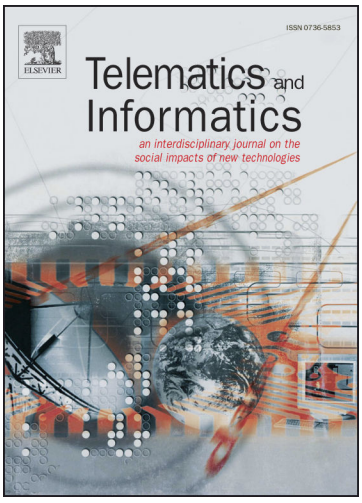
Journal Pre-proofs



Information-seeking vs. sharing: Which explains regional health? An analysis of Google Search and Twitter trends

Kokil Jaidka, Johannes Eichstaedt, Salvatore Giorgi, H. Andrew Schwartz, Niv Efron, Lyle H Ungar

PII: S0736-5853(20)30199-4
DOI: https://doi.org/10.1016/j.tele.2020.101540
Reference: TELE 101540

To appear in: *Telematics and Informatics*

Received Date: 6 April 2020
Revised Date: 9 November 2020
Accepted Date: 27 November 2020

Please cite this article as: Jaidka, K., Eichstaedt, J., Giorgi, S., Schwartz, H.A., Efron, N., Ungar, L.H., Information-seeking vs. sharing: Which explains regional health? An analysis of Google Search and Twitter trends, *Telematics and Informatics* (2020), doi: https://doi.org/10.1016/j.tele.2020.101540

Information-seeking vs. sharing: Which explains regional health? An analysis of Google Search and Twitter trends

Keywords: Google Trends, Twitter, topic modeling, machine learning, county health, information-seeking, search behavior

1. Introduction

Information-seeking behavior for health is an everyday activity among American Internet users and has been proven as a valuable resource for profiling some health conditions, as a complement to costly and time-consuming surveys and censuses [**?**]. At the regional level, a body of work has illustrated that trends in search behavior can be used to predict influenza epidemics [**?**], vaccination uptake [**?**] and dengue [**?**] outbreaks. The Pew Research Centre reported that[1] 73% of all Americans use the Internet, and 83% of all users reported to using Google most often [**?**]. Google is a search engine developed in 1998 and is the most popular internet search engine in the United States since 2004.

Americans also spend nearly a quarter of their time online on social networking platforms, where they can consume information about their friends, and share personal information about themselves. Social media platforms such as Twitter provide heightened interpersonal connectivity. Users are open to disclosing personal information online because they can communicate quickly and easily with their friends on social media. When users disclose their personal and health concerns online, it has been found to improve their online relationships with an increase in enacted and perceived social support [**?**], while also being

[1] Although newer studies report higher numbers of users, we refer to the survey conducted in 2011 because our analysis focuses on Google and Twitter trends from 2011-2013.

predictive of their health [**? ? ?**] and risky behavior, such as alcohol use [**?**].

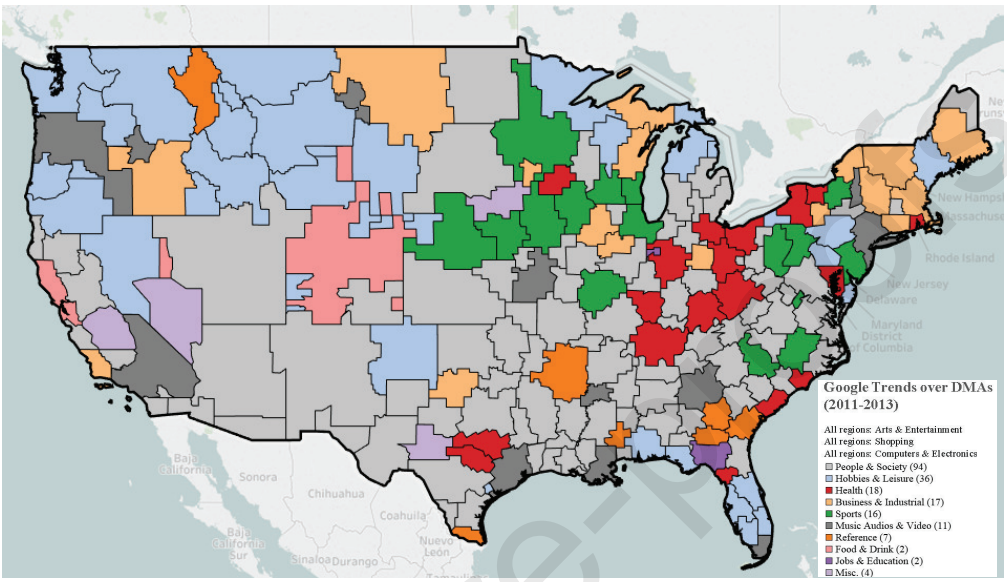


Figure 1: A map of USA. (Alaska and Hawaii are not pictured) showing the highest Google search categories for the period 2011-2013 at the DMA (Designated Market Area) level.

Previous studies suggest that the language of individuals can be aggregated to represent regions, and these provide valid estimates of community health [**? ? ? ?**] and region level well-being [**?**]. However, to our knowledge, there is no previous work that benchmarks the efficacy of predictive models based on information-sharing (i.e., the language of social media posts on Twitter or Facebook) against those based on information-seeking (i.e., the language of search queries from Google Search) for predicting the trends in regional health. We pose the following research questions:

- **RQ 1. How well does information-seeking behavior predict region-level health?**
- **RQ 1a. How does it compare with information-sharing behavior in terms of absolute predictive accuracy?**

- **RQ2. What are the kind of insights provided by information-sharing behavior about communities and their health?**
- **RQ3. What are the kind of insights provided by information-seeking behavior about communities and their health?**

We focus on a set of eight non-communicable diseases (NCDs), which account for a major proportion of medical spending [**?**]. We chose indicators of impending health risks such as poor health, physical inactivity, and alcohol and tobacco use. We also focused on diseases such as heart disease and strokes, diabetes, and obesity. Since both Google and Twitter elicit text-based input from users, we demonstrate that it is possible to train language models on the semantic features (Google entities vs. Twitter topics) of search volume and Twitter posts and compare their accuracy at predicting regional health. We also compare the language of significantly associated searching and posting behavior to identify how either format provides *different* insights into population health. All our analyses are conducted at the regional level; thus, we have only analyzed anonymized data aggregated according to the region where a Google search or a Twitter post was made.

