Journal of the American Medical Informatics Association, 25(11), 2018, 1444–1451
doi: 10.1093/jamia/ocy118
Research and Applications

Research and Applications

The therapy is making me sick: how online portal communications between breast cancer patients and physicians indicate medication discontinuation

Zhijun Yin,[1] Morgan Harrell,[2] Jeremy L Warner,[1,3] Qingxia Chen,[1,4] Daniel Fabbri,[1,5] and Bradley A Malin[1,4,5]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [2]Roam Analytics, San Mateo, California, USA, [3]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [4]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, and [5]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, Tennessee, USA

Corresponding Author: Zhijun Yin, PhD, Department of Biomedical Informatics, Vanderbilt University Medical Center, 2525 West End Avenue, Suite 1412A, Nashville, TN 37203, USA (zhijun.yin@vanderbilt.edu)

Received 7 April 2018; Revised 12 July 2018; Editorial Decision 9 August 2018; Accepted 10 August 2018

ABSTRACT

Objective: Online platforms have created a variety of opportunities for breast patients to discuss their hormonal therapy, a long-term adjuvant treatment to reduce the chance of breast cancer occurrence and mortality. The goal of this investigation is to ascertain the extent to which the messages breast cancer patients communicated through an online portal can indicate their potential for discontinuing hormonal therapy.

Materials and Methods: We studied the de-identified electronic medical records of 1106 breast cancer patients who were prescribed hormonal therapy at Vanderbilt University Medical Center over a 12-year period. We designed a data-driven approach to investigate patients' patterns of messaging with healthcare providers, the topics they communicated, and the extent to which these messaging behaviors associate with the likelihood that a patient will discontinue a prescribed 5-year regimen of therapy.

Results: The results indicates that messaging rate over time [*hazard ratio* (*HR*) = 1.373, $P = 0.002$], mentions of side effects ($HR = 1.214$, $P = 0.006$), and surgery-related topics ($HR = 1.170$, $P = 0.034$) were associated with increased risk of early medication discontinuation. In contrast, seeking professional suggestions ($HR = 0.766$, $P = 0.002$), expressing gratitude to healthcare providers ($HR = 0.872$, $P = 0.044$), and mentions of drugs used to treat side effects ($HR = 0.807$, $P = 0.013$) were associated with decreased risk of medication discontinuation.

Discussion and Conclusion: This investigation suggests that patient-generated content can inform the study of health-related behaviors. Given that approximately 50% of breast cancer patients do not complete a course of hormonal therapy as described, the identification of factors associated with medication discontinuation can facilitate real-time interventions to prevent early discontinuation.

Key words: patient portals, medication discontinuation, hormonal therapy, hierarchical clustering, survival analysis

BACKGROUND AND SIGNIFICANCE

User generated content (UGC) in online environments is increasingly supplementing traditional electronic medical records (EMRs) in biomedical research.[1] Prior studies have focused on public health surveillance,[2] patient privacy,[3,4] shifts in suicidal ideation,[5] medication discontinuation,[6,7] content mining,[8,9] sentiments and emotions,[6,10,11] and the impact of social support and influence among patients.[12–14] There is evidence that combining data streams can enable new types of investigations, such as early detection of adverse drug events.[15]

While social media dominates the field of UGC, another type of UGC are patient-generated messages (PGMs) communicated through the messaging service in online patient portals. Most research on PGM to date has focused on metadata such as messaging volume. For instance, it has been shown that frequent usage of a portal service is associated with better diabetic control[16–18] as well as reductions in outpatient and inpatient healthcare utilization.[19–21] More recently, there have been investigations into the content of PGMs.[22–24] However, these studies are limited in several important ways. First, they tend to rely on humans to manually review and annotate message instances,[22] a process that is both time consuming and lacks scalability. Alternatively, some investigations have applied supervised learning to classify the topics inherent in the messages, which addresses the issue of scalability, but fails to detect topics that are not defined *a priori* or are not found in training data.[23,24] To the best of our knowledge, there are no studies that focus on linking PGM content to health outcomes and behaviors. Since PGMs communicated through a patient portal may contain factors that are not captured in the structured portions of EMRs, they provide a novel opportunity to help learn the state of the patient.

