*, 9(1):58–70, 2014.

[49] David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

[50] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: an analysis of privacy leaks on Twitter. In *Proceedings of the* 10^{th} *Annual ACM Workshop on Privacy in the Electronic Society*, pages 1–12, 2011.

[51] Abhishek Gattani, Digvijay S Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. Entity extraction, linking, classification, and tagging for social media: a Wikipedia-based approach. *Proceedings of the VLDB Endowment*, 6(11):1126–1137, 2013.

[52] Shuang-Hong Yang, Alek Kolcz, Andy Schlaikjer, and Pankaj Gupta. Large-scale high-precision topic modeling on Twitter. In *Proceedings of the* 20^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1907–1916, 2014.

[53] Nilanjan Banerjee, Dipanjan Chakraborty, Koustuv Dasgupta, Sumit Mittal, Anupam Joshi, Seema Nagar, Angshu Rai, and Sameer Madan. User interests in social media sites: an exploration with micro-blogs. In *Proceedings of the* 18^{th} *ACM Conference on Information and Knowledge Management*, pages 1823–1826, 2009.

[54] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the* 14^{th} *Conference on Computational Natural Language Learning*, pages 107–116, 2010.

[55] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in Twitter to improve information filtering. In

Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 841–842, 2010.

[56] Nilanjan Banerjee, Dipanjan Chakraborty, Anupam Joshi, Sumit Mittal, Angshu Rai, and Balaraman Ravindran. Towards analyzing micro-blogs for detection and classification of real-time intentions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[57] Jordan Eschler, Zakariya Dehlawi, and Wanda Pratt. Self-characterized illness phase and information needs of participants in an online cancer forum. In *The International AAAI Conference on Web and Social Media*, pages 101–109, 2015.

[58] Funda Kivran-Swaine, Jeremy Ting, Jed R Brubaker, Rannie Teodoro, and Mor Naaman. Understanding loneliness in social awareness streams: Expressions and responses. In *The International AAAI Conference on Web and Social Media*, pages 256–265, 2014.

[59] Tawfiq Ammari, Meredith Ringel Morris, and Sarita Yardi Schoenebeck. Accessing social support and overcoming judgment on social media among parents of children with special needs. In *The International AAAI Conference on Web and Social Media*, pages 22–31, 2014.

[60] Tawfiq Ammari and Sarita Schoenebeck. Understanding and supporting fathers and fatherhood on social media sites. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1905–1914. ACM, 2015.

[61] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015.

[62] Acar Tamersoy, Munmun De Choudhury, and Duen Horng Chau. Characterizing smoking and drinking abstinence from social media. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 139–148. ACM, 2015.

[63] John A Naslund, Stuart W Grande, Kelly A Aschbrenner, and Glyn Elwyn. Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using YouTube. *PLoS One*, 10:e110171, 2014.

[64] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the* 23^{rd} *International Conference on Machine Learning*, pages 233–240, 2006.

[65] Daniel Cer, Marie-Catherine de Marneffe, Daniel Jurafsky, and Christopher D. Manning. Parsing to stanford dependencies: trade-offs between speed and accuracy. In *The* 7^{th} *International Conference on Language Resources and Evaluation (LREC 2010)*, 2010.

[66] Zhong-Yuan Zhang. Nonnegative matrix factorization: Models, algorithms and applications. In *Data Mining: Foundations and Intelligent Paradigms*, pages 99–134. Springer, 2012.

[67] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. Short and sparse text topic modeling via self-aggregation. In *International Joint Conferences on Artificial Intelligence*, pages 2270–2276, 2015.

[68] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1445–1456. ACM, 2013.

[69] Vivek Kumar Rangarajan Sridhar. Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 192–200, 2015.

[70] Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

[71] V Vydiswaran, Qiaozhu Mei, David A. Hanauer, and Kai Zheng. Mining consumer health vocabulary from community-generated text. In *Proceedings of the AMIA Annual Symposium*, page in press, 2014.

[72] Carol Magai, Nathan Consedine, Alfred I Neugut, et al. Common psychosocial factors underlying breast cancer screening and breast cancer treatment adherence: a conceptual review and synthesis. *Journal of Women's Health*, 16(1):11–23, 2007.

