Journal of Patient Experience
Volume 8: 1-7
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23743735211048058
journals.sagepub.com/home/jpx
SAGE

Voice Analysis of Cancer Experiences Among Patients With Breast Cancer: VOICE-BC

Ernest H Law[1], Maria J Auil[2,*], Patricia A Spears[3], Kiersten Berg[2,*], and Randall Winnette[1]

Abstract

Patient experience literature in early-stage breast cancer (eBC) is limited. This study used a mixed-methods approach to examine patient conversations from public online forums to identify and evaluate eBC-related themes. Among 60,000 eBC-related posts published September 2014–2019, text from a random subset of 15,000 posts was extracted and grouped into linguistically similar, mutually exclusive clusters using an advanced natural language processing (NLP) algorithm. Clusters were characterized using four quantitative metrics: *betweenness centrality* (linguistic similarity to other areas of the cluster network), *sentiment* (general attitude toward a topic), *recency* (average date of posts), and *volume* (total number of posts). This analysis represented 3906 unique users (67% and 33% obtained from cancer–specific and general health/nonhealth forums, respectively). Of the 27 clusters identified, most important were "discussing recurrence & progression," "understanding diagnosis & prognosis," and "understanding cancer, biomarkers, and treatments." Several major themes related to recurrence risk, diagnosis, monitoring, and treatment were identified. Additional emphasis on communicating the disease recurrence risk and shared decision-making could strengthen patient-clinician partnerships.

Keywords

cancer, patient perspectives/narratives, patient expectations, clinician-patient relationship

Introduction

In the United States, women have a 13% average lifetime risk of developing breast cancer (1). Stage I–III breast cancer accounted for 177,966 of the 196,094 total cases of invasive breast cancer reported between 2010–2013 (2). After 4 years of follow-up, survival rates were >95% in patients with stage I disease, >80% among those with stage II disease, and >50% among those with stage III disease (2). Cancer is associated with substantial burden of symptoms caused by the disease itself, comorbidities, and/or treatments for the disease (3). Treatment for early-stage breast cancer (eBC) improves disease-free survival, overall survival, and patients' quality of life (4).

Understanding the needs of a patient during their breast cancer journey is important to improve patient care and quality of life (5). Treatment plans for eBC can be highly individualized based on the patient's disease characteristics and needs (4,6). As such, patients' preferences regarding treatments and their engagement in the decision-making process are essential (6). The current understanding of patients' unmet needs is largely based on findings from studies with traditional designs and methods (eg, surveys, focus groups, interviews) (7,8).

The recent explosion of online resources, including online forums, represents a potential rich information source to complement the limited existing literature on the needs of patients with eBC. An estimated 56.3 million people in the United States use the internet to search for information about chronic disease (9). Moreover, patients' treatment choices are then influenced by the information they gather (10). Recent guidance from the US Food and Drug

[1] Patient & Health Impact, Pfizer Inc, New York, NY, USA
[2] Quid Inc., San Francisco, CA, USA
[3] Research Patient Advocate, Raleigh, NC, USA

*At the time of the study.

Corresponding Author:
Ernest H Law, PharmD, PhD, ACPR, Global Health Economics & Outcomes Research (Breast), Oncology, Patient & Health Impact | Pfizer Inc, 235 E 42nd Street, New York, NY 10017, USA.
Email: ernest.law@pfizer.com

Administration on patient-focused drug development cites research using social media tools as a potentially useful supplemental resource to traditional research tools (11) Online patient forums have become increasingly popular for patients to seek information and support (12). Using online forums, patients with cancer can discuss their experiences, interact with and support other patients, and express and address their concerns (13). These forum posts offer insights into the challenges that patients face throughout the patient journey of diagnosis, disease progression, and recovery (13).

This study aimed to assess the patient experience as described by patients with eBC in public conversations using advances in natural language processing (NLP) and open source online data to identify patient-reported unmet needs.

Methods

Study Design

A mixed-methods approach to a patient voice analysis was used to evaluate patient conversations via publicly available online forums to identify and assess themes related to the unmet needs and burdens of eBC disease. A multistage approach was used for data collection, data visualization and clustering analysis, and voice and engagement analysis.

