Learning When Communications Between Healthcare Providers Indicate Hormonal Therapy Medication Discontinuation

Zhijun Yin, PhD, Jeremy L. Warner, MD, MS, Bradley A. Malin, PhD
Vanderbilt University Medical Center, Nashville, Tennessee, USA

Abstract

Hormonal therapy is an effective yet challenging long-term treatment for patients with hormone receptor positive breast cancer. Understanding what factors indicate discontinuation of a recommended hormonal therapy medication can help improve treatment experience. To date, studies on medication discontinuation have focused on patient information gathered through questionnaires, structured electronic medical records and online discussion boards. However, there has been little investigation into the communications between healthcare providers, which may provide additional indicators of patients' medication discontinuation, particularly from a clinical perspective. In this paper, we investigate the relation between such communications and hormonal therapy medication discontinuation. We studied a cohort of 2,579 patients on hormonal therapy at the Vanderbilt University Medical Center over a 16-year period. We adopt a data-driven approach to investigate the clinical messages communicated by their healthcare providers, the messaging patterns, topics they communicated, and the extent to which these factors are affiliated with discontinuation to a recommended 5-year treatment protocol. Our findings suggest that notification of unread messages, plans for clinical trials and the occurrence of treatment-related complications are affiliated with an increased risk of medication discontinuation. By contrast, ordering prescriptions, making appointments, using positive communication verbs, and noting patients' stable health conditions are affiliated with a decreased risk of medication discontinuation.

Introduction

Hormone receptor (estrogen and/or progesterone) positive breast cancer makes up 80% of diagnosed cases[1]. Hormonal therapy is a treatment for this type of cancer that works by blocking either hormone production or a hormone's carcinogenic effects. In this treatment, surgically resected breast cancer patients are prescribed adjuvant medications (e.g., tamoxifen) for at least five years to effectively lower the risk of breast cancer recurrence and mortality. In spite of the recognized benefits, only 50% of breast cancer patients prescribed this treatment can complete the recommended 5-year treatment protocol[2]. This medication discontinuation behavior occurs due to many potential reasons such as high costs, side effects, and recurrence while taking medication[3–5]. As such, it is important to ascertain which factors are affiliated with medication discontinuation, so that treatment persistence for hormonal therapy can be improved.

To date, research on medication discontinuation has relied on 1) clinical factors that are documented in the structured sections of electronic medical records (EMRs) (e.g., prescribed medications, ordered laboratory tests, and insurance billing codes) or 2) feedback provided by patients through survey-like methods[6,7]. However, these approaches are often limited in that they either 1) lack information about a patient's treatment experience, and 2) are time consuming to implement and manage, thus limiting their scalability. More recently, patients have relied upon the Internet to facilitate discussion about their own, or inquire about others', treatment experience. This has created an opportunity to investigate additional factors that are related to hormonal therapy medication discontinuation. For example, it has been shown that knowledge extracted from discussion among patients in online environments can be applied to effectively predict when breast cancer patients are likely to discontinue their prescribed regimen[8–10]. While there is evidence showing that the clinical notes composed by healthcare providers can be effectively leveraged to predict health outcomes[11], there are few studies that investigate the associations of communications between healthcare providers and a patients behavior. In comparison to self-disclosed treatment experiences in online environments, our work is based on the hypothesis that the messages between healthcare providers, as documented in the EMR, can provide intuition into a patient's treatment discontinuation. This hypothesis stems from the fact that the latter environment brings a clinical perspective to the scenario.

Therefore, to examine this hypothesis, in this paper we investigate the association between what healthcare providers communicate through the messages they exchange in an EMR system and breast cancer patients' hormonal therapy medication discontinuation. Particularly, we study three related research questions: **1)** *What are the messaging patterns that healthcare providers exhibit with respect to their breast cancer patients?* **2)** *What are the topics communicated*

in these messages and do they change over the treatment timeline? (e.g., from cancer diagnosis date to the end of a 5-year treatment protocol) **3)** *How, and to what extent, do these messaging factors affiliated with the chance that a breast cancer patient will discontinue a cancer medication*? Investigating the first two research questions can provide intuition into the characteristics of communications between healthcare providers, while the third research question directly supports the evaluation of our hypothesis.
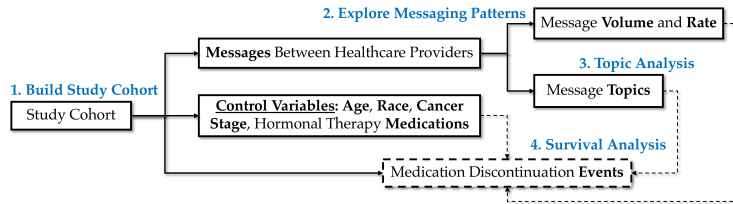

Figure 1: An illustration of the workflow of this study. After exploring messaging patterns and topics, we conduct a survival analysis to evaluate our hypothesis by controlling demographic and clinical factors.

