Breast Cancer Research

RESEARCH

Open Access

# Cancers missed, women dismissed yet persist: natural language processing of online forums

Kathleene T. Ulanday[1,2], Maxim Topaz[3], Jeanette Shekelle[1], Marley Gibbons[1,6], Desiree Walker[4], Paula M. Castaño[5], Amanda Nixon[4], Stacy Lewis[4], Mary Beth Terry[1,2,5] and Lauren C. Houghton[1,2,5*]

## Abstract

**Objective** To identify gaps and delays in the detection of early onset cancer.

**Methods** We examined firsthand experiences shared on an online discussion board hosted by the Young Survival Coalition—an advocacy group for young adults diagnosed with breast cancer—spanning the years 2009 to 2019. We used natural language processing to detect codes: "first signs and symptoms," "steps to diagnosis," "healthcare interactions," "patient-provider-system feelings," and "staging/type." In the training dataset, we used qualitative content analysis to code text from 750 of the forum's 571,914 posts. We developed and evaluated automated approaches to quantify the proportion of codes in all posts. Lastly, we qualitatively reviewed the classified posts to identify areas for improvement along the clinical pathway.

**Results** The vast majority (81%) of young adults self-detected their breast cancer rather than the cancer being detected through a clinical breast exam. Young adults (70%) were dissatisfied with their care because they encountered delays at three crossroads along the clinical pathway: 1) whether the clinician ordered tests or dismissed the individual as too young; 2) whether imaging modalities were sensitive or not; 3) whether a biopsy confirmed or missed the cancer. Mental health challenges and parenting pressures compounded these delays**.** True positive cases who experienced these delays strongly encouraged their peers to self-advocate, persist and insist on further testing until diagnosed accurately.

**Conclusion** Dismissal and delays in diagnosis of early onset breast cancer mean potentially worse prognosis since later stage cancers are more aggressive with fewer treatment options. The perspectives from survivors highlight the need for more research informing early detection in young adults by considering breast awareness, use of MRI and ultrasound, biopsy referrals for exhibited breast symptoms in the absence of positive imaging, and sociomedical support for individuals in their role as current or future parent.

## Background

Breast cancer is the most common non-skin cancer malignancy in U.S. women, and its incidence, particularly for non-localized disease, has increased alarmingly in women under age 40 years, with an annual 3.6% increase in risk. [1] Since 1996 [2], only 12% of younger adults have a positive family history of breast cancer, indicating a growing trend among those without a familial predisposition.

*Correspondence:
Lauren C. Houghton
lh2746@cumc.columbia.edu
[1] Columbia University Mailman School of Public Health, 722 West 168th Street, New York, NY 10032, USA
[2] Columbia University Herbert Irving Cancer Center, New York, NY, USA
[3] Columbia University School of Nursing, New York, NY, USA
[4] Young Survival Coalition, New York, NY, USA
[5] Columbia University Irving Medical Center, New York, NY, USA
[6] PATH, Washington DC, USA

Current screening strategies do not address early detection in individuals under 40, particularly if they do not have a family history of breast cancer. Yet with the increasing rates of early onset breast cancer, the proportion of young women without a family history will continue to grow. The U.S. Preventive Services Task Force (USPSTF) and the American Cancer Society (ACS) recommend that average-risk individuals start routine mammographic screening between ages 40 and 50. [3, 4] In 2024, the USPSTF decided to decrease the screening age from 50 to 40 years. [5] There are no other mammographic screening guidelines for younger adults, although the American College of Obstetrics and Gynecology (ACOG) recommends clinicians counsel individuals about breast awareness starting at age 25. [6] Without any screening in place for young adults and with rates among this age group on the rise, we need to develop new guidelines to detect early onset cancer. Learning from young adults affected with early onset breast cancer provides one helpful perspective to inform future guidelines.

To understand young adults' experiences navigating breast cancer, we analyzed an online community's [7–9] content using natural language processing and thematic coding to characterize the distribution of constructs along the clinical pathway ("first signs and symptoms," "steps to diagnosis," "healthcare interactions," "patient-provider feelings," and "staging/type") and to gain qualitative insight from their experiences using thematic coding. Identifying such patterns from young adults who have lived through the experience from detection to diagnosis is essential in identifying gaps in clinical care for young adults at risk for early onset cancer and moving towards better screening and diagnostics.

## Methods

### Study sample and data extraction

Young Survival Coalition (YSC, https://www.youngsurvival.org/) advocates for individuals diagnosed with breast cancer before age 40 across the U.S. and provides support through education and community building. Since 2009, YSC hosted an online discussion board in English where young adults posted questions and provided mutual support. The discussion board consisted of 571,914 posts in 43,112 threads, published between March 2009 and December 2019. The "Newbies" thread is where individuals joining the forum first told their breast cancer story.

The Columbia University Internal Review Board approved the ethical conduct of this study (Protocol # AAAS2862).

### Quantitative and qualitative methods

Figure 1 summarizes the methodological steps including manual coding, natural language processing (NLP) and machine learning classification (algorithm testing, parent, and child code classification), and the post-NLP qualitative review.