Here, we describe a study based on the hypothesis that patients' messaging behaviors, in terms of messaging rate and topics *discovered through unsupervised methods*, will associate with medication discontinuation. To investigate this hypothesis, we focused on a cohort of breast cancer patients prescribed hormonal therapy at Vanderbilt University Medical Center (VUMC). We focus on this subpopulation because, while hormonal therapy has been proven as an efficient long-term adjuvant treatment for patients with hormone receptor-positive breast cancer, rates of early discontinuation are as high as 50%.[25,26] While some factors that may trigger a patient to stop using a hormonal therapy medication, such as high cost or intolerable side effects, have been elucidated, there remains a major unmet need to recognize patients at risk of early discontinuation.[11,27–29]

Traditional investigations into hormonal therapy discontinuation tend to rely on information gathered through surveys with patients or data derived from EMRs.[28,30–32] However, as noted in a recent study,[6] these types of data sources are limited by the fact that they rely on non-scalable methods (eg, surveys) or lack information other than healthcare provider observations. In contrast, there is increasing evidence that UGC from online environments can be utilized to learn factors related to hormonal therapy discontinuation. For instance, we previously demonstrated that data from an online breast cancer forum suggested that patients who mentioned side effects, such as depression, were more likely to discontinue hormonal therapy.[6,7] Additionally, many potential barriers to breast cancer treatments (eg, cost and trust) can be detected in UGC from the online environment.[33] Yet, despite the opportunities these new domains support, there are no studies that examine the factors inferred from PGMs or their association with hormonal therapy discontinuation. This is notable because the PGMs in patient portals (which could be contributed by patients or their delegates, such as a spouse) can be linked to a patient's EMRs, thus making it feasible to directly investigate such associations.

METHODS

Data preparation

Study population

This study is based on de-identified data from the patient portal and EMRs of VUMC and was approved as non-human subject research by the Vanderbilt University institutional review board.[34] 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.

We focus on patients who were diagnosed with American Joint Committee on Cancer (AJCC) summary stage I to III invasive 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).[35] We constrain the study to breast cancer patients who commenced their treatment after the date of the first PGM sent by breast cancer patients, which took place in late 2005. Deceased patients before March 2017, the end of data collection window, were excluded due to their unresolved medication discontinuation status.

Determining the medication discontinuation events

Most guidelines specify AIs and/or SERMs for at least 5 years as an adjuvant treatment protocol;[36] we thus focus on discontinuation events within the first 5 years of hormonal therapy. Figure 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, as breast cancer patients prescribed hormonal therapy are expected to have a clinic encounter every 6 months in the first 5 years after diagnosis.[35] For patients with their $T_{max+0.5}$ beyond the data collection window, because we do not observe a full 5 years length of records, they are right-censored in the survival analysis (see below).

Using a 12-year data collection window, we obtained 245 (22.2%) right-censored patients, 478 (53.2%) patients finishing a 5-year protocol, and 383 (34.6%) patients who failed to complete the treatment regimen. This latter observation is consistent with a systematic review that indicated early discontinuation rates of hormonal therapy range from 31% to 73%.[37] Patients had an average age of 53.9 (SD ±11.1) at their cancer diagnosis; 91.3% were White, 5.9% African American, 1.9% Asian, and 0.9% Other race. 12.9% of the patients were in a relatively advanced cancer stage (ie, AJCC summary stage III), while 87.1% were in early cancer stages (ie, AJCC summary stages I and II). Among these patients, 52.4% were prescribed AI medications only, 8.6% were prescribed SERM medications only, and 39.0% were prescribed both AI and SERM medications during the course of their treatment. These patients generated 47 600 messages with a median (interquartile range) length of 39 words (19 to 71).

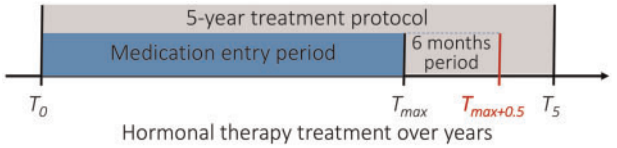

Figure 1. Illustration of the critical time points applied in the definition of medication discontinuation event within 5 years of treatment. T_0 refers to the first hormonal medication entry date, and T_{max} is the maximum medication entry date. $T_{max+0.5}$ re represents the 6 months after the T_{max} 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.

