Machine Learning Models for Predicting, Understanding, and Influencing Health Perception

ADA AKA AND SUDEEP BHATIA

ABSTRACT Lay perceptions of medical conditions and treatments determine people's health behaviors, guide biomedical research funding, and have important consequences for both individual and societal well-being. Yet it has been nearly impossible to quantitatively predict lay health perceptions for hundreds of everyday diseases due to the myriad psychological forces governing health-related attitudes and beliefs. Here we present a data-driven approach that uses text explanations on healthcare websites, combined with large-scale survey data, to train a machine learning model capable of predicting lay health perception. We use our model to analyze how language influences health perceptions, interpret the psychological underpinnings of health judgment, and quantify differences between different descriptions of disease states. Our model is accurate, cost-effective, and scalable and offers researchers and practitioners a new tool for studying health-related attitudes and beliefs.

In 2020 and 2021, COVID-19 disrupted billions of lives. Throughout the pandemic, medical authorities had stressed the importance of face masks as a precautionary measure for reducing the spread of the disease. However, many people rejected this advice, which aggravated the global health crisis and endangered the lives of others. Although there were many determinants of face mask avoidance and related behaviors, one key factor was the perceived severity of COVID-19 (Cheung 2020). Researchers found that understanding such health perceptions was necessary for influencing and improving behavior during the crisis (Van Bavel et al. 2020).

Of course, the importance of health perception in preventative health behavior extends beyond COVID-19. Significant previous work has shown that the perceived severity of a health problem and the perceived vulnerability of the decision maker are significant factors for determining precautionary behavior, in domains as diverse as vaccination and nutrition (Becker et al. 1977; Janz and Becker 1984; Van Der Pligt 1998; Weinstein 2000). More generally, there is a close relationship between perceptions of health outcomes (e.g., lung cancer), preventative health decisions (e.g., exercising regularly), and the likelihood of engaging in risky behaviors (e.g., smoking; Harrison, Mullen, and Green 1992; Brewer et al. 2007; Sheeran, Harris, and Epton 2014). Thus, to improve health decision making, it is essential to predict and understand how people perceive different diseases and other health-related outcomes (Li and Chapman 2013; Betsch, Bohm, and Chapman 2015). Modeling lay health perceptions can also help guide healthcare funding decisions, as patients and other stakeholders often have insights that complement those of healthcare professionals (Cornwall and Jewkes 1995; Entwistle et al. 1998).

How can we predict, understand, and influence people's health perceptions for common disease states? One approach is to use objective health outcomes that measure quality of life implications and assess the severity of different medical conditions associated with the disease state (Calvert and Freemantle 2003). One of the simplest of such measures is the count of raw mortality—the number of deaths caused by the health condition. A related measure is called the "years of life lost," which quantifies the number of years of life lost due to premature death. Other measures such as "disability-adjusted

Ada Aka (corresponding author: adaaka@wharton.upenn.edu) is a PhD candidate, and Sudeep Bhatia is an assistant professor in the Department of Psychology at the University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA. Funding was received from the National Science Foundation grant SES-1847794, the Alfred P. Sloan Foundation, Wharton Risk Management Center's Russell Ackoff Doctoral Student Fellowship, and Wharton Behavioral Lab. The authors thank Liat Gradstein for excellent research assistantship, and Russell Richie, Wanling Zou, and Grant Jirka for helpful discussion. Their data and models are accessible at https://osf.io/peym7/.

Issue Editors: Dipankar Chakravarti, Jian Ni, and Meng Zhu

Published online March 4, 2022.

Journal of the Association for Consumer Research, volume 7, number 2, April 2022. © 2022 Association for Consumer Research. All rights reserved. Published by The University of Chicago Press for the Association for Consumer Research. https://doi.org/10.1086/718456

life years" (DALYs) account for the quality of life lost due to health conditions that are not fatal.

These measures are useful for describing the global burden of diseases and for allocating limited medical resources (Arnesen and Kapiriri 2004). However, considerable work has found that such measures do not correlate with people's health state perceptions (Slovic and Peters 2006). That is, people are not actually good at evaluating the objective risk and severity of different health states or treatments (Lloyd 2001). Rather, their judgments rely on emotion, memory, language, and other psychological cues, which occasionally lead to perceptions that deviate from objective health outcomes like the mortality rates (Chapman and Elstein 2000; Slovic et al. 2002; Chapman and Coups 2006; Peters and Meilleur 2016). More generally, variables such as perceived pain, disability, physical distress, anxiety, depression, negative mood, functional inability, permanent state of ill health, and death are particularly strong predictors of (low) health perceptions (Garrity, Somes, and Marx 1978; Kaplan and Camacho 1983; Fýlkesnes and Førde 1992; Idler 1993; Krause and Jay 1994; Farmer and Ferraro 1997; Idler, Hudson, and Leventhal 1999).

The best way to predict health perceptions may thus be to understand the associations between a disease state and variables like perceived pain and disability, as well as the emotions and cognitions that these variables provoke. These associations depend critically on how information about the health state is obtained (Covello and Peters 2002; Vahabi 2007; Keller and Lehmann 2008), which increasingly is from health websites and other internet sources (Atkinson, Saperstein, and Pleis 2009). In fact, it has been argued that the internet forms the first opinion and doctor's advice is relegated to the second opinion for consumers of online health information (Gualtieri 2009). Importantly, much of the information obtained from the internet is textual (rather than quantitative or numerical), and the qualitative meaning (i.e., gist) of this information has a large role to play in risk perception (Reyna 2020).

The internet is not only a convenient tool for consumers. It also offers researchers easy access to the information sources that guide health perceptions. Thus, the internet can be used by researchers to predict, understand, and even influence health perceptions. Our goal in this article is to use information communicated on the internet to model health perceptions for disease states. For this purpose, we obtain textual descriptions of disease states on the National Health Service (NHS) website, which is one of the main online sources of health information in the United Kingdom (Powell et al. 2011). We additionally rely on recent advances in machine learning, known as word and sentence embeddings, which can quantify textual content for use in predictive analysis (Mikolov et al. 2013; Devlin et al. 2018; Bhatia 2019; Bhatia, Richie, and Zou 2019). Such methods have been shown to achieve state of the art results across different areas of natural language processing, including sentiment analysis and question answering. By using embedding methods to quantify the informational content of health descriptions on the NHS we can build a machine learning model that can predict health perceptions given a textual description of a health state.

In study 1, we illustrate this approach in a large-scale study involving participant judgments of 777 distinct disease states. Here we use participant data and text explanations to train our model and evaluate its predictions on held-out data (text explanations and participant evaluations on which the model has not been trained). We also compare the performance of our embedding model with other competing models, including those that rely only on objective health outcomes like mortality, and those that rely only on simpler text features like word concreteness. In study 2, we use our trained embedding models to interpret how information contained in online text explanations influences perceptions of the severity of health states. Finally, in study 3, we use our embedding models to predict how different descriptions of the same disease state can be associated with different health perceptions.