Data Sources

Data sources were publicly available online forums that were selected to account for different types of patients visiting distinct sources, representing a more accurate portrayal of online patient behaviors. The forums were classified as breast cancer–specific (BreastCancer.org), general cancer (American Cancer Society Cancer Survivors Network [CSN], Cancer Compass), general health (Inspire), and nonhealth-related (Reddit). Only 2018 and 2019 data were collected from the breast cancer–specific forums. Data from 2014 to 2019 were collected from the other types of forums.

All English-language posts published between September 2014 and September 2019 that referred to eBC (stages I–III) identified based on the formulated ontologies were included. Ontologies, a linguistic strategy used to organize language, were developed to identify posts relevant to eBC and its effect on the patient. The study team created a lexicon that referred to patients with eBC (Online Resource 1). Cancer subtype and/or stage of the patient were inferred from the content of their posts, with patients often self-identifying their stage and/or genetic mutation (eg, “I am stage 1a, HER 2+…” or “I was just diagnosed with stage 1”). The eBC population was defined in this study as patients with breast cancer stages I to III, with no exclusions made based on breast cancer subtype. To define this population, a 2-step search strategy was implemented by selecting self-identifying posts based on the ontology of terms that describe the eBC population and searching all posts from authors of self-identifying posts and excluding posts from anonymous authors.

Forum (name and URL) and post (title, date, URL, text) information from selected online threads were included. Posts by forum moderators, administrators, or sponsors were excluded, and no personally identifiable data were collected. A random sample of 100 screened posts was assessed by 2 reviewers on the study team who independently evaluated whether the post met the inclusion criteria to ensure that the ontologies were appropriately identifying eligible posts.

Cluster Identification

A total of 62,237 English-language posts published between September 2014 and September 2019 by patients with breast cancer referring to eBC (stages I–III) were extracted from 5 publicly available online forums. Of these, 15,000 posts were randomly selected, and post text and date were analyzed using an advanced NLP algorithm (Quid Inc., San Francisco, CA) to group linguistically similar posts into mutually exclusive clusters that exist within a network. Patient posts were assigned to a high-dimensional vector space, and linguistically similar communities that exist within those data points were identified. An interactive visual network grouped posts with similar language into clusters and mapped the relationship between clusters. Each cluster represented a patient topic of conversation. To detect themes in the data, the NLP algorithm used the Louvain method (14). This method grouped posts into common themes based on shared similar language and was performed numerous times to allow multiple levels of themes per cluster.

Cluster Naming

Clusters were assigned names based on keywords identified by the NLP algorithm and with input from the study team. The study investigators’ backgrounds are summarized in Online Resource 2. Unique concepts were prioritized (e.g., diagnosis). Words that frequently occurred but did not add meaning (eg, the, this, if) were ignored. The keywords suggested by the NLP algorithm were reviewed by the study team to refine names assigned to each cluster. The investigators evaluated a random sample of 30 original posts for each cluster to ensure that the algorithm appropriately identified the core themes.

Patient Engagement Metrics

Clusters were characterized by four scores representing *betweenness centrality* (degree to which language within a cluster is similar to other areas within the cluster network), *sentiment* (general attitude [positive to negative] toward a topic), *recency* (average date of posts), and *volume* (total number of posts within a cluster). The four scores were averaged to obtain a summary score that was used to rank each cluster by overall importance. Metrics were rescaled to allow assessment of relative importance, with the least important cluster rescaled to 1 and all other clusters assigned a relative score (Online Resource 3).

All analyses were conducted at the level of user posts. Results were summarized descriptively with mean values, medians, interquartile ranges, minimum/maximum ranges, and SDs of continuous variables of interest and frequency distribution for categorical variables.

Results

The 15,000 posts included in the analysis were represented by 3,906 unique users. Of these, 24% were obtained from breast cancer–specific forums (BreastCancer.org), 43% from general cancer forums (American Cancer Society CSN), and 33% from general health/nonhealth-related forums (Inspire and Reddit; Table 1). A total of 27 clusters were identified within the cluster network (Figure 1). The characteristics of forum posts generally were similar across cluster networks (Online Resource 4). Examples of quotes from forums in each cluster are shown in Online Resource 5.

Cluster themes were ranked by overall and component engagement scores (Table 2). Based on average scores across engagement metrics, the most important clusters overall were "discussing recurrence and progression" and "understanding diagnosis and prognosis." The clusters that were highest in volume were "emotional support from peers," "surgical procedures," and "understanding diagnosis and prognosis." "Discussing recurrence and progression," "anxiety before treatments," and "skepticism of healthcare system" were most central within the network. The clusters with the highest sentiment scores were "interpreting bloodwork," "understanding cancer, biomarkers, and treatments," and "experience with anastrozole." The clusters that had the highest recency scores were "discussing recurrence and progression" "family history and genetic testing," and "clinical trials and research."