1.1. The relationship of culture and lifestyle with health

Online behavior is reflective of users' innate characteristics, such as their society, lifestyle, culture, and personal health [**?**]. Socioeconomic status (SES) plays a vital role in determining one's flexibility in choosing their lifestyle, and its relationship with health is mediated by lifestyle characteristics such as budget and time constraints [**? ? ?**]. In studies of internet consumption, lifestyle is typically modeled along a few related dimensions – e.g., character, professional information and hobbies, and interests [**?**]. For health research, lifestyle is used to understand how likely people are to engage in certain health-risk behaviors – e.g., cigarette use and alcohol consumption – alternatively, how likely they are to engage in exercise or healthful living [**? ?**].

We are motivated by previous literature to focus on how digital traces that are representative of lifestyle, culture, and personal disclosures are predictive of health outcomes. Social media platforms such as Twitter would potentially afford the opportunity to examine personal disclosures when people share information with their friends. Realistically, we would not expect Twitter language to predict NCDs as well as Google searches, since these diseases usually affect an older, rural population as compared to Twitter's user base. Instead, they are likely to be indicative of the environmental and cultural differences that bring about the disparities in health.

On the other hand, seeking information online is likely to be an individual's first course towards an official diagnosis and treatment when they suspect that they are suffering from a health problem. Because search engines such as Google are typically used in a private setting, they are more likely to be devoid of social censoring and the social stigma that accompanies the disclosure of many health conditions [**?**]. They are expected to track socioeconomic variations more closely than Twitter by virtue of being more widely used by Americans.

2. Related work

Online websites provide different endpoints where people can seek or share information related to their life. It is thus imperative to understand how uncensored searches and public posts on social media compare in their association with regional trends in health. In simply understanding the difference between the two, the study by De Choudhury, Morris, and White [**?**] suggested that while seeking health information on search engines fulfills cognitive needs, sharing health information on social media is intended to garner emotional support.

2.1. Predicting regional health with information-seeking behavior

According to prior surveys, search engines act as the entry point into the web for 90% of those who go online: whether to read and send email (90%), seek directions (86%), look for medical information (80%), buy a product (75%),

seek the news (72%), visit a government website (66%), watch a video (62%) or other reasons [**?**]. Trends based on Google searches have been used in different fields of healthcare research to describe or monitor infectious diseases [**?** **?**], vaccination uptake [**?**], mental health and substance abuse [**?** **?**], NCDs [**?**] and other health factors [**?** **?**] (for a complete survey, see [**?**]). Studies have demonstrated the ability to track influenza epidemics [**?**], although such analyses are not without caveats [**?**]. The fact that most of these studies achieved a correlation of > 0.70 against a reference standard data source demonstrates the vast potential of Google search as a proxy for monitoring population health.

Previous work [**?** **?**] has approached the problem of modeling regional trends in health with the assumption that people with health conditions would search for their symptoms for the purpose of (a) self-diagnosis, (b) fulfilling an information need, and (c) identifying possibly courses of treatment [**?**]. They have accordingly mined the search volume distribution of a small set of seed words that are manually selected based on their semantic relevance to the diseases under consideration. Other studies [**?** **?** **?**] have also used search time series from such seed words to predict the different health factors for a subsequent year to a regional level. There is evidence to support that search trends are associated with the *lifestyles* of poor health [**?**]; for instance, search words about junk food ('McDonald's,' 'Dominos'), sedentary lifestyle ('weather,' 'books,' 'movies') or weight loss were able to predict over 93% of the Body Mass Index (BMI) average in a metro area. However, these studies rely heavily on the selection of variables; furthermore, they may miss an implicit signal from other co-associated search trends. In contrast, the present study obtained unfettered access to region-level Google Search trends, organized as an entity hierarchy[2]. Their predictive power was evaluated against a similar set of Twitter trends organized as a topic hierarchy. This offers a paradigm to compare searching and sharing behavior of communities and their explanatory power for region-

[2]https://goo.gl/LeJLgq

level health.

2.2. Predicting regional health with information-sharing behavior

Information-sharing on social media offers a platform to understand the social connectedness of communities or how they indulge in personal disclosure and social buffering. Previous studies have shown that the language of Twitter posts can be predictive of regional health statistics, such as the obesity [**?**], and age-adjusted mortality from atherosclerotic heart disease (AHD) in counties [**?**]. Furthermore, use of social media platforms and online self-disclosure is considered to have stress-buffering effects leading to enhanced well-being [**?**], and to alleviate symptoms of mental health in a young population [**?**]. They have reported that communities with poorer health and well-being are more likely to express negative emotions, boredom, and loneliness on social media [**?** **?**].

As compared to Google, Twitter's user base is skewed towards a younger demographic, because 36% of Americans 18-29 are on Twitter as compared to 22% of those aged 30-64 [**?**]. Non-white ethnicities, such as African-Americans and Hispanics, are over-represented on Twitter as compared to the actual population, which is considered a "black cultural outlet" [**?**]. The words people use on Twitter can be aggregated to region-level trends using geolocation information, to predict life satisfaction [**?**], physical fitness [**?**], flu epidemics [**?** **?**] and heart disease [**?**]. Culotta [**?**] explored the relationship of different health-related statistics such as obesity, health insurance coverage, and access to healthy foods, with the Twitter activity for the top 100 most populous counties in the US. The study found a significant correlation of Twitter language with six health statistics and demonstrated that Twitter-derived information could improve predictive accuracy for most of the statistics. Eichstaedt et al. [**?**] demonstrated that a cross-sectional regression model based only on Twitter language is better predictive of age-adjusted atherosclerotic heart disease (AHD) mortality than a model that combined 10 common demographic attributes, socioeconomic status, and co-morbid health risk factors.

Table 1: The health outcomes being investigated in this paper. Source: The County Health Rankings Project, 2011-2013

Acronym	County-level outcomes	Description
		Physical ill-health
PHY	% Physically inactive	Percent of adults that report no leisure time physical activity
FAI	% Fair/poor health	Percent of adults that report fair or poor health (age-adjusted)
		Chronic illnesses
OBE	% Obese	Percent of adults that report BMI >= 30
DIA	% Diabetic	Percent of adults that reported being diabetic
CHD	Coronary heart disease	Hospitalizations due to coronary heart disease, per 1000 individuals
STR	All strokes	Hospitalizations due to all coronary strokes, per 1000 individuals
		Alcohol & tobacco use
BIN	% Binge drinkers	Percent of adults reporting binge drinking in past 30 days
SMO	% Smokers	Percent of adults that reported currently smoking

3. Data

We focus on a set of eight non-communicable diseases (NCDs), which account for a major proportion of medical spending [**?**]. We chose diseases such as heart disease, diabetes, and obesity, as well as indicators of impending health risks such as poor health, physical inactivity, and alcohol and tobacco use.