STUDY 1: PREDICTING HEALTH JUDGMENTS

In study 1, we evaluated the accuracy of our approach for predicting lay health perceptions. In particular, we used sentence and word embeddings to quantify text descriptions and predict health ratings for hundreds of health states.

Data Sets and Overview of Variables

Health States and Their Online Text Explanations. We first scraped 777 unique health states and their text explanations and discussions from the NHS website. NHS is the publicly funded healthcare system of the United Kingdom, and its website is one of the most popular online resources for medical advice. Our analyses used only the summary information of the health state, which is presented as the very first paragraph and is thus the first thing that people see who visit the health state webpage.

Embeddings of Health States

Sentence Embeddings. We used a state-of-the-art language model called DistilBERT (Devlin et al. 2018; Sanh et al. 2019) to represent NHS text explanations as high-dimensional vectors. The DistilBERT model is deep neural network that was “pretrained” on very large corpora of text and is thus able to quantify sentences of text based on their meaning. By inputting each sentence in our health state text explanations into the network, and querying the resulting hidden layer representations in the network, we used DistilBERT to obtain a 768-dimensional vector representation for each sentence in each text. We averaged these sentence-level representations to obtain a single 768-dimensional vector for each health state.

Word Embeddings. We also used the Word2Vec model (Mikolov et al. 2013), another neural network trained on large amounts of text, that provides 300-dimensional vector representations for the meanings of individual words (rather than sentences). Previous work has demonstrated that Word2Vec vectors can successfully predict human judgment, and other related cognitive phenomena (Bhatia 2017, 2019; Richie et al. 2019; Bhatia et al. 2021; see Bhatia et al. [2019] for an overview). We obtained a 300-dimensional vector representation for each of our health states by averaging the Word2Vec vectors of their component words.

Deaths and DALYs. We obtained two different measures of objective health outcomes: mortality data and disability-adjusted life years (DALYs). These were taken from the Global Health Data Exchange website (http://ghdx.healthdata.org/gbd-results-tool), and we used statistics for all ages and genders from the year 2017. We used overlapping health states ($N = 77$ with mortality data and $N = 96$ with DALYs) between this data set and the NHS website to evaluate the extent to which easily quantifiable objective health outcomes predict health perceptions.

Search Frequency. We also measured the number of times a health state was searched in Google by querying the Google Trends data set (http://www.google.com/trends/). Similar to Searle et al. (2020), we concentrated on searches from the United Kingdom and only counted the total number of queries that happened within the 90-day period prior to our study. We were able to obtain such statistics for 463 of 777 of our health states, which were present in the Google Trends data set as matches under medically related categories.

Text Features. We measured 26 text features that have been shown to be important for the perception and interpretation of textual data, using the TextAnalyzer tool (Berger, Sherman, and Unger 2020). These features include the Flesch-Kincaid grade level (Kincaid et al. 1975), concreteness, valence, along with many others. TextAnalyzer scores are the weighted sums over the words in the text, with weights given by the relative frequencies of the words in a corresponding lexicon (Humphreys and Wang 2018; Berger et al. 2020).

Health States Ratings Survey

We used the variables described above as predictors in a regression model, with the outcome variable being lay participants’ ratings of the severity of health states. We obtained these human ratings in an online study.

Method. Our study was run on Prolific Academic. Participants ($N = 782$ UK residents; 484 female, 291 male, 7 other; mean age $= 35.12$) were asked to read NHS summaries for 10 randomly selected health states. Then they were asked to imagine that they were diagnosed with each of the health states and to report their evaluations of the health states using a 100-point scale.

At the end of our study, we also included an open-ended question to examine whether our participants actually used NHS.uk, or other similar online resources, to obtain health information.

Health States Familiarity Ratings Survey

We also wanted to test if our model’s performance depended on participants’ overall familiarity with the health state. For this purpose, we ran a separate online study to collect familiarity ratings on the 777 health states. Methods and results of this survey can be found in the appendix, available online.

General Statistical Methodology

First, we built a computational model to map the health state embeddings to people’s judgments. Here we used embeddings as predictors in a regression model and computed the model’s out-of-sample correlation using leave-one-out cross validation. Since the vector representations are high dimensional, we used a Ridge regression instead of a standard linear regression. We fit our Ridge regression using Python’s scikit learn library (Pedregosa et al. 2011) and adopted the default assumptions (including setting $\alpha = 1$ and allowing for a flexible additive intercept for predicting y_i).

We next built additional models to see whether other health-related measures, such as mortality rates, are good predictors of judgments of health. We used these measures as predictor variables in multiple regression models and once again computed each models' out-of-sample correlation using leave-one-out cross validation. In the appendix, we provide additional details on our study methods, including task instructions, exclusion criteria, details on how the Word2Vec and DistilBERT models were trained, and how the cross-validation method works.

Results

Overview of Data. In total, 75% of all study participants reported using NHS.uk (71%) and similar websites (4%) as one of their main sources of health information (if not the only source). The average judgments for each health state ranged from 1.82 to 90.89, with a mean of 47.39 and a standard deviation of 19.97. Figure 1*A* shows the distribution of mean health ratings for all health states. In figure 1*B*, we visualize the 10 lowest (*red arrows*) and 10 highest rated (*green arrows*) health states after applying multidimensional scaling to their sentence embeddings. Looking at the figure, we can see that the health states cluster based on their perceived severity. In addition, health states that are semantically similar to each other are also close to one another. Thus, our underlying vector space seems to accurately represent the similarity between health states as well as some variability in health state ratings.

Performance of Predictor Variables. We investigated the out-of-sample predictive power of the DistilBERT sentence embedding and Word2Vec word embedding models, using leave-one-out cross-validation. These models used a (Ridge) regression and attempted to predict the health rating for a held-out health state. In figures 2*A* and 2*B*, we plot the average participant judgments for each health state against the embedding models' out-of-sample prediction for that health state. As can be seen in these scatterplots, our model predictions had a large out-of-sample correlation with average health ratings ($r = 0.725$, $p < .001$, $R^2 = 0.525$ for DistilBERT and $r = 0.694$, $p < .001$, $R^2 = 0.482$ for Word2Vec). We also built predictive models using alternative machine learning algorithms, including Lasso, SVR, and Random Forest, as well as alternate hyperparameters. We report the accuracy rates of these models in appendix tables S1–S2.