Table 1. Characteristics of Online Forum Posts (N = 15,000).

Characteristic	Analytic Sample
Total unique users, n	3906
Unknown user, n	56
Online forum type, n (%)	
Breast cancer–specific	935 (24)
General cancer	1671 (43)
General health	586 (15)
Nonhealth	714 (18)
Post, y, n (%)	
2014	2663 (18)
2015	2189 (15)
2016	1500 (10)
2017	1413 (9)
2018	4254 (28)
2019	2981 (20)
Post text length, characters	
Mean (SD)	486 (393)
Median (IQR)	410 (369)
100–1000, n (%)	11,256 (75)
101–3000, n (%)	3426 (23)
>3000, n (%)	318 (2)

Abbreviations: IQR, interquartile range.

Based on engagement scores, posts that focused on "emotional support from peers" were approximately 5.3 times the volume of those related to "fear of chemotherapy" (Figure 2A). "Discussing recurrence and progression" was approximately 1.3 and 2.1 times more central to conversations within the network than "concerns with costs and insurance" and "menopause concerns and hot flashes," respectively (Figure 2B). On average, "sharing resources" was discussed with 3.3 times and 4.7 times more positive sentiment than "family history and genetic testing" and "interpreting bloodwork," respectively (Figure 2C). Compared with "inconvenience with ports" and "role of nutrition and weight," "discussing recurrence and progression" was approximately 5.3 and 2.2 times, respectively, more recently posted (Figure 2D).

Discussion

Using a novel application of NLP techniques to capture and process a large amount of patient-reported information from online forums, this study identified major themes discussed by patients with eBC. Overall, the most important clusters based on average scores across engagement metrics were "discussing recurrence and progression," "understanding diagnosis and prognosis," and "understanding cancer, biomarkers, and treatments." "Emotional support from peers," "surgical procedures," and "understanding diagnosis and prognosis" were highest in volume. "Discussing recurrence and progression," "anxiety before treatments," and "skepticism of healthcare system" were most central within the network. A strength of this study is that unsolicited information was collected from online patient forums without the intent of being used for research purposes, mitigating common research biases (eg, social desirability) that traditional qualitative research methods (eg, facilitated focus groups) introduce.

Literature evaluating the unmet needs of patients with eBC is sparse. A previous survey study conducted in the United Kingdom estimated the prevalence and nature of unmet needs in patients with eBC after they had completed initial treatment (15). Based on a questionnaire completed by patients, 61% of patients had ≥1 unmet need and 18% had ≥5 unmet needs (15). The most frequently reported unmet needs were physical (55%) and emotional (24%), with hot flashes being the most common unmet physical need and worry, fear, or anxiety being the most common emotional need; practical (6%), family (5%), and spiritual (4%) needs were reported less often (15). In contrast, "menopause concerns and hot flashes" was ranked 13th overall in importance in the current study. However, consistent with findings from the current study, this study showed that a substantial proportion of patients with eBC have needs that are unaddressed and suggested that providing patients with access to information and resources is important. Findings from the present study complement existing literature by summarizing large quantities of unsolicited information on