For each of the counties in the US, we collect demographic attributes, socioeconomic statistics, and eight health-related factors for the years 2011-2013 and aggregate them to the DMA-level. In this section, we describe our data collection method for (a) Google search volumes and (b) Twitter posts.

3.1. Health outcomes and demographic information

For the years 2011-2013, we identified eight health factors (Table 1) compiled by the US County Health Rankings and Roadmaps[3] from a wide range of sources including the Behavioral Risk Factor Surveillance System (BRFSS), American Communities Survey and the National Center for Health Statistics. Our selection of health factors is based on the NCD framework defined by WHO [**?**]. We use a population-weighted mean to aggregate all the statistics to the DMA-level.

[3] www.countyhealthrankings.org

Some of the chosen health variables are inter-correlated, and many are more likely to be prevalent in communities with older individuals and within certain races [**?**]. For instance, older populations are typically at greater risk for coronary heart diseases [**?**], while binge drinking is a behavior typically associated with young adults in the US [**?**]. Individuals with certain racial backgrounds such as African-Americans are, on average, more prone to diabetes and less prone to excessive drinking than Caucasians [**?**]. The pairwise correlations for our choice of variables are provided in the supplementary materials.

3.2. DMA-level Google search volumes

The Google Search Trends used in this study were provided to the authors as a normalized frequency count of search trends in 27 categories and their subcategories for the 210 Digital Marketing Areas (DMAs) for a calendar year. Google Search Trends are calculated by Google using deep learning to interpret users' original queries into an entity hierarchy of 27 categories and their 1188 subcategories, that comprise the Google Knowledge Graph. The Google Search Trends data includes search volumes on categories that cannot be explored on the Google Trends Explorer, such as 'Adult' and 'Sensitive Subjects' such as 'Death and Tragedy'.

The map in Figure 1 depicts the normalized Google Search trends for each DMA between the years 2011-2013 (Alaska and Hawaii are not pictured here). DMA segmentation was first established for market-based targeting by Nielson Media Research and is still widely used by companies for advertising and customer analytics. In the figure, DMAs are shaded according to some of the top Google Trend categories. The most popular high-level Google search categories for 2011-2013 across all DMAs in the US is Arts & Entertainment (*'Youtube', 'Netflix', 'Facebook'*) and Shopping (*'Amazon', 'eBay', 'Walmart' and 'shoes'*). We conducted a one-way ANOVA and found no significant differences across the year-on-year values for 2011-2013. We focused on the categories that were likely to be relevant to our research questions, as identified in previous work [**? ? ?**].

Table 2: Description of the selected Google Trend categories

Category name	**Example sub-categories**
	Health
Health	Drugs & Medication, Drugs & Alcohol testing
	Economic status
Arts & Entertainment	Country Music, Body Art, Religious Music
Automobiles	Chevrolet, Ford
Travel	Adventure travel, Hiking, Travel guides & travelogues
Jobs & Education	Classifieds
	Interests and Hobbies
Hobbies & Leisure	Weddings
Beauty & Fitness	Yoga & pilates, Running & walking
Shopping	Coupons, discounts & offers, Mass merchants department stores
Sports	Combat Sports, Hockey, Wrestling
Adult	Soft Porn, Gay Porn, Extreme Porn
Firearms & Weapons	Firearms & Weapons
	Food & Culture
Food	Fast food, Wine, Fruits & Vegetables
People & Society	Christianity, Buddhism, Spirituality
Death & Tragedy	Funeral & Bereavement

3.3. DMA-Level social media data

The Twitter data was obtained from the County Lexical Bank [**?**] and comprises 10% of the Twitter garden hose sample of public tweets collected over 2011-13. The data was prepared by following many steps for language filtering, geo-mapping approximately 20% of the tweets, and then tokenizing and aggregating the features to the user level, which are detailed in their work. We obtained their sample of normalized user frequencies for 2011-2013 and mapped the users into the respective DMAs. By aggregating the user-level features to the DMA-level, we obtained the DMA-level distribution of words across the 208 DMAs in the United States. These were further transformed into topic-level

Table 3: Description of the selected Twitter topics

Category name	**Example sub-categories**
	Health and Fitness
Exercise & diet	market, farmers, invest, stock, owned, produce, goods, owns
Hiking	creek, grand, hike, hiking, trail, pine, oak, cave, slam
Sleep & Tiredness	rested, ready, gotta, work, bed, nap, collapse, shower, home
	Economic status
Commuting	bus, taxi, driver, cab, buses, rides, ride, drivers, route
Jobs & job-seeking	management, engineering, development, communication, systems
Academic assignments	thesis, proposal, presentation, submitted, touches, research
Travel & Sightseeing	maps, directions, location, route, map, gps, china, directions
	Interests and Hobbies
Internet & online chatting	chrome, translate, google, search, typed, logo, images, wiki
Photography	cameras, camera, digital, photographer, photography, capture
The Universe	outer, aliens, space, alien, whale, rocket, launch, nasa
Solitary activities	watchin, chillin, sittin, home, tv, movies, drinkin, eatin
Sports & the Olympics	olympic, opening, ceremony, faces, places, closing
	Family, relationships and culture
Prayers & the Bible	thanking, prayed, bless, writes, allowing, god, praying, blessing
Family	flops, mommy, daddy, flip, mummy, daddy's, baka, sissy, dear
	Friends and social events
Friends	hangin, havin, bestie, besties, wit, chillin, wif, fam, kick
Social events	company, laughs, drinks, excellent, food, picnic, atmosphere
Other people	andy, somebody's, nothin, i'ma, heather, homie, nobody's, shawty
	Emotional expression
Positive emotion	adore, xoxo, admire, extraordinary, absolutly, love
Rumination	thinkin, talkin, bout, wonderin, layin, sittin, somethin, wonderin
Feelings & Reactions	adore, hurt, feelings, betrayed, ignore, apologize
Dishonesty	lie, cheat, steal, lying, kill, hate
Boredom	bored, hell, freaking, mind, ideas, usual

features representing each of the selected topics from the lexicon of Schwartz et al. [**?**].