We next examined the out-of-sample predictive power of several alternative variables, with individual regressions for each of the variables. These tests revealed an out-of-sample correlation of $r = -0.105$ ($p = .001$, $R^2 = 0.011$) for number of deaths, $r = 0.095$ ($p = .358$, $R^2 = 0.009$) for DALYs, and $r = 0.295$ ($p < .002$, $R^2 = 0.08$) for search frequency. By replicating prior work (Chapman and Elstein 2000; Slovic et al. 2002; Chapman and Coups 2006; Hertwig et al. 2008;

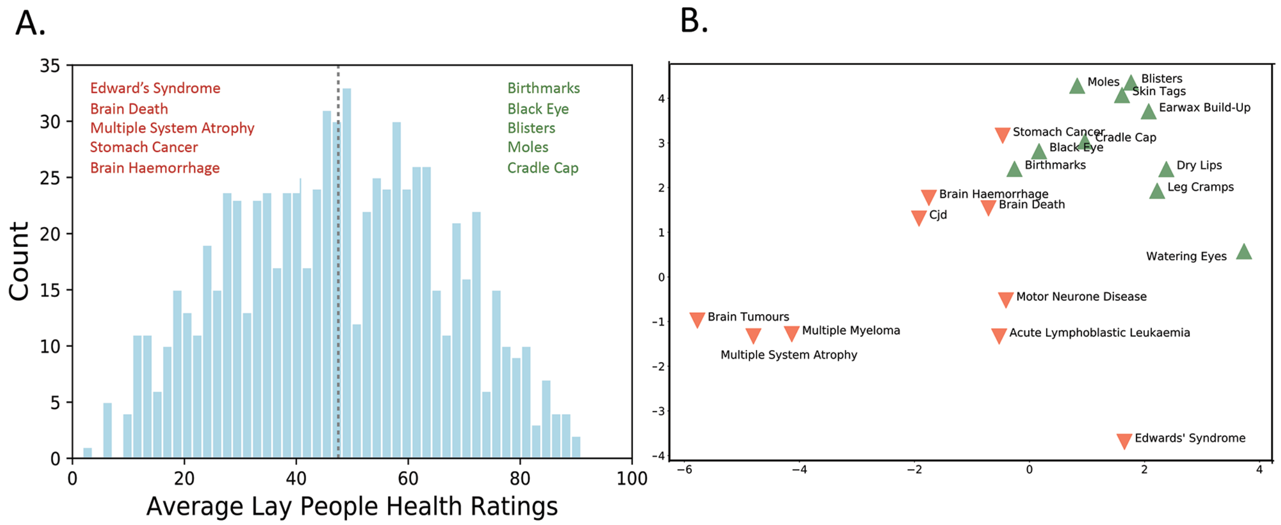

Figure 1. (*A*) Histogram of the average health ratings for all health states, and (*B*) a 2D multidimensional scaling solution on the underlying vector space . In panel *A*, the top five most and least severe health states are indicated in red and green. In addition, the gray dashed line represents the mean health rating across all health states. Unsurprisingly, conditions such as "brain hemorrhage" or "multiple system atrophy" were rated by our participants to lead to the worst health states imaginable, while others such as "black eyes" or "blisters" were predicted to lead to the best health state imaginable in this context. In panel *B*, the underlying vector space appears to accurately represent the similarity between health states and capture some variability in health state ratings. Health states are coded by the arrows based on their average ratings (red arrows show the most severe health states and green arrows show the least severe health states, as judged by participants).

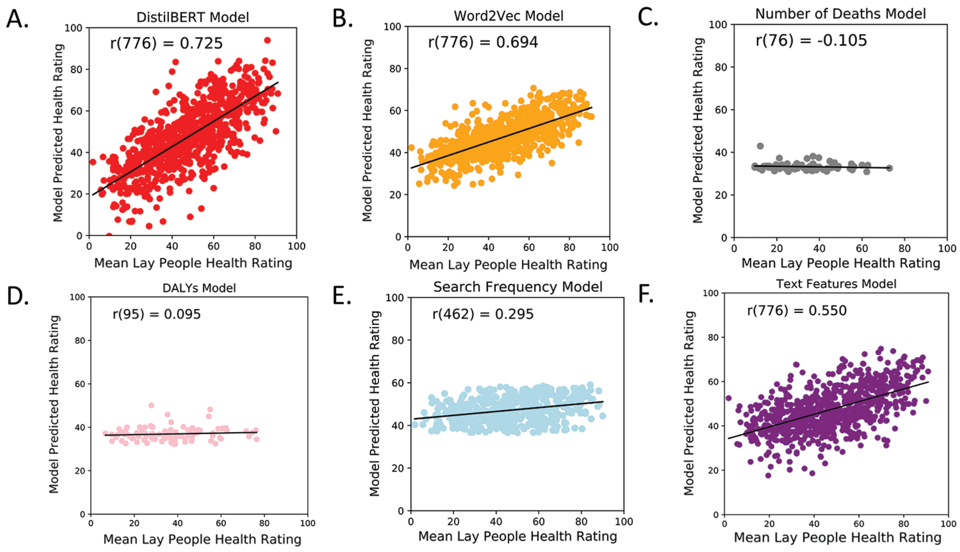

Figure 2. Leave-one-out cross-validation-based correlations for each of our predictive models. Each scatter plot shows the mean participant judgments for each health state against our models' out-of-sample predictions for that health state. Black lines show the best linear fit of the data. Standard parametric correlation between the two variables (*r*) is also presented.

Peters and Meilleur 2016), we found that health outcomes such as DALYs and death rates are not reliable predictors of human judgments of health severity, whereas search frequency is a more accurate predictor. These results are shown in figures 2*C*–2*E*, and additional results are reported in the "Comparison with Alternate Metrics and Models" section of the appendix.

We also attempted an out-of-sample predictive analysis of the 26 text features. This model performed fairly well, obtaining an out-of-sample correlation of $r = 0.550$ ($p < .001$, $R^2 = 0.303$; fig. 2*F*), although not as well our embedding models. We evaluated these correlations further by calculating alternative metrics such as split-half correlation. We found that split-half correlation was nearly identical to that obtained by our DistilBERT model, indicating that this model achieved high predictive accuracy. We also examined the effect of training data-set size on predictive accuracy using three of our best performing models (see app. fig. S1). These results are reported in the appendix.

In addition to these separate models, we also tested a combined model in which we concatenated all our predictor variables: the embeddings, number of deaths, DALYs, search frequency, and text features. Unsurprisingly, this combined model provided accurate out-of-sample predictions ($r = 0.729$, $p < .001$, $R^2 = 0.532$ for DistilBERT and $r = 0.689$, $p < .001$, $R^2 = 0.474$ for Word2Vec). However, these predictions did not differ significantly from the predictions of the embeddings alone in terms of a Steiger test for dependent correlations ($z = -1.055$, $p = .292$ for DistilBERT and $z = 0.484$, $p = .629$ for Word2Vec). Thus, it appears that the embeddings alone can capture most of the variability in health perception.

We also examined health states for which there were large differences (or "errors") between our models' predictions and human ratings. We found that "breast lumps," "children soiling their pants," and "broken or bruised ribs" were the three health states with the most positive (ranked) errors. These were health states that people tended to overestimate the severity of relative to our model. There are several reasons for these errors. For example, "breast lumps" are associated with cancer in people's minds, although most lumps are not cancerous, whereas "soiling" is not dangerous but quite an inconvenience for the child's parents. The histograms of all prediction errors (i.e., predicted health rating, observed health rating) can be found in appendix figure S2.

