Crowdsourcing Personalized Medicine: Applying Machine Learning to Unstructured Social Media Data to Improve Breast Cancer Patient Outcomes

Gregory Falco*, Shravan Ravi†, Modadeoluwa Ogunmuyiwa‡, Caleb Li§

*John Hopkins University, †University of Texas at Austin,
‡University of Pennsylvania, §Patient Journey Inc.
Email: falco@jhu.ed, shravanr@cs.utexas.edu, modade@sas.upenn.edu, caleb@openphr.org

***Abstract*—Millions of patients are leveraging online forums to discuss and share their patient journeys. This produces a wealth of unstructured data containing valuable patient journey's, including biomarkers, lines of treatment, and adverse events. However, such unstructured data has been challenging to gather insights without manual curation. We propose using a combination of sentiment analysis and unsupervised machine learning to generate real world data and personalized medicine at scale. The results provide evidence real world evidence for drug and treatment efficacy.**

***Index Terms*—Personalized Medicine, Breast Cancer, Patient Education, Patient Cohorting, Patient Self-Advocacy, Medical Machine Learning, Social Media Data**

I. Introduction

The increasing popularity of social media and online community forums has produced data on an extraordinary scale [12]. Hidden among this data exists population insights that can serve the public [8]. Patients with a variety of aliments have taken to social media to share their experiences with varying treatment plans. In particular, social media may have significant potential to help support breast cancer patients [2]. However, despite previous research on the usefulness of social media for breast cancer patients [1] there is a lack of longitudinal patient analysis with respect to the duration of their social media engagement and issues with the systematic digestion of the unstructured forum data [2]. Despite this shortcoming, there is considerable value that can be derived from such data with appropriate data cleansing measures in place. We propose a novel method of hierarchical data organization to help patients synthesize the copious amount of unstructured data found across social media breast cancer patient forums. Organizing volumes of social data in this way can be used to cohort patients and offer crowd-sourced personalized medicine treatment plan guidance for patients identified as belonging to the same cohort. Empowering patients with cleansed, structured social data could reduce treatment anxiety [7] by improving the accessibility of anecdotal, relevant guidance provided by others with shared experiences, in effect improving patient education [4], and by improving the reliability of internet sourced information thanks to automated data aggregation, classification and post-processing by researchers [6] [5].

II. Approach

To address the previously described shortcomings of engaging social media data to drive patient insight, we propose a novel method for end-to-end data curation, cleansing and analysis that will help make unstructured patient data useful to patients seeking information about others' experiences. This is completed over four phases. The phases include: (1) Data Aggregation, (2) Sentiment Analysis, (3) Clustering, and (4) Curated Patient-Reported Outcomes.

A. Phase 1: Data Aggregation

Patients often have access to social media and online forums to share their journey [3]. This produces several terabytes of unstructured data across all major social media platforms (Facebook, Reddit, Inspire, etc.) However, despite this wealth of available information, patients can find difficulty tracking relevant information in a consistent way as subthreads and posts can become confusing to track and be buried over time. To address this challenge, we identified and stored posts relating to stage 3 and 4 breast cancer across social platforms in a single data warehouse where it could be readily cleansed and analyzed. We then engaged a scraping algorithm that uses the Beautiful Soup library to store 50,000 social media posts from 1,200 in CSVs with author handle, location, age, hobbies and post content.

B. Phase 2: Sentiment Analysis

To ultimately arrive at personalized medicine recommendations for a given cohort, we sought to determine successful treatment plans using sentiment analysis. Sentiment analysis would enable us to automatically identify positive outcomes for cohorted patients. For those with positive outcomes that had similar treatment plans identified using cosine similarity, we could aggregate those results into patient journeys (further described below). However, using typical sentiment analysis dictionaries was not feasible given the unique language

requirements to assess the positive impact of given treatments. To develop the medical dictionary, we started with an open source pharmacologic dictionary complete with various treatments and associated side effects. A standard sentiment analysis algorithm was applied to each of the pharmacology definitions; yielding an aggregate sentiment for each drug composed of the sentiments of each word within the definition. The aggregate sentiment score of each definition became the sentiment applied to each drug. This scoring scheme for drugs was used to augment the typical sentiment analysis dictionary, which was then run across the posts.

