A Novel Pipeline for Targeting Breast Cancer Patients on Twitter for Clinical Trial Recruitment

Jon Sligh, Hamed Abedtash, Mengye Yang, Enming Zhang, Josette Jones
Department of BioHealth, School of Informatics and Computing, Indiana University-Purdue University Indianapolis

Background and Preliminary Exploration: Breast cancer is the leading form of cancer in women, estimated to reach the incidence rate of 246,660 in 2016 in the US population. Scientist have developed new therapies for mitigating the disease and side effects in recent years through conducting randomized clinical trials as the gold standard clinical research method. However, recruiting individuals into clinical trials including breast cancer patients has remained a significant challenge. Our preliminary analysis on ClinicalTrial.gov registry showed that the majority of terminated clinical trials were due to recruitment challenges. Out of 525 terminated trials on breast cancer patients registered in the database, 230 (43.8%) of the terminations happened due to low or slow accrual, 34 (6.5%) due to lack of funding, and 31 (5.9%) due to toxicity concerns.

Objectives: In this study, we developed and assess a scalable framework to identify Twitter users who have breast cancer based on personal health mentions on Twitter. In fact, we are looking for “fingerprints” of patients’ health status on Twitter, a microblogging social networking service. This method could provide a new avenue for contacting potential study candidates for recruitment.

Methods: We analyzed the tweets of users who were following at least one of the top 40 twitter accounts where breast cancer patients gather. The rationale behind this approach is that cancer patients are following certain Twitter accounts to access support from other patients, doctors, or healthcare institutions. Consequently, these top twitter accounts provide a central point in which to find actual patients with breast cancer. We retrieved users’ tweets from Twitter API, and processed through the framework to match cancer relevant words and phrases individually and in combinations (caner, benign, malignant, etc.), possessive terms (I, my, has, have, etc.), and supporting attributes (mass, tumor, hair loss, etc.) to determine if the user has been diagnosed with cancer. The performance of the pipeline was measured in terms of sensitivity and specificity of detecting actual breast cancer patients.

Results: We retrieved 25,870,106 tweets of 40 cancer community followers on Twitter. After excluding “retweets” and non-related breast cancer messages, we selected 81,429 tweets for further processing. The developed text processing pipeline could find total of 462 tweets based on the predefined sets of rules, representing 218 unique users. Our new method of Twitter data retrieval and text processing could identify breast cancer patients with remarkable sensitivity of 88.7% and specificity of 91.0%.