Using Twitter to Predict County-Level Cancer Mortality: Five Stages of Grief Framework

Emergent Research Forum (ERF)

Anton Ivanov  
University at Buffalo  
antoniva@buffalo.edu

Sanjukta Smith  
University at Buffalo  
sdsmith4@buffalo.edu

Raj Sharman  
University at Buffalo  
rsharman@buffalo.edu

Abstract

Cancer is the leading cause of morbidity and mortality in the United States, resulting in a high economic burden on individuals and the nation. Although community-level factors contribute significantly to health and disease, their assessment using traditional approaches of data collection (e.g., phone surveys or household visits) is costly and has limited spatial and temporal precision. Markedly, when people are affected by cancer, they are likely to use certain psychological language. Since the potential for cancer-related epidemiological discoveries remains largely unexplored, our study utilizes extensive unstructured data available from Twitter, which can be used instead of costly techniques for the purpose of population health assessment. In particular, we aim to answer the following research question: “Does psychological language used by people on the Internet provide an insight into cancer outcomes?”

To answer this question, we adopted the Kübler-Ross model, otherwise known as the five stages of grief, as our theoretical foundation. It postulates a progression of emotional states experienced by both terminally ill patients after diagnosis and by loved ones after a death. The model includes five grief stages: denial, anger, bargaining, depression, and acceptance—psychological effects associated with terminal disease (e.g., cancer). In this light, the language used on Twitter affords an opportunity to characterize community-level psychological correlates of age-adjusted mortality, and, therefore, might be used for population health detection and management. Moreover, given that cancer mortality rates are reported by counties and states with a time lag (currently available official public data sources provide cancer-related statistics for 2015 at the latest), analysis of readily available Twitter data can provide up-to-date estimates of cancer-related outcomes. Our motivation for this study was inspired by the work of Eichstaedt et al. (2015). However, our work is different in that we focus on cancer, adopt and follow a different modeling paradigm, and use newer data for twice as many counties.

To empirically test the proposed relationships, we collected longitudinal panel data ($N = 2827$) from (1) Centers for Disease Control and Prevention, (2) County Health Rankings & Roadmaps, and (3) Twitter (pre-processed 17 Terabytes of tweets). To extract different cues associated with stages of grief from the text, we used a vocabulary-based topic modeling approach, as implemented in Leximancer application. Our empirical design relied on the lagged effects modeling approach. To ensure robustness of our estimates to “bad leverage” points, bias associated with complex nonlinearities (or interactions), and heterogeneity of the marginal effects, we used Kernel-Based Regularized Least Squares (KRLS), a machine learning method.

Our results provide suggestive evidence that the anger, depression, and acceptance stages are significantly and positively associated with cancer mortality. In contrast, denial and bargain effects are insignificant. While denial, the first stage, is far from death, bargaining could be associated with openness to holistic treatment approaches that reduce mortality. The paper contributes to the medical informatics literature by providing conceptual clarifications regarding the association between psychological factors represented by stages of grief (found in unstructured text data) and cancer mortality rates. It also contributes to practice by providing suggestions on using our findings for the purposes of psychological segmentation and promoting health and wellness in the population.

---

"Psychological Language on Twitter Predicts County-Level Heart Disease Mortality," Psychological Science (26:2), pp. 159-160.