C. Phase 3: Clustering

Stage 3 and 4 breast cancer patients vary considerably, with their etiology ranging from genetic mutations to environmental factors. We sought to cohort those who posted about their experiences on the forums based on combinations of their age, family history described, location, lifestyle (extrapolated from hobbies) and genetic mutations. A benefit of using natural language processing is their ability to distill text data into numerical data for further analysis. Initial clustering was achieved using cosine similarity. Cosine similarity helps to measure the text similarity between two posts. After engaging this cursory means of establishing similarity, we engaged a k-means clustering algorithm to cohort the users.

D. Phase 4: Curated Patient-Reported Outcomes

Finally, for each cohort, the patient treatments with the highest sentiment were identified. Their posts were then strung together into what we call a patient journey. The patient journey acted as a narrative for those wishing to read through a patient's experience without sorting and scrolling through unstructured posts. After the patient journeys were developed, cosine similarity was applied across the patient journeys to assess those with similar experiences. The consistency of positive outcomes for treatments within a cohort was a statistically insignificant indicator that the treatment plan was effective. Those with similar treatments with positive outcomes in the same cohort were abstracted into a persona whose treatments and associated metadata was documented in an outcomes database. The outcomes database served as the foundation to our future personalized medicine recommendation portal.

III. Experiment

After sampling the available unstructured data from several social media sites, we were left with 50,000 social media posts from 1,200 users across Facebook, Reddit and Inspire, each of which are home to active breast cancer communities. In order to begin finding medically relevant information within this content, we utilized our medical semantic analysis model to transform the categorical information into numerical values. Our semantic analysis model leverages a pre-trained model that utilizes a Long-Short Term Memory (LSTM) neural network to identify key words and derive information from human language. We then provide a the medical dictionary mentioned above to augment the trained model. This allowed us to identify the presence of important medical information within the social media posts throughout the data set. After training, our model was able to effectively identify and rank posts based on their inclusion of keywords like drug names, "Herceptin", severity of disease, "Stage 4", and outcomes, "stable for 2 years". Using this derived information, we could then shortlist users with in-depth posts containing many drug names, several lines of treatments, and many patient reported outcomes. This provided us with vectors that summarized the pertinent information within each of the user's posts. Using these generating vectors, we were able prioritize posts with regard to their medical relevance and further refine the scope of our search. Additionally, because we were able to standardize the content and transform all the textual information into numbers, we were able to algorithmicaly calculate the relationships or trends between users, their lines of treatment, and outcomes.

To that effect, we look to identify cohorts within users by implementing a custom k-means clustering algorithm. This involved determining the number of clusters (cohorts) present within our data set of users. From that number 'n', we were able to place 'n' random centroids that were then adjusted to best fit the data set when trained. This "best fit" is calculated by determining the distance between the centroid and every data point for every feature inside our calculated vector. To train this model, we used the 11 distinct features captured from the vector derived from the social media posts and user metadata: 1.) Post Description, 2.) Post Text, 3.) Post Heading, 4.) User Age, 5.) Date Posted, 6.) User Community Involvement , 7.) User Interests, 8.) User Common Discussion Topics, 9.) User Marital Status, 10.) User Gender, and 11.) User Location. Using this algorithm, we were able to experimentally discover 15 distinct clusters in the data (lowest loss for several configurations of centroids). This application of machine learning to stratify users allowed us to identify trends between different treatments and outcomes within cohorts. This enabled us to bolster the confidence of a given treatment method within a cohort, because it allowed us to cross validate similar patient journey's and ensure that those with significant medical data were somewhat corroborated by others within our patient pool.

In this pilot case, we extracted our top 3 personas with the most supported and longest longitudinal medical history that could be extracted from users' social media activity. We then worked with clinicians to develop a detailed chronological patient journey with the resulting patient outcomes.

IV. Results

Below is a summary of the results of our initial work. This is a snapshot of a potential database that can be developed from the unstructured social media data based on aggregate successful treatment outcomes. As represented below, there are three distinct personas that were supported as meritorious treatment plans by both the amount of medical information within user posts, and the number of users who shared a similar journey.

