BUILDING A METADATA FRAMEWORK FOR SOCIAL MEDIA DATA REUSE IN MEDICAL RESEARCH BASED ON RESEARCH WORKFLOW

Background

The data-driven healthcare movement, with its premise that better data will lead to better quality of patient care, is changing the way medical professionals make decisions and take action. Health data comes from various sources in different forms. This current study is particularly aimed at understanding personal health data, especially the user-generated social health data. The goal of this study is to construct a descriptive metadata profile of social health data that may facilitate data curation, sharing, reuse, and eventually collaboration in medical research.

To achieve this goal, a systematic structured keyword search task was conducted on MEDLINE to identify peer-reviewed empirical studies that used social health data as the primary data source. 278 journal articles were identified and 42 were randomly selected and analyzed. All articles were published between 2010 to 2017. Through use of coding with descriptive labels, a complete description of the research process and behaviors related to personal health data reuse were constructed, and a profile that reflects the characteristics of user-generated health data can be modeled.

Research Methods

All articles titles and abstract retrieved in MEDLINE were screened to assess the relevance to the scope of the review using predefined inclusion criteria:

- Must be peer-reviewed and published in juried academic publications.
- Must be qualitative empirical studies reporting data derived from health-related social media data.
- Data analyzed in those past studies is in the form of pure text.

The “methods” section of each selected article was the unit of analysis, and the actual unit of observation for collecting meaningful data was a word, phrase, or a sentence pertaining to the description of data. Recurrent themes, topics, and relationships were defined through qualitative coding (Strauss & Corbin, 1998). Through use of coding with descriptive labels, a complete description of the research process and behaviors related to personal health data reuse were constructed, and a profile that reflects the characteristics of user-generated health data can be modeled.

Preliminary Findings

This study identifies 39 distinct descriptors pertaining to how medical researchers described the data use process, and they could be further grouped into 6 dimensions of characteristics.

DIMENSION // Descriptor
1. **Data Use** // Goal of data use
2. **Data Source** // Name of data site / Nature of the data site / Number of total user of the data collection site / Reason of selecting the data source / Coverage of the data site by user population / Data site function / Data types hosted by the data source / URL of data source site / Coverage of the data site by time span / Data sets volume
3. **Data Collection Process** // Time frame of data collection / Data search strategy / Exclusion criteria / Sampling strategy / Inclusion criteria / Data accessibility / Data recording and storage technique / Limitation of search criteria / Selection reason of data collection time frame / Search tool / Sampling intensity
4. **Nature of the Data** // Sample size / Population size / Data type / Data presentation (display) style
5. **Data Analysis Process** // Analysis method / Data analysis software / Unit of observation (coding unit) / Variables / Data analysis strategy / Data analysis procedures / Measurement of reliability / Reliability agreement rate / Data analysis framework / Theory-generating technique / Data compilation method
6. **Data Policy** // Privacy and confidentiality protection practice / Institutional permission

Implications

The descriptors schema may serve as alternative labels to better managing and curating social media data for the use in the healthcare setting. From the data reusers’ perspective, use of social media as research data in the medical setting is continuously on the rise as the number of social media services and users increase.

To choose the datasets for benchmark, baseline, comparison, and discovery, it requires a full scale of disclosure to the data reusers in terms of how the data is collected, organized and accessed. The descriptors may also serve as a transparent guide for data reusers to gain enough contextual information in the early data reuse process.