Messaging patterns

Messaging volume

To investigate the relationship between messaging volume and medication discontinuation, we focus on the patients who either 1) completed 5 years of treatment or 2) experienced a medication discontinuation event (excluding the right-censored patients). We order the patients according to their messaging volume and apply a moving average to estimate the probability that a patient discontinues his or her medication in each 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.

Messaging rate

To obtain an overall trend of messaging rates, we set the first hormonal therapy 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. In doing so, time period 0 corresponds to the first 6-month hormonal therapy treatment, while periods with negative (non-negative) index values correspond to the time before (after) hormonal therapy treatment. We calculate the messaging rate for each patient by counting the number of messages sent in each period.

Message topic inference

We infer the topics through applying a hierarchical clustering strategy directly on the words in the messages. By employing a word embedding technique, in the form of word2vec, we map words into a vector space of much lower dimensionality and calculate their semantic similarity in terms of cosine distance.[38] While topic modeling has been applied to personally contributed information,[39] our solution may generate more interpretable topics by directly combining words with similar semantics.

We use a skip-gram model with negative sampling as implemented in the *gensim* python package (version 0.13.1) to fit a word2vec model over all of the messages. We retain the words that appear at least 5 times, choose a sliding window of 5 words in length, and set the word vector length to 200. The fitted word2vec model contained over 13 000 distinct words, many of which were unlikely to assist in understanding messaging content. As a result, we chose the 2000 most frequent, and enlarged the selected word collection by incorporating words that exhibited a cosine distance greater than 0.6 with any of the selected words. This was expected to include more words due to misspelling and semantic similarity. This process recovered an additional 2010 words. We further removed several types of words that we suspected would contribute little towards constructing meaningful topics: 1) stop words (eg, *the* or *of*); 2) years; and 3) words with a cosine distance greater than 0.6 from the words noted above. This process yielded 3664 words.

We applied an agglomerative hierarchical clustering with complete linkage implemented in the *sklearn* python package (version 0.18.1) to extract topics. As such, we adopted a simple metric to help decide the number of clusters: the standard deviation of cluster size. When there are more than two clusters, the standard deviation of the cluster size decreases towards zero as the number of clusters grows towards the vocabulary size with a proper step (eg, 25 in our case). This is because a small step can induce in an unsmoothed 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 diminish. This indicated that the optimal number of clusters was 200. We refer the reader to Supplementary Appendices S1 and S2 for details on the topic extraction and analysis.

Survival analysis

Model

We apply a Cox proportional hazards regression model to evaluate for associations between messaging behaviors and medication discontinuation. There are 2 primary benefits in applying a Cox model instead of a model akin to logistic regression in this study. First, Cox is a semi-parametric model that does not assume any particular survival distribution. Second, the model incorporates right-censored patients. The hazard ratio (*HR*) at a significance level of 0.05 will be reported.

Control variables

We incorporated 4 additional variables into the model: age at diagnosis, race, cancer stage, and hormonal therapy medication. We imputed missing values for age (1.1%) with the average age and scale the variable into the (0, 1) range. We partitioned race into White and non-White (represented as 1/0). We further categorized cancer stage into advanced/early (represented as 1/0) cancer stage. The variable of AI medication (denoted as *taking AI*) is represented as a proportion of the number of periods on AIs divided by the number of periods on either AIs or SERMs.

Message related predictors

We constructed 2 types of predictors related to messages. First, we built topic predictors as follows: 1) for each patient, we aggregated all the messages sent after the breast cancer diagnosis date and before either: i) when the medication discontinuation event occurred or ii) the patient was right censored. As such, we modeled each patient as a collection of messages; 2) we replaced the words in each patient message collection with the corresponding topic numbers (if present); 3) we calculated the Term Frequency–Inverse Document Frequency values for each topic in each patient message collection, which we use as topic predictor values. Second, we included messaging rate (in terms of the average number of messages sent per 6-month period) as an additional variable. AS the messaging rate distribution is right-skewed (see Supplementary Appendix S3 for a histogram of messaging rate), we applied a log transform and scaled the data into a (0, 1) range before applying the Cox model.

