

Analysis of Self-Reported Health Outcomes Data from Web-Based Media Sources

Mark Wolff, Ph.D., SAS Institute, USA

Kenneth Lopiano, Ph.D., SAMSI, USA

Michael Wallis, SAS Institute, USA

ABSTRACT

The process of obtaining valuable information from free text fields and narratives is an important part of any analysis related to health care. Often important clinical outcomes and covariates are embedded within clinician narrative and need to be extracted using text mining tools. These data offer additional information to that captured in electronic health records. An ever-increasing number of individuals routinely contribute and consume solicited and unsolicited information about a variety of health related issues using a wide range of internet and social media channels. Personal web based reports on health status, symptoms, treatments and associated interventions (whether drug-related or behavioral-related) provide a rich source of data that has the potential to complement and inform structured (planned) and unstructured (spontaneous) observations surrounding a range of potential treatments, compounds, approaches and various environmental delivery mechanisms. Although the challenges of deriving information from these free text fields are similar to that of working with clinical notes, a unique challenge arises related to the applicability, utility and veracity of data collected from web-based media sources. Adoption of these data as a resource is hampered by concerns related to the accuracy and reliability of data. We believe that the ability to collect and evaluate these data for veracity would provide a benefit to public health and safety and ultimately contribute to improving healthcare outcomes. This brief abstract at the beginning summarizes and highlights the major points of your paper.

INTRODUCTION

The internet and its reach beyond the desktop via mobile devices, has become a dominant and ever present source of information for many around the world. The development and adoption of Web 2.0 technology and social media applications have further accelerated this trend. Vast amounts of data that are shared and accessed on the internet are unstructured. These data exist as documents, wikis, blogs, comments, tweets, Facebook status updates, crowdsourcing sites, etc. The wide spread popularity of photo sharing sites and the ability to tag images and video has further added to the available pool of unstructured data.

As more and more individuals, organizations and institutions rely on the internet for information to support important decision making, the question of data integrity and veracity becomes ever more critical. Whether a patient making an important personal health care decision or a pharmaceutical company developing and marketing a novel product or the government or other regulatory bodies, monitoring the internet has become part of our daily experience and has become a routine and important component of decision making.

Historically, self-reported clinical outcomes have been derived from direct surveys of patients and formal reports of physicians. In the internet age, many patients are using the internet, including web-based media, social media and web forums, to share information and opinions about healthcare outcomes and adverse effects from drugs and medical devices. These spontaneous personal reports, describing drug-related or health-related data, provide both structured and unstructured information yielding a source of feedback about treatments and patient outcomes. New methods of analysis must be developed in order to make use of this growing body of data. One key element of this is assessing the usefulness or reliability of web-based information. Tools from text mining can be used to gather, validate and analyze the text data to determine its accuracy and reliability. We propose methods to derive relevant information from text gathered from web forums.

Due to the increasing number of unstructured text data on web-based media related to healthcare and patient outcomes, text mining techniques have become very important for analyzing data related to adverse effects of pharmaceuticals and medical devices for postmarketing drug/device safety. Individuals are increasingly using the

PhUSE 2014

internet to find and share information related to healthcare. These data can be structured, but are often in the form of unstructured text. A key challenge in exploiting information from unsolicited web postings is the reliability and relevance of the data. Text posted on web forums may contain information that is inaccurate or biased. Therefore, before web-based information related to drugs and medical devices can be exploited to improve public health and patient safety, the challenge of assessing the accuracy and reliability of those data must be addressed.

ACKNOWLEDGMENTS

SAS Institute, Cary, NC USA

Statistical and Applied Mathematical Sciences Institute (SAMSI), RTP, NC USA

Fatena El-Masri, School of Physics, Astronomy, and Computational Sciences, George Mason University

Karianne Bergen, School of Physics, Astronomy, and Computational Sciences, George Mason University

Obeng Addai, School of Physics, Astronomy, and Computational Sciences, George Mason University

Piaomu Liu, Department of Statistics, University of South Carolina

Shrabanti Chowdhury, Department of Statistics, University of California Riverside

Xin Huang, Department of Mathematics, University of Texas at Dallas

RECOMMENDED READING

Recommended reading lists go after your acknowledgments. This section is not required.

CONTACT INFORMATION

Mark Wolff, Ph.D.

SAS Institute

SAS Campus Drive

Cary, North Carolina 27513 USA

919-531-1548

Mark.wolff@sas.com

:

Brand and product names are trademarks of their respective companies.