CALIFORNIA

Using Text Mining to Extract Systematic Treatment Information (Automation)

California Cancer Registry, Frances B. Maguire, Cyllene R. Morris, Arti Parikh-Patel, Rosemary D. Cress, Theresa H. M. Keegan, Chin-Shang Li, Patrick S. Lin, Kenneth W. Kizer

NATIONAL PROGRAM OF CANCER REGISTRIES SUCCESS STORY

SUMMARY: Surveillance of systemic treatment utilization at the population level can provide insight into dissemination of new or existing cancer treatments. Furthermore, survival outcomes by specific treatment type extend knowledge about the effectiveness of drug regimens among all patients, not just those eligible for clinical trials. Additionally, treatment disparities by source of health insurance, age, race/ethnicity, or socioeconomic status can be identified and addressed. The California Cancer Registry has specific treatment information for patients in an unstructured free-text format. Using SAS-based text mining, the CalCARES Program developed an algorithm to extract specific systemic treatment information for stage IV non-small cell lung cancer (NSCLC) from the free-text treatment text fields. Results were compared to a manual review of the same records. The methodology can be applied to other cancer sites.

CHALLENGES: Because specific treatment information for patients in the California Cancer Registry is contained in an unstructured free-text format and manual review is laborious, the information is infrequently used. Therefore the quality and completeness of the text-field records were unknown. CalCARES manually reviewed the treatment text fields first to serve as a gold standard for comparison to results from text mining.

SOLUTION: CalCARES developed a SAS-based text mining algorithm that searched for specific treatment drugs and classified them into six treatment groups that align with National Comprehensive Cancer Center (NCCN) guidelines for stage IV NSCLC. The SAS-based algorithm used Perl regular expressions and if/then logic in SAS 9.4. Perl regular expressions rely on text string matching and can be used by any SAS programmer and modified for other cancer sites and research questions.

RESULTS: Manual review of 24,845 text field records associated with 17,310 patients diagnosed with stage IV NSCLC from 2012 to 2014 found specific treatment information for 78% of patients. Percent agreement between SAS-based text mining and manual review ranged from 91.1% to 99.4% for the six treatment groups, unknown, and no treatment. Text mining used one-sixth of the time required for the manual abstraction of the same data. Findings indicated that there has been variable utilization of treatments and almost a third (31.7%) of patients did not receive systemic treatment. Disparities in treatment use by socioeconomic status (SES) were apparent.

SUSTAINING SUCCESS: The SAS-based text-mining algorithm is a viable alternative to natural language processing (NLP) which requires a significant amount of development, customization, and expertise and has implementation costs. It is CalCARES' intent to apply the text mining algorithm to other cancer sites, potentially maximizing the utility of extant information in the California Cancer Registry for comparative effectiveness research.