Disease and Disclosure: Investigating Health-Related Self-Disclosure Online

Dionna Taylor

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Arts

University of Washington

2026

Committee:

Benjamin Mako Hill

Carmen Gonzalez

Wang Liao

Program Authorized to Offer Degree:

Communication

University of Washington

Abstract

Disease and Disclosure: Investigating Health-Related Self-Disclosure Online

Dionna Taylor

Co-Chairs of the Supervisory Committee:

Benjamin Mako Hill

Carmen Gonzalez

Department of Communication

This study examines how stigma influences patterns of health-related self-disclosure online, alongside disease severity and prevalence. As digital platforms increasingly serve as spaces for health information-seeking, understanding what drives attention to specific diseases is critical. A disease list ($N = 166$) was drawn from the Global Burden of Disease dataset. Disability-adjusted life years (DALYs), prevalence, and demographic data were integrated with Reddit submissions identified using LLM-assisted search terms from 2005 to 2023. Stigma scores were computed using word embeddings to capture dimensions of disgust and judgment. Multiple linear regression models assessed associations between stigma, DALYs, prevalence, and discussion volume. While stigma and prevalence were positively associated with discussion, these effects were not significant. In contrast, DALYs significantly predicted discussion volume ($B = 0.45$, $p < 0.01$). These findings offer insight as to how public attention is distributed across diseases and inform strategies to better align health interventions with population needs.

Disease and Disclosure: Investigating Health-Related Self-Disclosure Online

The widespread connectivity of the Internet has fundamentally transformed how, when, and with whom people communicate about health. Discussions of disease that were once largely confined to clinical encounters or private interpersonal settings can now take place in computer-mediated environments, enabling individuals to share experiences, seek information, and connect with others at unprecedented scale. These digital spaces have also created new opportunities to discuss sensitive or stigmatized health conditions, which may have been difficult to disclose in offline contexts (Miller, 2020).

Prior research suggests that stigma can suppress communication (Saunders, 2014; Smith & Applegate, 2018; Zhu et al., 2017), whereas perceived anonymity in online environments has been shown to facilitate greater self-disclosure (Suler, 2004; Walther, 1996), particularly when individuals are interacting with strangers (Anonymous, 1998). The intersection of these two dynamics raises important questions about how individuals navigate health-related conversations in digital spaces.

Reddit provides a particularly useful environment for examining the emergence of online self-disclosure surrounding health-related stigma for several reasons. First, unlike many social media platforms, Reddit is organized around topic-based communities rather than pre-existing social networks, allowing users to engage with others based on shared interests or experiences. Second, with 116 million daily active users (Backlinko Team, 2026), users have access to a large-scale audience that increases their chances of connecting with others who have shared experiences (Baym, 2015). Finally, given the public nature of most subreddits, barriers to entry for support-seeking on Reddit are relatively low, promoting an ideal place for users to easily inquire about any topic at their leisure.

This study examines how stigma influences patterns of disease-related discussion in online communities by analyzing discourse across Reddit. Drawing on theories of stigma and anonymous communication, this work investigates whether stigmatized diseases are more or less likely to generate discussion in online spaces, and how these patterns compare to disease burden and prevalence.

Stigma

Stigma refers to an attribute or characteristic that is "deeply discrediting", marking certain individuals as undesirable or deviant (Goffman, 1986). Although stigma has historically been associated with visibly flawed physical attributes, Goffman argues that stigma does not exist inherently in a person or condition. Rather, stigma emerges relationally through social interaction, in which the perceived abnormality of one individual is defined against the assumed "normalcy" of another, thereby magnifying perceived differences between the two (Goffman, 1986).

Given health conditions can often be identifiable through physical characteristics or behavioral changes, and because perceptions of these indicators vary across cultures (Goffman, 1986), it is understandable that some conditions are more strongly associated with stigma than others. Health-related stigma can result in various negative outcomes for those who navigate the world with a disease. For example, Saunders (2014) links experiences of stigma to the perceived taboo surrounding symptoms of Inflammatory Bowel Disease (IBD), highlighting consequences such as non-disclosure and the concealment of the condition. Zhu and colleagues (2017) further support the idea that stigmatized experiences are kept private, identifying secrecy as a communicative response linked to higher levels of public stigma. Their findings suggest secrecy can be an appealing strategy to avoid negative judgment and loss of opportunity for those with a

stigmatized health condition. However, secrecy, non-disclosure, and concealment pertaining to health-related stigma typically further isolate individuals from support and resources. Smith and Applegate (2018) argue that while communication is essential for building social relationships, stigma compromises an individual's ability to build those very relationships.

Taken together, these findings suggest that stigma not only shapes how health conditions are experienced, but also constrains when, where, and how individuals are willing to discuss them. However, much of this research focuses on offline contexts, leaving the role of online environments in facilitating or constraining disclosure underexplored. This gap is particularly important given the unique affordances of digital spaces, where anonymity may enable individuals to disclose health-related stigma more freely.