#### *Manual coding*

Two study team members (JS and MG) manually coded text initially for four parent codes along the clinical pathway: "first signs and symptoms," "steps to diagnosis," "healthcare interactions," and "staging/type". We

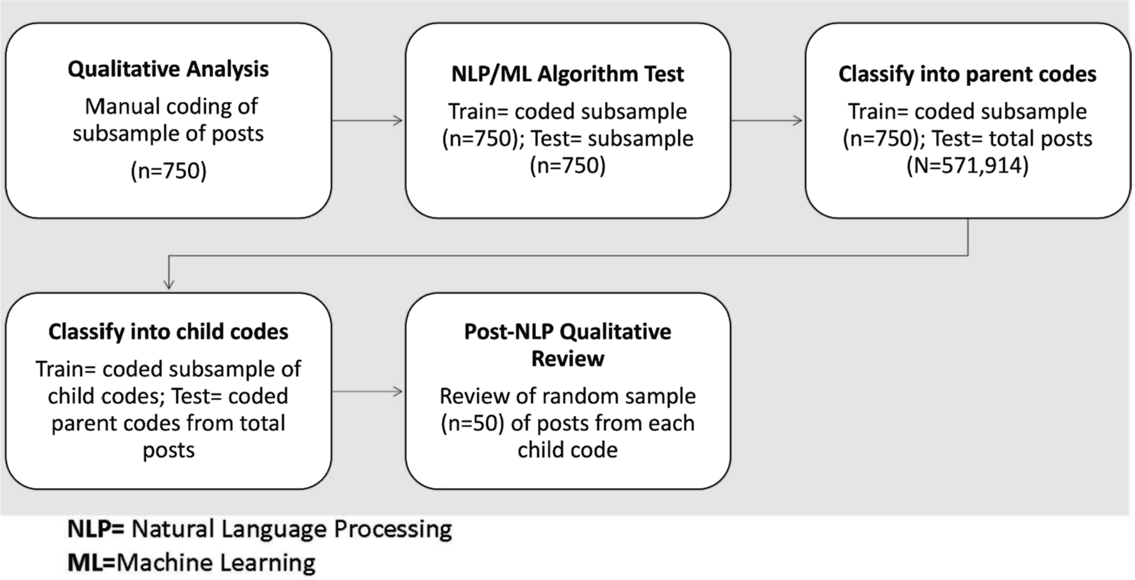
Qualitative Analysis
Manual coding of subsample of posts
(n=750)
NLP/ML Algorithm Test
Train= coded subsample (n=750); Test= subsample (n=750)
Classify into parent codes
Train= coded subsample (n=750); Test= total posts (N=571,914)
Classify into child codes
Train= coded subsample of child codes; Test= coded parent codes from total posts
Post-NLP Qualitative Review
Review of random sample (n=50) of posts from each child code
NLP= Natural Language Processing
ML=Machine Learning

**Fig. 1** Methods flow chart. NLP = Natural Language Processing, ML = Machine Learning

randomly selected one hundred posts from the "Newbies" thread to code sentence-by-sentence (approximately 1,370 sentences). Upon review, we added a fifth parent code: "patient-provider-system feelings" to capture how women felt about their healthcare interactions (examples of each code in Table 1a-e).

Next, we selected 650 more "Newbies" posts (total 750) for coding (approximately 6,000 sentences). Inter-annotator agreement was monitored throughout training, and training was terminated when a coder had achieved a 0.69 Kappa (agreement statistic). An adjudicator.

(LCH) reviewed all coded sentences, resolved differences between coders, and made final decisions over sentence codes. A fourth researcher (KU) further assessed sentences under each parent code for child and grandchild codes (Fig. 2). We then re-aggregated coded sentences into their original post by post ID.

### *Classification via natural language processing and machine learning*

We applied NLP and machine learning approaches to classify the dataset of 571,914 posts into topics. We used the open-source analytics platform KNIME for all classifications. [10] We used 750 coded posts to train and test machine learning algorithms. The preprocessing phase of our text data included tokenization, stemming, and removal of stop words, which are necessary to break down the text, streamline variant forms of a word, and reduce data dimensionality. The preprocessed data were then transformed into a 'bag of words'; this representation of text disregards sequence but captures word frequency, offering computational efficiency, while the latter encapsulates richer semantic information, capturing context and associations between terms. We then partitioned data into a training (70%) and test set (30%). Next, we implemented three machine learning models (i.e., support vector machine (SVM), random forest (RF), and decision tree (DT) to classify the data. For each model, we calculated the F-measure, the harmonic mean of the classification model's positive predictive value and sensitivity and averaged it across the five codes for the three machine learning models. We identified and selected the model with the best F-measure to classify our data.

We implemented the best-performing machine learning model to classify all posts for the presence of each of the five parent codes separately, using the manually coded data as the training set. We repeated the process to further classify each parent code into child codes.

### Statistical analysis

We applied the classification model to the entire dataset, then calculated the frequency and distribution of the parent codes. Next, we applied the classification model to each parent code dataset, then calculated the frequency of child codes. The natural language processing classified more posts under the parent codes than what were relevant to child codes and so the frequencies do not add to 100%. Moreover, the same post could be classified under different codes and so the classifications are not mutually exclusive.

### Post-natural language processing qualitative review

The NLP identified posts with relevant text for qualitative review. Two researchers (LH and KU) reviewed a random subset (n = 50) of posts under each parent code for further thematic coding, [11] which entails closely reviewing the text to identify common topics, ideas, and patterns of meaning that come up repeatedly.

## Results

In Table 1, we used joint display to present the distribution of parent and child codes with corresponding quotes that explain the underlying meaning of each code.

### Natural language processing classification

In the training sample, we manually classified 16% of posts for the presence of "first signs and symptoms," 25% for "steps to diagnosis," 39% for "healthcare interactions," 17% for "patient- provider feelings," and 48% for "stage at diagnosis" (Table 1). The average F-measure across codes was 79%, 77%, and 72% for the SVM, RF, and DT models, respectively; therefore, we used the SVM model to classify the training data and a subsequent larger dataset (Table 2). The most prevalent codes in the larger dataset of posts were "stage at diagnosis" (12.6%) and "healthcare interactions" (12.3%) followed by "patient-provider feelings" (5.6%), "steps to diagnosis (5.5%) and "first signs and symptoms" (1.5%) (Table 1).