4. Approach: Predictive models and insights

In the following paragraphs, we describe our approach to predicting spatial variations and then obtaining linguistic insights to answer our research questions. This study focuses on regional variations and not temporal variation, because the health factors are not expected to change significantly from year-to-year. This is because the Behavioral Risk Factor Surveillance System's (BRFSS) annual report of health factors typically comprises aggregated statistics collected over many previous years (e.g., % Obesity in the 2015 County Health Report comprises data from the year 2011).

4.1. Feature selection

In the present study, we focused on comparing a similar set of lifestyle indicators across Google Trends (Table 2) and Twitter topics (Table 3) in terms of their associations with population health. First, we selected trends related to searching and mentioning health and health conditions. Next, we expected that searches and mentions related to **automobiles, travel, jobs, and education** would be relevant to our analysis because they would reveal the average socioeconomic status of the community. The relationship between such indicators and health has been discussed elsewhere [**?**]. Searches and mentions about the common interests and hobbies of a region would provide cultural and lifestyle insights, e.g., **beauty and fitness, shopping, sports, adult entertainment, and firearms and weapons**, which are organized by the Google Knowledge Graph under 'Interests and Hobbies.' These, too, have been found to be associated with desirable cardiovascular indicators [**?**] and can be considered markers of culture [**?**]. Searches and mentions related to **food** would be expected to directly associate with health, since previous studies have shown that regions with higher searches for junk food were significantly more likely

to suffer from higher BMI [**?**]. Finally, previous work has shown that greater mentions of family and religion on Twitter were associated with higher diabetes and obesity [**?**]. Accordingly, we also included searches and mentions related to **people and society, death, and tragedy** as direct indicators of culture [**?**]. In Table 3, we identify the corresponding set of Twitter topics, which were selected out of a set of 2000 topics modeled from a large social media corpus in previous work [**?**]. Although there was no equivalent in the Google Knowledge Graph, we also included the Twitter topics which reflect personal disclosures – e.g., mentions of **feelings** [**?**], **loneliness** [**?**] and **stress** [**?**], which have been found in previous work to indicate poor personal health and regional health [**?**].

4.2. Predictive analysis

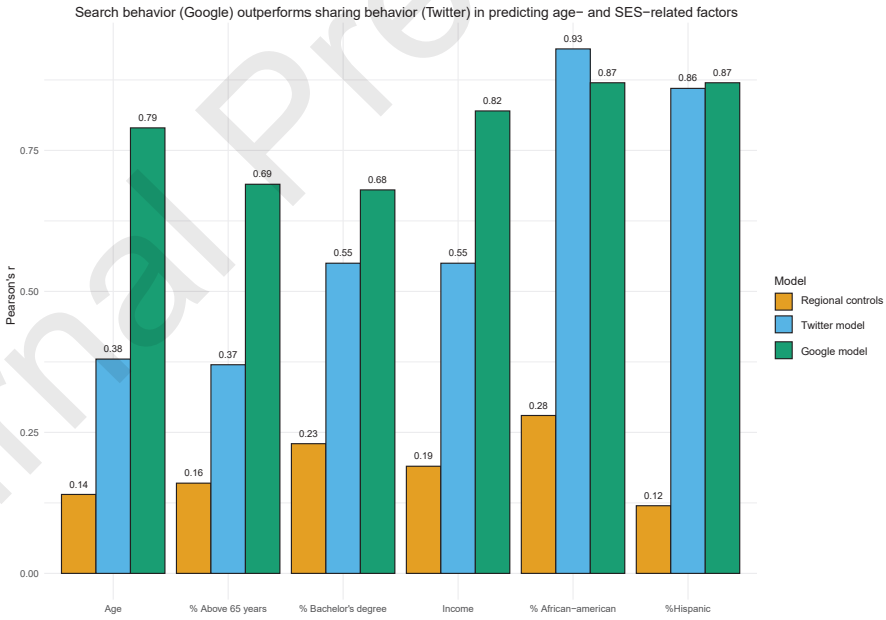
First, we wanted to understand whether Twitter or Google better explains demographic variance across the United States. To train our predictive models, we performed dimensionality reduction because we have thousands of Twitter and Google features corresponding to only 208 observations (DMAs). To reduce the list of predictors, we performed feature selection to discard those predictors, which were strongly inter-correlated (r >= 0.65) or were not significantly correlated with the dependent variable in a univariate analysis. We then performed a Principal Component Analysis to transform each of the large sets of Google and Twitter features into a smaller set of components. Next, we assigned DMAs into ten folds at random and conducted a ten-fold cross-validated linear regression for each of the dependent variables. We tested several regularization methods such as ridge and elastic-net. Finally, we reported the Pearson's r correlations for actual vs. predicted values on held out DMAs, using ridge regularization in linear regression models. There were no significant differences in performance among the ten folds.

Finally, we explored the linguistic indicators of health factors on Twitter and Google through an ordinary least squares regression analysis.

5. Results

5.1. Predictive modeling: Predicting spatial variations in regional demographics and SES

Exploratory analysis reported in Figure 2 shows that search behavior is better at profiling demographic features of a population such as the mean age ($r = .79$, $p < .001$), percentage of population over 65 ($r = .69$, $p < .001$), education ($r = .68$, $p < .001$) and income ($r = .82$, $p < .001$) as compared to sharing behavior. Search and sharing behavior were at par for predicting population ethnicities, such as the percentage of Hispanics and African-Americans in the population ($r = .55$, $p < .001$). All the correlations were significant at $p < 0.01$. These results support our expectation that in general, search behavior would be more representative of a population.



4.pdf 4.bb

Figure 2: Search behavior (Google) outperforms sharing behavior (Twitter) in predicting age- and SES-related factors in a ten-fold cross-validation across DMAs. * depicts that a one-way ANOVA revealed that the two models were significantly different from each other at the 0.01 level.

5.2. Predictive modeling: Predicting spatial variations in health factors

To answer **RQ1** and **RQ1a**, Figure 3 provides the predictive performances of models based on information-sharing (Twitter), and searching (Google) behavior on the eight health factors, reported in terms of the held out Pearson's r values. For the baseline model trained on demographic and SES variables, adding information behavior, whether seeking or sharing, always significantly improved predictive performance ($p < .001$).

We find that for all the models, the prediction of the Twitter and Google models are significantly correlated with the health statistic ($p< .001$). These Twitter results consider a different sample size as compared to the previous study by Culotta [**?**], where counties rather than DMAs were considered, and a different approach was followed. They show a performance improvement over those set of results. Results with regional controls and a combination of Twitter and Google signals are provided in the supplement.

