Patient Messaging Content Associated with Initiating Hormonal Therapy after a Breast Cancer Diagnosis

Zhijun Yin, PhD, Jeremy L. Warner, MD, MS, Qingxia Chen, PhD, Bradley A. Malin, PhD
Vanderbilt University, Nashville, TN

Abstract

Hormonal therapy is an effective, but challenging, long-term treatment for patients with hormone-receptor-positive breast cancer. Raising the rate of patients who initiate therapy may be possible by characterizing the factors that influence a patient's decision. We hypothesized that online patient portal messages convey such factors. To investigate this hypothesis, we focused on breast cancer patients who were prescribed hormonal therapy at Vanderbilt University Medical Center and sent messages through the portal between diagnosis and therapy initiation. We first conducted a topic modeling analysis to generate the main themes of portal messages. We subsequently applied survival analysis to learn the association between the factors conveyed in messages, in term of semantic word groups, and the time elapsed from diagnosis to therapy initiation. We found that consulting with healthcare providers increased the probability of therapy initiation, while mentions of symptoms or negative emotions exhibited a reduced probability.

Introduction

Approximately 12% of U.S. women will develop breast cancer over their lifetime[1]. It is the second most diagnosed cancer (after skin) and has the second highest cancer death rate (after lung) for U.S. women. It is estimated that there will be 268,600 new breast cancer cases in 2019 in the U.S. and over 41,000 women will die from the disease. Hormone-receptor-positive breast cancer makes up 80% of diagnosed cases. In hormone-receptor-positive breast cancer, the cancer cells grow and spread with the assistance of hormones (e.g., estrogen) in the blood. Hormonal therapy, which works by preventing estrogen from stimulating breast cancer cell growth, is an adjuvant (post-surgical) treatment for patients with this type of breast cancer[2]. Evidence suggests that taking hormonal therapy medications, such as tamoxifen, can reduce cancer mortality by one third[3]. As such, it is often recommended that patients take hormonal therapy medications for at least five years[2].

Adhering to hormonal therapy is not easy for a breast cancer patient. It is reported that nearly 50% of breast cancer patients prescribed hormonal therapy dropped off a regimen before completing a five-year treatment course[4]. There are many factors that may contribute to medication discontinuation behavior. For example, side effects (e.g., depression) can lead to medication discontinuation[1,5]. As a result, various studies have focused on learning the factors behind why breast cancer patients choose to stop taking a hormonal therapy medication. These studies can be roughly categorized into three classes based on the data that they investigate: 1) interview or survey[6], 2) structured electronic medical records (EMRs)[7], and 3) user generated content (UGC) in online environments[1,5,8,9]. The first two classes hold merit, but have notable limitations. Generally, studies based on interviews are often time consuming and not scalable to large study cohorts, while survey-based studies are often confined to the pre-defined questionnaires[1]. Studies based on structured EMRs are, on the other hand, limited in that they lack description of treatment experience (e.g., patient's feelings and emotions). By contrast, UGC has been shown to be an effective resource to learn about a patient's health related behaviors. For example, studies have shown that the messages that patients send to healthcare providers in a patient portal were indicative of the likelihood of discontinuing hormonal therapy medication[9].

However, few studies have focused on what factors affect the time at which a breast cancer patient initiates hormonal therapy, relative to their diagnosis. This information is important because it can provide insights into why a patient delays making a decision to start the therapy. While several studies investigated patient decision making, most relied on interviews or qualitative methods[10,11]. For example, Beryl et al. conducted a longitudinal series of interviews to identify the decision-making process of hormonal therapy. They found that most patients starting a therapy is not a single decision, but rather is a series of decisions[6]. More generally, Marla et al. pointed out that shared decision making needs to center on the person rather than the medical encounter[12], suggesting the importance of listening to the patient. Thus, in this study, we focused on the secure messages sent by patients to their healthcare providers, one particular type of UGC[13], in an online patient portal. There are several clinical factors that are likely to affect the decision to start hormonal therapy; e.g., undergoing additional surgery or an unplanned stay in the hospital. We hypothesized that the messages patients convey through online portals contain factors associated with the time from breast cancer diagnosis to hormonal therapy initiation. To investigate this hypothesis, we focused on the EMRs and portal

communications sent by breast cancer patients prescribed hormonal therapy at Vanderbilt University Medical Center (VUMC). Particularly, we studied patients who sent messages after their diagnosis date, but before taking a hormonal therapy medication. We applied topic modeling to infer the main themes that were discussed in these messages and performed a survival analysis to study the extent to which the themes were associated with the time that breast cancer patients started their treatment.

Methods

Data

This study used de-identified data from the VUMC EMR system[14] and was approved by the Vanderbilt University Institutional Review Board. In this setting, all patient identities were replaced with persistent pseudonyms by a third-party honest broker and all dates within a patient's records were consistently shifted by a random number of days that were uniformly sampled from the (-365, -1) range. We focused on the patients who were diagnosed with stage I to III breast cancer and were prescribed any of the following hormonal therapy medications: *anastrozole*, *exemestane*, *letrozole* (aromatase inhibitors; AIs), or *raloxifene*, *tamoxifen* (selective estrogen receptor modulators; SERMs)[9]. We restricted the cohort to those who sent messages between their diagnosis date and the first documentation of a hormonal therapy medication.