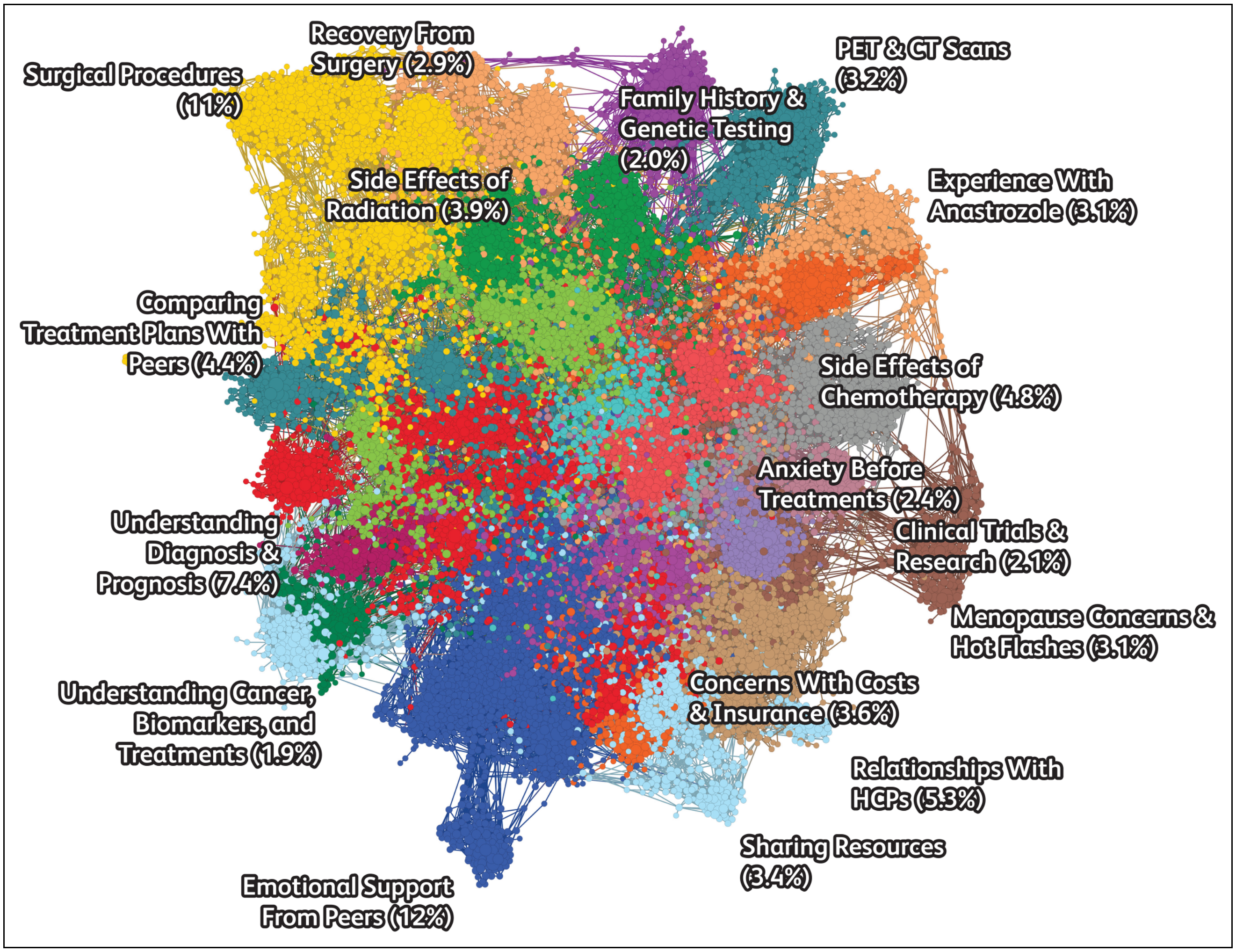

Figure 1. Visualization of cluster network. Each node represents 1 online forum post related to eBC experience and/or treatment; each unique color designates a different cluster, representing a group of posts that share similar language depicting a common theme; distance between clusters represents how closely the language is similar. Abbreviations: CT, computed tomography; eBC, early-stage breast cancer; HCP, healthcare professional; PET, positron-emission tomography.

patients' journey through eBC. Moreover, using NLP, an inductive approach was utilized, where themes could emerge organically through the identification of common linguistic patterns.

In contrast to the paucity of information on eBC, numerous studies exist assessing the needs of patients with breast cancer in general. A survey study of patients with breast cancer treated at either a university hospital or a rehabilitation center showed that more information on specific cancer-related issues, treatments, and prognosis were the most emphasized needs (16). Additionally, a survey study of patients with breast cancer across the first 5 years of diagnosis showed that the information needs of patients with breast cancer were high throughout the 5-year period, with information regarding the disease rated as most important (17). Similarly, another survey study showed that unmet information needs were high throughout the course of patients' cancer treatment, but the information patients' desired changed over the course of treatment (18). Unmet information needs related to "side effects and medication" and "medical examination results and treatment options and social issues" were highly reported among patients with newly diagnosed breast cancer, increased during follow-up treatment, and remained high during the posttreatment period (18). Of note, these previous studies assessing patients' unmet needs were all based on questionnaires.

