Understanding Ovarian Cancer Patients by Applying Text Mining to Online Health Community

Adriana Rush[1], Vivian Hui[2], Zhendong Wang[3], Young Ji Lee[2, 4]

[1]University of South Carolina School of Medicine Greenville, [2]University of Pittsburgh School of Nursing, [3]University of Pittsburgh School of Computing and Information, [4]University of Pittsburgh School of Medicine Department of Biomedical Informatics

INTRODUCTION

Unlike more prevalent cancers such as breast cancer or lung cancer, patients with ovarian cancer (OvCa) lack a plethora of resources to learn about their cancer and to seek community. OvCa patients often turn to online health communities (OHC) to seek information, address their concerns, and find community. We aimed to apply text mining using Python to explore contents being discussed in postings submitted to the National Ovarian Cancer Center's (NOCC) OHC.

OBJECTIVE

The aim of this study was to use use postings from OHC to determine which terms and phrases were most frequently used when discussing chemotherapy in the NOCC forum.

BACKGROUND

- Ovarian cancer (OvCa) is the deadliest gynecological cancer in the United States.
  - Estimated 21,750 new cases and 13,940 deaths expected in 2020
- More than half of patients with OvCa face unmet physical, social, or psychological needs related to their cancer care.
- How do we address these unmet needs?
  - In the modern era, people turn to the Internet for answers to questions that they do not know who or when to ask.

Ovarian carcinoma is the second most common gynecologic malignancy and the most common cause of gynecologic cancer death in the United States and other resource-abundant countries. The clinical presentation of EOC (ovarian cancers epithelial in origin) may be acute, subacute, or in some cases, found incidentally during examination, imaging, or surgery for another indication.

- EOCs are histologic diagnoses requiring biopsies for definitive diagnosis.

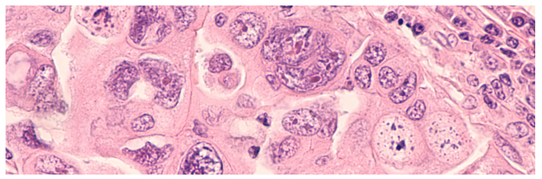
MATERIALS & METHODS

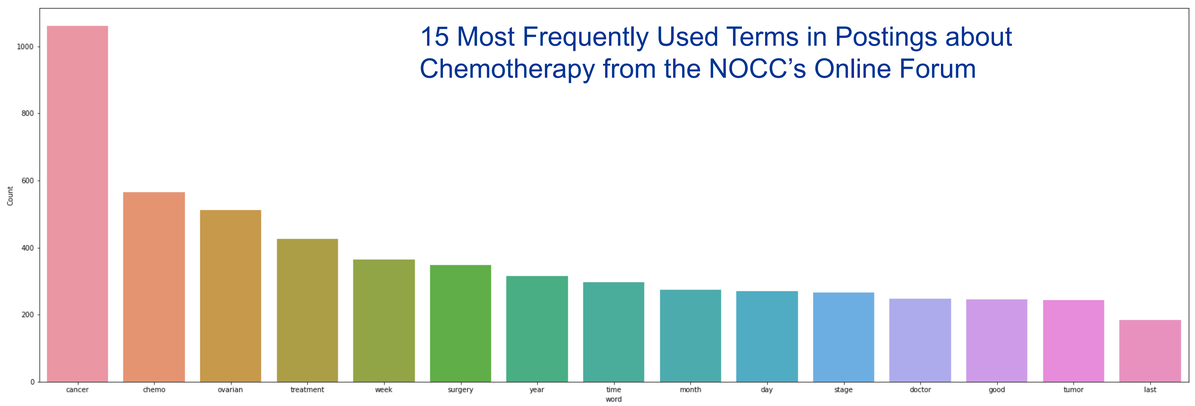
1. Online postings written by patients with ovarian cancer and their caregivers are pulled from the NOCC
2. Postings are annotated as one of fifteen needs (Caregiving, Communication, Complementary Therapy/ Integrative Medicine, Disease Management, Emotional Management, End of Life, Fertility, Genetics, Ovarian Cancer Organizations and Facilities, Practical Needs, Rehabilitation, Sexual Management, Sexuality, Symptom Management, Treatment.)
3. The postings annotated as 'Treatment: Chemotherapy' are then separated to be used separately of the other postings.
4. Annotated postings are then prepared by removing stop words and include synonyms.
5. Python is used to create a text mining resource to derive information from the data set, in this study, to determine frequency of each term or phrase
6. Another code in Python is created to visualize and plot the data and the text mining results