Among the "first signs and symptoms" ($n = 3,266$), 81% of posts were about self-detection of either lumps (56.5%) or other breast and health changes (25%), and about 17% of posts discussed provider-detected cancers (Table 1). Among posts coded for "Steps to diagnosis" ($n = 31,640$), a majority mentioned starting with imaging or clinical exams (66.5%). Out of 24,648 posts classified as "patient-provider feelings", 70% described having either neutral or negative feelings towards their providers. Posts about "stage at diagnoses" ($n = 71,879$) disclosed being at Stage 4 (7.3%), followed by Stage 0 (7.2%), Stages 2–3 (5.5%), and Stage 1 (4.8%), while others mentioned invasive cancer diagnoses (7.3%).

### Post-natural language processing qualitative review

#### *First signs and symptoms*

Many individuals began their initial posts of their breast cancer journey describing the first sign or symptom

**Table 1** Classification of young survival coalition online forum (2009–2019) into parent and child codes

| | **n** | **%** | **Quote** |
|---|---|---|---|
| **"Newbies" posts manully clasified into 5 parent codes (n = 750)** | | | |
| First signs and symptoms | 120 | 16 | a. My cancer was picked up by a pain I was having that felt like nursing too |
| Steps to diagnosis | 187.5 | 25 | b. An ultrasound, two biopsies, an MRI and a CT scan later I know that I have a 5.3 cm tumor in my right breast |
| Healthcare interactions | 292.5 | 39 | c. I feel very comfortable with my dr and just hope and pray that he is steering me in the right direction." |
| Patient-provider feelings | 127.5 | 17 | d. I remember my onc telling me that statistics don't really matter because I'm only concerned about the outcome of my one case, not all the cases that the statistics are based on |
| Stage at diagnosis | 360 | 48 | e. ER + PR +, HER2 -, BRCA1 & BRCA2 both negative YAY I think." |
| ***All YSC posts classified by support vector machine algorithm into parent codes (n = 571,914)*** | | | |
| First signs and symptoms | 8,605 | 1.5 | f. Well I just want to say that if I had not found the lump myself I would probably be stage IV by now. I just happened to be taking a shower one day and while washing I found the lump. It seemed to appear overnight (although I'm sure it didn't) one day nothing the next day large lump in my left breast. I think anything that discourages women from doing self exams is terrible. I was 34 when I found the lump not old enough to even qualify for a free yearly mammo |
| Steps to diagnosis | 31,640 | 5.5 | g. Found lump myself, was in GP's office the next day. Was told that it was probably a cyst but gave me a referral for an ultrasound to see if a solid mass. U/S 2 weeks later, tech scanned mass, then started checking my nodes. That's when I had an idea something was wrong. I was sent for a mammogram immediately, which they compared to my baseline mammo. Radiologist recommended core biopsy. Done 2 days later on a Friday—was told that 80% were benign…Monday—GP's office called and asked me to come in that day. At that point, I knew it was cancer as they don't have you come in for benign results |
| Healthcare interactions | 70,145 | 12.3 | h. Lump was tender and I had occasional shooting pains. Was also told by my gyn that cancer does not hurt, but luckily she also wanted to take it seriously, so I got the ultrasound. Since then, another doctor has said that medullary type tumors can hurt and he thinks that may be what I have. This diagnosis has a better prognosis than non-medullary, so since it doesn't currently make a difference to my treatment choices one way or another, this is what I am choosing to believe. So I welcome the pain! |
| Patient-provider feelings | 32,266 | 5.6 | i. I LOVE MY SURGEON…..he had cleared his appointments (had another doctor in the practice see them) and was waiting to talk to me….came in, sat down with us, and told me it was cancer, but I WOULD get through this and live a long life….I don't remember much after that….I remember asking "so I shouldn't loose sleep over this?" and he said "oh, you're going to loose sleep over this, you wouldn't be human if you didn't"….LOVE LOVE LOVE my surgeon. There's been several times that he's sat down with me to "talk", allowing me as much time as I need, despite his very busy schedule. I was his first "young breast cancer patient" (though sadly, not his last) and both he and his nurse say that my case changed how he handles young women with breast lumps |
| Stage at diagnosis (numbers not filtered) | 71,782 | 12.6 | |
| ***Parent codes classified by support vector machine algorithm into child codes*** | | | |
| First signs and symptoms (n = 3,266) | | | |
| Self-detected | 817 | 25 | j. Actually, I too had a brownish discharge (looked like dried blood) from my right nipple prior to diagnosis. I experienced it for about 3 weeks before making an appt. with my OB/GYN…he completely dismissed it…sent my soon to be husband and me home with a small petri dish and said if you can get anymore bring it in to me. … I know hindsight is 20/20, but I wish my husband and I would have been more diligent in finding out the answer for the nipple discharge…Ah, regardless I now tell my story to anyone that is interested to get the word out that cancer is not always a lump… |
| Lump present | 1,844 | 56.5 | k. i nursed while i had a 3 cm tumor that i was told was NOTHING. i had found a small lump while 2 mos pregnant and only 9 mm, after an ultrasound and needle biopsy i was told it was nothing and not to worry about it. i let it grow while my daughter grew inside me as well. after she was born she had no problems nursing, even with the 3 cm tumor, my milk was fine and breastfeeding was perfect!!! i did so till after my diagnosis and needed to stop for surgery |
| Provider-detected | 563 | 17.2 | l. I had a dimpling in my cancer breast (NOT that your is cancer) but my doctor noticed it when I was laying down…not really standing up. My tumor didn't show up on an ultrasound (my report is all clear) but the lump could be felt |
| *Steps to diagnosis (n = 31,640)* | | | |
| Biopsy | 7,368 | 23.3 | m. I was "old" at 42 when I was diagnosed. Eight months after a "clean" mammo (Boy you have dense breasts!) I found the lump and it wasn't until 3 months later that I had a biopsy done.—I was already Stage IV by the time I was staged. I pushed and pushed for my appointments but I was told that usually lumps are nothing and that if it is cancer it's been there for years and grows very slowly. Apparently my cancer had never heard that it was supposed to grow slowly… |
| | **n** | **%** | **Quote** |