Figure 1 illustrates the workflow that we use to investigate these research questions. Specifically, we focus on a de-identified co-hort of 2,579 breast cancer patients who were prescribed hormonal therapy between 2001 to 2017 at the Vanderbilt University Medical Center (VUMC). We identify all of the clinical communications between health-care providers through the messaging service of the EMR system. We first examine the messaging volume and the messaging rate along t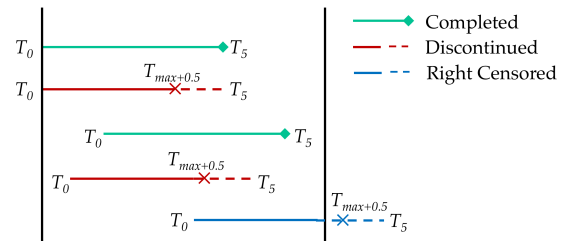he hormonal therapy timeline. Then, we extract the message topics through an unsupervised learning approach, as well as inspect the topic trends. Finally, we apply a survival analysis to investigate the extent to which the messaging rates and topics are affiliated with hormonal therapy medication discontinuation (after adjusting for clinical and demographic factors).

Methods

Data

Building the Study Cohort. To create the study cohort, we rely on the VUMC Cancer Registry. We select patients with documented information that satisfies the following criteria: 1) diagnosed with stage I to III breast cancer, 2) prescribed at least one of the following hormonal therapy medications: *Anastrozole*, *Exemestane*, *Letrozole*, which are known as Aromatase Inhibitors (AIs), *Raloxifene*, or *Tamoxifen*, which are known as Selective Estrogen Receptor Modulators (SERMs), 3) began hormonal therapy after the first message communicated between healthcare providers, and 4) were not reported as deceased. After filtering on these criteria, we obtained a cohort of 2,579 breast cancer patients prescribed hormonal therapy between 2001 and 2017. It should be noted that the data used in this study is de-identified and all dates with a patient record were consistently offset by a number of days uniformly sampled from a (-365, -1) range[12]. This study was approved by the Institutional Review Board of Vanderbilt University.

Medication Discontinuation Event. Figure 2 illustrates how medication discontinuation events within a 5-year treatment protocol are identified. To do so, we treat the first entry date of the aforementioned medications as the start of hormonal therapy (denoted by T_0). Since there is no information that explicitly documents whether a patient experienced such an event, we estimate it at 6 months after the last date of a medication entry (which we denote as $T_{max+0.5}$)[13]. This is due to the fact that, in practice, breast cancer patients prescribed hormonal therapy are expected to come in for a follow-up at least every six months during the first five years after diagnosis. We say a medication discontinuation is experienced at $T_{max+0.5}$ if 1) $T_{max+0.5}$ is smaller than T_5, the end of treatment, and 2) falls within the data collection window. The second criterion is applied to ensure that there is a full five-year worth of observations available for the patients, so that we can ascertain if a medication discontinuation event occurred. In the event that we cannot observe

Figure 2: An illustration of how discontinuation events are identified. The two black vertical lines represent the start and end of the data collection window. T_0 represents start of treatment, T_5 represents end of treatment, and $T_{max+0.5}$ represents six months after the latest date of a medication entry. The dashed lines represent unobserved medication entry records. The red lines represent mediation discontinuation events.

a patient for this length of time, the patient record is said to be right-censored and will be handled as described below. Based on this criteria, we obtained 314 (12.2%) right-censored patients, 1,082 (41.9%) patients who discontinued their medications, and 1,183 (45.9%) patients who completed a 5-year treatment protocol. The patients in this study had an average age of 55.5 with a standard deviation of 11.7 at breast cancer diagnosis, 86.3% are White, 7.8% are African American, 2.1% are Asian, and 2.2% are other races. There are 10.3% of the patients who were in an advanced cancer stage (i.e., stage IIIa, IIIb, or IIIc), while 89.7% were in early cancer stages (i.e., stage Ia, Ib, IIa, or IIb).

Messaging Patterns

To investigate what messaging patterns exist, we examine how messaging rate changes along the treatment timeline, and how it is correlated with medication discontinuation. In this analysis, we neglect the right-censored patients because their medication discontinuation status was not fully determined during the data collection window.

Trend of Messaging Rate. We 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 the time period with an index of 9 corresponds to the final 6-month period of the 5-year treatment protocol. We count the number of messages sent in each period, which we use to represent the messaging rate between healthcare providers about the patient. Given the messaging rates, we plot the LOWESS smooth curves and their 95% confidence intervals for 1) patients who completed the 5-year treatment protocol and 2) patients who experienced a medication discontinuation event. We compare the two smooth curves according to their difference along the treatment timeline.

