Linguistic Differences between Spanish and English Tweets that Mention Opioids

Michael Chary[1], Alex Manini[2]

[1]NewYork Presbyterian\Queens, Flushing, New York, USA, [2]Elmhurst Hospital Center, Elmhurst, New York, USA

Objective: Social media provides invaluable data for analyzing the nonmedical use of prescription opioids (NMUPO). Drug abuse research to date using social media has focused on English language communications despite the growing importance of Spanish language communications. English and Spanish tweets may discuss different aspects of the nonmedical use of prescription opioids (NMUPO).

Methods: This is a prospective study of publicly available tweets. We included tweets geocoded from the US that contained >1 Spanish/English keyword related to NMUPO. We used English keywords, or their Spanish translation, from previous work. We compared these tweets with tweets that contained >1 word in English or Spanish, not restricted to NMUPO key words. We compared Spanish and English tweets on lexical diversity, Flesh-Kincaid grade level, Jacquard similarity, and most common words.

Results: We acquired 64.909 tweets geotagged for the US over one month that mentioned English or Spanish NMUPD keywords. acquired tweets. Spanish and English NMUPD tweets had comparable lexical diversities (0.252 for English, 0.260 Spanish, p=0.773, two-tailed t-test). General English tweets had significantly greater lexical diversity than general Spanish tweets (0.930 for English, 0.468 for Spanish; p<0.0001 two-tailed t-test). General English tweets had significantly higher lexical diversity than NUMPD English tweets (0.93 vs 0.252; p<0.0001, two-tailed t-test), as did Spanish tweets (0.468 vs 0.26; p<0.0001, two-tailed t-test). Spanish tweets had greater Flesch-Kincaid Reading Ease than English tweets (96.48 vs 63.02 for NUMPD p<0.0001; two-tailed t-test, 93.18 vs 64.71, p<0.0001; two-tailed t-test). The Flesch-Kincaid Reading Ease of English NMUPD vs general tweets was not significantly different (63.02 vs 6.71; p=0.88, two-tailed t-test), nor was it for Spanish tweets (93.18 vs 94.68; p=0.63, two-tailed t-test). The Jaccard similarity between English and English-translated Spanish NMUPD tweets was not significantly different from 0 (0.026, p=0.88; two-sample Kolmogorov-Smirnov test).

Conclusion: Spanish and English NMUPO tweets use different vocabularies; English tweets use more unique words when discussing NMUPO than other topics and than Spanish NMUPO tweets. Spanish NMUPO tweets use fewer words with shorter syllables than English NMUPO tweets.