Ideally we would have fit our models individually on each participant. However, this would require hundreds of health ratings per participant, which, given that each health rating requires the participant to read and evaluate a paragraph of text, is not feasible. So to study the predictive accuracy of our models on an individual level, we correlated individual-level health ratings with the out-of-sample model-predicted average health ratings. This yielded an average correlation coefficient of 0.55 for the DistilBERT and 0.52 for the Word2Vec model (see app. fig. S3 for the histograms of the correlation coefficients). We also ran regression models for each individual where we predicted their individual-level health ratings using the model-predicted average health ratings. The average R^2 from these regression models was 0.394 (with an average mean absolute error of 19.90) for the DistilBERT model and 0.367 (with an average mean absolute error of 21.06) for the Word2Vec model. This indicates that we can predict individual-level ratings with moderate accuracy, by training a model to predict aggregate ratings (see the appendix section "Individual-Level Predictions" for more information).

Key Topics in Health State Overviews. We also used our models to measure the themes that were prevalent in health state overviews on https://www.NHS.uk. More specifically, we applied a K-means clustering algorithm with five clusters to the embeddings of the health states. Next, we took the overviews of the health states and computed the 20 relatively most frequent words in each predicted cluster, after removing the stop words and excluding words that appeared less than 15 times across the whole NHS.uk text data set. Looking at the relatively most frequent words in each predicted cluster, we can get a better sense of the types of words and categories that are present in the health overviews. Furthermore, we can also look at the average health ratings of the health states that belong in these predicted clusters to get an idea about how these cues relate to health judgment. This approach is similar to topic modeling but is better suited for tests such as ours, where text data are limited in quantity and pretrained embeddings that possess semantic knowledge about the world are readily available (Sia, Dalmia, and Mielke 2020).

In appendix figures S4–S5, we present word clouds showing prevalent themes in our data set. For example, we see we see words related to "gender," "personality," and "mental health" in one cluster and words signaling hormonal and reproductive conditions in a second cluster. In additional clusters, we have words relating to "acute" conditions like "leukemia," "easily" "spreading" conditions, and "anything" "harmless" that "might" "go away."

STUDY 2: UNDERSTANDING HEALTH JUDGMENTS

In the first part of our article, we predicted health state judgments using word and sentence embeddings of online health explanations. Our models performed much better than objective health outcomes (e.g., number of deaths, DALYs) and simple text features (e.g., concreteness, valence). At the end of study 1, we began investigating the key topics present in health state overviews to better understand the types of words that exist in these health overviews. Now we wish to interpret these words and categories based on established psychological variables. We do so using our word embedding model, which provides vector representations of individual words that preserve their meanings.

Keywords and Psychological Constructs

Linguistic Inquiry and Word Count. Linguistic inquiry and word count (LIWC) is an established language dictionary that is typically used to investigate the links between language and various psychological variables. It has several different higher and lower-level psychological categories and constructs, each consisting of hundreds of keywords. Traditionally, the keywords in each construct are used in automated text analysis to quantify the degree to which these constructs and concepts are reflected in a given text (Pennebaker et al. 2001; Pennebaker, Mehl, and Niederhoffer 2003; Humphreys and Wang 2018; Berger et al. 2020).

Health-Related Keywords Survey. We examined additional constructs and their keywords to confirm and build on top of the previous literature. For this purpose, we asked Prolific participants ($N = 30$ UK residents; 19 female, 11 male; mean age = 30.53) to generate keywords that reflected eight target constructs previously found to be at play in health judgment: death, functional inability, disability negative mood, anxiety, pain, physical distress, and depression (Idler 1993; Kaplan and Camacho 1983; Fýlkesnes and Førde 1992; Farmer and Ferraro 1997; Garrity, Somes, and Marx 1978). In the study, participants read the following prompt: "What are words that you would use to describe (or associate with) diseases with [CONSTRUCT]? Please write as many words as you can." Participants were paid at a rate equivalent to $9 an hour and did not receive any additional bonuses. Similarly, to the LIWC dictionary, this study yielded a set of keywords associated with various constructs, although the constructs used corresponded to health-relevant categories rather than general psychological categories.

Body Parts. Finally, we included a complete list of body parts since previous literature has emphasized the importance of physical functioning in health perception. These results can be seen in the appendix.

Computational Methods

We used the above lists to measure the degree to which different words and constructs map onto perceptions of disease severity. As in study 1, we trained our Word2Vec embeddings model on the full data set of health states and ratings using Ridge regression. Then, for a given target word (e.g., “ache”), we obtained the word’s 300-dimensional Word2Vec embedding. Finally, we passed this high-dimensional vector through our trained model to obtain a health state prediction for the word. Intuitively, this can be seen as a predicted health judgment for a text explanation consisting only of the target word. Words that are given high predictions are associated with high health ratings and less severe health states, whereas words that are given low predictions are associated with low ratings and very severe health states.

Results

Higher-Level LIWC Constructs. Figure 3*A* presents the health judgment predictions for the 14 higher-level LIWC constructs obtained through our Word2Vec embeddings model. In this figure, we can see that death-related (e.g., “coffin”) and work-related (e.g., “job”) constructs lead to the worst health states imaginable. While death-related constructs are likely to provoke negative affect, there is also an empirically documented positive relationship between hours of work and ill health (Sparks et al. 1997). By contrast, better health states are associated with informal language, and social and affective processes. Unsurprisingly, the use of informal language may signal that a health state is less important and thus lead to higher judgments of health. In addition, prior work has highlighted the role of both social and affective processes in risk perception and health behavior. For example, social cognitive theory (Luszczynska and Schwarzer 2015) emphasizes the importance of social and physical environment on self-efficacy and health behavior, and self-determination theory (Deci and Ryan 2012) concentrates on how social and cultural factors facilitate a sense of volition and initiative (Kasperson et al. 1988; Pechmann 2001). Likewise, the affect heuristic is strongly implicated in the perception of risk (Loewenstein et al. 2001; Slovic et al. 2007; Pachur, Hertwig, and Steinmann 2012; Van Cappellen et al. 2018). Here our interpretation of the LIWC analysis is rather speculative, as some LIWC categories are not informative or meaningful in the context of health judgment. However, it appears that patterns observed in our analysis confirm and strengthen important components of previous psychological theories of health judgment.