TABLE I
SAMPLE PATIENT-REPORTED OUTCOMES DATABASE

Personas	Joy	Mandy	Joan
Initial Diagnosis	Stage 3	Stage 4	Stage 4
Operation/Surgery Location	California	Ohio	New York / USA
Treatment 1	Chemotherapy and Radiotherapy	Tamoxifen and Herceptin	Letrozole and 10 sessions of Radiotherapy on the spine
Objective Response	Not Stated	Stable Disease	Stable Disease
Duration of Response	6 months	Tamoxifen - 2 years	3 years
Number of treatment related adverse events	Hot Flashes	Nausea, Increased Uterus Thickness, Pulmonary Embolism, Decreased Hemoglobin Level, Decreased Ejection Fraction	Not Applicable
Treatment 2	Letrozole	Radiotherapy, Tamoxifen, and Herceptin	6 cycles of Ibrance with Arimidex
Objective Response	Progressive Disease	Partial Response to Herceptin	Stable Disease
Duration of Response	2.5 years	Herceptin - 5 years	5.5 months
Number of treatment related adverse events	Hot Flashes	Nausea, Increased Uterus Thickness, Pulmonary Embolism, Decreased Hemoglobin Level, Decreased Ejection Fraction	Joint Pains, Elevated Liver Enzymes, Hair Thinning
Treatment 3	Faslodex	Armidex and Herceptin	Ibrance and Faslodex
Objective Response	Complete Response	Stable Disease	Stable Disease
Duration of Response	4 years No Evidence of Disease shortly after starting	2 years	1 month
Number of treatment related adverse events	Hot Flashes	Hot Flashes	Elevated Liver Enzymes, Hair Thinning
Treatment 4	Faslodex and Ibrance	Ibrance, Radiotherapy, and Herceptin	Kisqali and Faslodex
Objective Response	Complete Response	Stable Disease	Stable Disease
Duration of Response	3.5 years No Evidence of Disease shortly after starting	5 years	1.5 month
Number of treatment related adverse events	Hair Thinning, Hot Flashes Decreased ANC levels	Nosebleeds, Anemia	Hair Thinning, Nausea

V. Discussion

A. Approach Benefits

We learned that there are indeed benefits from employing our approach. First, there is considerable untapped potential in unstructured medical social media data. Organizing posts by users and then cohorting users can yield insights about patient populations and how they respond to given treatments. Consider that oncologists make recommendations to patients today based on their own experiences evaluating hundreds of patients over their period of practice. Nothing can replace this insight; however, recommendations from an oncologist are limited by their previous patient experiences or what they have learned from medical journals and continued education. Our approach offers oncologists additional data about patients far beyond the reaches of their individual experience.

Second, cancer patients are already attempting to find answers to their questions in forum posts scattered across the web. Many are scared and will cling to any guidance they can find in order to maintain hope that their condition will improve. Given that there is a great deal of scientifically untrue treatments noted on the Internet, our approach at least offers a curated guide to semi-validated treatment options. The information we surface is already available publicly, we simply help to organize it in a digestible manner. Through our implementation of persona identification with a combination of sentiment analysis and k-means clustering algorithms, we are able to gather similar personas, potentially 100s out of our 1,200 patients, and the personas could become real world data to understand the efficacy of drugs.

Finally, cleansing and refining publicly available unstructured data returns power to the patient in enabling their own education about what they could expect given their condition and placement in a given cohort. In effect, they can crowdsource their personalized medicine treatment plan based on what has worked for others like themselves.

B. Approach Limitations

This is, however, an idealistic view of how our proposed approach can help. There are considerable gaps and cautions to be taken when interpreting our results.

First, the initial quality of self-reported social media data is questionable. There is likely heavy bias inherent to each patient's posts that we capture. For example, there are reasons for patients to post overly optimistic views of their condition or a treatment's success. While we attempt to minimize this bias by only capturing treatment plans in the database that have been validated across multiple patients, it likely still exists.

Second, substantial data cleansing is required to achieve the ultimate clarity present in our database. Data cleansing leads to data loss. Nuanced experiences that may be pivotal aspects of treatment plans may become collateral damage omitted as "noise" for the sake of structure and clarity. The machine learning algorithms employed will be entirely ineffective arriving at insights without heavy cleansing, making some of this loss inevitable.

Third, patients may over-rely on the database and think that it is a replacement for their own oncologist or general practitioner's recommendations for treatment. Just because a computer says something does not mean the recommendations are correct. The database should be used to complement medical expert guidance.

Finally, machine learning models are known to be unreliable in many scenarios. It is feasible that the guidance our algorithms are providing is based on mislabeled training data or inherently biased algorithms. For our approach to be employed for clinical use, robust study will be needed on establishing assurance for these machine learning techniques.

C. Future Work

Despite the stated drawbacks of our approach and the concerns they present, we believe that further study could help to ameliorate some of these challenges. First, we aim to develop a patient portal to improve the usability of our database. The portal should allow patients to determine their cohort. Upon doing so, patients could receive only relevant treatment plans - yielding a truly personalized medicine experience.