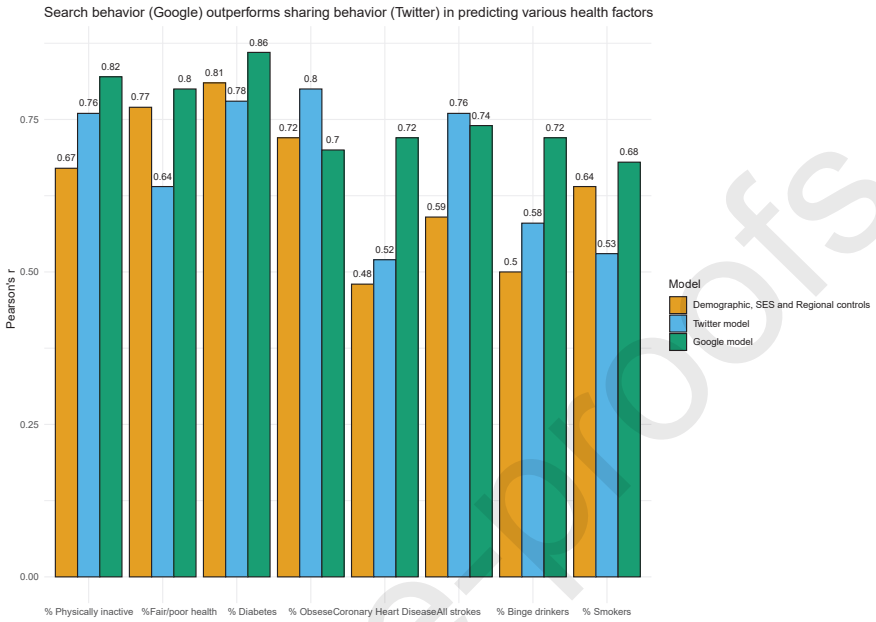
Google search significantly outperforms Twitter in 6 out of 8 metrics, with an average performance gain of over 15% over Twitter and an average gain of 19% over demographic and regional controls. The largest performance gain of Google over Twitter is observed for predicting fair/poor health ($r = .80$, $p < .001$) and coronary heart disease ($r = .72$, $p < .001$) – co-morbid conditions which are mostly reported by an older population. Twitter significantly outperforms Google in predicting % obese ($r = .80$, Performance gain = 14%).

6. Language insights

6.1. Language insights at the DMA level: Twitter topics vs. Google search categories

To answer **RQ2** and **RQ3**, we computed the univariate Simes-corrected[4] partial Pearson correlations between each of selected Twitter topics and Google

[4]We have presented all our results with Simes-correction. We found that the correlations remained significant with Bonferroni p-correction, which is a more conservative method.



5.pdf 5.bb

Figure 3: Predictive performance reported as average Pearson's r calculated on the held-out data over ten-fold cross-validation across DMAs. All the correlations are significant, Simes-corrected at p<.001. $*$ depicts that the model significantly improved upon the model on its left according to a one-way ANOVA.

search categories for 2011-2013, with eight health factors, controlling for our set of demographic and SES factors. We present our results in two groups: a. Physical ill-health and chronic illnesses, and b. Alcohol and tobacco use. Table 4a provides the Twitter topics, which are significantly positively (negatively) correlated with unhealthier(healthier) communities for the first six of our eight health factors. These diseases often have similar linguistic indicators on Twitter, associated with greater emotional expression in self-disclosure (.30<=r<=.53), mentions of family and friends (.32<=r<=.58), and sleep and tiredness (.34<=r<=.50). Table 4b provides the top Google search categories, which are significantly correlated with the same health factors. Regions where searches for fast food are high are also likely to have higher rates among these indicators of physical ill-health and chronic disease. Google Search trends are more

Table 4: Simes-corrected significant Pearson correlations ($p < .01$) of (a) Twitter topics and (b) Google search categories with physical ill-health and chronic illnesses controlling for demographic and SES factors. Topic labels are manually created. Disease acronyms are expanded in Table 1.

(a) Twitter

Twitter topic label	PHY	FAI	OBE	DIA	CHD	STR
Rumination	.60	-	.68	.54	-	.47
Family	.58	.32	.56	.56	.56	.37
Friends	.51	.52	.58	.42	-	.46
Sleep & tiredness	.44	.50	.60	.34	-	-
Feelings & reactions	.38	.49	.53	.30	-	-
Internet	-.46	-	-.49	-.37	-	-.44
Exercise & diet	-.42	-.37	-.41	-.34	-	-.45
Travel & sightseeing	-.40	-	-.48	-.32	-	-.38
The universe	-.39	-.34	-.47	-.34	-	-.40
Professional jobs & job-seeking	-.39	-.36	-.44	-.34	-	-.43
Academic assignments	-.34	-.46	-.30	-.31	-	-.47
Photography & films	-.34	-	-.44	-.27	-	-.28
Snow	-	-.55	-	-	-	-.41
Sports	-	-.55	-	-.13	-	-.35
Social events	-	-.44	-	-.23	-	-.51
Hiking	-	-.32	-	-.22	-	-.52

(b) Google

Google search category	PHY	FAI	OBE	DIA	CHD	STR
(Arts & Entertainment) Country music	.30	-	-	-	-	.18
(Death & Tragedy) Funeral & bereavement	.28	-	.33	-	-	-
(Automobiles) Chevrolet	.33	-	.28	-	.35	-
(Automobiles) Ford	.29	-	-	-	.32	.25
(Health) Drugs & medication	.23	-	-	-	-	.25
(Health) Drug & alcohol testing	.17	.12	.26	-	-	-
(Arts & Entertainment) Body art	.20	.33	-	-	-	-
(Sports) Wrestling	.37	-	.30	-	.36	.28
(Food and Drink) Fast food	.29	-	.23	.22	.39	.39
(People & Society) Christianity	.26	-	.19	.21	.24	-
(People & Society) Religious music	.23	-	-	-	-	.29
(Hobbies & Leisure) Weddings	.24	-	.18	.13	.26	-
(People & Society) Family & relationships	.23	-	-	-	.24	-
(Adult) Extreme porn	-	.21	-	-	-	-
(Adult) Gay porn	-	.20	-	-	-	-
(Shopping) Mass merchants & department stores	.25	-	-	-	.38	.25
(Shopping) Coupons, discounts & offers	.22	-	-	-	-	-
(Beauty & fitness) Yoga & pilates	-.50	-.29	-.43	-.26	-.43	-.36
(Beauty & fitness) Bicycles & accessories	-.43	-	-.30	-	-	-
(Travel) Hiking & camping	-.38	-	-	-	-	-
(Sports) Skiing & snowboarding	-.28	-.35	-	-	-	-.28
(Sports) Running & walking	-.14	-.31	-	-	-	-
(Travel) Ecology environment	-.37	-	-.26	-	-	-
(Travel) Specialty travel	-.32	-	-.27	-	-	-
(Travel) Travel guides & Travelogues	-.30	-	-.34	-	-.33	-.28
(Travel) Adventure travel	-.23	-	-	-	-	-
(Beauty & fitness) Massage therapy	-.33	-	-.25	-	-.37	-.28
(People & Society) Buddhism	-.42	-.25	-.24	-	-.35	-.25
(People & Society) Self-Help & motivational	-.31	-	-.32	-.24	-	-
(People & Society) Spirituality	-.30	-	-.30	-.23	-.33	-.28
(Food & Drink) Fruits & vegetables	-.28	-	-.21	-.23	-.33	-.28