Health-Related Keywords Survey. Mean predictions for each construct using our computational model, along with example participant generated keywords are shown in figure 3*B*. By confirming research findings from previous literature, we observed that death-related words received the lowest health judgments (mean prediction = 27.147), followed by functional inability (mean prediction = 28.845) and disability-related words (mean prediction = 33.838). However, pain (mean prediction = 43.970), physical distress

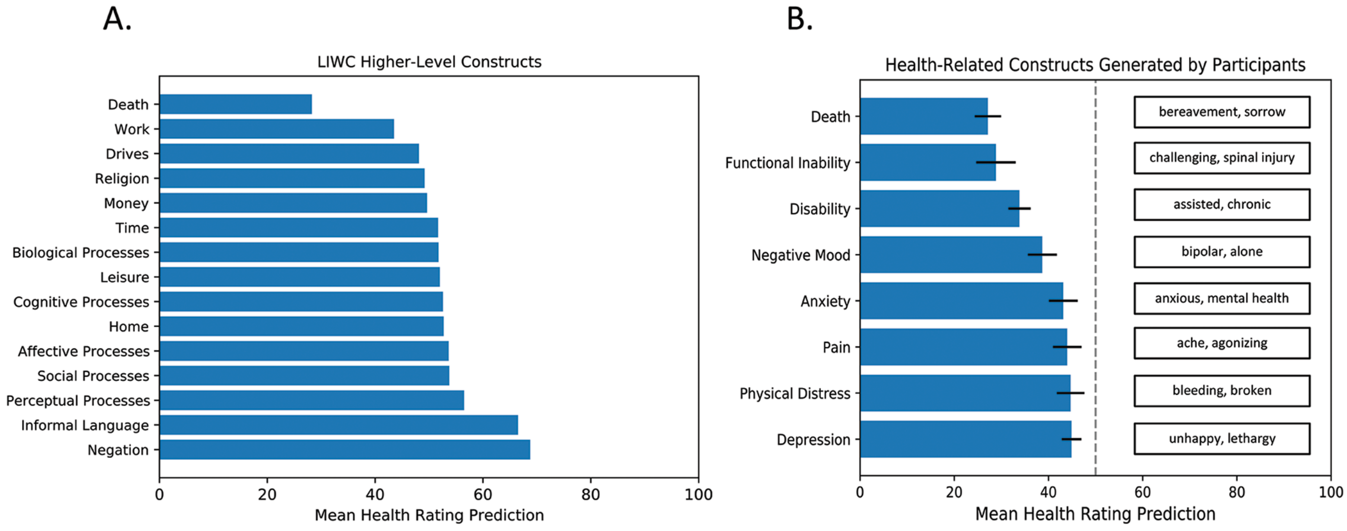

Figure 3. Higher-level linguistic inquiry and word count (LIWC) psychological constructs (*A*), human generated health-related constructs along with example keywords from each (*B*), and their health judgment predictions obtained through our Word2Vec embeddings model. Sorted from predicted to lead to “the worst health state imaginable” (or the most severe) to “the best health state imaginable” (or the least severe). We report the results from all LIWC constructs in appendix figure S6.

(mean prediction = 44.687), and depression-related words (mean prediction = 44.884) were perceived to be relatively less severe. While this may appear surprising at first glance, functional inability and disabilities are long-term poor conditions of health and may be judged to be very severe and disruptive to daily functioning. In contrast, pain may often be perceived as a temporary, ephemeral state that can be reduced with appropriate medical attention. Indeed, Goldstein, Siegel, and Boyer (1984) have found that acute conditions do not seem to have the expected negative impact on health judgment. Our model's predictions are consistent with this finding.

In appendix figures S7–S8, we present one more analysis using the 500 most common health explanation words and body parts to provide additional links with psychological constructs and health judgments.

STUDY 3: PREDICTING DIFFERENCES BETWEEN HEALTH EXPLANATIONS

The above studies have shown the utility of our approach for predicting and interpreting health state judgments. Next, we examine an application of this approach to health communication. Consider, for example, "breast cancer." There are overviews of "breast cancer" in different online health forums, including both the NHS and the American Breast Cancer Foundation (abcf.org) websites. However, these two websites describe "breast cancer" differently, and it is possible that the different explanations lead to different judgments about the severity of the disease. Since our approach in study 1 appears to be successful at describing health judgment across different health states, it may also be able to predict, in an a priori manner, the differences in judgments evoked by two explanations for the same health state. Such an a priori prediction is much harder to obtain but, if successful, can have many practical applications in health risk communication and perception.

Method

We identified 10 health states varying in severity that have explanations on both NHS.uk, as well as on other health-related websites. We extracted these explanations for each of the 10 health states, resulting in 20 different health explanations. Participants (N = 576 UK residents; 382 female, 186 male, 8 other; mean age = 37.10) were randomly assigned to one of the 20 health explanations and were asked rate the severity of the associated disease using the same rating scale as in study 1. While the survey sample was very similar in studies 1 and 3, study 1 was run earlier in the COVID-19 pandemic in August 2020, while study 3 was run in May 2021, later in the pandemic. Participants who completed study 1 were automatically excluded from study 3.

We matched the number of sentences in the NHS and non-NHS health overviews as closely as possible. In addition, we also investigated whether there textual or semantic differences between the overviews from the NHS and the other websites. A complete list of these overviews and results of the comparisons (using norms from existing studies and by asking a separate group of human participants) can be found in appendix tables S3–S5. Finally, we also asked participants in study 3 to rank the following items based on their frequency of usage: NHS.uk website, other health-related websites such as WebMD, asking a physician or medical professional, health-related books of published text, or other (please specify).

Results

A large majority of participants reported using the NHS.uk website (63%) and other health-related websites such as WebMD (31%) as their primary source when gathering health-related information. These text-based online resources were followed by "asking a physician or medical professional" (5%) response.

Next, for each health state overview in this study, we obtained model predictions from our Word2Vec embeddings model that was trained on full data from study 1. For the health states in this study, the Word2Vec model had higher predictive accuracy than DistilBERT, possibly because these health overviews had relatively few sentences each (we report model predictions from DistilBERT in appendix figures S9 and S10). First, we correlated human ratings for each health overview with the respective model predictions. This showed that our approach was able to achieve very high predictive accuracy ($r = 0.74$, $p < .001$, $R^2 = 0.541$). In figure 4*A*, we plot average participant judgments for each health state overview against the embedding models' prediction for the overview.

We more systematically examined the relationship between human participants' health ratings and model predicted ratings (presented in panel A) using a regression in which we controlled for disease specific effects. In this regression, our dependent variable was the individual-level health rating given by the participants. The model's predicted ratings for that health state served as the predictor variable. Our regressions permitted random effects for the ten diseases that were used in the study. In this regression, the coefficient for model predictions was positive and significant