To the authors' knowledge, only 1 previous study has used the specific NLP approach described in the this report. In that study, 500,000 comments from public online forums were analyzed from 2010 to 2018 to gain insight into the unmet needs of patients with various chronic diseases, including breast cancer (19). A total of 8 unmet patient needs were identified and ranked in the following order: "understanding medication side effects and impact on daily life," "coping with living with the condition," "understanding the disease and its causes," "identifying and mitigating

Table 2. Cluster Themes Ranked by Overall and Component Engagement Scores.

Cluster	Rank	Average	Volume	Centrality	Sentiment	Recency
Discussing recurrence & progression	1	0.908	−0.514	1.968	0.607	1.569
Understanding diagnosis & prognosis	2	0.536	1.351	−0.716	0.642	0.867
Understanding cancer, biomarkers, & treatments	3	0.484	−0.667	0.726	1.882	−0.006
Relationships with HCPs	4	0.472	0.569	1.077	−0.638	0.880
Emotional support from peers	5	0.472	3.327	0.118	−1.259	−0.300
Surgical procedures	6	0.436	2.676	−0.322	−0.195	−0.414
Interpreting bloodwork	7	0.416	−0.816	0.739	2.028	−0.288
Detection, early symptoms, & self-diagnosis	8	0.286	−0.334	0.772	0.919	−0.213
Comparing treatment plans with peers	9	0.261	0.251	1.104	−0.656	0.345
Side effects of radiation	10	0.242	0.054	0.063	−0.075	0.925
Side effects of chemotherapy	11	0.205	0.412	0.403	0.642	−0.636
Anxiety before treatments	12	0.144	−0.484	1.862	−1.373	0.570
Menopause concerns & hot flashes	13	0.107	−0.218	−0.482	0.718	0.409
Clinical trials & research	14	0.101	−0.595	−1.066	0.969	1.098
Skepticism of HC system	15	−0.022	−0.595	1.467	−1.127	0.167
Family history & genetic testing	16	−0.045	−0.625	−1.611	0.591	1.466
Experience with tamoxifen	17	−0.130	−0.448	−0.837	0.938	−0.172
Concerns with costs & insurance	18	−0.185	−0.043	0.990	−0.973	−0.716
Experience with AC-taxol regimen	19	−0.213	−0.434	−1.127	−0.227	0.936
Experience with anastrozole	20	−0.324	−0.218	−0.083	1.146	−2.140
Seeking second opinions	21	−0.343	−0.531	−0.054	−1.303	0.518
Recovery from surgery	22	−0.373	−0.281	−0.505	−0.114	−0.590
Role of nutrition & weight	23	−0.439	0.270	−0.790	0.146	−1.384
PET & CT scans	24	−0.487	−0.198	−1.307	−0.259	−0.183
Sharing resources	25	−0.569	−0.129	−0.794	−1.660	0.308
Fear of chemotherapy	26	−0.835	−0.941	−1.147	−1.002	−0.249
Inconveniences with ports	27	−1.105	−0.841	−0.449	−0.366	−2.765

Abbreviations: AC, adriamycin; CT, computed tomography; HC, healthcare; HCP, healthcare professional; PET, positron-emission tomography.
Cell shading denotes the scale of values where darker cells indicate a higher value for a specific engagement score.

