# Understanding health care demands through social media engagement

*Keywords:social media engagement, healthcare, personality, computational social science, reddit*

## Extended Abstract

### Introduction

4 in 5 US adults seek healthcare advice online, making social media a common tool for public health officials interested in reaching potential healthcare clients. Similarly, epidemiologists and computer scientists use social media data to study health conditions' prevalence, outbreaks, and evolution. The underlying assumption is that social media and official statistics assess the same concept and hence should converge. However, this is not always the case, with social media-based estimates regularly over- and underestimating certain conditions' prevalence, without there being a clear theory about the underlying reasons. Using the largest online derived medical taxonomy (derived from 130M social media posts by half a million users from the US) we show that there is implicit knowledge to be gained from looking at cases where official and social media data diverge (i.e., hypo and hyper engagement in online health discussions about conditions compared to their official prevalence). Specifically, we find that socio-demographic variables, while strongly related to real-life prevalence rates, are not strongly associated with people engaging in healthcare discussions online. Rather, health discussions are influenced by non-cognitive characteristics (i.e., personality). The divergence of real-life prevalence and social media-implied prevalence presents a unique chance for researchers and healthcare professionals alike as it allows to identify regions with underserved clients whose health-related concerns and questions are not fulfilled by traditional providers.

### Data collection and sample

Our work relies on four main data sources, i) social media data to derive health conditions prevalence ii) official prevalence data to benchmark our data and derive deviations iii) control variables that are likely to explain divergences, falling into a socio-demographic or non-cognitive (personality) domain.

i) Social media data: To obtain social media health indices for a wide set of conditions, we relied on the procedure previously developed to derive a health taxonomy from social media (Šćepanović et al., 2021). The first step is to extract all mentions of medical conditions and symptoms from the Reddit discussions by applying the Natural Language Processing medical entity extraction method MedDL (Šćepanović et al., 2020). The extracted medical mentions from Reddit were then connected into a network based on their co-occurrence in the same Reddit post. Such a co-occurrence network captures the semantic relatedness of those mentions: those that appeared often together are likely to describe the same condition. By applying the community detection algorithm Infomap (Rosvall, 2008) on the network the innate health taxonomy was uncovered.

ii) Official prevalence data: We collected official data from the Centers for Disease Control and Prevention (CDC) and from the Substance Abuse and Mental Health Services Administration (SAMHSA), both of which regularly publish health statistics in the U.S. To best match our categories, from CDC, we gathered state-level prevalence statistics for arthritis, and self-reported 'mentally unhealthy days'. From SAMHSA, we collected statistics on the

prevalence of: mental illnesses, abuse of different substances (e.g., heroin), conditions linked to metabolic syndrome (e.g., diabetes prevalence), and Sexually Transmitted Diseases (STDs). All SAMHSA statistics were compiled between 2017 and 2018. In total, we collected 17 health statistics.

iii) Control variables (personality): The most commonly used framework of personality is the Big Five (Soto et al., 2017). The lexical approach has been used successfully to predict political, economic, social and health-related outcomes. The 5 personality traits are 1) openness (tendency to be curious, creative and open-minded), 2) conscientiousness (tendency to be organised, responsible and self-disciplined), 3) extraversion (tendency to be sociable, assertive and energetic), 4) agreeableness (tendency to be compassionate, compliant and trusting) and 5) neuroticism (tendency to be anxious, fearful and emotionally unstable) (Soto et al., 2017; John et al, 1999). We use the US personality samples collected by the Gosling-Potter Internet Personality Project, comprising of 3.1M participants.

**Results**

In this paper, we drew from the largest social media-derived medical taxonomy to examine whether biases in the data could provide a window into healthcare-seeking behaviour. This paper offers several important conceptual and empirical contributions to the literature. Firstly, we highlight that there are regions where people seek healthcare advice at a higher rate than expected online. This indicates that traditional healthcare providers seems to not meet the required demand in those regions. Secondly, we empirically investigate likely reasons for this hyper engagement by condition. We find that four distinct health categories exhibit hyper/hypo engagement for different reasons. This changes our theoretical understanding about social media biases by highlighting that they do not only represent inaccuracies but rather windows into behaviour. Further, understanding the exact reasons for the biases is crucial for practice, as it informs classical as well as telehealth providers about how to reach unserved communities.

We first calculate the ranks for official and social media health scores across medical conditions. For illustration purposes, let's look at HIV as an example condition. The ranks for some states are consistent between the two measures (e.g., California and Florida rank highly on both measures) and for some states, the ranks differ significantly (e.g., Mississippi has a much better rank according to social media than officially). Subsequently, we investigate the degree to which socio-demographic, as well as non-cognitive attributes, are associated with hyper engagement, i.e., the degree to which social media discussions are prevalent above what we would expect based on the official health score. We find a positive relationship between hyper engagement and education, GDP, male-to-female ratio and openness, and a negative relationship (hypo engagement) with the rate of elderly, cultural tightness, religiosity, republican vote-share, agreeableness, conscientiousness, and neuroticism. To further understand if the associations are universal, or condition-specific, we investigate the structure of the conditions. The 17 conditions fall into four, theoretical medical categories, namely 1) metabolic conditions 2) sexually transmitted conditions, 3) mental health conditions and 4) substance (ab)use. While the previous results were focused on getting directionality and strength of associations, the degree to which each control variable explains variance in hyper engagement is of interest, too. We relied on dominance analyses, a permutation-based linear regression approach, running a total of 2244 models to determine variable importance. The results highlight the diverse nature of the conditions, with, e.g., male-to-female ratio being predictive of variance in sexual conditions, but not very much for most other conditions.

# References

References provided in-text

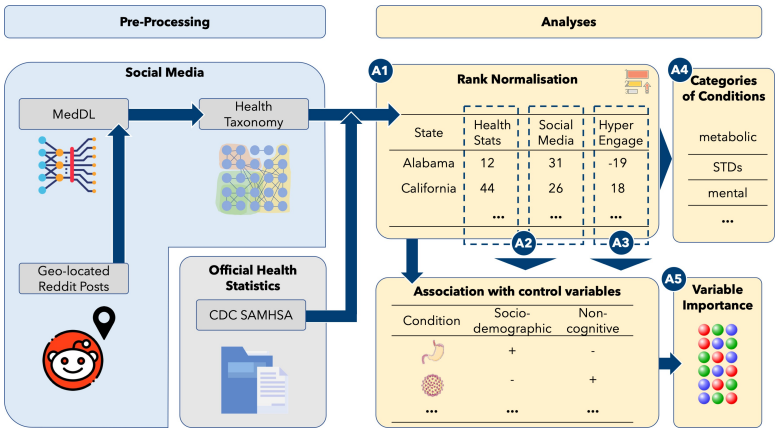


Figure 1. The overview of our methodology.