Anonymity and Self Disclosure

Anonymity has previously been linked with increased self-disclosure in offline contexts, particularly in interactions where individuals perceive a low risk of future consequences (Anonymous, 1998). However, the rise of the Internet has introduced new forms of mediated communication characterized by varying degrees of anonymity. Importantly, the extent to which individuals engage in self-disclosure online can be shaped by perceived anonymity (Anonymous, 1998; Clark-Gordon et al., 2019; Braun et al., 2021). Perceived anonymity can manifest in the absence of face-to-face interactions, even in contexts where true anonymity is limited. For example, online qualitative surveys have been used to study sensitive topics because they provide a "felt" sense of anonymity (Braun et al., 2021). Although such methods are not completely anonymous in practice, the lack of in-person interaction may reduce social desirability bias, facilitate participation, and encourage more candid responses. These approaches are particularly useful for individuals who are uncomfortable with face-to-face disclosure,

including but not limited to those with high social anxiety and those discussing topics that may feel sensitive, stigmatized, or emotionally difficult to disclose in person (Braun et al., 2021). This dynamic is especially relevant in discussions of health-related stigma, where concerns about judgment or social repercussions often inhibit disclosure in offline settings (Saunders, 2014; Smith & Applegate, 2018; Zhu et al., 2017).

Computer-mediated communication has long been associated with hyperpersonal communication, often resulting in higher levels of self-disclosure (Walther, 1996). The online disinhibition effect (Suler, 2004) offers a complementary framework, suggesting that individuals are more likely to express themselves openly and with fewer restraints when communicating on the Internet. This effect posits that individuals separate their online behavior from their offline life, reducing the perceived vulnerability associated with self-disclosure (Suler, 2004). Moreover, online spaces free individuals from immediate social circumstances, such as managing appearance, navigating real-time interactions, or interpreting others' reactions. Similarly, platforms that enable pseudonymous or anonymous participation, such as Reddit, may further facilitate open disclosure by reducing perceived vulnerability and social judgment (Luo & Hancock, 2020).

Together, perceived anonymity and the online disinhibition effect may play a key role in shaping how stigmatized health-related experiences are expressed and navigated. This is particularly relevant in digital spaces such as online communities and social media platforms, where individuals can discuss sensitive topics outside of traditional face-to-face constraints.

Online Communities

Online communities are spaces that foster group organization around a wide range of goals, providing opportunities for individuals to engage in discussions that may feel infeasible to

navigate offline (Kim & Dindia, 2011). These communities take many forms, including commons-based peer production (Benkler, 2006), in which members collaboratively produce shared resources (e.g., Wikipedia), and conversation-based communities, such as Reddit, where participants primarily engage in information-seeking and discussion (Hwang & Foote, 2021). Because these platforms are largely user-driven, they allow researchers to observe relatively unfiltered forms of communication and interaction among individuals who both seek and provide information. As a result, online communities offer a valuable window into social dynamics that are often difficult to capture in traditional research settings.

Information management is a crucial communicative process through which individuals navigate not only health conditions themselves, but also any uncertainty that often accompanies managing those conditions (Brashers et al., 2002). Mediated communication expands opportunities for this process by enabling both mass and interpersonal forms of health information and support seeking (Brashers et al., 2002). Within online communities, supportive interactions may take emotional, informational, appraisal, or instrumental forms (Malecki & Demaray, 2003). Among these forms, information support appears especially in stigmatized online communities, where individuals seek advice, experiential knowledge, and validation from others facing similar circumstances (Nobles et al., 2018). In this context, online platforms can function not only as spaces for self-disclosure, but also as accessible sources of health-related guidance and peer-based support.

Prior research suggests that traditional social networks may provide insufficient support for those coping with health-related stigma (Wright & Rains, 2013). However, online communities that foster opportunities for safe self-disclosure have been associated with providing supplemental support to compensate for limited offline resources (Miller, 2020). These

findings reinforce the idea that individuals often seek out others who share similar experiences as a form of coping and support-seeking (Zhu et al., 2017). Digital support allows for convenient connection with others who also have experiences with health-related stigma (Baym, 2015; Wright & Rains, 2013), by connecting those with rare conditions who may not know others in their immediate community who share the same condition (Brashers et al., 2002), providing an alternative for those who have limited offline support (Miller, 2020), and alleviating the pressure of face-to-face support groups for those with stigmatized conditions (Brashers et al., 2002). In such cases, participation in online communities can in and of itself function as a response to stigma, fostering solidarity among individuals with shared experiences (Zhu et al., 2017).

Peer production platforms, such as X (Twitter) and Reddit, enable nuanced interactions by enabling conversational, user-driven exchanges that differ from more visually oriented platforms, such as Instagram and TikTok. This type of exchange can be especially valuable for better understanding health-related stigma, as conversation-based platforms allow individuals to initiate and participate in discourse beyond their social spheres, illuminating how stigma is expressed, reinforced, or even resisted in online spaces. Additionally, the perceived safety of anonymity on some platforms creates a unique environment for information-seeking and exchange with minimal risk of identification.