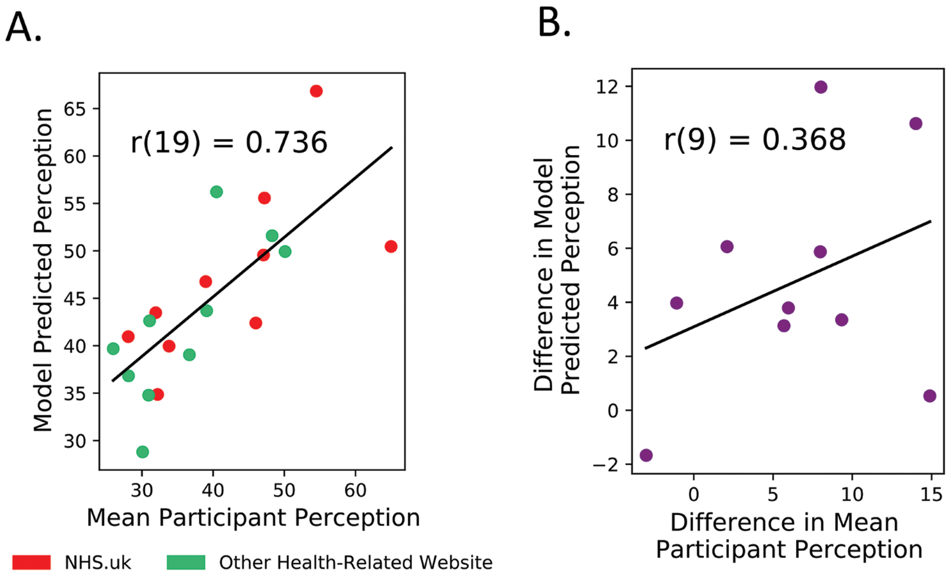

Figure 4. (*A*) Participant judgments plotted against the word embedding models' prediction for each health state overview. In this panel, red scatter points indicate the overviews from the National Health Service (NHS) website, and green scatter points indicate the overviews from the other health-related websites. Predicted and observed ratings differences between NHS.uk and other health-related website overviews for each health state are presented in panel *B*. Black lines show the best linear fit of the data. Standard parametric correlation between the two variables (*r*) is also presented.

($\beta = 0.726,\ p < .001,\ 95\%\ \mathrm{CI} = [0.466, 0.986]$), showing that overall, we were able to predict human participant ratings using the model predicted ratings, even after controlling for disease specific effects.

We also examined the model's ability to predict differences in ratings between NHS.uk overviews and other health-related website overviews for the ten health states. This is a more challenging task, as overviews for the same health state are likely to use similar language. In figure 4*B*, we plot the differences between NHS.uk and the other health-related website overviews as observed in human ratings and as predicted by the model. We do observe a positive correlation between these differences ($r = 0.37$, $p = .295$, $R^2 = 0.136$). Although the correlation is moderate in size, it fails to achieve significance, largely due to the small sample size (there are only ten health states). In figure 5, we illustrate the predictions and the corresponding human ratings. While different health state overviews led to different health ratings, our embeddings model was able to successfully predict the direction of these differences in nine of the ten health states. Note that a random model (that gives NHS or alternative health-related websites a higher rating with a probability of 50%) would get 9 out of 10 correct answers with only a 9/1,024 (or <1%) chance. Note that the one health state that was not correctly predicted was "hip pain." Here the alternative website overview was only one short sentence due to an experimenter error.

DISCUSSION

In this article, we have presented a novel machine learning approach that aims to predict lay people's health judgments. Using recent advances in machine learning, specifically sentence and word embedding models, we quantified the information contained in online text overviews of hundreds of different health states found on the NHS website. In study 1, we mapped these sentence and word embeddings onto lay health perceptions collected through an online study. Our machine learning approach was able to accurately predict how participants perceived different health states. Using the DistilBERT and Word2Vec embeddings, our models achieved high out-of-sample correlations between the predicted and observed health ratings. These accuracy rates were higher than those obtained from other predictors, such as number of deaths and DALYs, search frequency, and various simple text features, and they were also close to the split-half correlation. We also showed (in the appendix) how our approach continues to have high predictive accuracy for both familiar and unfamiliar health states. In study 2, we used our best-fit models to analyze the language present in these text overviews and understand the determinants of

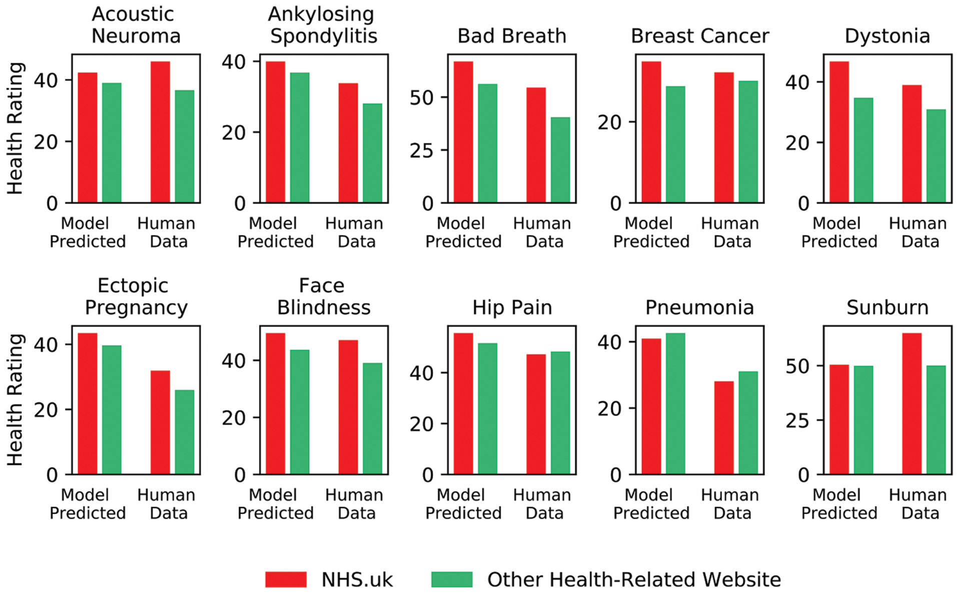

Figure 5. Model predictions and average participant ratings for each of the explanations. Explanations from NHS are indicated in red, whereas explanations from other health-related websites are indicated in green. For all 9 of the 10 health states, the direction of the difference observed in model predictions for the two websites (difference between red and green bars on the left side of the subplots) was also observed in average human participant ratings (difference between red and green bars on the right side of the subplots).

health state judgments. Our model replicated previous empirical and theoretical findings about the importance of constructs such as disability, death, and affective processes for health perception and extended them by discussing their respective contributions. Finally, in study 3, we demonstrated how our embeddings approach can be used to predict differences in health judgments associated with different explanations of a given health state.

Since self-assessed health is a complex, multidimensional concept (Simon et al. 2005), quantifying and understanding health judgment has been a challenging task for many years. In this article, we showcased a powerful and scalable new technique that can easily be adopted by researchers and policy makers to predict health perceptions. In addition to predicting health judgment, our method also offers many new applications and avenues for future research. For example, our approach can be used to study optimal health communication formats as well as predictions of judgment errors. From a methodological perspective, adopting cutting-edge technologies from data science and creating opportunities to further integrate these techniques with consumer health behavior research is essential for scalable health prediction. Thus, we believe that our approach is a beneficial first step for projects with important consumer behavior implications and for improving the well-being of people.

It is important to mention that our study also has some limitations. First, even though the media and internet have a substantial influence on how people acquire information and make decisions, these are not the only sources people can rely on. In fact, while we focused our analysis on NHS.uk health overviews, most people are exposed to a wide variety of information sources, including discussions with medical providers. While our participants typically reported using NHS.uk as a major resource, future work can try to obtain linguistic communication data sets involving other sources, and apply our embeddings approach to model individual differences in health behavior as well. Studying individual differences would also pave the way for targeted health communication and behavioral phenotyping. Second, our current approach does not analyze social network data, which is an important source for online health information (Gottlieb and Green 1984; Cohen-Cole and Fletcher 2008). Our methods can be used in conjunction with social network analysis to better predict lay health perception. Third, we did not consider any demographic or socioeconomic variables while explaining the variance in health judgment, but our methods can easily be applied to study differences in participant demographic and socioeconomic variables if the training data are collected from a target population. Future work can also concentrate on real medical forums where consumers ask medical questions and advise, and similar methodologies and models can easily be applied.