Messaging Rate and Likelihood of Discontinuation. We order the patients according to the messaging rate and apply a moving average to estimate the probability that a patient discontinued a medication in each sliding window. Each window includes 12.5% of the selected patients for this risk estimation. We also apply this moving average strategy to obtain the corresponding average log transformed messaging rates. In doing so, we examine the correlation between the log-transformed messaging rate and the likelihood of a medication discontinuation event.

Topic Analysis

Next, we investigate the semantic content of the messages communicated between healthcare providers. Specifically, we first identify latent topics and then compute the trends of the topics along the treatment timeline.

Topic Extraction. We rely on an unsupervised learning approach to obtain the topics. To do so, we first use all 303,042 messages communicated by healthcare providers to fit a word2vec model. This is because word2vec has been shown to be effective to learn a low-dimensional representation of terms in mining clinical notes[14]. We adopt a skip-gram model with negative sampling, as implemented in *gensim* python package (version 0.13.1) and retain words with a frequency of ten or greater. We apply a sliding window of 5 words and set the word vector length to 200. However, this parameterization yielded 18,124 distinct words in the fitted model, many of which, based on a cursory review, contributed little to the understanding of the message topics. Based on this observation, we narrow our focus to the 2000 most frequent words and systematically enlarge the set of selected words by incorporating those with a cosine distance greater than 0.75 from any of the selected words. This strategy enables us to incorporate misspellings and semantically similar words. After applying this process, the model consisted of a reduced vocabulary of 2,840 words.

We apply an agglomerative hierarchical clustering with complete linkage, as implemented in the *sklearn* python package (version 0.18.1), to extract topics. To determine the number of clusters, we rely on the standard deviation of the cluster size. This is because, in practice, we do not want to generate clusters that are extremely small or large size. While a large cluster might contain multiple topics, a small cluster might not have a sufficient amount of words to convey a meaningful topic. In our experiments, we observed that the standard deviation of the cluster size decreased towards zero as the number of clusters grows to the vocabulary size with a proper step (e.g., 25 in our case). Note that a small step may result in a jagged curve that is difficult to interpret for the ideal number of clusters. Based on this observation, we follow the elbow principle to locate the angle where the marginal gain in cluster size begins to decrease. The number of clusters that occur at this point is selected and applied in our analysis.

Temporal Topic Trends. To inspect how topics change along the treatment timeline, we first infer the topic distribution

for each patient. This is accomplished by calculating the percentage of topic words mentioned in a 6-month period. Next, we average the topic distribution across patients for each 6-month period. We then measure the significance of temporal topic trends by calculating their Spearman's rank-order correlation with the timeline at a significance level of 0.05. This analysis is expected to find the topics that correlate to patients' treatments, and evaluate the possibility of applying topics as indicators in the following survival analysis.

Survival Analysis

To perform the survival analysis, we investigate how the messaging rates and topics are affiliated with hormonal therapy medication discontinuation events.

Statistical Model. We apply a Cox proportional hazards regression model for the survival analysis. The Cox model is applied to investigate the effect of several independent variables with respect to the time when an event of interest occurs. There are two primary benefits in applying a Cox model 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 by incorporating both time (to when an event happened) and discontinuation status (i.e., patients who realized the events or were right censored) into the model. The risk for an independent variable 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 (smaller) than 1, then the variable is associated with an increased (a decreased) risk of discontinuing a prescribed medication. We used the implementation of the Cox model from the *lifelines* python package (version 0.9.4) to conduct survival analysis.

Variables. In this study, the dependent variable corresponds to a binary variable, where 1 indicates that a medication discontinuation event occurred and 0 indicates that either the patient is 1) right-censored or 2) completed a 5-year treatment protocol. We propose two types of messaging-related variables: 1) message topics and 2) messaging rate. The message topics are constructed as follows: 1) for each patient, we aggregate all of the messages sent between healthcare providers after the breast cancer diagnosis date and before either i) when medication discontinuation event occurred or ii) the patient was right-censored. As such, we represent each patient as a collection of messages; 2) we replace the words in each document with the corresponding topic numbers; and 3) we calculate the term frequency - inverse document frequency (TF-IDF) values for each topic in each document, which we invoke as topic variable values. The messaging rate is represented as the average number of secure messages sent per 6-month period. The messaging rate is standardized into (0, 1) scale before applying into Cox model.

In addition, we introduce four control variables into the model: 1) *age* at diagnosis, 2) *race* (as documented in the EMR), 3) cancer *stage*, and 4) hormonal therapy *medications*. We impute missing values for age (1.1% of instances) with the average age and scale the variable into the (0,1) range. We impute race with missing values (1.6%) with the most frequent category (White) and partition race into White and non-white two categories. We used 1 to encode an advanced cancer stage (and 0 for early cancer stage). The variable of hormonal therapy medications measures how often the patient was prescribed with AI medications. It is represented as the proportion of the number of periods on AI in the number of periods the patient was on either an AI or SERM medication.