Only recently have scholars begun to investigate online discourse trends pertaining to health-related stigma. For example, Brown and colleagues (2023) examined stigma communication in posts about groups of potentially stigmatized health conditions and disorders on Twitter (X), finding that while stigma was present, anti-stigma content was more prevalent than any other type of stigma communication. These findings extend prior research on how online communities foster supportive environments (Stellefson et al., 2018; Wright & Rains,

2013; Zhu et al., 2017) and play a role in combating health-related stigma messages. However, existing studies have primarily focused on identifying the presence of stigma within online communities, leaving open the question of how health-related stigma shapes the prevalence, attention, and engagement of disease self-disclosure online. As a result, further research is needed to better understand the relationship between stigma and online disease discussion.

Taken together, these works provide the foundation for this study. Prior research has shown that individuals with stigmatized conditions often conceal their experiences or avoid self-disclosure (Saunders, 2014; Zhu et al., 2017). In contrast, research on computer-mediated communication suggests that perceived anonymity in online environments can facilitate greater self-disclosure (Anonymous, 1998; Clark-Gordon et al., 2019; Braun et al., 2021). Building on this, online communities further support self-disclosure by connecting individuals with shared experiences and providing designated spaces for information exchange and social support, especially for those with stigmatized or rare conditions (Brashers et al., 2002). Collectively, these features suggest that online environments may be especially conducive to the disclosure of stigmatized health conditions in the absence of sufficient offline resources (Miller, 2020). Accordingly, this study examines whether stigma is associated with greater levels of online disease disclosure, proposing the following hypothesis:

H1: Diseases with higher levels of stigma will have greater discussion volume than diseases with lower levels of stigma in online communities.

Methods

To empirically test the proposed hypothesis, data were sourced from Reddit, a prominent computer-mediated communication platform frequently used to navigate sensitive health experiences (Nobles et al., 2018), seeking health information (Park & Conway, 2018), and

disclosing personal information (Miller, 2020). While health discourse is ubiquitous across digital spaces, Reddit's structure is uniquely well-suited for investigating disease-specific discussion for several reasons.

First, Reddit is organized around topics rather than interpersonal networks, meaning users subscribe to communities (subreddits) instead of individual profiles. This structure facilitates the aggregation of individuals around shared interests, rather than pre-existing social ties. Relatedly, health has been identified as one of the top audiences and community-related topics on Reddit, with more than 9 million patients, 6.2 million healthcare providers, and 9.1 million caregivers seeking support, advice, and community (Reddit Business, 2026). Second, this topic-based organization expands the audience size users can access (Baym, 2015), thereby increasing the likelihood that individuals with stigmatized diseases can connect with others who share their condition or experiences. In this way, information exchange is largely crowdsourced, allowing users to connect with others with similar questions or concerns. Prior work suggests that this form of exchange is not merely a substitute for professional knowledge, but a distinct form of patient expertise, as patients provide experiential and lived knowledge that clinicians may not offer, making patient-to-patient interaction especially valuable in health-related discussions (Hartzler & Pratt, 2011). Finally, because most subreddits are publicly accessible, barriers to participation are relatively low compared to private communities, which may restrict entry in ways that inadvertently exclude individuals seeking support for stigmatized diseases (Yeshua-Katz & Hård af Segerstad, 2020).

Curating a List of Diseases

In choosing diseases for this study, I began with all 463 health outcomes from the Global Burden of Disease (GBD) Collaborative Network Results (2025), which includes data from 204 countries and territories.

This dataset is organized into four hierarchical levels. Level 1 contains the most aggregated data, encompassing all corresponding level 2 diseases. Each level 2 disease aggregates its corresponding level 3 diseases, which in turn are made up of level 4 diseases, the most specific level in the hierarchy. Because each higher-level disease is composed of its subsequent diseases, including multiple levels simultaneously would create analytical redundancy and compromise the independence of the analysis. Restricting the dataset to a single hierarchical level therefore would ensure analytic clarity.

I downloaded the data in batches that corresponded to each disease level. Data nested under ‘Injuries’ were excluded from each batch, as they are outside the scope of this study.

The following parameters were used for each batch:

GBD Estimate: Cause of injury or death

Measure: DALYs, Prevalence

Metric: Rate

Location: Global

Age: All ages, <20 years, 20-54 years, 55+ years

Sex: Both, Female, Male

Year: 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023

The downloaded dataset did not include the corresponding hierarchy levels for each disease. Therefore, I used the GBD Collaborative Network (2025) Excel file to map the hierarchy levels onto my dataframe using the cause ID as the linking variable.

I excluded the level 1 (*N = 3*) and level 2 diseases (*N = 19)* as they were too broad on their own. For instance, the aggregated categories of level 1 diseases (Communicable, maternal, neonatal, and nutritional diseases; Non-communicable diseases; and Injuries) lacked the specificity required for meaningful differentiation amongst individual diseases.

However, when considering levels 3 and 4, each presented limitations if used independently. Level 3 diseases offered stronger conceptual organization across disease categories, but collapsed important distinctions between specific conditions, such as different sexually transmitted infections, types of drug use disorders, or types of congenital birth defects. In contrast, level 4 diseases were often too granular. Some included broad residual categories (e.g., ‘other mental disorders’ or ‘other unspecified infectious diseases’) that lacked sufficient specificity, while others divided conditions so finely that the distinctions in online discourse were unlikely to be detected. For these reasons, I derived the list of diseases from both levels 3 and 4, grouping diseases where necessary to reduce noise.