Topic Modeling and Word Clustering

Natural language clinical text has high dimensionality, but is also quite sparse. Thus, we to better summarize the content for inference purposes we reduced the dimensionality. There are two general types of methods to realize this goal, both of which we adopted for this project, as they serve different purposes: 1) topic modeling to identify the main messaging themes, 2) word clustering, based on a lower dimensional representation of words (e.g., word embedding terms of word2vec[15]), to create message content predictors for inference.

Topic modeling is a statistical approach for discovering latent topics in a collection of documents or messages. For example, Latent Dirichlet Allocation (LDA) is a generative statistical model that assumes each document can be represented by a small number of topics, where each word in the document can then be generated by one of these topics[16]. The inferred topics can be interpreted by their most relevant terms. Topic modeling is notable because it allows terms to be used in multiple topics and it has often been observed that it groups terms that are similar in their global context. In this study, we applied LDA, as implemented in *Mallet* Java package (version 2.0.8), to identify the main themes that were communicated in patient portal messages. Since LDA is an unsupervised technique, we adopted the coherence score to determine the optimal number of topics. The coherence score essentially measures the extent to which two high probability terms in a topic appear together in either external documents (e.g., in Wikipedia) or the modeling documents[17]. We selected the number of topics with the highest average coherence score across the proposed topics. This was accomplished by learning LDA models for 2 to 26 topics (with a step size of 1) over all of the messages. To mitigate word sparsity and ensure interpretability, we replaced each term with its lemma form and retained only nouns, verbs, adjectives and adverbs. We also generated bi-gram terms using the *genism* python package (version 3.6.0) to capture more meaningful phrases. We report on two aspects of this process in our experiments. First, we consider the most salient terms. *Saliency* of a term is a weighted term frequency that is introduced for better characterizing term importance in describing message topics[18]. The weight is defined as the sum of the Kullback-Leibler divergence between each marginal topic distribution and its conditional distribution on the given term. Second, we consider the topic distribution and sample relevant terms for each topic. The topic distribution is calculated after combining all of the messages into a single document. We display the most terms with the highest distribution in each topic.

Additionally, we applied word clustering to generate semantic word groups for inference purpose. Word clustering relies on a measure to calculate the similarity between two terms. For example, word2vec is a low dimensional representation technique that can be applied to measure semantic word similarity. It represents a group of shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words. The premise of word2vec is to apply each term to predict their adjacent neighbors, which represents the word similarity in a local context. If a standard clustering algorithm (e.g., *k*-means or hierarchical clustering) is applied, a term can only belong to a semantic word group. In this study, we observed that topics inferred from topic modeling techniques, in spite of good interpretability in summarizing the main themes, are poor in inference tasks. This might be due to the fact that the number of inferred topics is very small, which induces substantial information loss. However, increasing the number of topics would increase the possibility of conceptual overlap between topics, which reduces topic interpretability. Directly applying terms expressed in the messages as predictors is a simple method to avoid information loss. Yet, as

noted earlier, the high dimensionality and sparsity of natural language may not be beneficial for inference. For example, the number of unique terms can be larger than the number of messages. Furthermore, the correlations between terms will require additional regularization in the model, which makes inference more challenging.

In this respect, word clustering can reduce the correlation between term predictors and reduce the dimensionality, which leads to less information loss than topic modeling. To do so, we first trained a word2vec model based on all of the messages sent by breast cancer patients. We then ran an agglomerative hierarchical clustering algorithm with complete linkage, implemented in the *scikit-learn* python package (version 0.18.1) with 25 to 1000 groups (with a step size of 25). We applied the elbow rule to the standard deviation of the group size, such that we selected the number of groups at the angle where the marginal gain begins to diminish. This was done because, heuristically, a larger group is more likely to contain multiple meanings while a small group is more likely to have little contribution to dimensionality reduction.

Survival Analysis

To perform survival analysis, we investigate how the message content, measured by semantic word groups, is associated with the time from the diagnosis date until a breast cancer patient initiated a hormonal therapy medication.

Statistical Model. We applied a Cox proportional hazards regression model for the survival analysis. The Cox model is applied to investigate the effectiveness of independent predictors with respect to the time when an event of interest occurs. There are two benefits in applying a Cox model in this study. First, the Cox model is semi-parametric and does not assume any particular survival distribution. Second, the Cox model can make use of right-censored patients by incorporating both time (in terms of when an event happened or the latest time without event) and medication use status (e.g., patients who started taking hormonal therapy medications or did not by the end of follow-up) into the model. If the records showed that patients started taking hormonal therapy, then the event occurred. By contrast, if patients did not take any hormonal therapy medication by the end of data observation window, then right censoring occurred. The risk for an independent 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 bigger (smaller) than 1, then the variable is associated with an increased (a decreased) risk of taking a prescribed medication. We used the implementation of the Cox model from the *lifelines* python package (version 0.9.4) to conduct the survival analysis.