Results

In this section, we report messaging patterns, topics and the extent to which they indicate medication discontinuation.

Messaging Patterns

Figure 3a shows the histogram of the messaging volume between healthcare providers. The median number of messages was 86. Note that approximately 90% of the patients are affiliated with fewer than 250 messages. Figure 3b shows how messaging rates change along the treatment timeline. The red and green lines correspond to patients who discontinued medications and completed the treatment protocol, respectively. The leftmost point corresponds to 6 months before the start of hormonal therapy, while the rightmost point corresponds to the final 6 months of the treatment protocol. It should be noted that the median time interval between the date of diagnosis and the start of the therapy is 3 months. There are three important observations that we highlight here: 1) overall, the messaging rate for
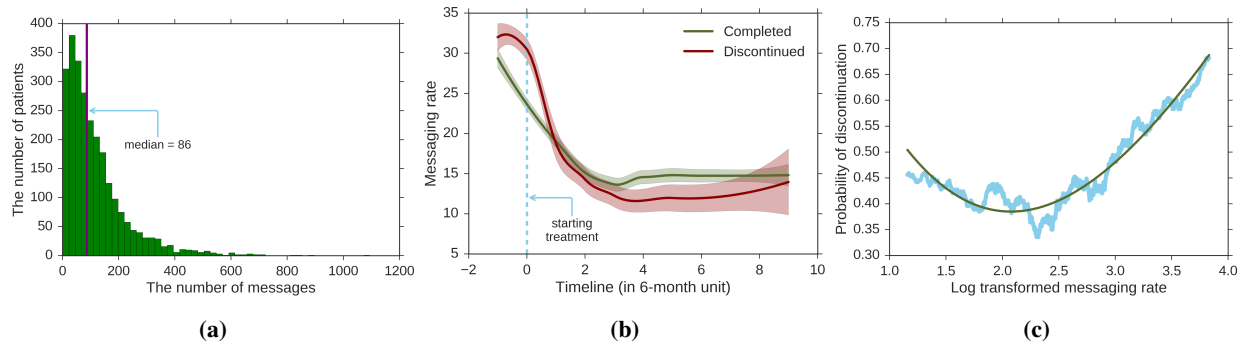
Figure 3: (a) A histogram of the messaging volume. The purple vertical line indicates the median number (86) of messages communicated between healthcare providers. (b) The LOWESS smoothed curves, along with the 95% confidence interval, of the number of messages sent per 6-month period (messaging rate) for patients who completed treatment protocol (green) and patients who discontinued medications (red), respectively. 25% of the data was applied to estimate each data point. (c) Log-transformed messaging rate with respect to the probability of a medication discontinuation event (after smoothing with a moving average).

both patient groups drops quickly between the date of diagnosis until 1.5 years after the start of treatment (index of 2), after which the messaging rate stabilizes, 2) before the first year of treatment (index of 1), patients with medication discontinuation events tend to affiliate with higher messaging rates between healthcare providers, and 3) after this point, these patients are affiliated with lower messaging rates.

Figure 3c shows the probability of a medication discontinuation with respect to the log-transformed messaging rate. The green line indicates a smoothing cubic spline fit. From this figure, it can be seen that the probability of medication discontinuation decreases as the messaging rate between healthcare providers increases. However, the trend of decreasing likelihood stops when the messaging rate grows to around 8 messages per 6-month on average. After this point the probability of discontinuing medication increases steadily. For example, the probability can reach 68.2% when the rate is 46 messages per 6-month on average (the rightmost data point in the figure).

Topic Analysis

Based on the elbow rule, we selected 150 topics for further consideration. Table 1 shows 10 topics that have statistically significant trends based on Spearman's rank-order correlation. For each topic in the table, we rank the words based on their average similarity with other words in the same topic. When the size of a topic is greater than 20 (10), we display the top 20 (10) words; otherwise, we show all words in the topic[1]. The table shows that surgery related topics (#65, #93, #108) lose popularity as the treatment progresses. By contrast, topics related to pharmacies (#27), lab tests (#9, #83), medications used to control for side effects (#29), and diabetes (#112) gain popularity as the treatment progresses. As messages below communicated:

- "*Dr. **NAME saw pt yesterday in clinic and has recommended a **segmental mastectomy**. We have her tentatively scheduled for **DATE but if you are possibly planning **surgery**, maybe a joint case could be arranged.*"
- "*Facility Name:**publix pharmacy** Please advise: Interaction between letrozole and activella (estradiol). Pharmacy sID***the last she got the activella was **DATE. She should discontinue the activella correct?*"

Survival Analysis

We fit a Cox model with a concordance of 0.736. Note that concordance is a variation of AUC that measures the proportion of patient pairs in which patients with a higher-risk variable discontinue medications before patients with the lower-risk variable. Among the four control variables, patient who were prescribed with AI ($HR = 0.812$, $P < 0.001$) have a decreased risk of medication discontinuation. After controlling for the four proposed predictors, the

---

[1]We follow the same rule when showing word samples in all following tables that contain topic related information.