To do so, I began with level 3 diseases, but when a level 3 category was too broad, I opted to use all of its nested level 4 diseases instead. Since level 3 data represent an aggregation of the corresponding level 4 causes, this substitution did not alter which diseases were included; rather, it provided greater specificity where it might otherwise have been lost. I took care to not include both a level 4 disease and its parent level 3 disease, as this would result in duplicate data.

The only exception to including a level 3 or nested level 4 disease was in cases where the category represented ‘other’ or unspecified conditions. These ‘other’ categories typically refer to

experiences that do not meet specific diagnostic criteria (e.g., ‘other sexually transmitted infections,’ ‘other neglected tropical diseases,’ or ‘other direct maternal disorders’) and therefore were excluded. There were six level 4 causes (Thalassemias, Thalassemias trait, Sickle cell disorders, Sickle cell trait, G6PD deficiency, and G6PD trait) that I felt were too granular and would be better represented as three distinct diseases. However, there were no corresponding level 3 categories to separate these groups; instead, they were all grouped together. Therefore, I opted to only include Thalassemias, Sickle cell disorders, and G6PD deficiency. My final list consisted of 183 causes.

Data Collection

The data for this study consisted of submissions from all subreddits on Reddit spanning from 2005 to 2023. Reddit data were collected from Pushshift.io (Baumgartner et al., 2020) via keywords. The keywords used as search strings for each disease were derived from a large language model (LLM). I supplied the list of diseases to the GPT-4.1 model via OpenAI’s ChatGPT API and prompted the LLM to generate a list of search terms that describe how someone might talk about a disease on Reddit. This yielded a total of 971 search terms across all 183 diseases. Because LLMs are trained on large, diverse corpora including informal and colloquial language, they are well-suited to bridging the gap between formal clinical terminology and the conversational language characteristic of online platforms like Reddit. By leveraging the model's ability to generate search terms, I was able to capture a broader range of community discussions that might otherwise be excluded by rigid keyword searches of just the clinical disease name.

After each submission was collected (N = 47,445,832), preliminary counts of submissions per disease were calculated. However, many keywords, particularly acronyms, can

refer to multiple concepts beyond the disease of interest. Given the breadth of topics discussed on Reddit, a term may reference a disease in one context but not in another. For example, the term “aids” in a hearing-related community is unlikely to refer to HIV/AIDS. Thus, determining how many submissions were inaccurately classified as referring to a disease in question was imperative.

To gain a more accurate understanding of Reddit discourse on each disease, I used an LLM-assisted framework to determine whether a given Reddit submission was talking about the disease in question. Prior research has demonstrated that LLMs can achieve moderate to substantial agreement with human coders in qualitative classification tasks (Hou et al., 2024), supporting their use as scalable alternatives for analyzing large text corpora. Because the full corpus of Reddit submissions was too large to classify in its entirety, random samples of submissions were collected for each disease to ensure representativeness. The central limit theorem supports drawing reliable inferences from these samples, provided they are sufficiently large (Kwak & Kim 2017). Thus, I took a random sample of 50 submissions per disease. The only time the sample size would be less than 50 was if there were less than 50 submissions for a given disease. In that case, every submission for that disease was classified. Then, I provided the sample submissions to the GPT-5.1 model via OpenAI’s ChatGPT API and requested binary responses (‘TRUE’ or ‘FALSE’) as to whether each submission was pertinent to the disease the post had been classified as. This process yields the rate of submissions that are inaccurately classified as referring to a given disease (the false-positive rate), enabling more accurate inference regarding the representativeness of the submission data.

I then calculated the false-positive rate per disease by dividing the false positives by the total number of submissions for each respective disease. 81% of the diseases had a false positive

rate of less than 75%, and 56% had a rate of less than 25%. Figure 1 illustrates the false-positive rates per disease. I then adjusted the total number of submissions per disease by multiplying the true positive rate by the initial total number of submissions, which yielded 20,003,751 submissions across all 183 diseases.

Evaluation of the model's performance, based on a random sample of at most five submissions per disease ($N = 899$), was conducted by comparing the classifications produced by a human for the same sample with those produced by the LLM. There was 93% agreement between the researcher and the LLM, calculated from binary coding of whether a submission was about the disease it claimed to be about.

Measurement

Building a Stigma Classifier

Stigma shapes whether, and under what conditions, individuals are willing to disclose health-related information. My research aims to investigate the social influence of stigma. As a result, I strive to avoid imposing any cultural assumptions about the degree to which certain diseases may be considered stigmatized. To empirically capture disease stigma, I adopted a word embedding approach used by Best and Arseniev-Koehler (2023). Word embeddings are a text analysis approach that capture word-based associations determined by the context in which they appear within a given corpus. Here, word embeddings are used to measure the degree in which a given disease is associated with stigmatizing language, such that diseases with stronger such associations receive higher stigma scores and those with weaker associations receive lower scores.