Variables. We defined an observation window of one and a half years from the diagnosis date. The dependent variable was binary, where 1 indicated that a breast cancer patient started taking the medication and 0 indicated that the patient is right censored. The time between the diagnosis date and the start of medication use (or the end of the observation window) was measured in 6-month blocks. The message content variables were constructed as follows: 1) for each patient, we aggregated all the messages sent between the diagnosis date and either the start of medication use or the end of the observation window. As such, we represented each patient as a document of messages; 2) we replaced the words in each message document with their corresponding semantic word group numbers; and 3) we calculated the term frequency – inverse document frequency (TF-IDF) values for each semantic word group in each document, which we subsequently applied as the values of semantic word group variables.

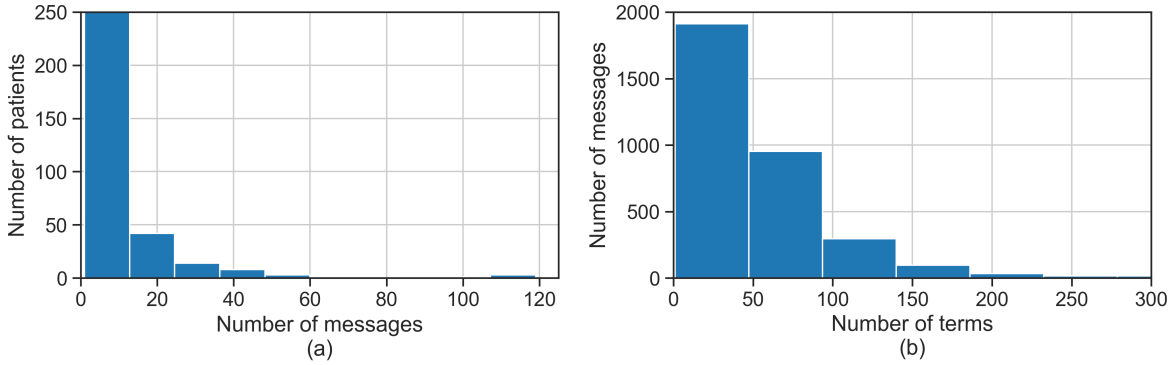

Figure 1. A summary of the patient messages in this study, shown as a histogram of the number of (a) messages sent per patient and (b) terms expressed per message.

Additionally, we introduced four independent variables: 1) the *age* at diagnosis, 2) *race* as documented in the EMR, 3) *cancer stage* at diagnosis, and 4) the *number* of Current Procedural Terminology (CPT) codes. We dichotomized

the race into Caucasian and non-Caucasion categories. We encoded advanced cancer stage (stage III) with 1 and early cancer stages (stage I and II) with 0. We included the number of CPT codes because, intuitively, the more procedures a patient is affiliated with, the greater the likelihood that they may delay the start of a hormonal therapy regimen. We report the statistically significant features at the 0.05 significance level.

We also applied the Kaplan–Meier (KM) estimator to estimate how the probability of starting hormonal therapy changes over time. In clinical studies, this estimator is often relied upon to measure the proportion of patients who live for a certain amount of time after treatment. We use the implementation of KM in the *lifeline* python package (version 0.9.4) to calculate and visualize the proportion of breast cancer patients who begin hormonal therapy along the treatment timeline.

Results

Cohort Summary

The study cohort consists of 336 breast cancer patients prescribed hormonal therapy medications between the years 2005 to 2017, which consisted of 30% of all the breast cancer patients with early stages. The average patient age is 54.8 (with a range of [22, 88]). Based on race in the EMR, 90.1% of the cohort is Caucasian (303). With respect to the disease, 89.3% of the cohort were diagnosed with early stage breast cancer (300). The patients generated a total of 3,329 messages. Figure 1(a) depicts the frequency distribution of the number of messages generated per patient. It can be seen that most patients posted a limited number of messages. For example, 71.7% of the patients sent less than 10 messages during this period. By contrast, only 11.7% of the patients sent at least 20 messages. Figure 1(b) summarizes the frequency distribution of the distinct terms per message. It can be seen that most of the messages contained only a small number of terms. For example, 58.9% of the messages contained no more than 50 terms. By contrast, only 12.0% of messages contained at least 100 terms.

