International Conference on Digital Disease Detection
#DigDisDet

Michael Paul
Johns Hopkins University and Microsoft Research (internship)
Joint work with Eric Horvitz and Ryen White (Microsoft Research)
Understanding Cancer Patients through Search Engine Query Logs

- What are the information needs of a person diagnosed with cancer?

- How do these needs change over time?
Understanding Cancer Patients through Search Engine Query Logs

• What can we learn about a cancer patient from their search queries?

• Insights into what a person is thinking and planning over the course of an illness.
Search Logs

• **12 TB** of search query logs
  – from Bing + Internet Explorer & toolbar
  – **18 months** of history

• Logs are anonymized; users consent to share
Search Logs

- 140,000 users searched “breast cancer” at least 3 times

- Collected queries containing relevant terms
Ontology of Search Terms

Treatment

treatment
treatments
medication
medications
...
Ontology of Search Terms

Treatment

Chemotherapy
- Side Effects
  - hair loss
  - hair falling
  - lose your hair
...

Radiation
- Side Effects
  - fatigue
  - feel tired
  - feeling tired
...

Alternative
Identifying Likely Cancer Patients

• Looked at partial search histories of 480 users

• Identified 107 users who appeared to have been recently diagnosed
Automatic Classification

• Supervised machine learning

• Trained models using 480 annotated users
Automatic Classification

- Cross-validation performance:
  - 94% precision
  - 29% recall

- Over 2,000 users identified by high-precision classifier as likely to have cancer
Identifying Day of Diagnosis

• Annotated the timelines of the 107 users identified as likely recent cancer patients

• Identified the plausible day of diagnosis
Automatic Classification

• Cross-validation performance:
  – 41% exactly correct
  – 75% within 7 days
Search Volume Over Time

• Can align all user histories around the day of diagnosis

• Analyzed the information users search for before and after diagnosis
Search Volume Over Time

![Chart showing search volume over time]

- **Y-axis**: Average Proportion of Queries
- **X-axis**: Days Since Diagnosis

Legend:
- Other Queries
- Medical Queries
Search Volume Over Time

![Graph showing search volume over time with two lines representing different types of queries. The x-axis is labeled "Days Since Diagnosis" and the y-axis is labeled "Average Proportion of Queries." The graph indicates a decrease in other queries and an increase in medical queries over time.](image-url)
Search Volume Over Time

![Graph showing search volume over time for various categories: Mammography & Biopsy, Symptoms/Concern, Chemotherapy & Radiation, General Healthcare. The x-axis represents days since diagnosis, and the y-axis represents the average proportion of queries.](image-url)
Search Volume Over Time

![Search Volume Over Time Graph](image)

- Mammography & Biopsy
- Symptoms/Concern
- Chemotherapy & Radiation
- General Healthcare

Average Proportion of Queries

Days Since Diagnosis

-300 to 400
Search Volume Over Time

![Graph showing search volume over time for lumpectomy & mastectomy and chemotherapy & radiation. The x-axis represents days since diagnosis, and the y-axis represents average number of queries. The graph has two curves: one for lumpectomy & mastectomy (dashed black line) and one for chemotherapy & radiation (green dotted line). The highest search volume occurs around the time of diagnosis.]
Search Volume Over Time

- Lumpectomy & Mastectomy
- Chemotherapy & Radiation

Average # of Queries vs Days Since Diagnosis
Future Directions

• Deeper analysis of search queries during key points in history

• Break down search histories based on age of user, stage of cancer, or other attributes