Drawing on existing theory, Best and Arseniev-Koehler (2023) derived four dimensions of stigma: danger, disgust, immorality, and negative personality traits. Each dimension is then

assigned anchor words that represent a spectrum from least to most similar in connotation, with the most similar connotation being the stigmatized end of the dimension. For instance, the anchor word “dangerous” would represent the stigmatized end of the danger dimension, whereas “safe” would represent the opposite end. The word embedding will determine how closely a given disease is toward the stigmatized end of each dimension. While Best and Arseniev-Koehler created their own embeddings from media articles, I opted to use a pre-trained Word2Vec model to generate scores across the stigma dimensions. Following their framework, I operationalized stigma through two dimensions, disgust and judgment, where judgment was computed as the mean of immorality and negative personality trait scores for each disease. A stigma score was then generated for each disease by averaging the disgust and judgment dimensions. However, stigma scores could not be calculated for 17 diseases; therefore, my final sample consisted of 166 diseases.

Self-Disclosure

Although Reddit interactions can be operationalized in multiple ways (e.g., submissions, comments, and votes), this study focuses on the number of submissions as the primary indicator of self-disclosure of a given disease. While comment counts may reflect engagement or popularity, they can also capture the extent to which a topic reflects a shared experience. Given the sensitive nature of some of the diseases examined, lower levels of comment activity on individual posts are not unexpected. Accordingly, restricting the analysis to submissions allows for a more direct measure of users’ willingness to initiate self-disclosure, capturing the emergence of disease discussion rather than the subsequent level of interaction they generate.

Control Variables

Because the self-disclosure of a disease online may reflect factors beyond stigma alone, this study controls for disease prevalence and burden. These variables account for the possibility that some diseases may generate greater discussion due to the scale of their impact on the population. For instance, since asthma has a higher prevalence than measles, it would be reasonable for more people to discuss it online. Similarly, just as more common conditions have larger pools of potentially affected individuals who may seek information online, more burdensome diseases may prompt greater disclosure due to the severity and complexity of the experiences they come with. Without accommodating these factors, any observed relationship between stigma and self-disclosure would likely reflect differences in how common or serious a disease is, rather than the influence of stigma itself.

Prevalence, operationalized as the number of cases per 100,000 population, captures the reach of a disease within the population and accounts for the likelihood that a larger affected population produces a greater volume of online self-disclosure. Disease severity is then operationalized using Disability-Adjusted Life Years (DALYs) per 100,000 population, a measure that captures overall disease burden by combining years of life lost due to premature mortality with years lived with disability. DALYs are widely used in public health research to inform which health conditions carry the greatest burden, where to allocate resources to help people manage these conditions, and how to promote equitable and accessible healthcare (Murray & Acharya, 1997). Within the scope of this study, DALYs serve as an indicator of the overall severity a disease has on an individual, as conditions with higher DALYs may generate more discussion simply because they are more debilitating or life-altering, independent of stigma. Data for both measures were obtained from the Global Burden of Disease (GBD) Collaborative Network (2025).

Statistical Analysis

To test H1, which states that diseases with higher levels of stigma will have greater discussion volume, I fit a multiple linear regression model with Reddit submissions as the dependent variable, and stigma as the independent variable. In this model, I included disease prevalence and severity as controls. I log-transformed submissions, prevalence, and DALYs to account for the right-skewed nature of these measures.

$$log(submissions) = B_0 + B_1 stigma + B_2 log(prevalence) + B_3 log(DALYs) + \epsilon$$

Results

The linear regression results to test H1 are presented in Table 1; additionally, Figure 2 illustrates the regression of submissions on stigma. Contrary to H1, which predicted that higher levels of stigma would be associated with greater discussion volume, there was no significant relationship between stigma and submissions in my sample ($B = 0.12$, $p > 0.1$). Similarly, log-transformed mean prevalence was not significantly associated with submissions ($B = 0.10$, $p > 0.1$), as shown in Figure 3. Together, these findings suggest that neither stigma nor prevalence meaningfully predicts variation in submission volume.

Mean DALYs, however, were positively associated with submissions ($B = 0.45$, $p < 0.01$), indicating that for every 10% increase in disease burden, submissions increased by approximately 4.5% in my sample (Figure 4). This result indicates a potential link between disease burden and submission patterns that warrants further investigation. All estimates can be found in Table 1.

Following the same data transformation used in the logistic regression model, the Pearson correlation between the log-transformed prevalence and DALYs ($r = 0.44$) was also moderately correlated, as was the correlation between disease severity and prevalence.

Discussion

This study sheds light on how self-disclosure varies across diseases and can inform future health communication strategies to better align intervention with population needs. The findings suggest that structural indicators of disease burden, particularly DALYs, are more strongly associated with online disease discussion than stigma or prevalence. This highlights the importance of disease burden in shaping online health discourse.