[73] Paul S Ciechanowski, Wayne J Katon, Joan E Russo, et al. The patient-provider relationship: attachment theory and adherence to treatment in diabetes. *American Journal of Psychiatry*, 158(1):29–35, 2001.

[74] Jeffrey S Gonzalez, Frank J Penedo, Michael H Antoni, et al. Social support, positive states of mind, and hiv treatment adherence in men and women living with HIV/AIDS. *Health Psychology*, 23(4):413, 2004.

[75] Caitlin C Murphy, L Kay Bartholomew, Melissa Y Carpentier, Shirley M Bluethmann, and Sally W Vernon. Adherence to adjuvant hormonal therapy among breast cancer survivors in clinical practice: a systematic review. *Breast Cancer Research and Treatment*, 134(2):459–478, 2012.

[76] 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.

[77] 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.

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

[79] Sumita S Bhatta, Ningqi Hou, Zakiya N Moton, Blase N Polite, Gini F Fleming, Olufunmilayo I Olopade, Dezheng Huo, and Susan Hong. Factors associated with compliance to adjuvant hormone therapy in black and white women with breast cancer. *SpringerPlus*, 2(1):1, 2013.

[80] P Wuensch, A Hahne, R Haidinger, K Meißler, B Tenter, C Stoll, B Senf, and J Huebner. Discontinuation and non-adherence to endocrine therapy in breast cancer patients: is lack of communication the decisive factor? *Journal of Cancer Research and Clinical Oncology*, 141(1):55–60, 2015.

[81] Louise L Beryl, Katharine AS Rendle, Meghan C Halley, Katherine A Gillespie, Suepattra G May, Jennifer Glover, Peter Yu, Runi Chattopadhyay, and Dominick L Frosch. Mapping the decision-making process for adjuvant endocrine therapy for breast cancer the role of decisional resolve. *Medical Decision Making*, page 0272989X16640488, 2016.

[82] Annette Wigertz, Johan Ahlgren, Marit Holmqvist, Tommy Fornander, Jan Adolfsson, Henrik Lindman, Leif Bergkvist, and Mats Lambe. Adherence and discontinuation of adjuvant hormonal therapy in breast cancer patients: a population-based study. *Breast Cancer Research and Treatment*, 133(1):367–373, 2012.

[83] B Makubate, PT Donnan, JA Dewar, AM Thompson, and C McCowan. Cohort study of adherence to adjuvant endocrine therapy, breast cancer recurrence and mortality. *British Journal of Cancer*, 108(7):1515–1524, 2013.

[84] Claudia Brito, Margareth Crisóstomo Portela, and Mauricio Teixeira Leite de Vas-

concellos. Adherence to hormone therapy among women with breast cancer. *BMC Cancer*, 14(1):1, 2014.

[85] Jasmin A Tiro, Joanne M Sanders, L Aubree Shay, Caitlin C Murphy, Heidi A Hamann, L Kay Bartholomew, Lara S Savas, and Sally W Vernon. Validation of self-reported post-treatment mammography surveillance among breast cancer survivors by electronic medical record extraction method. *Breast Cancer Research and Treatment*, 151(2):427–434, 2015.

[86] Ethan Basch, Bryce B Reeve, Sandra A Mitchell, , et al. Development of the national cancer institutes patient-reported outcomes version of the common terminology criteria for adverse events (pro-ctcae). *Journal of the National Cancer Institute*, 106(9):244, 2014.

[87] Jun Wu and Z Kevin Lu. Hormone therapy adherence and costs in women with breast cancer. *The American Journal of Pharmacy Benefits*, 5(2):65–70, 2013.

[88] Alfred I Neugut, Xiaobo Zhong, Jason D Wright, Melissa Accordino, Jingyan Yang, and Dawn L Hershman. Nonadherence to medications for chronic conditions and nonadherence to adjuvant hormonal therapy in women with breast cancer. *JAMA Oncology*, 2016.

[89] Nina Schmidt, Karel Kostev, Achim Jockwig, Iannis Kyvernitakis, Ute-Susann Albert, and Peyman Hadji. Treatment persistence evaluation of tamoxifen and aromatase inhibitors in breast cancer patients in early and late stage disease. *International Journal of Clinical Pharmacology and Therapeutics*, 52(11):933–939, 2014.

