Yael Beshaw
GVPT728
Final Project Memo

Investigating Predictors of Health Information Seeking Behavior

Introduction

Health information refers to the data regarding your personal health including information about symptoms and/or outcomes that may be relevant to you. This data can come in the form of Electronic Health Records (EHR), results for lab tests and more. While physicians and clinicians are being trained in how to utilize these technologies, there is an increasing need for patients to be able to access this information in order to make personal health decisions. The COVID-19 pandemic highlighted this urgency as many were forced to utilize telehealth in lieu of seeing their physicians in person and relied on information from the news and social media to assess their risk of getting COVID-19. Now more than ever, it is imperative that healthcare, public health, and public policy professionals are able to create and enact interventions that address widespread and personal health concerns and outcomes. In order to do so, we require an understanding of how people interact with health information- both personal and general.

Research Question and Theory

In order to assess how people interact with health information, we consider how people utilize the internet and technology. According to a Pew Research Center survey, 95% of Americans utilize the internet, with about 80% utilizing Youtube (Gottfried, 2023). The report goes on to list various social media platforms that U.S. adults use, including Facebook, Instagram, Pinterest, and Tiktok among others. As this number increases across platforms, Americans find themselves connected to the internet in a variety of ways through social media; watching long form versus short form media, reading or writing blogs and captions, or editing pictures and videos. As a result, we have now coined the term “screen-time” to encapsulate all the things that could be done on the internet, with most complaints about high screen time relating to social media use. Studies acknowledge the impact that this has on younger generations but seem to ignore that a majority of Americans utilize the same media platforms.

We argue that social media has become a mechanism through which users learn and practice digital literacy. Studies have linked technical competencies to successful use of different social media platforms (Polanco-Levicán & Salvo-Garrido, 2022). Thus, frequency of social media utilization can indicate a comfortability with technology, especially if this utilization is cross-sectional. Therefore, my research question becomes: **does the frequency of internet utilization by way of social media impact one’s likelihood of seeking health information online?**

Yael Beshaw
GVPT728
Final Project Memo

Hypothesis

Based on the information provided in the above section, we hypothesize that an increase in internet utilization (by way of social media usage frequency) will result in an increased likelihood of seeking health information online. We assert that as social media utilization increases, one's comfortability with technical competencies does as well, resulting in the internet being the most intuitive way to look for information.

Data Collection and Variable Selection

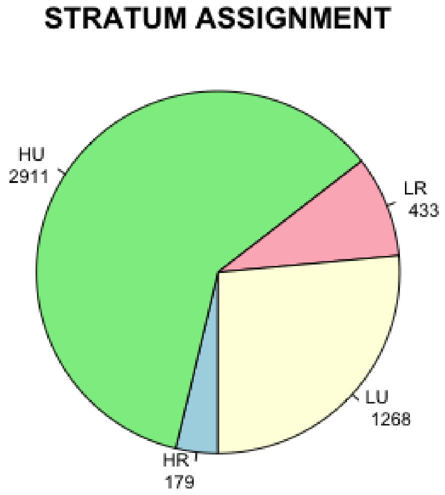
In order to test this hypothesis, we employ the Health Information National Trends Survey (HINTS) by the National Cancer Institute. This is a repeated cross-sectional survey that has collected data from 2003 to 2024, with response rates between 20-40%. We utilized the HINTS6 dataset which collected data between March through November of 2022, in a two-stage design stratification method. The first stage required a stratified sample of residential addresses; stratified by rural versus urban and low minority versus high minority. Then, one adult (U.S. citizen, 18+, non-incarcerated) was randomly selected from each sampled household.

In order to conduct statistical analysis, we considered a variety of IV's and controls that could impact whether or not an individual would utilize the internet to look for health information. The dependent variable utilized in this study is a binary variable asking respondents "In the past 12 months, have you used the internet to look for health or medical information?". The independent variable is an ordinal variable asking respondents "In the last 12 months, how often did you visit a social media site?". For controls, we consider three distinct groups; health information related, demographics, and economic factors.