Stigma communication theory predicts that individuals with stigmatized conditions are more likely to engage in non-disclosure (Saunders, 2014; Zhu et al., 2017), while anonymity theory suggests that online environments may reduce these barriers by making individuals more comfortable sharing sensitive information with strangers (Anonymous, 1998; Suler, 2004). However, the findings of this study do not support H1, which proposed that higher levels of stigma would be associated with greater online discussion volume. This suggests that although anonymity may reduce some of the constraints on disclosure observed in offline contexts, it may not be sufficient to override other determinants of health-related engagement in online settings. Another explanation for this finding may be due to a small sample size. A sensitivity analysis indicated that the study was powered to detect small-to-moderate effects, but may not have been sensitive to very small associations; therefore, subtle effects of stigma cannot be ruled out. Furthermore, a methodological consideration is that the analysis focused only on Reddit submissions, excluding comments, which may have contained unaccounted indicators of self-disclosure. More broadly, despite these findings suggesting that stigma may not strongly influence the volume of online health discussion, further research exploring whether stigma shapes the nature of online self-disclosure that occurs would be fruitful.

DALYs emerged as a significant predictor of online discussion volume, suggesting that conditions associated with greater burden tend to generate higher levels of self-disclosure in digital spaces. DALYs capture the total loss of healthy life associated with a condition by combining years of life lost due to premature mortality and years lived with disability or poor health. As such, they provide a comprehensive population-level estimate of disease impact that extends beyond standard measures of prevalence or mortality.

One unexpected finding was that disease prevalence was not a significant predictor of discussion volume. The initial rationale for including prevalence was the assumption that more common diseases would generate more online self-disclosure, given more people online are more likely to be affected by them. While the estimated effect was positive, its magnitude was small and did not reach statistical significance. In contrast, when considered alongside the results for DALYs, these findings suggest that disease burden is a stronger driver of online health discussion than prevalence alone. In other words, diseases that are less common but more severe may attract disproportionately high levels of self-disclosure online.

Overall, the findings from this study suggest that health-related self-disclosure online is not primarily driven by the degree of stigma or how widespread a disease is, but rather by indicators of overall burden. Disease severity appears to play a more central role in shaping whether conditions are discussed, potentially outweighing the influence of both stigma and prevalence. Notably, a comparison of the top 10 conditions ranked by DALYs (Figure 5) and by stigma (Figure 6) reveals no overlap between the two groups. Highly stigmatized conditions in this sample, such as miscarriage/abortion, acne, and syphilis, do not appear among those considered to harbor the greatest burden, which instead includes conditions such as HIV, malaria, tuberculosis, and stroke. This stark difference illustrates how stigma and disease severity capture

fundamentally different dimensions of health, which has crucial implications for health communication. Diseases that are highly stigmatized but lower in overall burden may be systematically underrepresented in online discourse, despite the significant social, psychological, and interpersonal challenges they pose for affected individuals. As a result, efforts to raise awareness or foster engagement around these conditions may require more intentional and targeted communication strategies, rather than relying on organic patterns of discussion that appear to be more closely aligned with disease burden.

Future research should move beyond measuring the volume of disease-related discussion to examine its nature, including the types of support, information sharing, or personal experiences that emerge. Moreover, analyzing comment-level data, rather than focusing solely on submissions, may provide a more nuanced understanding of engagement and reveal patterns of participation not captured through submission volume alone. Another important direction for future work is to investigate the role of disease burden in driving online engagement. At the time of this study, no other work explicitly examines DALYs as a predictor of activity within online communities. Future studies could extend this work by exploring whether the relationship between DALYs and discussion volume holds across different platforms, cultural contexts, or disease categories. Finally, expanding the scope of analysis to include additional platforms would help determine whether these patterns are consistent across digital spaces. Such work would contribute to a more comprehensive understanding of how public attention to health issues is distributed online and how it may be leveraged to better align communication strategies with public health priorities.

Limitations

The findings of this study should be considered in light of several limitations. First, demographic detail from the Global Burden of Disease (GBD) data was aggregated, limiting the ability to examine age- and sex-specific variation in disease burden. As a result, conditions that disproportionately affect specific populations (e.g., cervical or testicular cancer) may be obscured, potentially masking relationships between disease burden and online discussion that are more pronounced within particular demographic groups.

Second, both DALYs and prevalence were averaged over the 2005-2023 period, while Reddit submission counts were aggregated over the same time frame. This temporal aggregation reduces the sensitivity of year-to-year variation in disease burden, prevalence, and online discussion patterns, and may veil informative short-term associations between changes in disease impact and shifts in public attention. Relatedly, while DALYs and prevalence estimates are derived from global data, Reddit's user base is not globally representative. The platform's weekly viewership is concentrated in the USA, UK, Canada, Germany, and Australia (Reddit Business, 2026), and access inherently requires both an electronic device and reliable internet connectivity, factors that skew representation. This geographic mismatch has two implications for interpretation. First, disease burden metrics for conditions with highly regionalized distributions, such as Malaria, in which the last case in the United States pre-dated 2005, may not reflect the experiences of the users discussing them. Additionally, restricting DALYs and prevalence estimates to Reddit-dominant regions would have excluded meaningful global burden data and rendered analysis of diseases such as malaria impossible. Therefore, global estimates remain as the most comprehensive available proxy, with the acknowledgment that the relationship between online discourse and disease burden may be blurred for geographically concentrated conditions.