Topic	Word Samples	Coeff.	p
28	*assist, counseling, counselor, counselors, offer, arranged, assistance, arrange, worker, coordinate*	-0.918	< 0.001
65	*surgery, operation, procedures, procedure, possibly, possibility, dental, dentist, possible*	-0.900	< 0.001
93	*segmental, sln, supraclavicular, lumpectomy, axillary, mrm, seg, partial, sentinel, axilla, mast, masectomy, mastectomy, retroareolar, margin, ax, ln, dissection, excision, te*	-0.891	< 0.001
108	*discoloration, incision, redness, reddened, tenderness, reddness, warmth, wounds, induration, compress, erythema, bruising, heat, incisions, soreness, warm, area, soaks, skin, sites*	-0.882	< 0.001
15	*performed, indicated, initial, perform, recent, prior, following, previous, additional, post*	-0.873	< 0.001
27	*costco, walgreens, krogers, walmart, publix, walid, cvs, kroger, walgreen, wg, medco, eckerd, kmart, rx, rxs, script, levoxyl, levothyroxine, pharmacy, prescription*	0.827	0.002
83	*lipids, lfts, cpk, hga1c, tsh, hemoglobin, electrolytes, a1c, hba1c, bmp, hgba1c, flp, level, albumin, protein, levels, labs, labwork, bloodwork, count*	0.855	0.001
29	*cipro, macrobid, ceftin, augmentin, keflex, levaquin, flagyl, amoxicillin, doxycycline, cephalexin, bactrim, azithromycin, clindamycin, levoquin, septra, culture, ds, ua, cultures, cx*	0.873	< 0.001
9	*carotid, duplex, doppler, dopplers, echocardiogram, ecg, pulmonary, ekg, renal, cardiac, pelvic, transvaginal, egd, dvt, cxr, pulm, vascular, echo, colonscopy, stress*	0.882	< 0.001
112	*sweets, diet, exercise, dm, wt, diabetes, weight, diabetic, htn, regular*	0.964	< 0.001

Table 1: Topics with statistically significant temporal trends (at the 0.05 level). The topics are sorted according to the coefficients of their Spearman's rank-order correlation. Positive (negative) coefficients suggest increasing (decreasing) temporal topic trends.

average messaging rate was found to have an increased risk of medication discontinuation ($HR = 1.482$, $P < 0.001$). This confirms our messaging pattern analysis that the messaging rate can be indicators of medication discontinuation.

Table 2 shows the nine topics with an increased risk of medication discontinuation. The topics are sorted according to their HRs in an decreasing order. Here, there are several results worth noting.

First, when there are messages sent by the healthcare providers that go unread by the patients, an automated notification is generated to require healthcare providers' attention (#66, $HR = 1.172$). One such example is shown as below:

- "****Automated** notification*: A message sent to a patient via **MyHealthAtVanderbilt** more than **120** hours ago had not yet been read by the patient as of **DATE. Please consider whether further action is necessary to notify the patient. The original **text** of the message (which is still available to the patient) is: ...*"

Second, the original plans for participation in a clinical trials may not be successful in the end (#126, $HR = 1.097$; #21, $HR = 1.090$; #101, $HR = 1.077$):

- "*I spoke with **NAME regarding her head CT results. Unfortunately she is not **eligible** for the clinical **trial**. Please cancel the PET-CT tomorrow and all other **study** related appointments. I will coordinate her follow-up from here.*"
- "*she is not having this biopsy. She had **planned** on doing a research biopsy and then the axillary biopsy at the same time. She is no longer **interested** in the research **study** and therefore they will check her axillary nodes during surgery. Please confirm with clinic. Thank you.*"

Third, occurrence of treatment-related complications may have led to medication discontinuation (#127, $HR = 1.092$).