PHY: % Physically inactive, Percent of adults that report no leisure time physical activity

FAI: % Fair/poor health, Percent of adults that report fair or poor health (age-adjusted)

OBE: % Obese, Percent of adults that report BMI >= 30

DIA: % Diabetic, Percent of adults that reported being diabetic

CHD: Coronary heart disease, Hospitalizations due to coronary heart disease, per 1000 individuals

STR: All strokes, Hospitalizations due to all coronary strokes, per 1000 individuals

diverse and disease specific: searches of funerals ($.28<=r<=.33$), and medicines (drugs) ($.23<=r<=.25$) are higher in physically inactive and obese regions. A

cultural pattern that also emerged from the data was the prominence of searches for Christianity (.19<=r<=.26) and weddings (.13<=r<=.26) among regions with higher rates of NCDs. Note also, that these regions have higher mentions of family and friends on Twitter, but not on Google Search.

Table 5: (a) Simes-corrected Pearson correlations of (a) Twitter topics and (b) Google search categories with alcohol and tobacco use when controlling for demographic and socioeconomic factors. All correlations are significant, Simes-corrected at $p < .01$, two-tailed t-test.

(a) Twitter

Twitter topic label	% Binge Drinkers	% Smokers
Rumination	-	.47
Commuting	-	.65
Getting ready	-	.80
Friends	-	.54
Dishonesty	-	.56
Sleep & tiredness	-	.47
Other people	-	.58
Feelings & reactions	-	.57
Solitary activities	-	.59
Prayers & the Bible	-.43	-
Family & relationships	-.46	.64
Positive emotion	-.40	-
Interjections	-.36	-

(b) Google

Google search category	% Binge drinkers	% Smokers
(Sports) Hockey	.45	-
(Sports) Team sports	.30	-
(Sports) Boating	.27	-
(Sports) Ice skating	.25	-
(Food & Drink) Fast food	-	.34
(Automobiles) Chevrolet	-	.27
(Shopping) Coupons, discounts & offers	-	.24
(Health) Drugs & medications	-	.24
(Sports) Motor sports		.23
(Games) Massive multiplayer	-	.22
(People & Society) Body art		.21
Death & tragedy	.34	-
(Arts & Entertainment) Country music	-	.42
Firearms & weapons	-	.40
(People & Society) Religious music	-.54	-
(People & Society) Christianity	-.55	-
(People & Society) Classifieds	-.28	-
(Travel) Theme Parks	-.26	-
(Sports) Cheerleading	-.25	-
(Beauty & Fitness) Yoga & pilates	-	-.36
(Sports) Winter sports	-	-.28
(Sports) Cycling	-	-.29
(People & Society) Buddhism	-	-.30
(People & Society) Spirituality	-	-.22

BIN: Binge drinkers, Percent of adults reporting binge drinking in past 30 days

SMO: Smokers, Percent of adults that reported currently smoking

In Table 5a, we observe that there were no Twitter topics that positively correlated with % of binge drinkers. Table 5b shows that regions with a larger population of binge drinkers was also more likely to search for sporting events ($r = .25 < \mathrm{r} < .45$). The results appear face valid, as regions with a higher percentage of binge drinkers are likely to be college football towns [**?**]. Regions with a large population of smokers, on the other hand, were also more likely to have Twitter posts about work (getting ready ($r = .80$), commuting

(r = .65)), introspection (rumination (r = .47), and feelings and reactions (r = .57)). Once again, note how Twitter and Google trends offer very different insights into regions with a higher percentage of smokers. Regions with more smokers are likely indicative of a more stressful and solitary lifestyle of an older population, where people are more likely to ruminate on social media, search for and potentially participate in online gaming communities (massive multiplayer r = .22), and search for firearms (r = 0.40).

The information-seeking and sharing behavior of healthier communities, on the other hand, tends to converge. The search behavior has references to exercise (yoga & pilates), outdoor activities (running & walking, bicycles & accessories, hiking & camping), travel (specialty travel, adventure travel) and spirituality (self-help & motivational, Buddhism). Correspondingly, the social media posts are more likely to mention travel, exercise, jobs and school, hobbies (films, photography), leisure activities (socializing), and sports (hockey).

6.2. Signals of socioeconomic and cultural variance

Our findings suggest that digital traces are effective tools for diagnostics. Information behavior captures the cultural and socioeconomic contexts of their respective regions. Delving deeper into the results, we repeated the main analysis at a word- and query-level. We also found instances of cultural and socioeconomic variances when, for instance, African-American linguistic signals and country music correlated with NCDs, or search trends about tanning lotions and rap songs were predictive of higher diabetes. A majority of the linguistic signals on Twitter were directly representative of psychological insights, while those from Google search queries were directly representative of medicines, diseases and symptoms. The social media language of communities with higher coronary heart disease and strokes refers to tiredness, boredom and fatigue, feelings, and rumination, which also matches previous findings as the topics most correlated with county-level AHD mortality [**?**].

7. Discussion

This is the first study to offer a comparison between the predictive efficacy of public vs. private information behavior on the internet, and it uses the aggregate data from billions of individual signals of behavior to evaluate the predictive efficacy of such signals for a spatial analysis. While correlation should not be mistaken for causation, our findings are useful in comparing how information-sharing and seeking behavior manifested through linguistic patterns, across an extended time, can help to understand regional health. The online information behavior of communities reflects their processes of cognitive reflection and affective expression, and are further indicative of their general health. These signals are stronger than those obtained by looking at the demographic constitution of the communities.