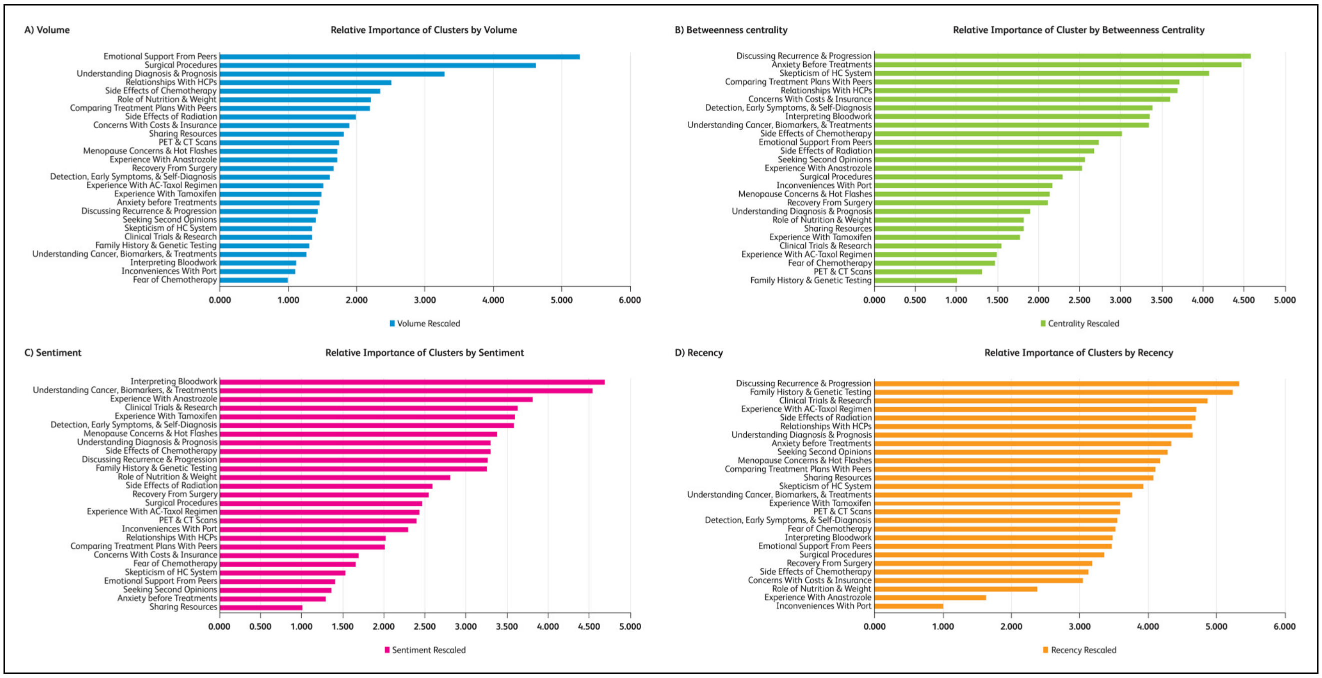

Figure 2. Relative importance of cluster themes by engagement scores. Abbreviations: CT, computed tomography; HCP, healthcare professional; PET, positron-emission tomography.

symptoms," "connect with other patients to share," "managing disease progression," "more and better treatment options," and "understanding diagnostic tests, procedures, and results" (19). Overall, these needs were consistent with the most important clusters identified in the present study, which also focused on understanding the disease, including recurrence, progression, diagnosis, prognosis, biomarkers, and treatments. Together, these results highlight the opportunity for healthcare providers to collaborate with their patients to address these unmet needs.

Given the nature of qualitative analyses, this study had several limitations. This analysis included a large study size and number of patient conversations, but had the potential for selection bias. Individuals who are predisposed to express their experience online may be systematically different from the broader population of patients with breast cancer, limiting the generalizability of these results. Additionally, clinical aspects (eg, disease stage, treatments) of the disease were self-reported by patients and could not be evaluated for accuracy. The interpretation and naming of patient topics/clusters were subject to the perspectives of the study team members. An assumption was made for this study that the posts in the online forums were from real patients diagnosed with breast cancer. Moreover, some posts may have been counted several times in the descriptive analyses because patients may have discussed multiple phases of their disease journey in a single post. Because of the volume of data collected, the level of assessment that is often seen in traditional mixed-methods or qualitative research was not feasible. However, to ensure transparent and systematic handling of the data, a combination of NLP algorithms and researcher assessment/data validation was utilized.

Conclusions

This study represents a novel application of NLP techniques to capture and process a large amount of patient-reported information surrounding experiences in eBC. Several major themes in the patient experience were identified, including the importance of understanding disease recurrence risk, diagnosis, monitoring, and treatment options. These findings suggest that the patient-physician partnership may be strengthened by shared decision making and a greater emphasis on communicating the risk of breast cancer recurrence.

Acknowledgments

This study was sponsored by Pfizer Inc. Editorial support was provided by Anny Wu, PharmD, of ICON plc (North Wales, PA, USA) and was funded by Pfizer Inc.

Authors' contributions

All authors provided substantial contributions to the conception or design of the work, or the acquisition, analysis, or interpretation of data for the work, drafted the work or revised it critically for important intellectual content, provided final approval of the version to be published, and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Availability of data and material