Topic	Word Samples	HR	95% CI	p
66	*text, myhealthatvanderbilt, automated, original, 120, physican, callers, advises*	1.172	(1.040, 1.321)	0.009
126	*trial, eligible, study, trials, phase, studies, screen*	1.097	(1.031, 1.166)	0.003
127	*inpatient, discharged, discharging, admitted, admission, admit, hospital, nursing, nhc, hosp*	1.092	(1.018, 1.170)	0.013
21	*neoadjuvant, planning, considering, adjuvant, planned, interested, neo, endocrine, radiation, xrt*	1.090	(1.021, 1.163)	0.009
49	*images, discs, disk, pacs, imported, cd, impax, films, film, slides, disc, osf, jacket, csi, radiology, outside, records, norma, si, wmc*	1.087	(1.012, 1.167)	0.023
97	*chief, complaint, relationship, caller, reason, established*	1.085	(1.021, 1.153)	0.008
10	*speaking, contacted, spoken, spoke, talked, speak, talk, contact, talking, contacting*	1.084	(1.003, 1.172)	0.041
101	*declined, cancelled, offered, canceled, transportation, refused, ride, missed, moving, keeping*	1.077	(1.024, 1.132)	0.004
61	*clearence, necessity, neccessity, clearance, transplant, status*	1.064	(1.008, 1.124)	0.025

Table 2: Topics that are positively affiliated with a medication discontinuation event and are statistically significant at the 0.05 level. The topics are sorted according to their HRs. A larger HR suggests an more increased risk of medication discontinuation.

As a healthcare provider wrote:

- "*Dr. **NAME, Ms. **NAME and her supportive daughter have been to the ED several times in the last few months and has even required **admission** for her severe case of acid reflux. We are trying to reduce her admissions and risk for exposure to MDR bacteria in the **hospital** and would like to request for more follow up with either you or a colleague to help address her GERD. ...*"

Fourth, patients' complaints are positively affiliated with medication discontinuation (#97, **HR** $= 1.085$):

- "*As you know, she recently started an aromatase inhibitor Arimidex and believes this may be the reason why her BP has been so poorly controlled lately requiring increases of her BP meds. She also **complains** of weight gain, SOB, and fatigue. While some of this may be due to the Arimidex, her fatigue in particular is very likely due to recently completely radiation therapy.*"

Finally, a patient who is new to the practice or leaving the practice (#49, $HR = 1.087$) could be positively affiliated with medication discontinuation.

- "*By report it is somewhat similar with the report from **DATE (outside scan). However, we will need to get a **CD** scanned in **PACS** with the **images** of the **DATE scans for a fair comparison. Please order the **CD** as soon as possible, and have **radiology** issue an amendment on the current scan once they have a chance to compare.*"

Table 3 shows the topics that are significantly affiliated with a decreased risk of medication discontinuation. The main findings in this table can be summarized as follows:

First, topics related to prescriptions (#71, $HR = 0.658$; #131, $HR = 0.861$; #0, $HR = 0.869$; #135, $HR = 0.909$) suggest that regular treatment activities are negatively affiliated with medication discontinuation. These topics include healthcare providers' preparation of prescriptions, and patients' request of medications used for high blood pressure, diabetes and other diseases. As the messages below communicated:

- "***DATE appt with Dr **NAME . script sent to your **task** list / ready for you to review, edit (add or subtract refills , change dispense amount if needed) and fax to the pharmacy, thank you*"
- "*dr **NAME- med refills were called in to vandy pharm ***PHONE. One month was called in but she needed 3 mo. **cozaar** 100 mlg one daily, levoxyl 25 mcg one daily, gen maxide 75 mlg one daily, minocycline 50 mlg*"

Topic	Word Samples	HR	95% CI	p
42	*spotting, heavier, began, heavy, periods, episode, vag, episodes, bleeding, light, noticed, developed, started, period, vaginal, usual, approx, past, ago, completely*	0.918	(0.852, 0.989)	0.025
46	*50s, 30s, 40s, 60s, age, based, old, indicates, including, according*	0.916	(0.848, 0.989)	0.024
135	*requisition, forms, paperwork, form, papers, order, orders, requistion, slip*	0.909	(0.840, 0.984)	0.018
2	*want, willing, prefer, wish, wants, need, rather, like, asked, prefers, told, wishes, going, encourage, wanting, encouraged, needs, lets, advised, wanted*	0.901	(0.830, 0.979)	0.013
90	*confirm, inform, verify, remind, ask, tell, advise, notify, know*	0.897	(0.825, 0.976)	0.012
104	*unchanged, stable, shows, showed, normal, wnl, compared, showing, appear, looked*	0.896	(0.826, 0.972)	0.008
8	*423, 865, 256, 502, 606, 731, 931, 618, 270, 615, beeper, pager, phone, cell, cb, ph, callback, reached, paged, robin*	0.879	(0.804, 0.961)	0.005
17	*icd9, icd10, codes, billable, icd, code, 174, dx, diagnosis, lipoprotein*	0.877	(0.807, 0.953)	0.002
0	*cozaar, spironolactone, benicar, amlodipine, losartan, avapro, chlorthalidone, metoprolol, norvasc, ramipril, glipizide, atenolol, dyazide, diovan, verapamil, micardis, torsemide, toprol, coreg, maxide*	0.869	(0.790, 0.956)	0.004
131	*200mg, 400mg, 150mg, 300mg, 15mg, 100mg, 50mg, 30mg, 60mg, qam, 2mg, 80mg, 120mg, 180mg, 160mg, 240mg, 4mg, qpm, 10mg, 75mg*	0.861	(0.782, 0.948)	0.002
59	*10am, 9am, 11am, 8am, 1pm, 2pm, 22, 3pm, 19, 26, 1130, 1100, 1030, 12pm, 40am, 0830, 0900, 27, 00, 20am*	0.799	(0.728, 0.877)	< 0.001
19	*schedulers, scheduler, office, asap, patient, tomorrow, psr, nurse, clinic, today*	0.751	(0.688, 0.819)	< 0.001
71	*electronically, task, providers, nc, notification*	0.658	(0.537, 0.807)	< 0.001