**Table 1** (continued)

| | | | |
|---|---|---|---|
| ***Parent codes classified by support vector machine algorithm into child codes*** | | | |
| Biomarker test | 319 | 1 | n. Ask for a second opinion on your pathology at a lab outside your current hospital. er/pr staining is actually pretty subjective so you want to make sure you're getting the right percentage. current standard is er/pre positive is considered if you're 5% or more. and you should get two different numbers—one for er and one for pr… if either is over 5%, then most would rec hormonal therapy. The testing was done at a independant lab outside of the hospital. Pr was completely negative. Her2 was negative. Pr was 1 to 4% |
| Imaging and clinical exams | 21,043 | 66.5 | o. i had several radiologists say that my mri was fine. fortunately my primary radiologist is awesome and aggressive and sent my images to another colleague who suggested a core biopsy. without it they would have missed more DCIS that runs from my chest wall to my nipple |
| ***Patient-provider-system feelings (n = 24,648)*** | | | |
| Positive | 9,082 | 36.8 | p. My Mom just passed away in January of metastatic breast cancer... Afterwards I realized that I had forgotten to go to the gyno so I made an appt when I got back home. She did the usual breast exam and thought she might have felt something but wasn't convinced. However, given my history (my maternal grandmother died of breast cancer as well) she wanted me to get a mammogram. So I went, continuing to self exam and not feeling anything anymore, I wasn't worried. I had the mammo two weeks ago and they found two "calcification" clusters. They're like the size of a pin head! But now I have to get them biopsied. I can't even begin to tell you how frightened I am. The doctors haven't filled me with hope because of my history, but they rattle off the statistics..."70% chance of it being benign, etc." |
| Negative | 17,337 | 70.3 | q. I am 37 yrs old from California. I was diagnosed with DICS and Invasive breast cancer on December... Summer, I found my lump and was dismiss my former gyno thinking it was only a cyst. She drained it twice but did not send it to pathology. I had a mammogram and ultrasound I was told it was not cancer. Contact gyno again because I felt the lump getting bigger. She insisted that it was a cyst. Informed her that my grandmother had died of breast cancer. She told me not be concerned that it would be a concerned if my mother had cancer. She also told me that no surgeon would touch me. I insisted and finally got a referral to meet with a surgeon. I met with my surgeon in Oct. Finally someone listen to me. She was wonderful and supportive and agreed that it would be removed and send to pathology. My lump was 4.8 cm. Everything happened so fast for me. Had lump removed and was diagnosed. Worst day of my life!!! |
| Neutral | 18,808 | 76.3 | r. Hi...was diagnosed in Novemeber..originally told squamous cell carcinoma of the right breast..very rare.. went to cancer center for 2nd opinion..diagnosed IDC right breast...did FNA on left breast.."funny" looking cells. Had right breast mastectomy in December w/SNB (2 sentinel nodes one was actually a cluster of 2.. so I guess 3 sentinel nodes) + 6 additional nodes taken.originally told "Clean" also excisional biopsy on left breast - B9. Later told 12 cancer cells found in the 1st of the 2 clustered SN..barely positive but am being treated as lymph node positive. Also had recon w/tissue expanders...last fill in January. Started chemo yesterday..not feeling that bad. Had AC+Avastin...on clinical trial. Will follow 4DD AC+Avastin followed by 4DD Abraxane (Form of Taxol) + Avastin then continue Avastin for 12 more cycles every 3 weeks. Seemed best option as I am triple neg. Not sure yet if I have to do rads. Will schedule implant exchange for July. Should also note that lump 1st felt in July before going on vacation to Hilton Head...had lumps before..had mammos before was always NOTHING..lump did not concern me as no one in my family has had breast cancer.. or any other cancer for that matter and that I always had "lumpy" breasts. |
| ***Stage at diagnosis (n=71,879)*** | | | |
| Stage 0 | 5,141 | 7.2 | I was dx w/ DCIS. ER/PR+. 1.2cm. no lymph node involvement |
| Stage 1 | 3,419 | 4.8 | I had surgery to remove stage 1 bc p/g-. |
| Stage 2-3 | 3,954 | 5.5 | I was diagnosed with Stage 3 ductal carcinoma with node involvement. |
| Stage 4 | 5,247 | 7.3 | I was newly diagnosed at stage IV with bone mets to the spine. |
| Invasive (Stages 2 to 4) | 5,223 | 7.3 | I had an ILC tumor last summer, and I would like to get a roll call of other YSC members who had only ILC tumors |
| ***Inductive codes*** | | | |
| Mental health | – | – | s. Does anyone know if there is a psychiatrist/psychologist that specializes in dealing with cancer patients? I am really starting to fall into a very dark place and I need some help... I too am feeling numb and am withdrawing from my life. I need someone to help me fight this from completely taking over my life, but don't currently have anyone to do so...I really need this because I am beginning to feel like a prisoner of war. I have 1 of 2 choices. Either give up and die or fight and be absolutely miserable, hurt all the time, and generally have a crappy life until the cancer (or the treatment for it..which I think is as bad if not worse than the cancer at times) ultimately drains the life out of me. I don't know how to go on anymore. |
| Fertility | – | – | t. My surgical oncologist has suggested that I do my chemo first, to hopefully shrink the lump before the lumpectomy. So my boyfriend and I are hoping to harvest eggs before we start chemo. My consultation with the fertility specialist is in January. It feels so far away just to find out what's involved. And then there's the whole process itself. I do have a question for ya. My cancer like yours is hormone receptor positive. I understood my onc to say that getting pregnant would increase my risk of having any precancerous to become cancerous. So we're considering surrogacy. |