Top Salient Terms

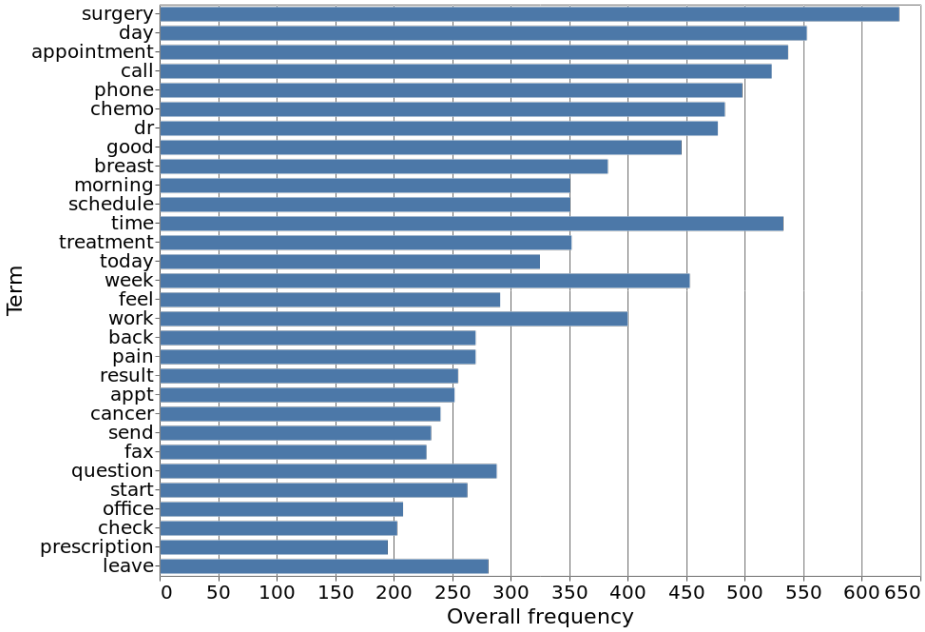

Figure 2. The 30 most salient terms in the patient portal messages. The terms are ordered in a descending rank according to their saliency.

Figure 2 demonstrates the 30 most salient terms (from top to bottom) in the portal messages. In most cases, the more frequent the word, the higher the saliency; however, there are some exceptions. For example, the terms *time*, *week*, *work*, *question*, *start* and *leave* have relatively higher frequency but lower saliency, suggesting saliency might be more effective in describing the importance of a term. Figure 2 further illustrates that *surgery*, *appointment* and *chemo* are among the most salient words, while other terms such as *back pain*, *result*, and *prescription* have relatively lower saliency.

Message Topics

Table 1. The most relevant terms for each of the six inferred topics. The topic distribution is shown in the rightmost column. The terms are ordered based on their probabilities in each topic in a descending manner.

#Topic	Top Most Relevant Terms	Dist.
1	feel, pain, leave, night, med, bad, arm, area, lot, continue, normal, sore, sleep, issue, friday, skin, nausea, red, side, fever, hand, infection, hurt, cough, run, swell, taxol, eat, head, chest	17.46%
2	chemo, breast, time, treatment, week, cancer, start, find, biopsy, infusion, report, year, ve, thing, center, put, chemotherapy, month, recommend, give, begin, mention, read, show, yesterday, drug, post, question, end	16.87%
3	surgery, appointment, dr, schedule, question, week, time, follow, dateth, radiation, wait, change, visit, reconstruction, plan, long, meet, remove, understand, discuss, make, mastectomy, set, decide, procedure, cancel, thought, lumpectomy, reschedule, lymph_node	16.72%
4	day, good, morning, back, result, test, blood, hope, great, lab, hour, concern, pm, mg, advise, tomorrow, problem, stop, foot, drain, forget, dear, low, tablet, blood_pressure, daily, weekend, clinic, guess, level	16.44%
5	phone, work, send, fax, office, prescription, receive, doctor, place, number, pharmacy, information, refill, email, institution, form, return, fmla, order, request, fill, pick, insurance, leave, medicine, hospital, medication, medical, update, letter	16.33%
6	call, today, appt, make, check, give, yesterday, home, message, thursday, contact, work, port, friday, talk, monday, tuesday, speak, regard, cell, care, place, nurse, husband, hear, tomorrow, vanderbilt, mom, afternoon, move	16.18%

Table 1 shows the six topics that were learned from the messages. Here, we take a moment to provide some intuition into the topics. First, the most distributed (or mentioned) topic (#1) is about *symptoms* (17.46%). Second, the next most distributed topic (#2) is about *chemotherapy* (16.87%). Third, the next most distributed topic (#3) is about *surgery*. Fourth, the next most distributed topic (#4) is about *laboratory tests*. Finally, and the last two topics (#5, 16.33% and #6, 16.18%) are about *prescriptions* and *communications*, respectively. Note that the topic distribution is calculated after combining all messages into a single document.

Model Fitting

Our approach finally generated 150 semantic word groups, which served as message content predictors in the survival analysis. Our Cox model presented a concordance of 0.872, suggesting a good fit. With respect to the four control variables, it was found that the advanced cancer stage indicated a reduced probability of patients initiating hormonal therapy (HR = 0.851, P = 0.001). This suggests that patients with early stage cancer are more likely to begin hormonal therapy. It was also found that the number of CPT codes correlated with a decreased probability of starting hormonal therapy (HR = 0.983, P < 0.001). Neither the age at diagnosis nor the race were found to be statistically significant.