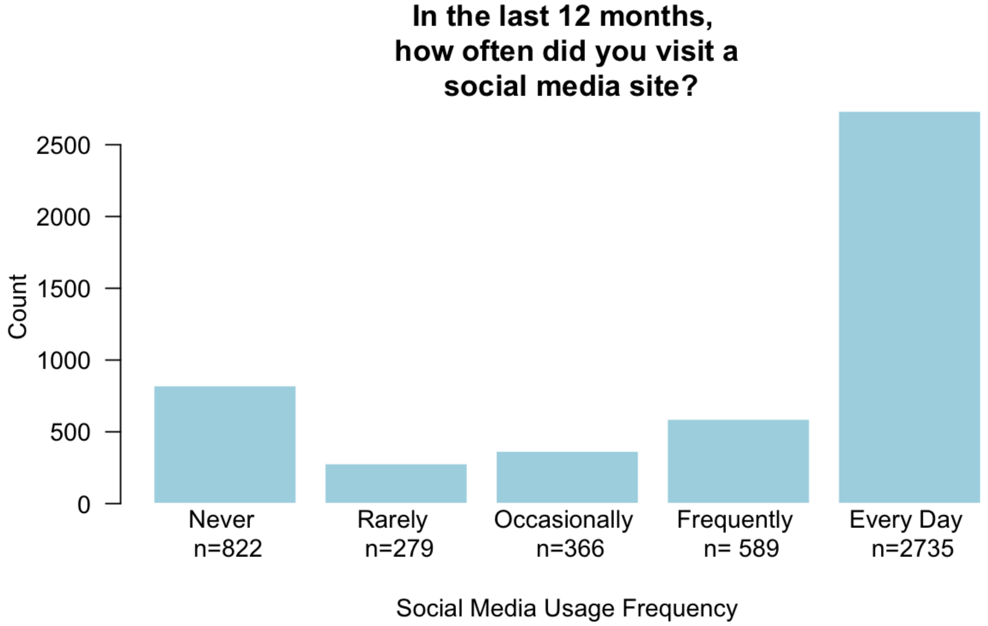
Descriptive Statistics

This dataset includes responses from 6,252. After adjusting the dataset to include our variables of interest and omit any observations with missing data, we have 4,971 observations. The sample is predominantly white (58.4%), male (59.5%), with a median age of 56 years old and with almost half holding a college degree or above (49.9%). Interestingly, the majority (68%) of our sample either has cancer and/or another chronic condition (Diabetes, High BP, Heart Condition, Lung Disease, and/or Depression). Thus it is imperative that this was utilized as a control given that having a low health status may likely require one to utilize the internet to seek health information compared to their healthier counterparts. This sample has a median household income at or above $75,000 USD and a majority work full time, if not part time, with no children under the age of 18 (75.2%). A majority do not live in rural areas, however, as the sample was stratified by urbanization and high/low minority, we assessed the stratum classifications in which a majority of the sample was in high minority urban areas (62%).

The pie chart below depicts this breakdown with HU= High Minority and Urban, HR = High Minority and Rural, LU = Low Minority and Urban, and LR = Low Minority and Rural.



In terms of health information-seeking behaviors, 76.2% of the sample has utilized the internet to look for health or medical information in the past 12 months, with 87.9% using the internet in general. We see that 57.1% of respondents utilize social media almost everyday.



Despite this, only 12% are completely confident in their ability to find helpful health resources on the internet and close to 70% believe that the health information they see on social media is false or misleading. This alone provides evidence that addressing health literacy online and through social media could aid in the public's general sentiment towards health information and utilization. Furthermore, as a majority of the sample has more than one type of access to devices and internet connection (60% and 67%, respectively), providing avenues for accessing health information can also come in the form of diversifying how health information mediums (health portals, journals, infographics) could aid in the dissemination of this information.