Postings from 2010 through 2020 were extracted for annotation with various topics, such as chemotherapy, symptoms, etc. Postings annotated as 'Treatment: Chemotherapy' were selected and pre-processed for the analysis. Then, we analyzed with Python using text mining to determine which words and phrases appeared most frequently in these postings.

RESULTS

We analyzed 147 postings annotated as 'Treatment: Chemotherapy'. We identified the fifteen most frequently used terms, and these terms included "time", "week", "surgery", and "year" after removing the obvious words/ phrases (e.g. ovarian cancer, chemotherapy, treatment, etc.).





DISCUSSION

Our result suggests that many patients are curious about treatment timelines, with terms "week", "year", "month", and "day" being in the top fifteen used terms. Further, many postings were conversations of how patients decided between "chemotherapy", "surgery", or both. Using computational analysis for these postings can provide new insights for personalized patient care and allow clinicians to learn about the unmet needs of patients with OvCa from the patients themselves. Further study would include the entire postings with advanced techniques such as natural language processing.

LIMITATIONS

Because deidentified postings were used, a major limitation of this work is not knowing the identifiable factors about the patients who posted in these forums (age, gender identity, etc.) Further, the sheer amount of postings did not allow for every single posting from study's time period (2010 – 2020) to be annotated and thus, unannotated postings were not included in this study. Finally, posting to online forum requires technology and a desire to do so that not every patient with ovarian cancer has. Thus, the postings are not entirely reflective of the entire population of patients with OvCa.

REFERENCES

- Lee, Young Ji PhD, MSN, RN; Jang, Hyeju PhD; Campbell, Grace PhD, MSW, RN; Carenini, Giuseppe PhD; Thomas, Teresa PhD, RN; Donovan, Heidi PhD, RN Identifying Language Features Associated With Needs of Ovarian Cancer Patients and Caregivers Using Social Media, Cancer Nursing: February 12, 2021 - Volume Publish Ahead of Print - Issue - doi: 10.1097/NCC.0000000000000928
- Lee, Young Ji PhD, MSN, RN; Park, Albert PhD; Roberge, Mary RN; Donovan, Heidi PhD, RN What Can Social Media Tell Us About Patient Symptoms, Cancer Nursing: July 8, 2020 - Volume Publish Ahead of Print - Issue - doi: 10.1097/NCC.0000000000000860
- Chen L-may, Berek JS. Epithelial carcinoma of the ovary, fallopian tube, and peritoneum: Clinical features and diagnosis Chakrabarti A, editor. UpToDate. 2021 Jun 3.
- Kapadia S. Topic Modeling in Python: Latent Dirichlet Allocation (LDA). Medium. 2019 Apr 14.

ACKNOWLEDGEMENTS and FUNDING

- Many thanks to the research team on the HELPeR project at the University of Pittsburgh (Daqing He, Heidi Donovan, Peter Brusilovisky, Mary Lou Klem, Khushboo Thaker, Yu Chi, and Zhimeng Lou) all of the School of Nursing and/ or the School of Computing and Information, as well as David Boone and Harry Hochheiser of the University of Pittsburgh School of Medicine.
- Financial support was provided by the National Library of Medicine 1R01LM013038-01A1 and T15 grants for the development and implementation of a Health e-Librarian with Personalized Recommender (HELPeR) 5R01LM013038

CONTACT

Adriana Rush aerush@email.sc.edu

Young Ji Lee leeyoung@pitt.edu