**Table 1** (continued)

| ***Inductive codes*** | | | |
|---|---|---|---|
| Menstruation | – | – | u. Somehow I thought that my period would return in May if I didn't have my scheduled Lupron shot in April. Sounded logical to me! Nice and tidy. I had a plan, but alas... It's like crickets chirping around here...not a drop, a sore boob, nothing. A little heavy feeling in the uterus area but I'm sure it's all in my mind...Some neighbors said something about an ovulation kit, I guess in theory of the hopes that I'm ovulating but not menstruating?...Who wants any of that???... I'll just wait for Flo to show. If I were 18 and had never had cancer, this would be a lot easier. |
| Childrearing | – | – | v. So here's the deal... I'm losing my battle with cancer...My heart break is leaving my son. He's 20, an only child... But if I wasn't here, he wouldn't have made rent this month!!!! And it's stuff like that. I'm not going to be here to help him. He's going to be on his own. He has his father...But his dad rarely ever see's him and lives 7 hours away. I'm really close with my son. We talk everyday, even since he's moved out. The thought of leaving him so early in life |
| Advocacy | – | – | w. I know SOFT has had a terrible time enrolling women in the numbers needed, for a lot of demographic and other good reasons. I get it. But I'm trying to enroll and can't seem to interest the hospital trials administrator! I spoke with her in person on Tuesday and she offered, half-heartedly, to see if I'm eligible. When I hadn't heard from her, I called her yesterday. She didn't remember who I was or that we met, took my information a second time, and said she wouldn't get to it until next week, though she didn't sound very interested...But this is an important study (not to be obscure: it looks at various hormonal treatments for premenopausal bc patients with ER positive disease-- Tamoxifen alone, ovarian suppression + Tamox, or OS + AI-- information we need!). A couple other national trials in this same category failed because they couldn't enroll enough women. I love the hospital where I've been treated. Their patient care has to be among the best anywhere, but maybe this is the downside: research doesn't seem to be a priority. |

that made them think something was wrong. Individuals talked about finding their breast cancer by accident while in the shower (Table 1f), exercising, or during their honeymoon, for example. Others found their breast cancer while pregnant or breastfeeding (Table 1a and k). A majority of these first signs and symptoms were lumps in the breast, although individuals also discussed nipple discharge and bleeding (Table 1j), dimpling and redness, and general pain and fevers. For some individuals, their doctors noticed the first signs and symptoms of breast cancer while screening for patients with a family history or routine clinical breast exam (Table 1l).

### *Steps to diagnosis*

Individuals described various paths and crossroads to diagnosis. Once they suspected cancer according to first

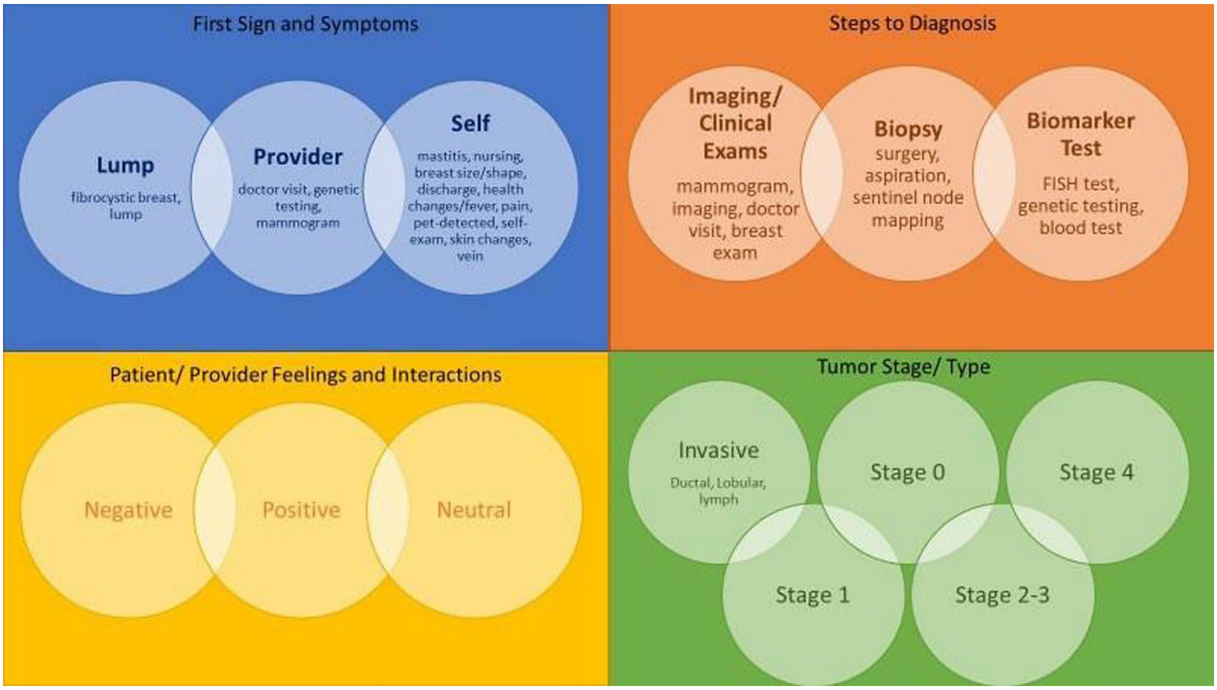
First Sign and Symptoms
Lump
fibrocystic breast, lump
Provider
doctor visit, genetic testing, mammogram
Self
mastitis, nursing, breast size/shape, discharge, health changes/fever, pain, pet-detected, self-exam, skin changes, vein
Steps to Diagnosis
Imaging/ Clinical Exams
mammogram, imaging, doctor visit, breast exam
Biopsy
surgery, aspiration, sentinel node mapping
Biomarker Test
FISH test, genetic testing, blood test
Patient/ Provider Feelings and Interactions
Negative
Positive
Neutral
Tumor Stage/ Type
Invasive
Ductal, Lobular, lymph
Stage 0
Stage 4
Stage 1
Stage 2-3