Findings should be interpreted with this in mind, particularly for diseases with distinct regional variation in prevalence or severity.

Third, the use of submission counts as a proxy for self-disclosure is inherently imperfect. The rationale for privileging submissions over comments is that submissions may better indicate the emergence of discussion, whereas comments are typically elicited responses prompted by an initial post. However, submissions themselves may reflect a range of communicative purposes beyond self-disclosure. For example, users may initiate discussions about medications or disease-related statistics, post on behalf of friends or family members, or describe observations of another individual's experience coping with a disease. As such, submission frequency alone cannot definitively distinguish personal self-disclosure from other forms of engagement. Future research would therefore benefit from content analysis of submissions to better assess the nature of self-disclosure online.

Fourth, the analysis relied on an LLM-assisted filtering process to assess the relevance of Reddit submissions to each respective disease. Because submissions were evaluated outside of context, misclassification is possible, particularly in distinguishing whether a submission meaningfully pertains to a disease topic. While this approach enabled large-scale data processing, classification error may introduce noise into the measurement of discussion volume.

Finally, this study focuses exclusively on Reddit as a platform for online health discussion. Reddit's user base, norms, and affordances, such as perceived anonymity and topic-based structure, may not be representative of other online spaces. As a result, the generalizability of these findings to other platforms (e.g., more identity-linked or visually oriented social media) remains uncertain.

Conclusion

This study examined how stigma, prevalence, and disease burden relate to patterns of health-related self-disclosure in online communities. Contrary to expectations derived from stigma communication and online anonymity theories, stigma was not a significant predictor of discussion volume. Similarly, disease prevalence was not strongly associated with online self-disclosure. Instead, measures of disease burden emerged as the most consistent predictor of discussion volume, suggesting that conditions associated with greater overall impact on health generate more engagement in digital spaces.

These findings highlight a critical disconnect between different dimensions of health, as conditions that are highly stigmatized are not the same as those that have the greatest overall disease burden. The absence of overlap between these groups suggests that self-disclosure is not evenly distributed across conditions, but instead follows patterns more closely aligned with severity than with social experience. As a result, highly stigmatized but lower-burden conditions may remain comparatively underrepresented in digital spaces, despite the substantial challenges they pose for affected individuals. As digital platforms continue to play a central role in health communication, addressing this imbalance will be essential to ensure that targeted communication intervention more equitably reflects both population health priorities and the lived experiences of those navigating stigma.

References

Anonymous. (1998). To reveal or not to reveal: A theoretical model of anonymous communication. *Communication Theory, 8*(4), 381-407. https://doi.org/10.1111/j.1468-2885.1998.tb00226.x

Backlinko Team. (2026). *Reddit User and Growth Stats.* Backlinko. Available at: https://backlinko.com/reddit-users (accessed 2 April 2026).

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media 14*, 830–839. https://doi.org/10.48550/arXiv.2001.08435

Baym, N. K. (2015). *Personal Connections in the Digital Age* (Second edition). Polity Press.

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale University Press.

Best, R. K., & Arseniev-Koehler, A. (2023). The stigma of diseases: Unequal burden, uneven decline. *American Sociological Review, 88*(5), 938–969. https://doi.org/10.1177/00031224231197436

Brashers, D. E., Goldsmith, D. J., & Hsieh, E. (2002). Information seeking and avoiding in health contexts. *Human Communication Research*, *28*(2), 258–271. https://doi.org/10.1111/j.1468-2958.2002.tb00807.x

Braun, V., Clarke, V., Boulton, E., Davey, L., & McEvoy, C. (2021). The online survey as a qualitative research tool. *International Journal of Social Research Methodology, 24*(6), 641–654. https://doi.org/10.1080/13645579.2020.1805550

Brown, R., Sillence, E., Coventry, L., Branley-Bell, D., Murphy-Morgan, C., & Durrant, A. C. (2023). Health stigma on Twitter: Investigating the prevalence and type of stigma

communication in tweets about different conditions and disorders. *Frontiers in Communication*, *8*. https://doi.org/10.3389/fcomm.2023.1264373

Clark-Gordon, C. V., Bowman, N. D., Goodboy, A. K., & Wright, A. (2019). Anonymity and Online Self-Disclosure: A Meta-Analysis. *Communication Reports, 32*(2), 98–111. https://doi.org/10.1080/08934215.2019.1607516

Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2023 (GBD 2023) Causes of Death and Nonfatal Causes Mapped to ICD Codes. *Seattle, United States of America: Institute for Health Metrics and Evaluation (IHME)*, 2025. https://ghdx.healthdata.org/record/ihme-data/gbd-2023-cause-icd-code-mappings

Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2023 (GBD 2023) Results. *Seattle, United States: Institute for Health Metrics and Evaluation (IHME)*, 2024. Available from https://vizhub.healthdata.org/gbd-results/.

Goffman, E. (1986). *Stigma: Notes on the management of spoiled identity* (1st ed.). Touchstone.