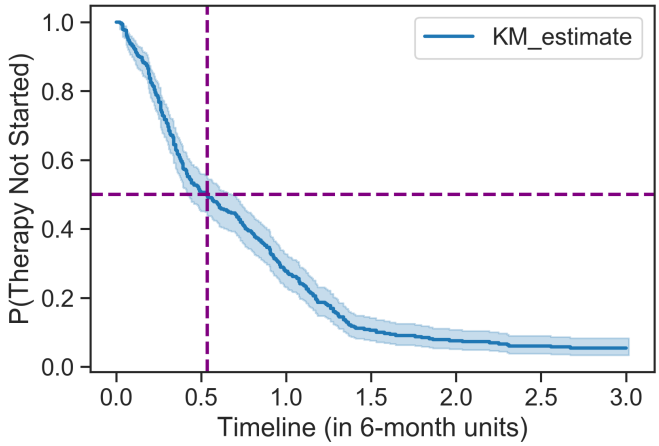

Figure 3. The KM curve for the probability that patients did not initiate hormonal therapy. The intersection of the dashed lines indicates that 50% of patients began hormonal therapy 3 months after diagnosis.

Figure 3 shows the KM curve for the probability of not starting hormonal therapy as a function of time. The horizontal dashed line indicates where the probability is 0.5. The dashed lines illustrate that 50% of the patients initiated hormonal therapy at 3.24 months (0.54 × 6-month unit). It can further be seen that approximately 80% of breast cancer patients began hormonal therapy within one and a half years of the diagnosis.

Semantic Word Groups

Table 2. The semantic word groups that are positively associated with hormonal therapy initiation. The terms are ordered based on their similarity with the center of the group in a descending manner.

Group #	Word Samples	HR	95% CI	p
12	understand, think, suggest, say, recall, agree, remember, know, decide, realize, knew, consider, assume, recommend, believe, guess, choose, forget, thought, determine	1.964	(1.537, 2.510)	<.001
40	purple, yellowish, bumpy, whelps, raw, thick, electrodes, phlegm, puffy, colored, raised, weeping, scalp, scabs, mucous, corners, runny, peeling, pulsing, bruised	1.928	(1.542, 2.410)	<.001
131	upcoming, reconstructive, cataract, pre, post, rescheduling, prior, canceled, scheduling	1.764	(1.444, 2.154)	<.001
59	later, earlier, sooner, maybe, possibly	1.599	(1.328, 1.924)	<.001
106	sugar, glucose, sugars, systolic, creatinine, pcv, ferritin, count, counts, soaked, platelets, consistently, pressures, diastolic, rate, heartrate, dried, values, 101, 99	1.457	(1.179, 1.800)	<.001
24	ideas, insights, suggestions, cancellations, openings, word, idea, cure, conclusions	1.408	(1.162, 1.707)	<.001
39	meet, speak, confer, consult, communicate, talk, discuss, share, speaking, consulting	1.405	(1.152, 1.712)	0.001
100	baptist, gateway, institution, centennial, vandy, summit, vu, oho, premier, dayani, vumc, regional, umc, vanderbilt, ctr, center, main, campus, location, l00	1.391	(1.126, 1.718)	0.002
37	inconvenience, inconvience, confusion, missed, notice, missing	1.379	(1.100, 1.730)	0.005
99	confirm, clarify, verify, remind, inform, explain, notify, cancel, reschedule, tell, ask, make, schedule, change, mention, postpone, switch, book, miss, attend	1.370	(1.131, 1.659)	0.001
79	throwing, backed, picking, threw, stirred, screwed, messed, flared, built, lit, clears, picks, woken, speed, build, lined, chalked, clearing, messes, stood	1.360	(1.109, 1.668)	0.003
120	papers, documents, paperwork, forms, form, paper, release, ticket, parking, intermittent	1.336	(1.079, 1.654)	0.008
32	hcl, 1mg, 100mg, 50mg, 500mg, 150mg, 40mg, 25mg, 200mg, 4mg, 2mg, 60mg, amlodipine, 75mg, qd, 80mg, bid, tid, 30mg, glipizide	1.326	(1.072, 1.639)	0.009
21	important, beneficial, safe, alright, confusing, acceptable, frustrating, true, effective	1.306	(1.065, 1.602)	0.010
27	discontinued, steriods, cipro, flagyl, decadron, stopped, initially, previously, finished	1.297	(1.050, 1.602)	0.016
135	whatsoever, injuries, masses, avail, worries, evidence, friction, complaints, apparent, ink, hurry, regurgitation, signs, luck, improvement, complications, success, rush, longer	1.279	(1.062, 1.540)	0.009
3	sunday, saturday, evening, tonight, morning, afternoon, today, yesterday, fell, fall	1.243	(1.014, 1.523)	0.036
138	avastin, initiation, recon, inhibitors, cytoxan, goserelin, radation, laparoscopic, stereotactic, tc, minimally, taxotere, shrink, cosmetic, perjeta, subsequent, 2mm, invasive, carpal, navelbine	1.238	(1.026, 1.495)	0.026
57	sleeve, compression, lymphedema, glove, garment, gauntlet, garments, massage, lymphadema, bra	1.232	(1.016, 1.494)	0.033
140	hospital, room, rehab, clinic, emergency, er	1.224	(1.009, 1.486)	0.040
36	handle, figure, imagine, find, tolerate, manage, afford, pass, watch, run	1.215	(1.012, 1.458)	0.036