Yael Beshaw
GVPT728
Final Project Memo

Methods

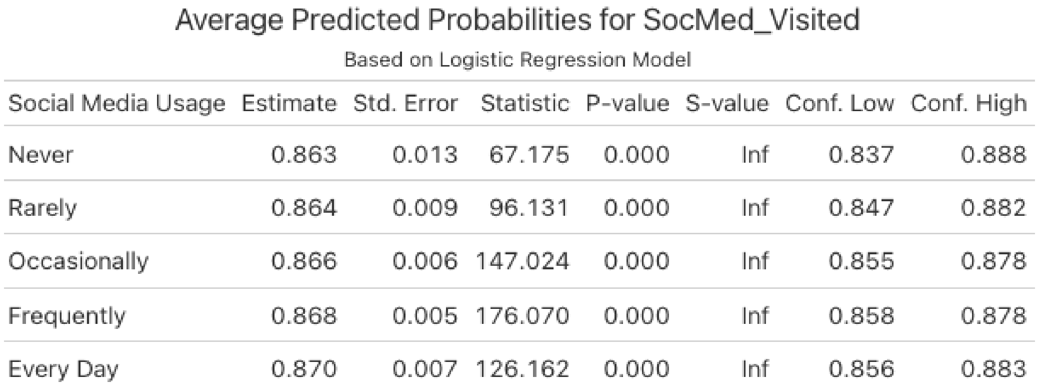
As our outcome is binary, we conducted four logistic regression models, sequentially adding control variables to evaluate the factors influencing the likelihood of using the internet for health information. Which variables we re-coded and how can be found in our markdown file attached to this memo.

- **Model 1 (Baseline):** Included only the independent and dependent variables.
- **Model 2: Added health-related controls** (Internet Access Type, Device Type, Confidence in Accessing Health Information, and Proportion of Misleading/False Information on Social Media).
- **Model 3: Incorporated demographic controls** (Health Status, Age, Gender, Race, Education, and Rural/Urban Residence).
- **Model 4: Included all controls, adding economic factors** (Household Income and Full-Time Employment).

Based on the AIC and BIC across the four models, Model Four provided the best fit. Additionally, after conducting a global F test, we found that Model Four performed better than Model Three with a low and significant residual deviance, resulting in our utilization of this model to discuss our key findings and to perform our robustness checks later on. Summaries of the models and model comparisons can also be found in the markdown file attached to this memo.

Key Findings

Our logistic regression models show that social media usage is positively associated with the likelihood of using the internet for health information, though the effect is small and not statistically significant in Model 4. As controls are added, the effect diminishes but remains positive, with significant controls including Internet Access Type, Device Type, Confidence in Health Information Access, Distrust in Social Media Information, Health Status, Gender, Race, Education, and Household Income.