To conclude, we have presented a novel, generalizable, and high-performing method to predict and understand health judgment. This method maps embeddings from online text explanations and discussions of different health states to laypeople's perceptions of health. Our method combines insights from data science and health behavior research and opens up many potential research questions with substantial real-world applications. We hope that researchers, policy makers, and medical professionals will use this approach to investigate many other exciting research questions related to judgments of health.

REFERENCES

Arnesen, Trude, and Lydia Kapiriri (2004), "Can the Value Choices in DALYs Influence Global Priority-Setting?" *Health Policy*, 70 (2), 137–49.

Atkinson, Nancy, Sandra Saperstein, and John Pleis (2009), "Using the Internet for Health-Related Activities: Findings from a National Probability Sample," *Journal of Medical Internet Research*, 11 (1), e5.

Becker, Marshall H., Lois A. Maiman, John P. Kirscht, Don P. Haefner, and Robert H. Drachman (1977), "The Health Belief Model and Prediction of Dietary Compliance: A Field Experiment," *Journal of Health and Social Behavior*, 18 (4), 348–66.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer, and David A. Schweidel (2020), "Uniting the Tribes: Using Text for Marketing Insight," *Journal of Marketing*, 84 (1), 1–25.

Berger, Jonah, Garrick Sherman, and Lyle Ungar (2020), "TextAnalyzer," https://textanalyzer.org.

Betsch, Cornelia, Robert Böhm, and Gretchen B. Chapman (2015), "Using Behavioral Insights to Increase Vaccination Policy Effectiveness," *Policy Insights from the Behavioral and Brain Sciences*, 2 (1), 61–73.

Bhatia, Sudeep (2017), "Associative Judgment and Vector Space Semantics," *Psychological Review*, 124 (1), 1–20.

——— (2019), "Predicting Risk Perception: New Insights from Data Science," *Management Science*, 65 (8), 3800–3823.

Bhatia, Sudeep, Christopher Y. Olivola, Nazlı Bhatia, and Amnah Ameen (2021), "Predicting Leadership Perception with Large-Scale Natural Language Data," *Leadership Quarterly*, 101535, https://doi.org/10.1016/j.leaqua.2021.101535.

Bhatia, Sudeep, Russell Richie, and Wanling Zou (2019), "Distributed Semantic Representations for Modelling Human Judgment," *Current Opinion in Behavioral Sciences*, 29 (October), 31–36.

Brewer, Noel T., Gretchen B. Chapman, Frederick X. Gibbons, Meg Gerrard, Kevin D. McCaul, and Neil D. Weinstein (2007), "Meta-Analysis of the Relationship between Risk Perception and Health Behavior: The Example of Vaccination," *Health Psychology*, 26 (2), 136–45.

Calvert, Melanie J., and N. Freemantle (2003), "Use of Health-Related Quality of Life in Prescribing Research. Part 1: Why Evaluate Health-related Quality of Life?" *Journal of Clinical Pharmacy and Therapeutics*, 28 (6), 513–21.

Chapman, Gretchen B., and Elliot J. Coups (2006), "Emotions and Preventive Health Behavior: Worry, Regret, and Influenza Vaccination," *Health Psychology*, 25 (1), 82–90.

Chapman, Gretchen B., and Arthur S. Elstein (2000), "Cognitive Processes and Biases in Medical Decision Making," in *Decision Making in Health Care: Theory, Psychology, and Applications*, ed. G. B. Chapman and F. A. Sonnenburg, Cambridge: Cambridge University Press, 183–210.

Cheung, Helier (2020), "Coronavirus: Why Attitudes to Masks Have Changed around the World," *BBC News*, https://www.bbc.com/news/world-53394525.

Cohen-Cole, Ethan, and Jason M. Fletcher (2008), "Is Obesity Contagious? Social Networks vs. Environmental Factors in the Obesity Epidemic," *Journal of Health Economics*, 27 (5), 1382–87.

Cornwall, Andrea, and Rachel Jewkes (1995), "What Is Participatory Research?" *Social Science and Medicine*, 41 (12), 1667–76.

Covello, Vincent T., and Richard G. Peters (2002), "Women's Perceptions of the Risks of Age-Related Diseases, Including Breast Cancer: Reports from a 3-Year Research Study," *Health Communication*, 14 (3), 377–95.

Deci, E. L., and R. M. Ryan (2012), "Self-Determination Theory," in *Handbook of Theories of Social Psychology*, ed. P. A. M. Van Lange, A. W. Kruglanski, and E. T. Higgins, Los Angeles: Sage, 416–36.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018), "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, ed. J. Burstein, C. Doran, and T. Solorio, New York: ACL Press, arXiv preprint: 1810.04805.

Entwistle, Vikki A., Mary J. Renfrew, Steven Yearley, John Forrester, and Tara Lamont (1998), "Lay Perspectives: Advantages for Health Research," *BMJ*, 316 (7129), 463–66.

Farmer, Melissa M., and Kenneth F. Ferraro (1997), "Distress and Perceived Health: Mechanisms of Health Decline," *Journal of Health and Social Behavior*, 38 (3), 298–311.

Fýlkesnes, Knut, and Olav Helge Førde (1992), "Determinants and Dimensions Involved in Self-Evaluation of Health," *Social Science and Medicine*, 35 (3), 271–79.

Garrity, Thomas F., Grant W. Somes, and Martin B. Marx (1978), "Factors Influencing Self-Assessment of Health," *Social Science and Medicine*, 12 (2), 77–81.

Goldstein, Michael S., Judith M. Siegel, and Richard Boyer (1984), "Predicting Changes in Perceived Health Status," *American Journal of Public Health*, 74 (6), 611–14.

Gottlieb, Nell H., and Lawrence W. Green (1984), "Life Events, Social Network, Life-Style, and Health: An Analysis of the 1979 National Survey of Personal Health Practices and Consequences," *Health Education Quarterly*, 11 (1), 91–105.

Gualtieri, Lisa Neal (2009), "The Doctor as the Second Opinion and the Internet as the First," in *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009*, ed. D. R. Olson, R. Arthur, and SIGCHI Group, New York: ACM Press, 2489–98.

Harrison, Joel A., Patricia D. Mullen, and Lawrence W. Green (1992), "A Meta-Analysis of Studies of the Health Belief Model with Adults," *Health Education Research*, 7 (1), 107–16.

Hertwig, Ralph, Stefan M. Herzog, Lael J. Schooler, and Torsten Reimer (2008), "Fluency Heuristic: A Model of How the Mind Exploits a By-Product of Information Retrieval," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34 (5), 1191–206.

Humphreys, Ashlee, and Rebecca Jen-Hui Wang (2018), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–306.