[90] Annette L Stanton, Keith J Petrie, and Ann H Partridge. Contributors to nonadherence and nonpersistence with endocrine therapy in breast cancer survivors recruited from an online research registry. *Breast Cancer Research and Treatment*, 145(2):525–534, 2014.

[91] Hayley E Walker, Shoshana M Rosenberg, Annette L Stanton, Keith J Petrie, and Ann H Partridge. Perceptions, attributions, and emotions toward endocrine therapy in young women with breast cancer. *Journal of Adolescent and Young Adult Oncology*, 5(1):16–23, 2016.

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

[93] Gi Woong Yun, Morin David, Sanghee Park, Claire Youngnyo Joa, Brett Labbe, Jongsoo Lim, Sooyoung Lee, and Daewon Hyun. Social media and flu: Media twitter accounts as agenda setters. *International Journal of Medical Informatics*, 91:67–73, 2016.

[94] Noémie Elhadad, Shaodian Zhang, Patricia Driscoll, and Samuel Brody. Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *Proc AMIA Annual Fall Symposium*, 2014.

[95] 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.

[96] 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.

[97] Shaodian Zhang, Erin O'Carroll Bantum, Jason Owen, Suzanne Bakken, and Noémie Elhadad. Online cancer communities as informatics intervention for social support: conceptualization, characterization, and impact. *Journal of the American Medical Informatics Association*, page ocw093, 2016.

[98] Sarah A Marshall, Christopher C Yang, Qing Ping, Mengnan Zhao, Nancy E Avis, and Edward H Ip. Symptom clusters in women with breast cancer: an analysis of

data from social media and a research study. *Quality of Life Research*, 25(3):547–557, 2016.

[99] Deanna J Attai, Michael S Cowher, Mohammed Al-Hamadani, Jody M Schoger, Alicia C Staley, and Jeffrey Landercasper. Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey. *Journal of Medical Internet Research*, 17(7), 2015.

[100] Kenneth Portier, Greta E Greer, Lior Rokach, Nir Ofek, Yafei Wang, Prakhar Biyani, Mo Yu, Siddhartha Banerjee, Kang Zhao, Prasenjit Mitra, et al. Understanding topics and sentiment in an online cancer survivor community. *JNCI Monographs*, 47:195–198, 2013.

[101] David C Mohr, Michelle Nicole Burns, Stephen M Schueller, Gregory Clarke, and Michael Klinkman. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. *General Hospital Psychiatry*, 35(4):332–338, 2013.

[102] Keith J Horvath, J Michael Oakes, BR Simon Rosser, Gene Danilenko, Heather Vezina, K Rivet Amico, Mark L Williams, and Jane Simoni. Feasibility, acceptability and preliminary efficacy of an online peer-to-peer social support art adherence intervention. *AIDS and Behavior*, 17(6):2031–2044, 2013.

[103] 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 Research and Treatment*, 158(2):395–405, 2016.

[104] Jun J Mao, Annie Chung, Adrian Benton, Shawndra Hill, Lyle Ungar, Charles E Leonard, Sean Hennessy, and John H Holmes. Online discussion of drug side effects

and discontinuation among breast cancer survivors. *Pharmacoepidemiology and drug safety*, 22(3):256–262, 2013.

[105] Salma Begum and Ramazan S Aygun. Greedy hierarchical binary classifiers for multi-class classification of biological data. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 3(1):1–15, 2014.

[106] Dewan Md Farid, Li Zhang, Chowdhury Mofizur Rahman, M Alamgir Hossain, and Rebecca Strachan. Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4):1937–1946, 2014.

[107] T Mikolov and J Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013.

[108] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[109] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[110] Sarath Chirayil Subhash. *Personality Analysing on Watson Cloud by tracking the digital footprints of the user*. PhD thesis, Dublin, National College of Ireland, 2015.

[111] Mohamed Mostafa, Tom Crick, Ana C Calderon, et al. Incorporating emotion and personality-based analysis in user-centered modelling. In *Research and Development in Intelligent Systems XXXIII: Incorporating Applications and Innovations in Intelligent Systems XXIV*, pages 383–389. Springer, 2016.