**Fig. 2** Parent, child, and grandchild codes

**Table 2** Accuracy statistics of three predictive models to classify training data set (n = 750)

| | Support vector machine | Random forest | Decision tree |
|---|---|---|---|
| Accuracy statistic | % | % | % |
| Recall | 78.4 | 63.2 | 76.8 |
| Precision | 80.5 | 84.8 | 76.6 |
| F-measure | 79.3 | 72.2 | 76.6 |
| Accuracy | 85.4 | 80.1 | 82.9 |

signs and symptoms and their doctors agreed to order tests, either willingly or after much patient persistence, imaging, either by ultrasound or mammography and less so MRI, was usually the first step to a diagnosis. When imaging detected possible tumors, biopsies quickly followed (Table 1g). Other individuals discussed needing to advocate for additional testing either because the imaging results were equivocal or they did not detect anything, and they or a doctor knew to keep asking for further testing (Table 1m and o).

Individuals received biopsies including "core," "excisional," and "fine needle" types, which either confirmed cancer or led to false negatives, leading patients to advocate for further testing. Once the cancer was confirmed, individuals wrote about genetic testing of both germline and tumor cells to help guide treatment decisions. In terms of tumor genetics, there was much discussion about "tumor markers" and specifically FISH testing and the quality of such tests in terms of reproducibility, sensitivity, and specificity (Table 1n).

The steps to diagnosis were less clear for adults who found their cancer while pregnant. One woman spoke of being at a complete stalemate as further testing for her was being weighed in relation to the health of the fetus. Another woman wrote of delayed detection, "I went and had a mammogram, it came back with nothing, I got pregnant the next month, so I couldn't have an MRI. 2 months after my baby was born, I found a lump."

### *Healthcare interaction between patients and providers*

Posts about interactions with providers and the healthcare system were mostly negative or neutral, but some were positive. Some providers dismissed individuals' concerns saying they were "too young" to have breast cancer, or dismissed pain because "cancer doesn't hurt" or lumps because they were "only a cyst" (Table 1q). Other providers acknowledged these as low risk factors, but took action:

> *I saw my GP because I had found a lump right on my cleavage (what there is/was of it) the week before and it hurt when I rubbed it. He said, you're too young for B/C and there's no history in your family. Besides, sore is good. Cancer hardly ever hurts. In the next breath he said but let's do a mammo and get you a surgeon consult just to be sure. THANK GOD!! I had my mammo done on Dec. 31st - Happy New Year! At this point I still haven't told anyone what's going on. In Jan, I saw the surgeon. He hadn't gotten the report because of the holidays so I get there, he examines me and says the same thing the GP says. Too young, no history and sore is good. But... let's do a mammo."*

Additional positive interactions included having a knowledgeable provider (Table 1p) or one that changed their practice after having their first young case (Table 1i). Lastly, neutral interactions talked about working with their team of doctors throughout the process of being diagnosed without indications of having either a positive or negative experience with them or the overall healthcare system (Table 1r).

### *Inductive themes*

While connecting and telling their stories, young adults wrote about how having breast cancer affected other aspects of health and well-being, unique to young people. Posts mentioned mental health, either seeking help to handle depression while handling the stress of a diagnosis (Table 1s) or out of fear of their treatment triggering previous depression.

The reproductive decisions for young adults changed after a diagnosis. People without children considered whether to preserve their eggs prior to treatment (Table 1t), and after treatment, young adults considered whether having children would increase their risk for a reoccurrence or increase the risk of their children not having a mother. Parents with hindsight after deciding to have children, offered much advice. One parent emphatically wrote that having children is "absolutely worth the risk," even with a BRCA mutation. People also shared fears of having breast cancer while being a parent. Parents of small children spoke of handling their diagnosis and treatment during the day-to-day tasks of caring for small children; parents of older children feared not being able to support adult children financially and emotionally (Table 1v).

Individuals also spoke of the effect of breast treatments on their menstrual cycles, some welcoming not having periods anymore and others missing their periods (Table 1u).

Messages about the need to self-advocate and to be persistent permeated all the deductive and inductive codes. Some individuals saw the structural barriers in healthcare systems and called for collective action to participate in research. They shared scientific articles and abstracts with commentary explaining the findings. They shared how to interpret test results and, depending on the results, what to expect next. Information also included clinical trial opportunities (Table 1w).

## Discussion

Using NLP to aid qualitative review of over 500,000 posts from 10 years, we identified, from the perspective of young adults diagnosed with breast cancer, four areas where we need more evidence-based guidelines for screening and diagnostics for early onset breast cancer.

First, our results confirm the importance of breast awareness counseling as a precursor to screening because the vast majority of young adults self-detected their breast cancer rather than the cancer being detected through a clinical breast exam at an annual gynecology appointment. Some women were dismissed by their gynecologists and had to insist on further work-up of cysts and other palpable tumors. Our results are in line with studies that found clinical breast exams detected few additional cancers in adults under the screening age. [12] Second, we need more evidence and implementation of guidelines that indicate when mammograms, ultrasound and/or MRI should be used when young adults suspect breast cancer. Third, we need evidence-based guidelines to determine when biopsy should follow exhibited breast symptoms referrals in young adults, even when imaging results are negative. Lastly, healthcare services for young breast cancer patients should include support for individuals in their role as a parent or their childbearing goals.