RESULTS

Messaging patterns

Messaging volume

Figure 2 depicts a log–log plot of the messaging frequency distribution. 10% of the patients sent only 1 message during the data collection window. In contrast, 0.2% of the patients sent more than 500 messages. Figure 3 illustrates the probability of discontinuing medication against the log transformed messaging volume. The figure shows that patients are less likely to discontinue medication as the number of messages they send increases. However, this trend falters when the messaging volume grows to around 3 messages on average. At this point, patients begin to exhibit an increasing risk of discontinuing their treatment regimen until the messaging volume reaches another inflection point at around 20 messages. After this point, the probability of discontinuing medication rapidly decreases.

Messaging rate

Figure 4 depicts the LOWESS smooth curve, along with its 95% confidence interval, for messaging rates along the treatment timeline. There are 3 clinically important dated events highlighted in the figure, as indicated with vertical dotted lines: *left*) the diagnosis of breast cancer (index = -1.6, 90% percentile), *middle*) the start of hormonal therapy (beginning of the period with index = 0), and *right*) completion of a 5-year treatment protocol (beginning of the period with index = 10). The figure shows that breast cancer patients send messages at an increasing rate before the disease is diagnosed, suggesting increased coordination of care in the period of workup prompted by an abnormal mammogram or palpable breast lump. The messaging rate reaches its maximum just before the start of the hormonal therapy, then drops quickly in the initial 2.5 years of treatment, and holds steady until the end of the 5-year regimen. These rate changes show that breast cancer patients tend to have less communications with healthcare providers as their hormonal therapy progresses.

Survival analysis

Model fitting

We fit a Cox model with a concordance of 0.753 based on the *lifelines* python package (version 0.9.4). With respect to the 4 control variables, it was found that 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 ($HR = 0.715$, $P < 0.001$).

Messaging rate

After controlling for the 4 variables, the average messaging rate was positively associated with discontinuation ($HR = 1.373$, $P < 0.002$). In addition, many topics were found to be statistically significant. This suggests that topics contributed additional information above and beyond the messaging rate for explaining medication discontinuation. We summarized these topics, based on their increased or decreased risk, as follows.

Topics with increased risk

Table 1 summarizes the inferred topics that were associated with an increased risk ($HR > 1$) of medication discontinuation. In Table 1, 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

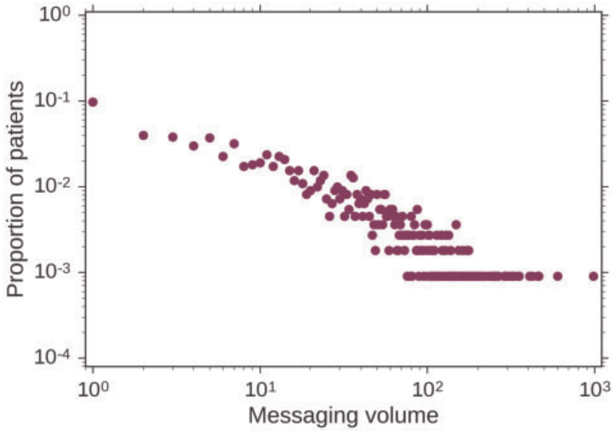

Figure 2. Log–log plot of the number of messages sent by patients through the patient portal.

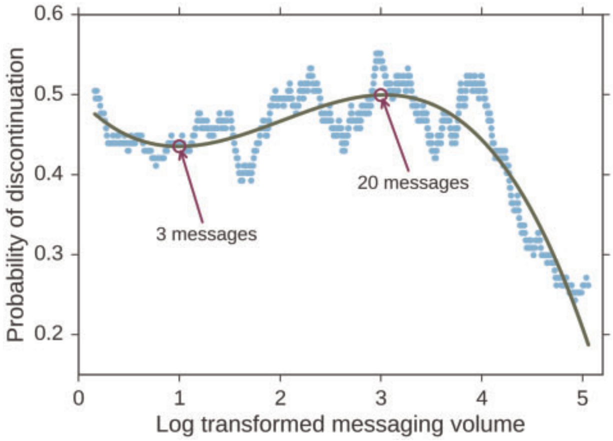

Figure 3. Log transformed messaging volume with respect to the probability of medication discontinuation (after smoothing with a moving average).