Our study has two major takeaways. First, our findings suggest the potential of information-seeking behavior as a signal for understanding regional populations which is less likely to suffer from social desirability or social censoring biases. Information-seeking behavior is relatively less explored as a data source, but it appears to track regional sociodemographics better than social media datasets. Communities with higher % physically inactive and % obese are more likely to search for co-morbid symptoms. Communities with higher % smokers are more likely to search for symptoms and guns. Unhealthier communities are more likely to search for death and drugs. All of these topics are not visible in their social media vocabulary, and therein lies the advantage of using search engine logs.

Our second takeaway is that online information-sharing vs. seeking offer complementary signals to better understand the information behavior of ill health. It is much harder to distinguish how the psychological insights from information-sharing could correspond to different health outcomes, as it generally reflects a low mood, unhappiness, and tiredness across all outcomes. Information-seeking offers rich context to health problems; for instance, communities with poorer health are more likely to search for medicines, daily sedentary

activities such as watching TV, or unhealthy eating habits such as fast food.

An underlying reason why models based on information-seeking perform better than information-sharing would be the differences in their user bases, since Twitter users are typically middle-aged or younger professionals in urban areas. Fortunately, such biases can be corrected posthoc. A recent study demonstrated that by correcting the common sociodemographic biases in the data [**?**], Twitter language models can improve on the county-level prediction of % fair/poor health by two percentage points to $r = 0.78$.

The insights from our study have many policy implications. Firstly, they can be used to design culturally-conscious interventions and awareness programs. We recommend targeted interventions for at-risk populations, which may work better than traditional counterparts based on socioeconomic conditions. As of August 2017, Google search redirects users searching for depression to an online questionnaire to help them diagnose themselves and seek professional help if needed [**?**]. Linguistic insights could also help in the better design of health websites that motivate self-awareness and disclosure, facilitating timely help and outreach for undiagnosed patients in a cost-effective manner [**?**]. An understanding of the cultural distinctions of different diseases in different communities can help to adapt health messages to the cultural markers of the target audience [**?**], plan focused interventions, and design new public health campaigns. For instance, public health messages about healthy eating can target communities with higher % diabetics and %physically inactive by advertising on religious websites and distributing flyers offline after church masses.

Secondly, we also recommend that digital traces of information behavior could be used for health risk management. Individuals can act more responsible when they are at higher risk as a community [**?**]. Insights from information-seeking and searching have the potential to detect risky behavior – e.g., alcohol use, drug use, or smoking – and motivate a confrontational style to generate awareness in a region. Previous work has recommended that such insights are best implemented on social media platforms themselves, where specific groups can be targeted, rather than a general-purpose information awareness campaign

in schools, public squares, or community centers. Health advisories for different age groups can be published at online and offline venues which are are the most frequented by them. For instance, advisories against binge drinking could be positioned at websites that are popular with the youth, such as sports websites (ESPN, NBA, NFL, and the Superbowl).

Finally, the psychological insights from into different NCDs can also help medical organizations in planning online awareness and counseling services, and aid hospitals in recommending mental health assessments for patients with chronic illnesses in order to help them better manage their physical health and recovery.

Like other cross-sectional studies of its kind, this study does have a few limitations. Firstly, as stated above, it makes no claim to causation, and instead provides a way to understand digital audiences in the United States based on their online behavior. Secondly, studies based on correlation of trends are wont to suffer from omitted variable biases. Thirdly, both Google search and Twitter data would not be representative of the general US population. However, recent studies suggest that even stratifying digital traces to make them more representative of the US census yields similar patterns as those obtained before stratification. Fourthly, while the Google variables were based on 100% of the searches conducted in the United States, the Twitter data is based on a random 1% data stream of tweets.

[1] S. Sarigul, H. Rui, et al., Nowcasting obesity in the us using google search volume data, in: 2014 AAEA/EAAE/CAES Joint Symposium: Social Networks, Social Media and the Economics of Food, May 29-30, 2014, Montreal, Canada, no. 166113, Agricultural and Applied Economics Association & Canadian Agricultural Economics Society & European Association of Agricultural Economists, 2014.

[2] A. Culotta, Towards detecting influenza epidemics by analyzing twitter messages, in: Proceedings of the first workshop on social media analytics, ACM, 2010, pp. 115–122.

[3] N. Dalum Hansen, C. Lioma, K. Mølbak, Ensemble learned vaccination uptake prediction using web search queries, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, ACM, 2016, pp. 1953–1956.

[4] E. H. Chan, V. Sahai, C. Conrad, J. S. Brownstein, Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance, PLoS Negl Trop Dis 5 (5) (2011) e1206.

[5] K. Purcell, J. Brenner, L. Rainie, Search engine use 2012.

[6] R. Zhang, The stress-buffering effect of self-disclosure on facebook: an examination of stressful life events, social support, and mental health among college students, Computers in Human Behavior 75 (2017) 527–537.

[7] L. Manikonda, M. De Choudhury, Modeling and understanding visual attributes of mental health disclosures in social media, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, 2017, pp. 170–181.

[8] H. A. Schwartz, J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, L. Ungar, Towards assessing changes in degree of depression through facebook, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 118–125.

[9] T. E. Simoncic, K. R. Kuhlman, I. Vargas, S. Houchins, N. L. Lopez-Duran, Facebook use and depressive symptomatology: Investigating the role of neuroticism and extraversion in youth, Computers in human behavior 40 (2014) 1–5.

[10] J. J. van Hoof, J. Bekkers, M. van Vuuren, Son, you're smoking on facebook! college students' disclosures on social networking sites as indicators of real-life risk behaviors, Computers in human behavior 34 (2014) 249–257.

[11] S. Giorgi, V. Lynn, S. Matz, L. Ungar, H. A. Schwartz, Correcting sociodemographic selection biases for accurate population prediction from social media, arXiv preprint arXiv:1911.03855.

[12] S. Giorgi, D. Preoţiuc-Pietro, A. Buffone, D. Rieman, L. Ungar, H. A. Schwartz, The remarkable benefit of user-level aggregation for lexical-based population-level predictions, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1167–1172.

[13] A. Culotta, Estimating county health statistics with twitter, in: Proceedings of the 32nd annual ACM conference on Human factors in computing systems, ACM, 2014, pp. 1335–1344.

[14] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, L. Weeg, L. H. Ungar, M. E. Seligman, Psychological language on twitter predicts county-level heart disease mortality, Psychological science 26 (2) (2015) 159–169.