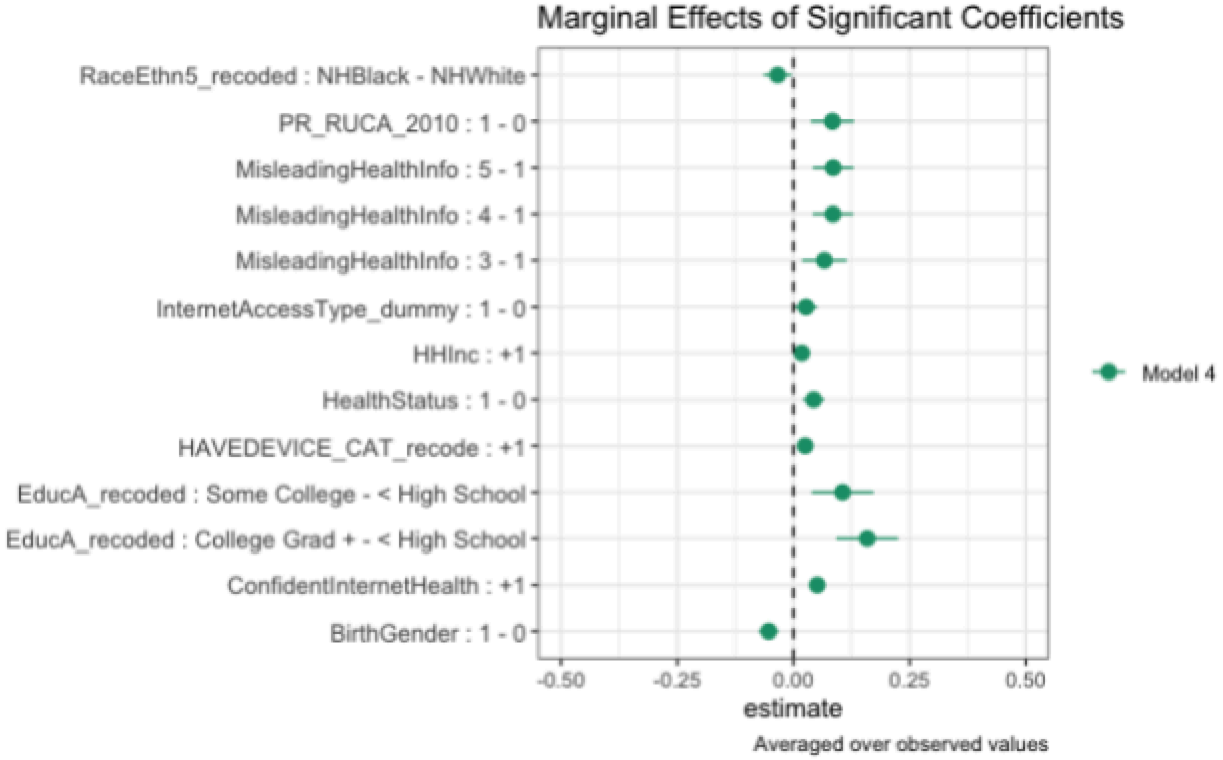


Average Predicted Probabilities for SocMed_Visited

Based on Logistic Regression Model

Social Media Usage	Estimate	Std. Error	Statistic	P-value	S-value	Conf. Low	Conf. High
Never	0.863	0.013	67.175	0.000	Inf	0.837	0.888
Rarely	0.864	0.009	96.131	0.000	Inf	0.847	0.882
Occasionally	0.866	0.006	147.024	0.000	Inf	0.855	0.878
Frequently	0.868	0.005	176.070	0.000	Inf	0.858	0.878
Every Day	0.870	0.007	126.162	0.000	Inf	0.856	0.883

Using average predicted probabilities above, we find that the likelihood of using the internet for health information increases as social media usage frequency rises, from 0.863 for "Never" to 0.870 for "Every Day," with this trend being statistically significant across all categories. Despite non-significance in the logit model, we can confirm that there is a substantive significance that may be overpowered by our controls. Using the observed values/counterfactual approach, we find that all significant controls, except being Black or Female, increase the likelihood of using the internet for health information. The marginal effect of our controls on this probability is 95%, regardless of high social media frequency, reflecting their strength.



Diagnostics and Robustness Checks

Main concerns surrounded multicollinearity, however the highest VIF in our model is 1.33, with others between 1 and 1.33, indicating no significant multicollinearity. Fixed and random effects were not implemented, as our data is neither panel data nor time-dependent, and is not clustered by county, region, or state. Model diagnostics show heteroskedasticity and deviations from normality, but due to the binary and categorical nature of our variables, we were able to proceed with the analysis. While we couldn't use Cluster Robust Standard Errors, assessing the robust standard errors alone confirms the model's robustness after adjusting for heteroskedasticity. Future work could aim to explore difference-in-differences and Instrumental Variables to assess the impact of Cell Phone Network Access.

References

Gottfried, J. (2024). Americans' social media use. Pew Research Center. https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/

Polanco-Levicán, K., & Salvo-Garrido, S. (2022). Understanding Social Media Literacy: A Systematic Review of the Concept and Its Competences. International journal of environmental research and public health, 19(14), 8807. https://doi.org/10.3390/ijerph19148807