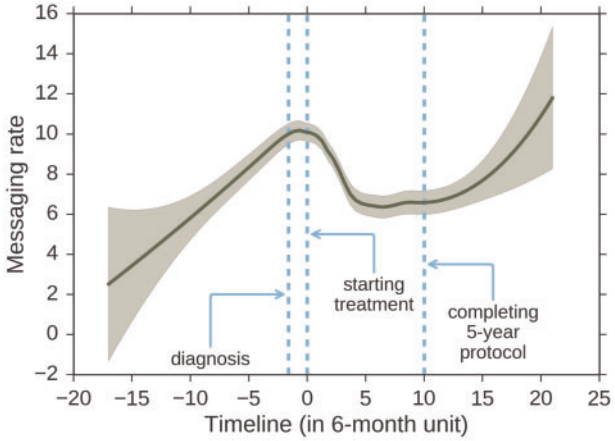

Figure 4. LOWESS smoothed curve and 95% confidence interval of the number of messages sent per 6-month period. 25% of the data is applied to estimate each data point. The 3 dotted lines from left to right indicate: *left*) the 90th percentile of diagnosis dates (index = -1.6), *middle*) the therapy start date (index = 0), and *right*) the end of the 5-year protocol (index = 10).

Table 1. Topics that are positively associated with a medication discontinuation event (statistically significant at the 0.05 level)

Topic	Word samples	HR	95% CI	p
122	*xray, marrow, ekg, echo, abd, density, mets, emg, tail, echocardiogram, scann, spur, egd, endoscopy, scan, exray, scans, ultra*	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, list*	1.171	(1.044, 1.313)	0.007
44	*reconstruction, lumpectomy, mastectomy, diep, flap, bilateral, hysterectomy, surgery, procedure, ovaries, port, replacement, surgeon, uterus, iud, plastic*	1.17	(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)	0.02
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*	1.13	(1.008, 1.267)	0.036
192	*suppression, blocking, hormonal, induced, bkm120, osteoporosis, estrogen, nutrition, diabetes, non, hormone, holistic, bc, specifically*	1.129	(1.003, 1.272)	0.044

Note: The topics are sorted according to *HR*. A larger *HR* suggests a greater risk of medication discontinuation. 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, we display the top 20 words; otherwise, we show all of the words in the topic.

increased risk of discontinuing medications. As one patient communicated when taking tamoxifen,

- "*I am having really bad* ***mood*** *... 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 surgery (#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*, ie, *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 [sic]? It has been almost a week since my surgery.*"

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

- "*I read* ***online*** *that if you are allergic to iron oxide you shouldn't take it. I think I am allergic to carbide. Would that be a problem?*"

Topics with decreased risk

Table 2 summarizes the topics with a decreased risk ($HR < 0$) 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 Femara on Saturday and* ***restarted*** *Arimidex instead.*"
- "*He saw my eye issues as* ***indicative*** *of paraeoplastic syndrome.*"

Patients may also mention terms related to decision making (#187, $HR = 0.833$) or expressing a 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* [sic] *for the next five years.*"

While there are 2 statistically significant topics (#109, #127) for which it is difficult to explain their association with medication discontinuation, most of our findings are in accordance with the literature, as discussed in the following section.

DISCUSSION

The findings of this study have several notable implications. First, the topics communicated by breast cancer patients appear to be effective indicators of medication discontinuation. In particular, we

Table 2. Topics that are negatively associated with a medication discontinuation event (statistically significant at the 0.05 level)

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

Note: The topics are sorted according to their HR. A smaller HR suggests a lower risk of medication discontinuation.

discover several topics that positively associate with medication discontinuation events and are supported by evidence in the literature. For example, gastrointestinal reactions such as nausea and vomiting (topic #13) have been shown to be risk factors of hormone therapy discontinuation.[32] Additionally, an echocardiogram (echo) or electrocardiogram (EKG) (topic #122) that was mentioned more by patients with medication discontinuation may be due to the fact that cardiac complications are also recognized as a severe side effects of AI.[40]

We further observe that patients who request professional suggestions (topic #136) or express gratitude to healthcare providers (topic #105) are less likely to discontinue hormonal therapy. This finding aligns with evidence that indicates respect for the advice of their caring physicians and family members can drive breast cancer patients to adhere to prescribed treatments.[41] Furthermore, there are 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.[42] Meanwhile, it has been shown that managing side effects can prevent discontinuation.[43] This may explain why mentions of drugs for treating side effects were associated with decreased risk of medication discontinuation.