[15] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. Seligman, et al., Characterizing geographic variation in well-being using tweets, in: Proceedings of the International AAAI Conference on Web and Social Media, ICWSM, 2013, pp. 583–591.

[16] S. S. Lim, T. Vos, A. D. Flaxman, G. Danaei, K. Shibuya, H. Adair-Rohani, M. A. AlMazroa, M. Amann, H. R. Anderson, K. G. Andrews, et al., A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010, The lancet 380 (9859) (2012) 2224–2260.

[17] P. J. Rentfrow, S. D. Gosling, M. Jokela, D. J. Stillwell, M. Kosinski, J. Potter, Divided we stand: Three psychological regions of the united

states and their political, economic, social, and health correlates., Journal of Personality and Social Psychology 105 (6) (2013) 996.

[18] P. Contoyannis, A. M. Jones, Socio-economic status, health and lifestyle, Journal of health economics 23 (5) (2004) 965–995.

[19] C. K. Coursaris, W. Van Osch, Lifestyle-technology fit: Theorizing the role of self-identity in is research, Computers in Human Behavior 49 (2015) 460–476.

[20] A. Giddens, Modernity and self-identity: Self and society in the late modern age, Stanford university press, 1991.

[21] M. de Reuver, H. Bouwman, Explaining mobile internet services adoption by context-of-use and lifestyle, in: Proceedings of the Ninth International Conference on Mobile Business and 2010 Ninth Global Mobility Roundtable (ICMB-GMR), IEEE, 2010, pp. 176–182.

[22] A. Glendinning, L. Hendry, J. Shucksmith, Lifestyle, health and social class in adolescence, Social science & medicine 41 (2) (1995) 235–248.

[23] M. De Choudhury, M. R. Morris, R. W. White, Seeking and sharing health information online: comparing search engines and social media, in: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, 2014, pp. 1365–1376.

[24] J. Sydney, S. Fox, Generations online in 2009, Pew Internet & American Life Project.

[25] L. J. Carr, S. I. Dunsiger, Search query data to monitor interest in behavior change: application for public health, PloS one 7 (10) (2012) e48158.

[26] T. Nguyen, T. Tran, W. Luo, S. Gupta, S. Rana, D. Phung, M. Nichols, L. Millar, S. Venkatesh, S. Allender, Web search activity data accurately predict population chronic disease risk in the usa, J Epidemiol Community Health 69 (7) (2015) 693–699.

[27] J. Ojala, E. Zagheni, F. C. Billari, I. Weber, Fertility and its meaning: Evidence from search behavior, arXiv preprint arXiv:1703.03935 (2017) 640–643.

[28] C. F. Ricketts, C. G. Silva, An analysis of morbidity and mortality using google trends, Journal of Human Behavior in the Social Environment (2017) 1–12.

[29] S. V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R. P. Dreyer, S. I. Chen, K. Murugiah, The use of google trends in health care research: a systematic review, PloS one 9 (10) (2014) e109583.

[30] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (7232) (2009) 1012–1014.

[31] D. Lazer, R. Kennedy, G. King, A. Vespignani, The parable of google flu: traps in big data analysis, Science 343 (6176) (2014) 1203–1205.

[32] M. Pittman, B. Reich, Social media and loneliness: Why an instagram picture may be worth more than a thousand twitter words, Computers in Human Behavior 62 (2016) 155–167.

[33] A. Perrin, M. Duggan, Americans' internet access: 2000-2015, Tech. Rep. 6, Pew Research Center (2015).

[34] A. Brock, From the blackhand side: Twitter as a cultural conversation, Journal of Broadcasting & Electronic Media 56 (4) (2012) 529–549.

[35] V. Lampos, T. De Bie, N. Cristianini, Flu detector-tracking epidemics on twitter, Machine Learning and Knowledge Discovery in Databases (2010) 599–602.

[36] D. M. Rabi, A. L. Edwards, D. A. Southern, L. W. Svenson, P. M. Sargious, P. Norton, E. T. Larsen, W. A. Ghali, Association of socio-economic status

with diabetes prevalence and utilization of diabetes care services, BMC Health Services Research 6 (1) (2006) 124.

[37] J. L. Rodgers, J. Jones, S. I. Bolleddu, S. Vanthenapalli, L. E. Rodgers, K. Shah, K. Karia, S. K. Panguluri, Cardiovascular risks associated with gender and aging, Journal of cardiovascular development and disease 6 (2) (2019) 19.

[38] T. S. Naimi, R. D. Brewer, A. Mokdad, C. Denny, M. K. Serdula, J. S. Marks, Binge drinking among us adults, Jama 289 (1) (2003) 70–75.

[39] R. A. Bell, E. J. Mayer-Davis, J. W. Beyer, R. B. D'agostino, J. M. Lawrence, B. Linder, L. L. Liu, S. M. Marcovina, B. L. Rodriguez, D. Williams, et al., Diabetes in non-hispanic white youth prevalence, incidence, and clinical characteristics: the search for diabetes in youth study, Diabetes Care 32 (Supplement 2) (2009) S102–S111.

[40] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al., Personality, gender, and age in the language of social media: The open-vocabulary approach, PloS one 8 (9) (2013) e73791.

[41] J. S. Feinstein, The relationship between socioeconomic status and health: a review of the literature, The Milbank Quarterly (1993) 279–322.

[42] K. Saihara, S. Hamasaki, S. Ishida, T. Kataoka, A. Yoshikawa, K. Orihara, M. Ogawa, N. Oketani, T. Fukudome, N. Atsuchi, et al., Enjoying hobbies is related to desirable cardiovascular effects, Heart and vessels 25 (2) (2010) 113–120.

[43] S. M. Gelber, Hobbies: Leisure and the culture of work in America, Columbia University Press, 1999.

[44] S. C. Guntuku, A. Buffone, K. Jaidka, J. C. Eichstaedt, L. H. Ungar, Understanding and measuring psychological stress using social media, in:

Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13, 2019, pp. 214–225.

[45] NIAAA, College drinking fact sheet.

[46] M. Gilberti, Learning more about clinical depression with the phq-9 questionnaire, The Keyword.

[47] Y. J. Sah, W. Peng, Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website, Computers in Human Behavior 45 (2015) 392–401.

[48] M. J. Dutta, Communicating about culture and health: Theorizing culture-centered and cultural sensitivity approaches, Communication Theory 17 (3) (2007) 304–328.