Our results reinforce the importance of breast self-awareness [13], however only a few governing bodies recommend clinicians counsel individuals about breast awareness. [6, 14] Screening guidelines require evidence and years to change, so the concept of breast awareness recognizes that until there is more reliable early detection, young adults should be familiar with their breasts, be able to detect any change from the norm, and if they find an irregularity, then insist on a complete clinical evaluation. How knowledgeable of and adherent to these guidelines clinicians are warrants separate investigation because advocating for breast self- awareness is only effective if clinicians are receptive to young adults' concerns about any irregularities they detect. There is debate as to whether age or having symptoms is the stronger predictor of delays in detection, however these studies and ours point to the importance of the implementation of breast awareness in clinical practice as a guideline not only for patients [15] but also for clinicians [16], who need guidance on what imaging and biopsy tests they should order when patients present with symptoms. [17] These stories also provide insights into what healthcare systems need to consider during the pre-diagnosis period of the clinical pathway.

Our findings suggest that young adults encountered three crossroads along the clinical pathway between the initial clinical visit and the formal diagnosis stages, when there was delay in treatment (Fig. 3). The first crossroad occurred at the initial visit when the clinician either was convinced enough to order tests or dismissed the individual as too young or the symptoms as not concerning. The second crossroad was whether the imaging led to additional testing. Individuals discussed the inadequacies of the available imaging methodologies and access to alternative imaging, raising the need for more evidence and implementation of the most accurate imaging modality for young adults with breast symptoms. The third crossroad was whether a biopsy confirme dor missed the cancer. This third crossroad was greatly dependent on whether imaging in crossroad number two was positive.

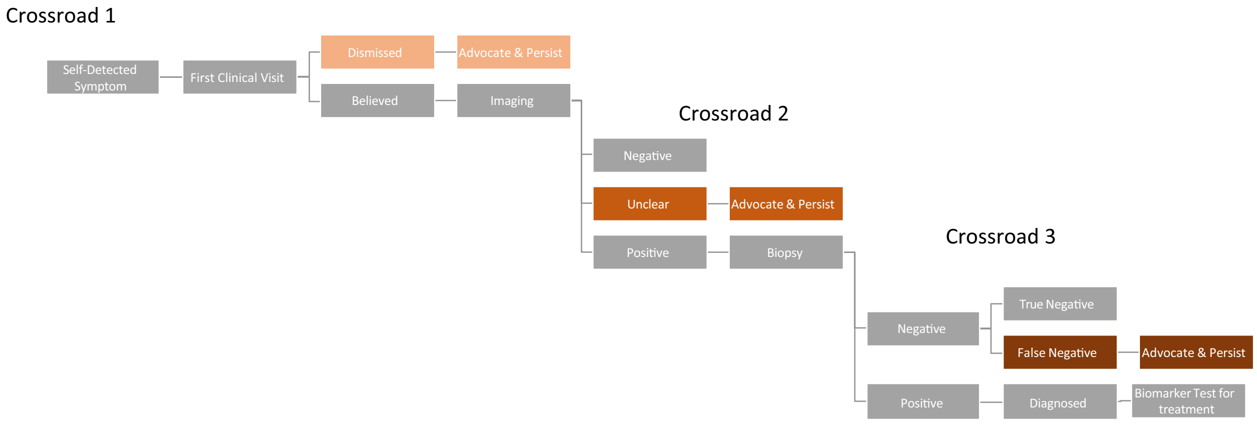

**Fig. 3** Crossroads during the clinical pathway and delays in diagnosis

Therefore, we need more research if biopsies should follow symptomatic cases even when imagining results are negative. Mental health challenges and the choices and demands of parenting while living with breast cancer compounded these delays**.** True positive cases who experienced these delays strongly encouraged those behind them in the journey to self-advocate and persist and insist on further testing until diagnosed accurately. Delays in diagnosis also mean potentially worse prognosis since later stage cancers are more aggressive with fewer treatment options. Breast cancer mortality rates for young adults are low and the rate has slowed down over the last 10 years thanks to advancements in treatment. [18] However, for late stage and more aggressive cancers that are harder to treat, earlier detection is still the best option.

Our study has strengths and limitations. Our methodology combined traditional thematic coding with natural language processing to gain an in-depth emic, i.e., insider, perspective of young adults' experiences with a breast cancer diagnosis and leverage the breadth of a large dataset of over 500,000 online forum posts over 10 years. Online forums provide a natural space for honesty about people's experiences that they may not share with their provider or on a questionnaire. On the other hand, the sample does not capture the potentially negative experiences of patients who have undergone invasive procedures with normal results. Furthermore, we did not review the full breadth of experiences existing in the dataset because the number of relevant posts was large, and we only reviewed a small proportion of posts within each parent code. To first train the algorithm, we selected posts from the forum entitled "newbies," as we assumed this is where most young adults would share stories that captured the parent codes. Our algorithm and results may have differed if we used posts from another forum. It is possible that our sample or the overall forum itself selected for young adults with negative experiences because writing at the time of a recent diagnosis was more likely to be during a negative experience. As a limitation, our training data set had class proportions that were imbalanced. This imbalanced data may have prevented the machine learning algorithms from learning effectively, where the performance of the classifying algorithm was biased towards the majority class. [19] To mitigate some bias, we first trained multiple classifying machine learning algorithms on the manually coded dataset, then chose the best-model based on the F-measure rather than other evaluation metrics such as accuracy. [20, 21] While we were able to calculate the distribution of our codes, including steps to diagnosis, we could not quantify the length of delay nor decipher between the source of delay (self vs care) as other studies have quantified [22], but our qualitative findings complement and provide greater insight into the experiences and drivers of previously quantified self- or care- delay. Finally, we did not have demographic data on the forum participants, and all were English- speaking, limiting generalizability or capturing experiences of specific communities. Furthermore, we did not know extensive family history of breast cancer among individuals and so cannot draw conclusions about screening of young adults with a strong family history of cancer. However, this is a very large sample to qualitatively understand the journey from first symptom to diagnosis for young adults with breast cancer to identify gaps in the clinical pathway in order to improve from the patient's perspective.