Table 3: Topics that are negatively affiliated with a medication discontinuation event and are statistically significant at the 0.05 level. The topics are sorted based on their HRs. A smaller HR suggests a more decreased risk of medication discontinuation.

> *takes twice daily rx faxed to tvc pharm per rx star*"
- "*DSR-pharm called req refill on **metoprolol 50mg** #180 last refill **DATE pharm# ***PHONE*"
- "*I spoke with pt, and relayed the noted recommendations. Pt voiced understanding, and agreed with plan of care. See RxStar for medication **order** details.*"

Second, topics related to appointments (#59, $HR = 0.799$; #19, $HR = 0.751$; #8, $HR = 0.879$) have a decreased risk of mediation discontinuation. As the messages below communicated:

- "*Dr. **NAME- the patient called me today asking me to make an appointment to see you on Monday since that is what you told her regarding her bolsters, so I made a nurse visit on **DATE at **10am** and for you to be paged when she gets there. Please let me know if this is ok or if I need to change it. Thanks!*"
- "*Please call in arimdex 1 mg po qd , disp 30 with 12 refills to **Place. Please also send Breast core biopsy for oncotype DX testing. Clinical stage IIA. The **schedulers** are working on scheduling her CT and bone scans.*"

Third, topics related to positive communications verbs have a decreased risk of medication discontinuation (#90, $HR = 0.897$; #2, $HR = 0.901$). As the messages below communicated:

- "*Pt wanted to **confirm** that she is going to stay on Arimidex and call psychiatrist about psych medications.*"

- "*I **encouraged** pt that these were normal results, but that the lightheadedness was not normal. I **advised** her to increase protein at breakfast for "lasting energy" to get her through her workout, such as adding boiled eggs, lean cheese, etc. She agreed to try this on Friday morning, then report back with her glucose results pre and post workout.*"

Fourth, topics related to stable health conditions (#104, $HR = 0.896$) have a decreased risk of medication discontinuation. As a healthcare provider stated in a message:

- "*let her know bone density is **unchanged**, very **stable** and **encourage** her to sign up for mhav.*" (*mhav* represents *MyHealthAtVanderbilt*, a patient portal in VUMC)

Discussion

Findings and Implications. There are two notable findings from this study. First, we observe that the average messaging rate between healthcare providers are affiliated with an increased risk of hormonal therapy medication discontinuation. Furthermore, our messaging pattern analysis indicates a local minimum messaging rate that corresponds to the lowest discontinuation probability. This suggests that patients with a high rate of messages (between their care providers) are likely to be complex medically or psychosocially, which would naturally correlate with medication discontinuation[15]. Conversely, patients with a low rate of messages may have too little contact with the medical ecosystem and are at high risk of slipping through the cracks. These implications are substantiated by findings from previous studies, which specifically found that patients who had more consultations with a breast specialist and clinical oncologist, as well as maintained a good relationship with their healthcare providers, were more likely to adhere to the recommended treatment[4, 16].

Second, our topic analysis shows that the messages between healthcare providers are related to patients' treatments: before taking hormonal therapy medications, breast cancer patients generally have other treatments such as surgery. This indicates that the message content can be applied to study patients' medication discontinuation as well. A further survival analysis confirms this conjecture. Particularly, it shows that message topics indicative of a notification about unread messages, plans for clinical trials and the occurrence of treatment-related complications are affiliated with an increased risk of discontinuing treatment. By contrast, message topics related to ordering prescriptions, making appointments, voicing verbs indicating positive communications, and stable health conditions for the patient are affiliated with a decreased risk of discontinuing treatment. These topic-related findings suggest that, in comparison to self-reported treatment experiences, where rich emotions, feelings and detailed side effects are described[8], messages between healthcare providers contain factors mainly from the perspectives of the clinical domain. These factors reflect how healthcare providers discuss a patients health condition over the long arc of their treatment. We believe these factors may be useful (particularly if combined with structured data from a patients EMRs) for building prediction models that sense when a patient is exhibiting the potential for medication discontinuation. Still, further research into how such a predictive model could support interventions that lead to decreased rates of medication discontinuation is an open research direction.