In addition, we find that the average messaging rate correlates with an increased risk of medication discontinuation. While our messaging volume analysis shows that the probability of discontinuing medication rapidly decreases after 20 messages, as 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 of activity. This still suggests that messaging rates may be a useful indicator for predicting medication discontinuation.

Finally, it should be recognized that our data-driven approach requires minimal human efforts to analyze patient portal messages. For example, we can identify common usages of the messaging service in patient portals, which are in alignment with the findings of previous studies that rely on manually explored message content.[22] While it is appealing to apply a pre-trained word2vec model (eg, fit a model using a public dataset, such as the Google news corpus) for topic extraction, our experience with data from an online breast cancer forum suggest that training a specific word2vec model from a particular context is more effective.[6,7] Part of the reason is that many words applied in the clinical setting are unlikely to exist in the pre-trained corpus. By invoking unsupervised methods, it may be possible to design and deploy an automated alert system to monitor patients' messages. Such a system could lower the burden of healthcare providers, as well as assist in providing interventions that assist patients in 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 population was derived from a single institution, which may limit the generalizability of our findings. Second, we did not distinguish between medication discontinuations that were consultative (ie, a decision made at the recommendation of the prescribing clinician, with or without shared decision making) or autonomous (ie, a patient's independent decision). This is notable because the factors influencing these scenarios may differ. As such, we believe that a natural next step to this line of research is to determine the context around a discontinuation event (if such information is available) and incorporate the scenario as a factor in the model. Third, our medication discontinuation events were determined based on medication history recorded in the EMR system, which would probably result in some degree of immortal time bias.[44] Future work could consider incorporating prescription data and patient self-report to mitigate this time lag. Fourth, because of the heavy tailed distribution of message volume, our investigation may be biased towards users who more frequently use the messaging service in patient portals.

Other future work could investigate the effectiveness of other clustering algorithms (eg, spectral clustering[45]), as well as topic modeling[46] for topic extraction. While spectral clustering is more suitable for data that consist of connected components that are well separated (which is likely for natural language terms), topic modeling allows for a term to belong to multiple topics. We also 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. We suspect this approach may be feasible, as it has shown to be beneficial in automated phenotype discovery.[47,48] Furthermore, the message features (eg, messaging rates and topics) can be constructed as time-dependent covariates in survival analysis in order to capture their time-varying patterns and effects. Finally, we believe a predictive model can be built upon a combination of structured EMR data and PGMs. However, to make it practical for clinical decision support, further research efforts should evaluate 1) the effectiveness of combining these resources, 2) the extent to which care providers would accept such a model, and 3) what they could do with the information to impact change.

CONCLUSIONS

This investigation demonstrated the potential merits of patient-contributed information for learning medication discontinuation. Specifically, this study focused on the online portal messaging patterns and inferred topics from the messages communicated by breast cancer patients to care providers. We showed that there are associations between these factors and hormonal therapy medication discontinuation. The findings of this investigation suggest that PGMs can assist in predicting a breast cancer patients' likelihood of medication discontinuation and could serve as the basis of automated decision support tools to recognize patients at risk for discontinuation of a prescribed regimen.

FUNDING

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

CONTRIBUTORS

ZY and BM contributed to the idea of the work. ZY and MH performed the data collection. ZY, BM, and QC designed the methods. ZY carried out experiments. ZY, JW, and BM interpreted the results, drafted the paper, and edited, reviewed, and approved the final manuscript. QC and DF contributed to editing and reviewing the manuscript.