Second, we aim to continue refining the medical dictionary and sentiment analysis described. This is a critical component of identifying successful treatment plans and increasing its robustness will considerably improve our recommendations.

Finally, we will seek to include medical practitioners on the team to help refine the machine learning algorithm's accuracy. Currently, there is relatively limited expert judgement present in our 4-phase process.

VI. Conclusion

Our study furnished preliminary evidence that social media data can be used to provide guidance on lines of treatment, disease progression, and adverse events. Patients, loved ones and physicians can leverage our database to improve their understanding of what their future could hold and the options available to them.

Unstructured data on social media contains hidden insight to patient outcomes and treatment plans. While there is still considerable work to do, our approach can help to create personalized treatment plans for patients seeking guidance on their condition, based on others' experiences. We hope that our approach will empower patients to make informed decisions about their treatment while improving their outcomes in the process.

Appendix A
Full Patient Journey

In **June 2014**, when Sarah had just turned **45**, she felt a lump in her right breast. She thought it was just a benign one, similar to that of her mother and aunt. However, when she consulted her doctor, she was advised to go for a **mammogram**. However, both her **mammogram** and **ultrasounds** came out to be normal. Sarah still felt that there was something wrong with her. However, it eventually proved to be quite a difficult task to convince her doctor to conduct a **biopsy**.

According to Sarah, "I persisted and was eventually diagnosed with a 2.8cm IDC with micropapillary features, lymphovascular and perineural invasion."

Finally, Sarah was diagnosed with bony metastasis and visceral metastasis to lung and liver, and she was already at **Stage 4**.

At first, Sarah had to undergo a **bilateral mastectomy**, which was later accompanied with all sorts of complications including an MRSA infection. Sarah had 4 cycles of **Taxotere** and **Cytoxan**.

Sarah described her reaction to chemotherapy as, "The previous chemotherapy (**Taxotere** and **Cytoxan**) nearly killed me (for real)."

Sarah suffered from skin rashes and high fever, after her chemotherapy. Since her liver enzymes were at an elevated level, her chemotherapy got delayed. Moreover, her **alk phos** was **801** and **ALT/AST** was **5-6** times normal.

Unfortunately, cancer did not respond to chemotherapy and it failed to shrink or regress. Then Sarah was started on **hormone therapy** with **Letrozole**, but it resulted in her cholesterol levels increasing dramatically (e.g., **LDL 220** and then **290**).

"I had miserable hot flashes, night sweats and all the signs of a second 'menopause'".

In **July 2016**, the radiologist identified enlarging intrathoracic lymph nodes. Technically, the lymph nodes were "borderline" enlarged. Her doctor recommended **bronchoscopy** with biopsy of a lymph node, and the pathology report came back as **metastatic breast cancer**, still 100% ER+, 20% PR+, **Her2 negative**. The treatment with **Letrozole** was continued for almost **2 years**.

As a result, **Letrozole** was discontinued and she was put on **Faslodex** and **Ibrance**. Since Sarah was already suffering from osteoporosis, therefore, the doctor started **Xgeva**. Sarah was given **Xgeva**, as an injection under the skin, after every **four weeks** but it made her feel miserable for several days. The most side effects that Sarah experienced after having **Xgeva** shots were **jaw pain, shortness of breath and unusual muscle or bone pain**. Within 3 months, only cancer evident on CT scans was an 8 mm lesion in her thoracic spine. Sarah was able to stay on the highest dose of **Ibrance** for about **11 months**. Scans remained stable and her liver function tests also returned to normal. Even **transaminases** normalized after **3 months**. Her **alk phos** levels also dropped to 128 after **two months** and they continued to show a considerable decrease over the next **two months**. **Faslodex**, **Ibrance** and **Xgeva** were continued for more than a year.

The doctor told Sarah that the hot flashes and "menopausal" symptoms were due to hormone therapy (**Letrozole** followed by **Faslodex**). Although Sarah had shown great progress with **Ibrance**, she experienced some side effects as well. These side effects included some **hair thinning**, **fever**, and **mild fatigue**.

Moreover, it was observed that **both her platelet counts and white blood cell counts (especially neutrophils) had lowered. Ibrance** also elevated liver enzymes. Sarah managed to stay on **Ibrance**for **11 months** and avoided bone marrow toxicity. She had a reduced dose of **Ibrance** 100 mg.