Upon request, and subject to certain criteria, conditions and exceptions (see https://www.pfizer.com/science/clinical-trials/trial-data-and-results for more information), Pfizer will provide access to individual de-identified participant data from Pfizer-sponsored global interventional clinical studies conducted for medicines, vaccines and medical devices (1) for indications that have been approved in the US and/or EU or (2) in programs that have been terminated (i.e., development for all indications has been discontinued). Pfizer will also consider requests for the protocol, data dictionary, and statistical analysis plan. Data may be requested from Pfizer trials 24 months after study completion. The de-identified participant data will be made available to researchers whose proposals meet the research criteria and other conditions, and for which an exception does not apply, via a secure portal. To gain access, data requestors must enter into a data access agreement with Pfizer.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Yes, EH Law and R Winnette are employees of Pfizer Inc and have stock options in Pfizer Inc. MJ Auil and K Berg have no conflicts of interest to disclose. PA Spears is an advisor/consultant for Pfizer Inc.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Pfizer

ORCID iD

Ernest H Law https://orcid.org/0000-0002-6111-8008

Supplemental Material

Supplemental material for this article is available online.

References

1. American Cancer Society. About breast cancer, 2020. Accessed 28 September, 2020. https://www.cancer.org/cancer/breast-cancer/about.html
2. Howlader N, Cronin KA, Kurian AW, Andridge R. Differences in breast cancer survival by molecular subtypes in the United States. Cancer Epidemiol Biomarkers Prev 2018;27(6):619-26.
3. Cleeland CS. Symptom burden: multiple symptoms and their impact as patient-reported outcomes. J Natl Cancer Inst Monogr 2007(37):16-21.
4. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: eSMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 2019;30(8):1674.
5. Wen KY, Gustafson DH. Needs assessment for cancer patients and their families. Health Qual Life Outcomes 2004;2(1):11.

6. Burstein HJ, Curigliano G, Loibl S, Dubsky P, Gnant M, Poortmans P, et al. Estimating the benefits of therapy for early-stage breast cancer: the St. Gallen international consensus guidelines for the primary therapy of early breast cancer 2019. Ann Oncol 2019;30(10):1541-57.
7. Todd BL, Feuerstein M, Gehrke A, Hydeman J, Beaupin L. Identifying the unmet needs of breast cancer patients post-primary treatment: the cancer survivor profile (CSPro). J Cancer Surviv. 2015;9(2):137-60; quiz 151–160.
8. Ellegaard MBB, Grau C, Zachariae R, Bonde Jensen A. Fear of cancer recurrence and unmet needs among breast cancer survivors in the first five years. A cross-sectional study. Acta Oncol. 2017;56(2):314-20.
9. Meier A, Lyons E, Frydman G, Forlenza M, Rimer B. How cancer survivors provide support on cancer-related internet mailing lists. J Med Internet Res. 2007;9(2):e12.
10. Meric F, Bernstam EV, Mirza NQ, Hunt KK, Ames FC, Ross MI, et al. Breast cancer on the world wide web: cross sectional survey of quality of information and popularity of websites. Br Med J. 2002;324(7337):577-81.
11. Administration UFaD. Patient-focused drug development: collecting comprehensive and representative input. 2020. Accessed 16 November, 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-collecting-comprehensive-and-representative-input
12. Elhadad N, Zhang S, Driscoll P, Brody S. Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. AMIA Annu Symp Proc 2014;2014:516-25.
13. Harkin LJ, Beaver K, Dey P, Choong K. Navigating cancer using online communities: a grounded theory of survivor and family experiences. J Cancer Surviv 2017;11(6):658-69.
14. Quid Inc. Summarized network graph for network graph for semantic similarity graphs of large. 2016.
15. Capelan M, Battisti NML, McLoughlin A, Maidens V, Snuggs N, Slyk P, et al. The prevalence of unmet needs in 625 women living beyond a diagnosis of early breast cancer. Br J Cancer 2017;117(8):1113-20.
16. Salminen E, Vire J, Poussa T, Knifsund S. Unmet needs in information flow between breast cancer patients, their spouses, and physicians. Support Care Cancer 2004;12(9):663-8.
17. Sheehy EM, Lehane E, Quinn E, Livingstone V, Redmond HP, Corrigan MA. Information needs of patients with breast cancer at years one, three, and five after diagnosis. Clin Breast Cancer 2018;18(6):e1269-75.
18. Halbach SM, Ernstmann N, Kowalski C, Pfaff H, Pfoertner TK, Wesselmann S, et al. Unmet information needs and limited health literacy in newly diagnosed breast cancer patients over the course of cancer treatment. Patient Educ Couns 2016;99(9):1511-8.
19. Tewarie B, Bailey V, Rebarber M, Xu J. Unmet needs: hearing the challenges of chronic patients with artificial intelligence. NEJM Catalyst 2019;9(1).