[112] Ferdinand Thies, Michael Wessel, Jan Rudolph, et al. Personality matters: How signaling personality traits can influence the adoption and diffusion of crowdfunding campaigns. Technical report, Darmstadt Technical University, Department of Business Administration, Economics and Law, Institute for Business Studies, 2016.

[113] Oliver P John, Laura P Naumann, and Christopher J Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3:114–158, 2008.

[114] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, et al. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110. ACM, 2016.

[115] Marianne Amir and Alona Ramati. Post-traumatic symptoms, emotional distress and quality of life in long-term survivors of breast cancer: a preliminary research. *Journal of Anxiety Disorders*, 16(2):191–206, 2002.

[116] Alice B Kornblith, James E Herndon, Raymond B Weiss, et al. Long-term adjustment of survivors of early-stage breast carcinoma, 20 years after adjuvant chemotherapy. *Cancer*, 98(4):679–689, 2003.

[117] Nicole L Cohen, Erin C Ross, R Michael Bagby, Peter Farvolden, and Sidney H Kennedy. The 5-factor model of personality and antidepressant medication compliance. *The Canadian Journal of Psychiatry*, 49(2):106–113, 2004.

[118] Kerry S Courneya, Christine M Friedenreich, Rami A Sela, et al. Correlates of adherence and contamination in a randomized controlled trial of exercise in cancer survivors: an application of the theory of planned behavior and the five factor model of personality. *Annals of Behavioral Medicine*, 24(4):257–268, 2002.

[119] Malin Axelsson, Eva Brink, Jesper Lundgren, and Jan Lötvall. The influence of personality traits on reported adherence to medication in individuals with chronic disease: an epidemiological study in west sweden. *PLoS One*, 6(3):e18241, 2011.

[120] Jared M Bruce, Laura M Hancock, Peter Arnett, et al. Treatment adherence in multiple sclerosis: association with emotional status, personality, and cognition. *Journal of behavioral medicine*, 33(3):219–227, 2010.

[121] Malin Axelsson, Christina Cliffordson, Bo Lundback, et al. The function of medication beliefs as mediators between personality traits and adherence behavior in people with asthma. *Patient Prefer Adherence*, 7:1101–1109, 2013.

[122] George S Alexopoulos, Jo Anne Sirey, Samprit Banerjee, et al. Two behavioral interventions for patients with major depression and severe copd. *The American Journal of Geriatric Psychiatry*, 24(11):964–974, 2016.

[123] Adewale B Ganiyu, Langalibalele H Mabuza, Nomsa H Malete, et al. Non-adherence to diet and exercise recommendations amongst patients with type 2 diabetes mellitus attending extension ii clinic in botswana. *African journal of primary health care & family medicine*, 5(1), 2013.

[124] Mark Dredze, Renyuan Cheng, Michael J Paul, et al. Healthtweets. org: a platform for public health surveillance using twitter. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*, pages 593–596, 2014.

[125] Zhijun Yin, Bradley Malin, Jeremy Warner, Pei-Yun Hsueh, and Ching-Hua Chen. The power of the patient voice: Learning indicators of treatment adherence from an online breast cancer forum. In *Eleventh International AAAI Conference on Web and Social Media*, pages 337–346, 2017.

[126] Kang Zhao, John Yen, Greta Greer, Baojun Qiu, et al. Finding influential users of

online health communities: a new metric based on sentiment influence. *Journal of the American Medical Informatics Association*, 21(e2):e212–e218, 2014.

[127] Zhijun Yin, Lijun Song, and Bradley Malin. Reciprocity and its association with treatment adherence in an online breast cancer forum. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 618–623, 2017.

[128] Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, et al. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212, 2015.

[129] Kathleen I Pritchard, Daniel F Hayes, 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.

[130] Lina Sulieman, Daniel Fabbri, Fei Wang, Jianying Hu, and Bradley A Malin. Predicting negative events: Using post-discharge data to detect high-risk patients. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1169. American Medical Informatics Association, 2016.

