

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

Mark Wolff (SAS) and Kenneth Lopiano (SAMSI)

The process of obtaining valuable information from free text fields and narratives is an important part of any analysis related to healthcare. 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.