Table 2 shows the semantic word groups that are positively associated with the start of hormonal therapy. Based on this result, there appear to be four types:

- Mentions of semantic word groups related to cognitive processes (#12, HR = 1.964, P < 0.001), suggestion-related nouns (#24, HR1.408, P<0.001), and communication or consulting related verbs (#39, HR = 1.405 P = 0.001).
- Mentions of semantic word groups related to schedules (#131, HR = 1.764, P < 0.001; #99, HR = 1.370, P = 0.001; #37, HR=1.379, P=0.005), locations (#100, HR = 1.391, P = 0.002; #140, HR = 1.224, P = 0.040), and paperwork (#120, HR=1.336, P=0.008).
- Mentions of semantic word groups related to symptoms (#40, HR = 1.928, P < 0.001), vital and test related words (#106, HR=1.457, P<0.001), medication dosage (#32, HR = 1.326, P = 0.009), discontinuation of certain medications (#27, HR=1.297, P=0.016), medications for chemotherapy and radiation (#138, HR = 1.238, P=0.026), and verbs related to manage health conditions (#36, HR=1.215, P = 0.036).
- Other semantic word groups, such as time related words (#59, HR = 1.599, P < 0.001; #3, HR = 1.243, P = 0.026), and adjectives (#21, HR = 1.306, P = 0.010).

Table 3. The semantic word groups that are negatively associated with hormonal therapy initiation. The terms are ordered based on their similarity with the center of the group in a descending order.

Group #	Word Samples	HR	95% CI	p
65	watery, soft, irritable, excessive, sweating, caffeine, loose, feverish, wheezing, breaths, achy, shaking, soaking, stools, dry, fullness, severely, sensitivity, heartbeat	0.596	(0.466, 0.761)	<.001
82	difficult, ill, hard, sick, quickly, easy, busy, long, much, rough	0.631	(0.516, 0.771)	<.001
26	referal, referral, favor, personally, explaining, answered	0.635	(0.528, 0.763)	<.001
149	blood, bood, lab, hydrocephalus, dental, excuse	0.645	(0.502, 0.830)	0.001
33	group, hospice, private, provider, services, program, ymca, coordinator, facility, assisted, department, living, abc, critical, tenn, tri, university, lives, fitness	0.647	(0.504, 0.831)	0.001
35	abundance, estimate, invoice, opthamologist, oversight, official, integral, itemized, eco, exception, appoinment, appeal, indigent, mir, extension, aflac, expert, eob, endo, error	0.652	(0.533, 0.798)	<.001
98	cervical, grafting, sparing, areola, dissection, silicone, radiated, lateral, prolapse, scar, tissue, lumbar, ovarian, surrounding, lat, reduction, implant, partial, uterine, expanders	0.659	(0.537, 0.810)	<.001
10	mile, bed, class, miles, chair, hrs, hour, car, couch, hours	0.668	(0.540, 0.827)	<.001
17	dexa, cat, muga, ct, pet, echo, ekg, xray, density, emg, marrow, pap, dye, bone, contrast, echocardiogram, smear, thyrogen, tail, hida	0.696	(0.560, 0.866)	0.001
29	thigh, soreness, calf, upper, elbow, numbness, tingling, tenderness, sensation, stiffness, wrist, forearm, abdomen, ankle, weakness, neck, thumb, tightness, shoulder	0.721	(0.581, 0.895)	0.003
123	reconstruction, mastectomy, lumpectomy, diep, flap, bilateral, procedure, hysterectomy, surgery, surgical, surgeries, mammograms, ovaries, expansion, mris, port, operation, uterus, final, cath	0.73	(0.595, 0.894)	0.002
126	phone, telephone, number, cell, address	0.737	(0.606, 0.896)	0.002
81	wks, seconds, nights, days, weeks, rounds, months, sessions, occasions, years	0.748	(0.605, 0.925)	0.007
133	expected, involved, likely, present, normal, open, available	0.753	(0.615, 0.921)	0.006
103	2nd, 1st, 3rd, third, second, preop, tentatively	0.755	(0.620, 0.918)	0.005
20	gyn, urologist, cardiologist, neurologist, dermatologist, gynecologist, oncologist, specialist, ob, neuro, obgyn, onc, rheumatologist, surgeon, doc, internist, psychiatrist	0.769	(0.628, 0.943)	0.012
89	affects, effect, effects, right, left	0.779	(0.622, 0.976)	0.030
92	single, ie, multiple, plus, various, including, entire, numerous, whole, throughout	0.788	(0.653, 0.950)	0.013
41	contacting, emailing, asking, writing, calling, hearing, posted, waiting, hear, heard	0.796	(0.653, 0.971)	0.024
130	bruises, fingernails, cases, parts, instances, areas, flagged, toenails, spots, organs, bumps, lumps, blockages, sites, separated, cancers, places, layers, factors, situations	0.797	(0.641, 0.992)	0.042
60	need, want, needed, needs, necessary, wants, possible, plan, decided	0.803	(0.646, 0.998)	0.048