Her isoenzymes were elevated for liver, bone, and intestine in **2014**, but after treatment with **Letrozole, Ibrance, Faslodex and Xgeva**, there was a tremendous decrease in her isoenzymes levels in **2017**. Sarah is really happy now to have a normal alk phos and she is even hopeful that the normal value means her bone disease is stable and dormant.

Apart from the aggressive cancer treatment, Sarah was also prescribed to have a healthy and balanced diet, with **fruits and vegetables and yogurt**. The diet itself made Sarah feel a whole lot better.

REFERENCES

[1] A. Koskan, L. Klasko, S. N. Davis, C. K. Gwede, K. J. Wells, A. Kumar, N. Lopez, and C. D. Meade, "Use and Taxonomy of Social Media in Cancer-Related Research: A Systematic Review," American Journal of Public Health, vol. 104, no. 7, 2014.

[2] A. L. Falisi, K. P. Wiseman, A. Gaysynsky, J. K. Scheideler, D. A. Ramin, and W.-ying S. Chou, "Social media for breast cancer survivors: a literature review," Journal of Cancer Survivorship, vol. 11, no. 6, pp. 808–821, 2017.

[3] D. J. Attai, M. S. Cowher, M. Al-Hamadani, J. M. Schoger, A. C. Staley, and J. Landercasper, "Twitter Social Media is an Effective Tool for Breast Cancer Patient Education and Support: Patient-Reported Outcomes by Survey," Journal of Medical Internet Research, vol. 17, no. 7, 2015.

[4] D. Stacey, F. Légaré, N. F. Col, C. L. Bennett, M. J. Barry, K. B. Eden, M. Holmes-Rovner, H. Llewellyn-Thomas, A. Lyddiatt, R. Thomson, L. Trevena, and J. H. C. Wu, "Decision aids for people facing health treatment or screening decisions," Cochrane Database of Systematic Reviews, 2014.

[5] E. C. Dee and G. Lee, "Adverse Effects of Radiotherapy and Chemotherapy for Common Malignancies: What Is the Quality of Information Patients Are Finding Online?," Journal of Cancer Education, 02-Sep-2019. [Online]. Available: https://link.springer.com/article/10.1007/s13187-019-01614-2. [Accessed: 08-Jul-2021].

[6] E. M. Quinn, M. A. Corrigan, S. M. McHugh, D. Murphy, J. O'Mullane, A. D. K. Hill, and H. P. Redmond, "Breast cancer information on the internet: Analysis of accessibility and accuracy," The Breast, vol. 21, no. 4, pp. 514–517, 2012.

[7] J. F. Wagner, D. Lüdders, F. Hoellen, A. Rody, and C. Banz-Jansen, "Treatment anxiety in breast cancer patients," Archives of Gynecology and Obstetrics, 22-Jan-2019. [Online]. Available: https://link.springer.com/article/10.1007/s00404-018-05038-z. [Accessed: 08-Jul-2021].

[8] J. G. Chase, J.-C. Preiser, J. L. Dickson, A. Pironet, Y. S. Chiew, C. G. Pretty, G. M. Shaw, B. Benyo, K. Moeller, S. Safaei, M. Tawhai, P. Hunter, and T. Desaive, "Next-generation, personalised, model-based critical care medicine: a state-of-the art review of in silico virtual patient models, methods, and cohorts, and how to validation them," BioMedical Engineering OnLine, vol. 17, no. 1, 2018.

[9] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories," ACM SIGKDD Explorations Newsletter, vol. 15, no. 2, pp. 20–29, 2014.

[10] M. A. Stewart, "Effective Physician-Patient Communication Heath Outcomes ...," Canadian Medical Association, 1995. [Online]. Available: https://msurgery.ie/wp-content/uploads/2019/09/Effective-Physician-Patient-Communication.pdf. [Accessed: 08-Jul-2021].

[11] M. Naaman, "Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications," Multimedia Tools and Applications, vol. 56, no. 1, pp. 9–34, 2010.

[12] R. Conte, N. Gilbert, G. Bonelli, C. Cioffi-Revilla, G. Deffuant, J. Kertesz, V. Loreto, S. Moat, J.-P. Nadal, A. Sanchez, A. Nowak, A. Flache, M. S. Miguel, and D. Helbing, "Manifesto of computational social science," The European Physical Journal Special Topics, 05-Dec-2012. [Online]. Available: https://link.springer.com/article/10.1140/epjst/e2012-01697-8. [Accessed: 08-Jul-2021].