Conflict of interest statement. None declared.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

REFERENCES

1. Hull S, Warner J. Redefining our picture of health: towards a person-centered integrated care, research, wellness, and community ecosystem. In: *AMIA Informatics Educators Forum*. June 19, 2018; New Orleans, LA.
2. Dredze M, Cheng R, Paul MJ, *et al*. Healthtweets. org: a platform for public health surveillance using twitter. In: *AAAI Workshop on the World Wide Web and Public Health Intelligence*. July 27, 2014: 593–6; Quebec city, Canada.
3. Yin Z, Fabbri D, Rosenbloom ST, *et al*. A scalable framework to detect personal health mentions on Twitter. *J Med Internet Res* 2015; 17 (6): e138.
4. Yin Z, Chen Y, Fabbri D, *et al*. # PrayFordad: learning the semantics behind why social media users disclose health information. In: *Proceedings of the Twelth International AAAI Conference on Web and Social Media*. May 18, 2016: 456–65; Cologne, Germany.
5. De Choudhury M, Kiciman E, Dredze M, *et al*. Discovering shifts to suicidal ideation from mental health content in social media. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. May 7, 2016: 2098–110; San Jose, CA.
6. 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: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. May 16, 2017: 337–46; Montreal, Canada.
7. Yin Z, Xie W, Malin BA. 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.
8. VanDam C, Kanthawala S, Pratt W, *et al*. Detecting clinically related content in online patient posts. *J Biomed Inform* 2017; 75: 96–106.
9. Zhang S, Grave E, Sklar E, *et al*. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *J Biomed Inform* 2017; 69: 1–9.
10. Zhang S, Bantum E, Owen J, *et al*. Does sustained participation in an online health community affect sentiment? *AMIA Annu Symp Proc* 2014; 2014: 1970–9.
11. Qiu B, Zhao K, Mitra P, *et al*. Get online support, feel better-sentiment analysis and dynamics in an online cancer survivor community. In: *Proceedings of IEEE International Conference on Privacy, Security, Risk and Trust*. 2011: 274–81.
12. Zhao K, Yen J, Greer G, *et al*. Finding influential users of online health communities: a new metric based on sentiment influence. *J Am Med Inform Assoc* 2014; 21 (e2): e212–8.
13. Yin Z, Song L, Malin B. Reciprocity and its association with treatment adherence in an online breast cancer forum. In: *Proceedings of IEEE Symposium on Computer-Based Medical Systems*. June 22, 2017: 618–23; Thessaloniki, Greece.
14. Zhang S, O'Carroll Bantum E, Owen J, *et al*. Online cancer communities as informatics intervention for social support: conceptualization, characterization, and impact. *J Am Med Inform Assoc* 2017; 24 (2): 451–9.
15. Sarker A, Ginn R, Nikfarjam A, *et al*. Utilizing social media data for pharmacovigilance: a review. *J Biomed Inform* 2015; 54: 202–12.
16. Harris LT, Haneuse SJ, Martin DP, *et al*. Diabetes quality of care and outpatient utilization associated with electronic patient-provider messaging: a cross-sectional analysis. *Diabetes Care* 2009; 32 (7): 1182.
17. Harris LT, Koepsell TD, Haneuse SJ, *et al*. Glycemic control associated with secure patient-provider messaging within a shared electronic medical record. *Diabetes Care* 2013; 36 (9): 2726–33.
18. Price-Haywood EG, Luo Q, Monlezun D. Dose effect of patient–care team communication via secure portal messaging on glucose and blood pressure control. *J Am Med Inform Assoc* 2018; 25: 702–8.
19. Zhou YY, Garrido T, Chin HL, *et al*. Patient access to an electronic health record with secure messaging: impact on primary care utilization. *Am J Manag Care* 2007; 13 (7): 418–24.
20. Shimada SL, Hogan TP, Rao SR, *et al*. Patient-provider secure messaging in VA: variations in adoption and association with urgent care utilization. *Med Care* 2013; 51 (3 Suppl 1): S21–8.
21. Dumitrascu AG, Burton MC, Dawson NL, *et al*. Patient portal use and hospital outcomes. *J Am Med Inform Assoc* 2018; 25 (4): 447–53.
22. Shimada SL, Petrakis BA, Rothendler JA, *et al*. An analysis of patient-provider secure messaging at two Veterans Health Administration medical centers: message content and resolution through secure messaging. *J Am Med Inform Assoc* 2017; 24: 942–9.
23. Cronin RM, Fabbri D, Denny JC, *et al*. A comparison of rule-based and machine learning approaches for classifying patient portal messages. *Int J Med Inform* 2017; 105: 110–20.
24. Sulieman L, Gilmore D, French C, *et al*. Classifying patient portal messages using convolutional neural networks. *J Biomed Inform* 2017; 74: 59–70.
25. Pritchard KI, Hayes DF, Vora SR. Adjuvant endocrine therapy for non-metastatic, *hormone receptor-positive breast cancer*. September 6, 2018.