52	mood, depression, intestinal, appetite, vision, balance, migraines, activity, anxiety, inflammation, ibs, neuropathy, overall, swallowing, activities, peripheral, intake, hair	0.811	(0.668, 0.983)	0.033
61	leptomeningeal, receptors, enhanced, hardening, scientific, atherosclerosis, fdg, distant, cardiovascular, lobular, metastatic, enhancing, degenerative, foci,	0.820	(0.677, 0.993)	0.042

Table 3 shows the semantic word groups that are negatively associated with the start of hormonal therapy. Based on this result, there appear to be three types:

- Mentions of symptoms (#65, HR = 0.596, P < 0.001; #82, HR = 0.631, P < 0.001; #29, HR = 0.721, P = 0.003; #149, HR=0.645, P=0.001), different physician roles (#20, HR = 0.769, P = 0.012), reconstruction and other surgery (#123, HR=0.730, P=0.002), body components (#130, HR = 0.797, P = 0.042), negative emotions (#52, HR=0.811, P=0.033), x-ray related examinations (#17, HR = 0.696, P = 0.001).
- Mentions of fitness service (#33, HR = 0.647, P = 0.001), distances and hours (#10, HR=0.668, P<0.001; #81, HR=0.748, P=0.007) have decreased probability of taking hormonal therapy.
- Mentions of communication related terms (#126, HR = 0.737, P = 0.002; #41, HR = 0.796, P = 0.024).

Discussion and Conclusion

Primary Findings

This study investigated the messages sent by patients through the VUMC patient portal after their breast cancer diagnosis and before the start of hormonal therapy. We characterized how certain factors mentioned by the patients correlated with when they initiated a regimen of hormonal therapy medications. There were several notable findings. First, our exploratory analysis showed most patients sent only a limited number of messages. This is a similar observation to those made about individuals who published posts in online health communities. Second, our topic analysis suggested that patients were primarily communicating about their symptoms, chemotherapy, surgery, laboratory tests and prescriptions. This has face validity because surgery and chemotherapy are two of the most common interventions that take place before hormonal therapy is prescribed. These two interventions often generate side effects that require additional medications to manage. Third, our survival analysis suggested that breast cancer patients with advanced cancer stages and a larger number of procedures (as documented by CPT codes) were associated with a decreased probability of hormonal therapy initiation. This confirms our expectation that CPT codes should be control variables in the survival analysis. The KM curve indicated that most breast cancer patients who initiated hormonal therapy did so within 6 months after their diagnosis.

There are several notable findings from a comparison of the semantic word groups that increased and decreased the probability of hormonal therapy medication initiation. It should be recognized that it is the factors that are mentioned in the messages (e.g., the act of taking surgery), instead of the mentions themselves, that are material to this investigation. First, we observed that patients who mentioned consulting-related words and chemotherapy were more likely to initiate hormonal therapy. This suggests that obtaining suggestions from healthcare providers may help realize a smooth transition from diagnosis to taking hormonal therapy. By contrast, patients who mentioned symptoms and surgeries were less likely to start hormonal therapy, suggesting that complex health conditions or significant procedures (e.g., surgery) may delay the start of taking hormonal therapy. In particular, women who must undergo repeated breast conserving surgeries due to positive margins, as well as those who develop post-surgical complications (e.g., infection), may be less likely to initiate hormonal therapy; this phenomenon has been studied in the context of adjuvant radiotherapy[19]. It should be noted that, while symptoms-related semantic groups were found to be both positively and negatively associated with the start of hormonal therapy, they are referring the different symptoms (e.g., #40 against #29). Second, it was observed that patients who mentioned negative emotions (e.g., depression and anxiety) were less likely to initiate the therapy. By contrast, patients who mentioned *afford* related verbs (#36) or *acceptable* related adjectives (#21) were more likely to start therapy. This emphasizes the importance of complex health conditions in affecting the start of therapy. Finally, we found that patients who mentioned scheduling related terms are more likely to start therapy. This suggested that actively managing appointments with healthcare providers is beneficial to start therapy.