## Conclusion

Prior studies have examined the breast cancer clinical pathway starting with diagnosis and also through qualitative interviews. [23] Our study captured earlier steps in the journey, including the "prelude" when the patient notices something is wrong, followed by the "warning" stage when something triggers a visit to the doctor. Most young adults self-detected their cancer by feeling a lump. This is not a surprise given there is no -population-based screening for this age group, however, the descriptions of the surrounding context when they first detected their cancer, such as being on honeymoon, or breastfeeding, amplifies how much a diagnosis under age 40 violates widely held, aged-based assumptions that breast cancer is a disease of the elderly.

Examining the pre-diagnosis experiences of early-onset breast cancer patients identifies potential research gaps and areas where clinical practice needs improvement. Our findings support ACOG's and YSC's recommendations for breast awareness counseling and highlight the urgent need for early detection in young adults. [6, 24] We need new guidelines to ensure that young adults with breast concerns receive complete evaluations with minor delay along the clinical pathway in terms of imaging and biopsy. Healthcare services for young breast cancer patients should include support for individuals in their role as a parent or considering their childbearing goals.

**Acknowledgements**

Thank you to the young adults who contributed their stories of being diagnosed with breast cancer to the Young Survival Coalition online forums.

**Author contributions**

Data acquisition: D.W, A.N, S.L Data analysis: K.T.U, M.T, J.S, M.G Drafting manuscript: K.T.U, L.H All authors revised the manuscript.

**Funding**

Irving Institute for Clinical and Translational Research MS-POR # UL1TR001873.

**Data availability**

We obtained a de-identified dataset from the Young Survival Coalition.

## Declarations

### Conflict of interest

The authors declare no competing interests.

Received: 23 September 2024 Accepted: 21 February 2025
Published online: 13 May 2025

### References

1. Johnson RH, Chien FL, Bleyer A. Incidence of breast cancer with distant involvement among women in the United States, 1976 to 2009. JAMA. 2013;309(8):800–5.
2. Shiyanbola OO, Arao RF, Miglioretti DL, Sprague BL, Hampton JM, Stout NK, et al. Emerging trends in family history of breast cancer and associated risk. Cancer Epidemiol Biomark Prev. 2017;26(12):1753–60.
3. Smith RA, Andrews KS, Brooks D, Fedewa SA, Manassaram-Baptiste D, Saslow D, et al. Cancer screening in the United States, 2019: a review of current american cancer society guidelines and current issues in cancer screening. CA Cancer J Clin. 2019;69(3):184–210.
4. USPSTF. Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement. Ann Intern Med. 2005;143(5):355–61.
5. USPSTF. Breast Cancer: Screening 2024 [Available from: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening#:~:text=The%20Task%20Force%20now%20recommends%20that%20all%20women%20start%20getting,screened%20when%20they%20turn%2040.
6. Mango V, Bryce Y, Morris EA, Gianotti E, Pinker K. Commentary ACOG practice bulletin July 2017: breast cancer risk assessment and screening in average-risk women. Br J Radiol. 2018;91(1090):20170907.
7. Zhang S, Bantum E, Owen J, Elhadad N. Does sustained participation in an online health community affect sentiment? AMIA Annu Symp Proc. 2014;2014:1970–9.
8. 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.
9. Zhang S, Grave E, Sklar E, Elhadad N. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. J Biomed Inform. 2017;69:1–9.
10. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B, et al. Data analysis, machine learning and applications. In: Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al., editors. KNIME: The Konstanz Information Miner 2008. Berlin Heidelberg: Springer; 2008.
11. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3(2):77–101.
12. Menes TS, Coster D, Shenhar-Tsarfaty S. Contribution of clinical breast exam to cancer detection in women participating in a modern screening program. BMC Womens Health. 2021;21(1):368.
13. 2017 Practice Bulletin Number 179: Breast Cancer Risk Assessment and Screening in Average-Risk Women. Obstet Gynecol. 130 (1): e1-e16.
14. Bevers TB, Helvie M, Bonaccio E, Calhoun KE, Daly MB, Farrar WB, et al. Breast cancer screening and diagnosis, version 32018, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw. 2018;16(11):1362–89.
15. Hindmarch S, Gorman L, Hawkes RE, Howell SJ, French DP. "I don't know what I'm feeling for": young women's beliefs about breast cancer risk and experiences of breast awareness. BMC Womens Health. 2023;23(1):312.
16. Partridge AH, Hughes ME, Ottesen RA, Wong YN, Edge SB, Theriault RL, et al. The effect of age on delay in diagnosis and stage of breast cancer. Oncologist. 2012;17(6):775–82.
17. Costa L, Kumar R, Villarreal-Garza C, Sinha S, Saini S, Semwal J, et al. Diagnostic delays in breast cancer among young women: an emphasis on healthcare providers. Breast. 2024;73: 103623.
18. Ervik M LF LM, Ferlay J, Bray F. Global Cancer Observatory: Cancer Over Time. International Agency for Research on Cancer [Available from: https://gco.iarc.fr/overtime.
19. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv. 2019;52(4):1–36.
20. Justin L. Just into Data [Internet]2021. [cited 2023]. Available from: https://www.justintodata.com/imbalanced-data-machine-learning-classification/#:~:text=Within%20it%2C%20we%20have%20imbalanced,is%20a%20highly%20imbalanced%20dataset.
21. Google. Machine Learning Concepts: Datasets: Imbalanced data [Available from: https://developers.google.com/machine-learning/data-prep/construct/sampling- splitting/imbalanced-data.
22. Ruddy KJ, Gelber S, Tamimi RM, Schapira L, Come SE, Meyer ME, et al. Breast cancer presentation and diagnostic delays in young women. Cancer. 2014;120(1):20–5.
23. Lindop E, Cannon S. Experiences of women with a diagnosis of breast cancer: a clinical pathway approach. Eur J Oncol Nurs. 2001;5(2):91–9.
24. Position on breast self examination (BSE) and early detection [press release]. 2006.

## Publisher's Note