http://www.uptodate.com/contents/adjuvant-endocrine-therapy-for-non-metastatic-hormone-receptor-positive-breast-cancer. Accessed October 6, 2018.

26. Gotay C, Dunn J. Adherence to long-term adjuvant hormonal therapy for breast cancer. *Expert Rev Pharmacoecon Outcomes Res* 2011; 11 (6): 709–15.
27. Aiello Bowles EJ, Boudreau DM, Chubak J, *et al*. Patient-reported discontinuation of endocrine therapy and related adverse effects among women with early-stage breast cancer. *J Oncol Pract* 2012; 8 (6): e149–57.
28. Wu J, Kevin Z. Hormone therapy adherence and costs in women with breast cancer. *Am J Pharm Benefits* 2013; 5: 65–70.
29. Kemp A, Preen DB, Saunders C, *et al*. Early discontinuation of endocrine therapy for breast cancer: who is at risk in clinical practice? *Springerplus* 2014; 3 (1): 282.
30. Hershman DL, Kushi LH, Shao T, *et al*. Early discontinuation and nonadherence to adjuvant hormonal therapy in a cohort of 8, 769 early-stage breast cancer patients. *J Clin Oncol* 2010; 28 (27): 4120–8.
31. Bluethmann S, Murphy C, Tiro J, *et al*. Deconstructing decisions to initiate, maintain, or discontinue adjuvant endocrine therapy in breast cancer survivors: a mixed-methods study. *Oncol Nurs Forum* 2017; 44 (3): E101–10.
32. He W, Fang F, Varnum C, *et al*. Predictors of discontinuation of adjuvant hormone therapy in patients with breast cancer. *J Clin Oncol* 2015; 33 (20): 2262–9.
33. Freedman RA, Viswanath K, Vaz-Luis I, *et al*. Learning from social media: utilizing advanced data extraction techniques to understand barriers to breast cancer treatment. *Breast Cancer Res Treat* 2016; 158 (2): 395–405.
34. 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): 362–9.
35. 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.
36. Davies C, Pan H, Godwin J, *et al*. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. *Lancet* 2013; 381 (9869): 805–16.
37. 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): 459–78.
38. Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. In: *Adv Neural Inf Process Syst*. December 5, 2013: 3111–9; Stateline, NV.
39. Paul MJ, Dredze M. Discovering health topics in social media using topic models. *PLoS One* 2014; 9: e103408.
40. Valachis A, Nilsson C. Cardiac risk in the treatment of breast cancer: assessment and management. *Breast Cancer (Dove Med Press)* 2015; 7: 21.
41. Ell K, Vourlekis B, Xie B, *et al*. Cancer treatment adherence among low-income women with breast or gynecologic cancer. *Cancer* 2009; 115 (19): 4606–15.
42. 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.
43. Davey MP. Oral therapy: managing side effects can aid adherence. *Oncol Nurse Advis* 2012: 24–31.
44. Lévesque LE, Hanley JA, Kezouh A, *et al*. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010; 340: 907–11.
45. Schwartz HA, Eichstaedt JC, Kern ML, *et al*. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One* 2013; 8: e73791.
46. Blei DM, Edu BB, Ng AY, *et al*. Latent dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
47. Wang Y, Chen R, Ghosh J, *et al*. Rubik: knowledge guided tensor factorization and completion for health data analytics categories and subject descriptors. In: *Proceedings of ACM SIGKDD Int Conf Knowl Discov Data Min*. August 10, 2015: 1265–74; Sydney, Australia.
48. Kim Y, El-Kareh R, Sun J, *et al*. Discriminative and distinct phenotyping by constrained tensor factorization. *Sci Rep* 2017; 7: 1114.