Limitations and Future Work

Despite these findings, we wish to highlight several limitations, which serve as the basis of future research. First, our findings were generated from a cohort that were predominantly Caucasion and diagnosed with early stage breast

cancer at a single medical center, which may limit their generality. Second, we are unable to discern whether hormonal therapy was prescribed in the adjuvant or neo-adjuvant (pre-surgical) setting. Neoadjuvant hormonal therapy was intensively studied in the early 2000's but then fell out of favor; recently it has been used more often in the context of "window of opportunity" trials[20]. This is an important distinction, since the initiation rate of neoadjuvant hormonal therapy is expected to be close to 100% because it is a prerequisite to the surgery. Third, the cohort consisted of the patients who were prescribed hormonal therapy medication; however, it would be useful to compare this population with patients who had the same disease but did not undergo hormonal therapy. Such a comparison could provide greater insight into a patient's decision-making process with respect to this treatment. Fourth, while our study demonstrated the effectiveness of applying word clustering to group single terms, it is limited in that there were still several groups that were difficult to interpret their meanings. It will be interesting to investigate the extent to which including human judgment in clustering process can help improve the interpretability of word groups. Further, we identified the start of the hormonal therapy based on the medication entry date recorded in EMR system. It will be very interesting to investigate the intended start of the therapy. Future work can also consider incorporate insurance status, income and healthcare provider characteristics into the model, as well as investigate the influence that patient have on their first hormonal therapy prescription. Finally, another useful future direction is investigating the extent to which hormonal therapy initiation is associated with hormonal therapy medication discontinuation.

Acknowledgment

This work was supported by the National Science Foundation grant number IIS1418504.

References

1. Yin Z, Malin B, Warner J, Hsueh PY, Chen CH. The power of the patient voice: learning indicators of treatment adherence from an online breast cancer forum. In: Proceedings of the 11th International AAAI Conference on Web and Social Media 2017:2017:337-46.
2. Early Breast Cancer Trialists' Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet 2005:365:1687–717.
3. Davies C, Pan H, Godwin J, Gray R, Arriagada R, 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 2013:381: 805–16.
4. Chlebowski RT, Kim J, Haque R. Adherence to endocrine therapy in breast cancer adjuvant and prevention settings. Cancer Prev Res (Phila. Pa.) 2014:7:378–87.
5. Yin Z, Xie W, Malin B. Talking about my care: detecting mentions of hormonal therapy adherence behavior in an online breast cancer community. AMIA Annu Symp Proc. 2017:2017:1868–77.
6. Beryl LL. et al. Mapping the decision-making process for adjuvant endocrine therapy for breast cancer: the role of decisional resolve. Med. Decis. Mak. Int. J Soc Med Decis Mak. 2017:37:79–90.
7. Weaver KE, Camacho F, Hwang W, Anderson R, Kimmick G. Adherence to adjuvant hormonal therapy and its relationship to breast cancer recurrence and survival among low-income women. Am J Clin Oncol. 2013:36: 181–7.
8. Yin Z, Warner J, Malin B. Learning when communications between healthcare providers indicate hormonal therapy medication discontinuation. AMIA Annu Symp Proc. 2018:2018:1591–1600.
9. Yin Z, Harrell M, Warner J, Chen Q, Malin B. The therapy is making me sick: how online portal communications between breast cancer patients and physicians indicate medication discontinuation. J Am Med Inform Assoc. 2018:25:1444–51.
10. Eraso Y. Factors influencing oncologists' prescribing hormonal therapy in women with breast cancer: a qualitative study in Córdoba, Argentina. Int J Equity Health 2019:18:35.
11. Velikova G, Fallowfield L, Younger J, Board RE, Selby P. Problem solving in patient-centred and integrated cancer care. EBN Health, an imprint of Evidence-based Networks Ltd, 2018.
12. Clayman ML, Gulbrandsen P, Morris MA. A patient in the clinic; a person in the world. Why shared decision making needs to center on the person rather than the medical encounter. Patient Educ Couns. 2017:100:600–4.
13. Yin Z. Automated learning of health behaviors through consumer authored natural language text, PhD Dissertation 2018. Available at: https://etd.library.vanderbilt.edu//available/etd-02112018-221351/. (Accessed: 13rd March 2019)
14. Roden DM, Pulley JM, Basford MA, Bernard Gr, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008:84:362–369.

15. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013:2:3111–9.
16. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003:3:993–1022.
17. Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D. Exploring topic coherence over many models and many topics. in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012:952–61.
18. Chuang J, Manning CD, Heer J. Termite: visualization techniques for assessing textual topic models. In Proceedings of the International Working Conference on Advanced Visual Interfaces. 2012:74-7.
19. Hershman DL, Wang X, McBride R, Jacobson JS, Grann VR, et al. Delay in initiating adjuvant radiotherapy following breast conservation surgery and its impact on survival. Int J Radiat Oncol. 2006:65:1353–60.
20. Niraula S, Dowling RJ, Ennis M, Chang MC, Done SJ, et al. Metformin in early breast cancer: a prospective window of opportunity neoadjuvant study. Breast Cancer Res. Treat 2012:135:821–30.