Limitation and Future Work. There are several limitations that we want to highlight, which can serve as the basis of future work. First, the data in this study is derived from a single healthcare organization, which may limit the generality of our findings. Second, we adopt a pure data-driven approach for topic extraction and association study. In spite of its efficacy, this approach may lead to some factors that are difficult to interpret. We believe that this problem may be mitigated by incorporating additional guidance from domain experts, but this is outside the scope of this investigation. Since topic modeling allows for a term to belong to multiple topics, future investigations could also consider the effectiveness of topic modeling (e.g., Latent Dirichlet Allocation[17]) for topic extraction. Finally, we only consider the clinical communications between healthcare providers. We believe that it would be useful to integrate communications between healthcare providers and patients into medication discontinuation investigation.

Conclusion

In this paper, we investigated how clinical communications between healthcare providers are affiliated with 2,579 breast cancer patients prescribed hormonal therapy across a 16-year period in VUMC. We explored the patterns of clinical messages communicated between their healthcare providers and extracted topics from these messages through

an unsupervised learning approach. We conducted a survival analysis to investigate the effects of these messaging factors on the chance that a patient would discontinue a particular treatment (in the form of recommended hormone replacement medication). We demonstrated that both messaging rate and topics are significantly affiliated with medication discontinuation. This study lends credibility to the hypothesis that communications between healthcare providers can be effectively applied to ascertain if a breast cancer patients' hormonal therapy treatment is likely to be successful.

Acknowledgments

References

[1] Kohler BA, Sherman RL, Howlader N, et al. Annual report to the nation on the status of cancer, 1975-2011, featuring incidence of breast cancer subtypes by race/ethnicity, poverty, and state. J Natl Cancer Inst. 2015;107(6):djv048.

[2] Chlebowski RT, Kim J, Haque R. Adherence to endocrine therapy in breast cancer adjuvant and prevention settings. Cancer Prev Res. 2014;7(4):378–87.

[3] Neugut AI, Subar M, Wilde ET, et al. Association between prescription co-payment amount and compliance with adjuvant hormonal therapy in women with early-stage breast cancer. J Clin Oncol. 2011;29(18):2534–542.

[4] Murphy CC, Bartholomew LK, Carpentier MY, et al. Adherence to adjuvant hormonal therapy among breast cancer survivors in clinical practice: a systematic review. Breast Cancer Res Treat. 2012;134(2):459–478.

[5] Kuba S, Ishida M, Nakamura Y, et al. Persistence and discontinuation of adjuvant endocrine therapy in women with breast cancer. Breast Cancer. 2016;23(1):128–33.

[6] Weaver KE, Camacho F, Hwang W, et al. Adherence to adjuvant hormonal therapy and its relationship to breast cancer recurrence and survival among low income women. Am J Clin Oncol. 2013;36(2):181–7.

[7] Beryl LL, Rendle KA, Halley MC, et al. Mapping the decision-making process for adjuvant endocrine therapy for breast cancer the role of decisional resolve. Med Decis Making. 2016;37(1):79–90.

[8] Yin Z, Malin B, Warner J, et al. The power of the patient voice: learning indicators of treatment adherence from an online breast cancer forum. In: International AAAI Conference on Web and Social Media; 2017. 337–46.

[9] Yin Z, Song L, Malin B. Reciprocity and its association with treatment adherence in an online breast cancer forum. In: Computer-Based Medical Systems; 2017. 618–23.

[10] Yin Z, Xie W, Malin B. Talking about my care: detecting mentions of hormonal therapy adherence behavior in an online breast cancer community. In: AMIA Annu Symp Proc.; 2017. 1868–77.

[11] Hubbard R, Ungar L, Greene C, et al. Inclusion of unstructured text data from clinical notes improves early prediction of death or prolonged icu stay among hospitalized patients. Am J Respir Crit Care Med. 2017;195:A1084.

[12] Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther. 2008;84(3):362–69.

[13] Harrell M, Fabbri D, Levy M. Analysis of adjuvant endocrine therapy in practice from electronic health record data of patients with breast cancer. JCO Clin Cancer Inform. 2017;1:1–8.

[14] Liu Y, Ge T, Mathews K, et al. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. Proceedings of BioNLP. 2015; 92–97.

[15] Martin LR, Williams SL, Haskard KB, et al. The challenge of patient adherence. Ther Clin Risk Manag. 2005;1(3):189–99.

[16] Moon Z, Moss-Morris R, Hunter MS, et al. Barriers and facilitators of adjuvant hormone therapy adherence and persistence in women with breast cancer: a systematic review. Patient Prefer Adherence. 2017;11:305–22.

[17] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3(Jan):993–022.