[131] Stephanie L Shimada, Beth Ann Petrakis, James A Rothendler, Maryan Zirkle, et al. An analysis of patient-provider secure messaging at two veterans health administration medical centers: message content and resolution through secure messaging. *Journal of the American Medical Informatics Association*, page ocx021, 2017.

[132] Lynne T Harris, Sebastien J Haneuse, Diane P Martin, and James D Ralston. Diabetes quality of care and outpatient utilization associated with electronic patient-provider messaging: a cross-sectional analysis. *Diabetes Care*, 32(7):1182–1187, 2009.

[133] Lynne T. Harris, Thomas D. Koepsell, Sebastien J. Haneuse, Diane P. Martin, et al. Glycemic control associated with secure patient-provider messaging within a shared electronic medical record. *Diabetes Care*, 36(9):2726–2733, apr 2013.

[134] YY Zhou, T Garrido, HL Chin, AM Wiesenthal, et al. Patient access to an electronic health record with secure messaging: impact on primary care utilization. *The American Journal of Managed Care*, 13(7):e418–e424, 2007.

[135] Stephanie L Shimada, Timothy P Hogan, Sowmya R Rao, Jeroan J Allison, et al. Patient-provider secure messaging in va: variations in adoption and association with urgent care utilization. *Medical Care*, 51:S21–S28, 2013.

[136] Robert M Cronin, Daniel Fabbri, Joshua C Denny, S Trent Rosenbloom, et al. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *International Journal of Medical Informatics*, 105:110–120, 2017.

[137] Lina Sulieman, David Gilmore, Christi French, Robert M Cronin, et al. Classifying patient portal messages using convolutional neural networks. *Journal of Biomedical Informatics*, 74:59–70, 2017.

[138] Morgan Harrell, Daniel Fabbri, and Mia Levy. Analysis of adjuvant endocrine therapy in practice from electronic health record data of patients with breast cancer. *JCO Clinical Cancer Informatics*, 1:1–8, 2017.

[139] 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. *The Lancet*, 381(9869):805–816, 2013.

[140] Bradley Carron-Arthur, John A Cunningham, and Kathleen M Griffiths. Describing the distribution of engagement in an internet support group by post frequency: A comparison of the 90-9-1 principle and zipf's law. *Internet Interventions*, 1(4):165–168, 2014.

[141] Breastcancer.org. Hormonal therapy side effects comparison chart. http://www.

breastcancer.org/treatment/hormonal/comp_chart, 2018. [Online; accessed Jan. 21, 2018].

[142] Wei He, Fang Fang, Catherine Varnum, et al. Predictors of discontinuation of adjuvant hormone therapy in patients with breast cancer. *Journal of Clinical Oncology*, 33(20):2262–2269, 2015.

[143] Antonis Valachis and Cecilia Nilsson. Cardiac risk in the treatment of breast cancer: assessment and management. *Breast Cancer: Targets and Therapy*, 7:21, 2015.

[144] Dawn L Hershman, Lawrence H Kushi, Theresa Shao, et al. Early discontinuation and nonadherence to adjuvant hormonal therapy in a cohort of 8,769 early-stage breast cancer patients. *Journal of Clinical Oncology*, 28(27):4120–4128, 2010.

[145] Kathleen Ell, Betsy Vourlekis, Bin Xie, Frances R Nedjat-Haiem, et al. Cancer treatment adherence among low-income women with breast or gynecologic cancer. *Cancer*, 115(19):4606–4615, 2009.

[146] Zoe Moon, Rona Moss-Morris, Myra S Hunter, et al. Barriers and facilitators of adjuvant hormone therapy adherence and persistence in women with breast cancer: a systematic review. *Patient Prefer Adherence*, 11:305, 2017.

[147] MP Davey. Oral therapy: managing side effects can aid adherence. *Oncol Nurse Advis*, pages 24–31, 2012.

[148] Carlos Castillo, Fernando Diaz, Emre Kıcıman, and Alexandra Olteanu. A critical review of online social data: Limitations, ethical challenges, and current solutions. at the 10th International AAAI Conference on Web and Social Media (ICWSM16), 2016. Organizers appear in alphabetical order.

[149] Leanne Townsend and Claire Wallace. Social media research: A guide to ethics. *University of Aberdeen*, 2016.