Title: Evaluating the Extent to which Posted Personally Identifying Information Jeopardizes the Privacy of Users of Breastcancer.org

Authors: Ayush Sangari 1, Anish Sangari 2, Aditya Sood 3, Nitish Sood 2
Author Affiliations: 1 Renaissance School of Medicine at Stony Brook University, 2 Medical College of Georgia, 3 Emory School of Medicine

Abstract

Background: Prior literature has elucidated the therapeutic benefits for breast cancer patients to engage in global digital communities and support groups before, during, and after treatment which may include surgery. Breast cancer patients post on these forums on the condition of anonymity; however, content posted by themselves may expose their identity. The authors conducted a study to examine the possibility of re-identifying users based on the protected health information (PHI) and personal-identifiable information (PII) shared on breastcancer.org.

Methods: 2,781 posts from 64 users between 2018 and 2020 were analyzed from the thread “Topic: Chemo Starting April 2018”. Each post was analyzed on its content, signature, and location tags using Philter, an open source clinical text de-identification tool developed at UCSF. The amount of PII or PHI related to the disclosure of dates or times, locations, persons, and indications of nationality, religion, or politics in each post was calculated using Microsoft Presidio, a tool designed for scrubbing sensitive text.

Results: Each post content contained an average of 3.96 pieces of PII. 45 (70.3%) of these 64 users had a signature and 29 (45.3%) had a location tag associated with each post. Signatures on average had 4.44 pieces of PHI and location tags had an average of 0.33 pieces of PHI. When including the location tag and signature, each post had an average of 9.52 pieces of PII.

Conclusions: These results suggest that surgical patients are threatened by the possibility of re-identification based on their posts in purported anonymous medical forums.