Idler, Ellen L. (1993), "Perceptions of Pain and Perceptions of Health," *Motivation and Emotion*, 17 (3), 205–24.

Idler, Ellen L., Shawna V. Hudson, and Howard Leventhal (1999), "The Meanings of Self-Ratings of Health: A Qualitative and Quantitative Approach," *Research on Aging*, 21 (3), 458–76.

Janz, Nancy K., and Marshall H. Becker (1984), "The Health Belief Model: A Decade Later," *Health Education Quarterly*, 11 (1), 1–47.

Kaplan, George A., and Terry Camacho (1983), "Perceived Health and Mortality: A Nine-Year Follow-up of the Human Population Laboratory Cohort," *American Journal of Epidemiology*, 117 (3), 292–304.

Kasperson, Roger E., Ortwin Renn, Paul Slovic, Halina S. Brown, Jacque Emel, Robert Goble, Jeanne X. Kasperson, and Samuel Ratick (1988), "The Social Amplification of Risk: A Conceptual Framework," *Risk Analysis*, 8 (2), 177–87.

Keller, Punam Anand, and Donald R. Lehmann (2008), "Designing Effective Health Communications: A Meta-Analysis," *Journal of Public Policy and Marketing*, 27 (2), 117–30.

Kincaid, J. P., R. P. Fishburne, R. L. Rogers, and B. S. Chissom (1975), "Flesch-Kincaid Grade Level," Memphis, TN, United States Navy.

Krause, Neal M., and Gina M. Jay (1994), "What Do Global Self-Rated Health Items Measure?" *Medical Care*, 32 (9), 930–42.

Li, Meng, and Gretchen B. Chapman (2013), "Nudge to Health: Harnessing Decision Research to Promote Health Behavior," *Social and Personality Psychology Compass*, 7 (3), 187–98.

Lloyd, A. J. (2001), "The Extent of Patients' Understanding of the Risk of Treatments," *BMJ Quality and Safety*, 10 (Suppl. 1), i14–i18.

Loewenstein, G., H. Hsee, E. U. Weber, and N. Welch (2001), "Risk as Feeling," *Psychological Bulletin*, 127 (2), 267–86.

Luszczynska, Aleksandra, and Ralf Schwarzer (2015), "Social Cognitive Theory," in *Predicting Health Behaviour*, 2nd ed., ed. M. Conner and P. Norman, Buckingham: Open University Press, 127-69.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean (2013), "Distributed Representations of Words and Phrases and Their Compositionality," *Advances in Neural Information Processing Systems*, 26 (October), 3111–19.

Pachur, T., Hertwig, R., and Steinmann, F. (2012), "How Do People Judge Risks: Availability Heuristic, Affect Heuristic, or Both?" *Journal of Experimental Psychology: Applied*, 18 (3), 314–30.

Pechmann, Cornelia (2001),"A Comparison of Health Communication Models: Risk Learning versus Stereotype Priming," *Media Psychology*, 3 (2), 189–210.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau (2011), "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12 (2011), 2825–30.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth (2001), *Linguistic Inquiry and Word Count: LIWC 2001*, Austin, TX: Erlbaum.

Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer (2003), "Psychological Aspects of Natural Language Use: Our Words, Our Selves," *Annual Review of Psychology*, 54 (1), 547–77.

Peters, Ellen, and Louise Meilleur (2016), "The Influence of Affect on Health Decisions," in *Handbook of Health Decision Science*, ed. M. A. Diefenbach, S. Miller-Halegoua, and D. J. Bowen, London: Springer, 109–20.

Powell, John, Nadia Inglis, Jennifer Ronnie, and Shirley Large (2011), "The Characteristics and Motivations of Online Health Information Seekers: Cross-Sectional Survey and Qualitative Interview Study," *Journal of Medical Internet Research*, 13 (1), e20.

Reyna, Valerie F. (2020), "A Scientific Theory of Gist Communication and Misinformation Resistance, with Implications for Health, Education, and Policy," *Proceedings of the National Academy of Sciences*, 118 (15), 20192441, https://doi.org/10.1073/pnas.1912441117.

Richie, Russell, Wanling Zou, Sudeep Bhatia, and Simine Vazire (2019), "Predicting High-Level Human Judgment across Diverse Behavioral Domains," *Collabra: Psychology*, 5 (1), 50, https://doi.org/10.1525/collabra.282.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019), "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv preprint:1910.01108.

Searle, Tamara., Al-Niaimi, Firas, and Ali, Faisal R. (2020), "Dermatological Insights from Google Trends: What Does the Public Think Is Important during COVID-19 Lockdown?" *Clinical and Experimental Dermatology*, 45 (7), 898–900.

Sheeran, Paschal, Peter R. Harris, and Tracy Epton (2014), "Does Heightening Risk Appraisals Change People's Intentions and Behavior? A Meta-Analysis of Experimental Studies," *Psychological Bulletin*, 140 (2), 511–43.

Sia, Suzanna, Ayush Dalmia, and Sabrina J. Mielke (2020),"Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics Too!" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, New York: ACL Press.

Simon, Jeanette G., J. B. De Boer, I. M. A. Joung, Hans Bosma, and J. P. Mackenbach (2005), "How Is Your Health in General? A Qualitative Study on Self-Assessed Health," *European Journal of Public Health*, 15 (2), 200–208.

Slovic, Paul, Melissa Finucane, Ellen Peters, and Donald G. MacGregor (2002), "Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics," *Journal of Socioeconomics*, 31 (4), 329–42.

——— (2007), "The Affect Heuristic," *European Journal of Operational Research*, 177 (3), 1333–52.

Slovic, Paul, and Ellen Peters (2006), "Risk Perception and Affect," *Current Directions in Psychological Science*, 15 (6), 322–25.

Sparks, Kate, Cary Cooper, Yitzhak Fried, and Arie Shirom (1997), "The Effects of Hours of Work on Health: A Meta-Analytic Review," *Journal of Occupational and Organizational Psychology*, 70 (4), 391–408.

Vahabi, Mandana (2007), "The Impact of Health Communication on Health-Related Decision Making," *Health Education*, 107 (1), 27–41.

Van Bavel, Jay J., Katherine Baicker, Paulo S. Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J. Crockett, Alia J. Crum, Karen M. Douglas, and James N. Druckman (2020), "Using Social and Behavioural Science to Support COVID-19 Pandemic Response," *Nature Human Behaviour*, 4 (5), 460–71.

Van Cappellen, Patty, Elise L. Rice, Lahnna I. Catalino, and Barbara L. Fredrickson (2018), "Positive Affective Processes Underlie Positive Health Behaviour Change," *Psychology and Health*, 33 (1), 77–97.

Van Der Pligt, Joop (1998), "Perceived Risk and Vulnerability as Predictors of Precautionary Behaviour," *British Journal of Health Psychology*, 3 (1), 1–14.

Weinstein, Neil D. (2000), "Perceived Probability, Perceived Severity, and Health-Protective Behavior," *Health Psychology*, 19 (1), 65–74.