Title: A pilot of a cognitive computing system to analyze immunization data

CDC Authors: Stacie Greby, Alexandra Bhatti, Allison Fisher, Yoonjae Kang, Cynthia Knighton, Pamela Srivastava

Abt Authors: Sarah Ball, William Campbell, Alison Thaung, Sarah Hamad

Temple Authors: Marija Stanojevic, Fang Zhou, Zoran Obradovic

Session type: Sharing Session – 60 minutes

- Presentations, panel discussions, or hands-on activities.
    - This session will engage participants in hands-on activities.
- Maximum of 3 speakers, plus 1 moderator allowed.

Session information:

Topic 5: Informatics-Based Solutions to Improving Population Health
Specific topics may include: approaches to addressing problems affecting population health pertaining to a variety of topic areas (e.g., chronic diseases, communicable diseases, natural disasters, preparedness, opioids, immunization, health inequities); conducting and translating research; and accessing, analyzing, and visualizing data across different sectors.

Session Description for Review: (Limit 500 words/4,000 characters, including spaces) (496/3,334)

The goal of immunization programs is to maintain or improve vaccination coverage to prevent diseases. While quantitative and qualitative data are available to support programs, qualitative data are not always actionable because of labor intensive analytic methods. The lack of qualitative data analysis can result in a lack of context for quantitative data, with changes in vaccination coverage not linked to changes in program implementation. Cognitive computing systems, routinely used in business settings, can address this public health challenge by leveraging automation for quantitative and qualitative data analysis. In this session we will describe the development of an immunization lexicon, or local language, and an immunization program cognitive computing system (IPCCS) to analyze immunization data. The session includes three presentations: the development of the lexicon and IPCCS, the challenges that came up during development and solutions used by the team to address the issues, and a demonstration of the IPCCS.

Text data from a variety of formal and informal sources were collected to develop the lexicon. Formal data included immunization program reports; vaccine-related scientific documents, journals, websites, and books; and current state vaccination-related laws.. Informal data included Sysomos searches from November 2016 to May 2018 of Twitter, online forums, and news , and social media feeds.

Challenges were identified that researchers may encounter while developing similar systems, including:

1. Data collection:
    a. The qualitative data was not was not collected as systematically or formatted as well as quantitative data.
    b. Much of the formal data on this project did not include several years of similar data.
    c. The formal data, unlike the Twitter data, did not have the same granular level of geographic data, making spatial and temporal identification difficult.
    d. To overcome these challenges, the team collected a much higher volume of data than initially expected.
2. Data cleaning and pre-processing:

    a. Qualitative data may be ambiguous when a single term has different meanings. For example, “MMR” was a query term that could refer to measles, mumps, rubella or “Match Making Rank”, a rating system used in online gaming.
    b. Qualitative data contained references to external documents, non-relevant data, jargon, typos, and misspellings.
    c. To address these challenges, the data were iteratively and thoroughly cleaned and filtered.
3. Creation and evaluation of an effective cognitive computing system:
    a. Customized features for the lexicon needed to be developed to ensure that the results of the cognitive computing system could be useful to researchers, including classification of whether a portion of text was vaccine hesitant or vaccine confident, aggregation of text by state, and evaluation of the cognitive computing system’s different algorithms

The pilot IPCCS can quickly search formal and informal textual immunization data. The formal data generated by the programs may be helpful in identifying activities associated with changes in vaccination coverage. The informal data from online sources may be useful in assessing what information is being shared during an outbreak or other emergencies. The lexicon databases will need regular updating to continue to be relevant.

Session Description for Publication: (Limit 50 words/350 characters, including spaces) (44/347)

Immunization programs maintain/improve vaccination coverage to prevent diseases. While quantitative and qualitative data are available to support programs, qualitative data are not always actionable because analysis is labor intensive. We describe development of an immunization lexicon and cognitive computing system to analyze immunization data.

Learning Objectives:

1. Identify at least one new or innovative technology, strategy, or use case that exemplifies the successful application of public health informatics and population health improvement.
2. Describe at least one approach, challenge, or success to enabling or enhancing electronic information exchange.
3. Describe at least one current barrier or facilitator to leveraging electronic information for population health.