Hartzler, A., & Pratt, W. (2011). Managing the personal side of health: How patient expertise differs from the expertise of clinicians. *Journal of Medical Internet Research, 13*(3), e62. https://doi.org/10.2196/jmir.1728

Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., & Ker, C. L. (2024). Prompt-based and fine-tuned GPT models for context-dependent and -independent deductive coding in social annotation. In: *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 518–528.

Hwang, S., & Foote, J. D. (2021). Why do people participate in small online communities? *Proceedings of the ACM on Human‑Computer Interaction, 5*(CSCW2), Article 462. https://doi.org/10.1145/3479606

Kim, J., & Dindia, K. (2011). Online self-disclosure: A review of research. In K. B. Wright & L. M. Webb (Eds.), *Computer-Mediated Communication in Personal Relationships* (pp. 156-180). New York: Peter Lang Publishing.

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology, 70*(2), 144. https://doi.org/10.4097/kjae.2017.70.2.144

Luo, M., & Hancock, J. T. (2020). Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current Opinion in Psychology, 31*, 110–115. https://doi.org/10.1016/j.copsyc.2019.08.019

Malecki, C. K., & Demaray, M. K. (2003). What type of support do they need? Investigating student adjustment as related to emotional, informational, appraisal, and instrumental support. *School Psychology Quarterly*, *18*(3), 231–252. https://doi.org/10.1521/scpq.18.3.231.22576

Miller, B. (2020). Investigating Reddit self-disclosure and confessions in relation to connectedness, social support, and life satisfaction. *The Journal of Social Media in Society*, *9*(1), 39-62.

Murray, C. J. L., & Acharya, A. K. (1997). Understanding DALYs. *Journal of Health Economics, 16*(6), 703–730. https://doi.org/10.1016/S0167-6296(97)00004-0

Nobles, A., Dreisbach, C., Keim-Malpass, J., & Barnes, L. (2018). "Is this an STD? Please help!": Online information seeking for sexually transmitted diseases on Reddit. *Proceedings of the International AAAI Conference on Web and Social Media, 12*(1). https://doi.org/10.1609/icwsm.v12i1.15050

Park, A., & Conway, M. (2017). Tracking health related discussions on Reddit for public health applications. *AMIA Annual Symposium Proceedings*, 1362–1371.

Reddit Business. (2026). *Audience Insights.* Reddit. Available at: https://www.business.reddit.com/audience-insights (accessed 13 May 2026).

Saunders, B. (2014). Stigma, deviance and morality in young adults' accounts of inflammatory bowel disease. *Sociology of Health & Illness, 36*(7), 1020–1036. https://doi.org/10.1111/1467-9566.12148

Smith, R. A. (2011). Stigma, communication, and health. In T. L. Thompson, R. Parrott, & J. F. Nussbaum (Eds.), *The Routledge Handbook of Health Communication* (pp. 480-493). Routledge. https://doi.org/10.4324/9780203846063.ch29

Smith, R. A., & Applegate, A. (2018). Mental health stigma and communication and their intersections with education. *Communication Education, 67*(3), 382-393. https://doi.org/10.1080/03634523.2018.1465988

Stellefson, M. L., Shuster, J. J., Chaney, B. H., Paige, S. R., Alber, J. M., Chaney, J. D., & Sriram, P. S. (2018). Web-based health information seeking and eHealth literacy among patients living with chronic obstructive pulmonary disease (COPD). *Health Communication*, *33*(12), 1410–1424. https://doi.org/10.1080/10410236.2017.1353868

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior, 7*(3), 321-326. https://doi.org/10.1089/1094931041291295

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal and hyperpersonal interaction. *Communication Research, 23*, 3–43. https://doi.org/10.1177/009365096023001001

Wright, K. B., & Rains, S. A. (2013). Weak-tie support network preference, health-related stigma, and health outcomes in computer-mediated support groups. *Journal of Applied*

Communication Research, *41*(3), 309–324. https://doi.org/10.1080/00909882.2013.792435

Yeshua-Katz, D., & Hård af Segerstad, Y. (2020). Catch 22: The paradox of social media affordances and stigmatized online support groups. *Social Media + Society, 6*(4). https://doi.org/10.1177/2056305120984476

Zhu, X., Smith, R. A., & Parrott, R. L. (2017). Living with a rare health condition: The influence of a support community and public stigma on communication, stress, and available support. *Journal of Applied Communication Research, 45*(2), 179–198. https://doi.org/10.1080/00909882.2017.1288292

Appendix

Table 1. Linear Regression of Submissions on Stigma

Variables	β	SE	95% CI	p
(Intercept)	6.371^{***}	(0.816)	$[4.914, 7.828]$	< 0.001
Stigma	0.120	(0.164)	$[-0.205, 0.444]$	0.468
Logged Mean Prev	0.096	(0.117)	$[-0.098, 0.290]$	0.330
Logged Mean DALYs	0.452^{**}	(0.185)	$[0.129, 0.774]$	0.006
R^2	0.084			
Adj. R^2	0.067			
Num. obs.	166			

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Figure 1

Distribution of false positive rates per disease

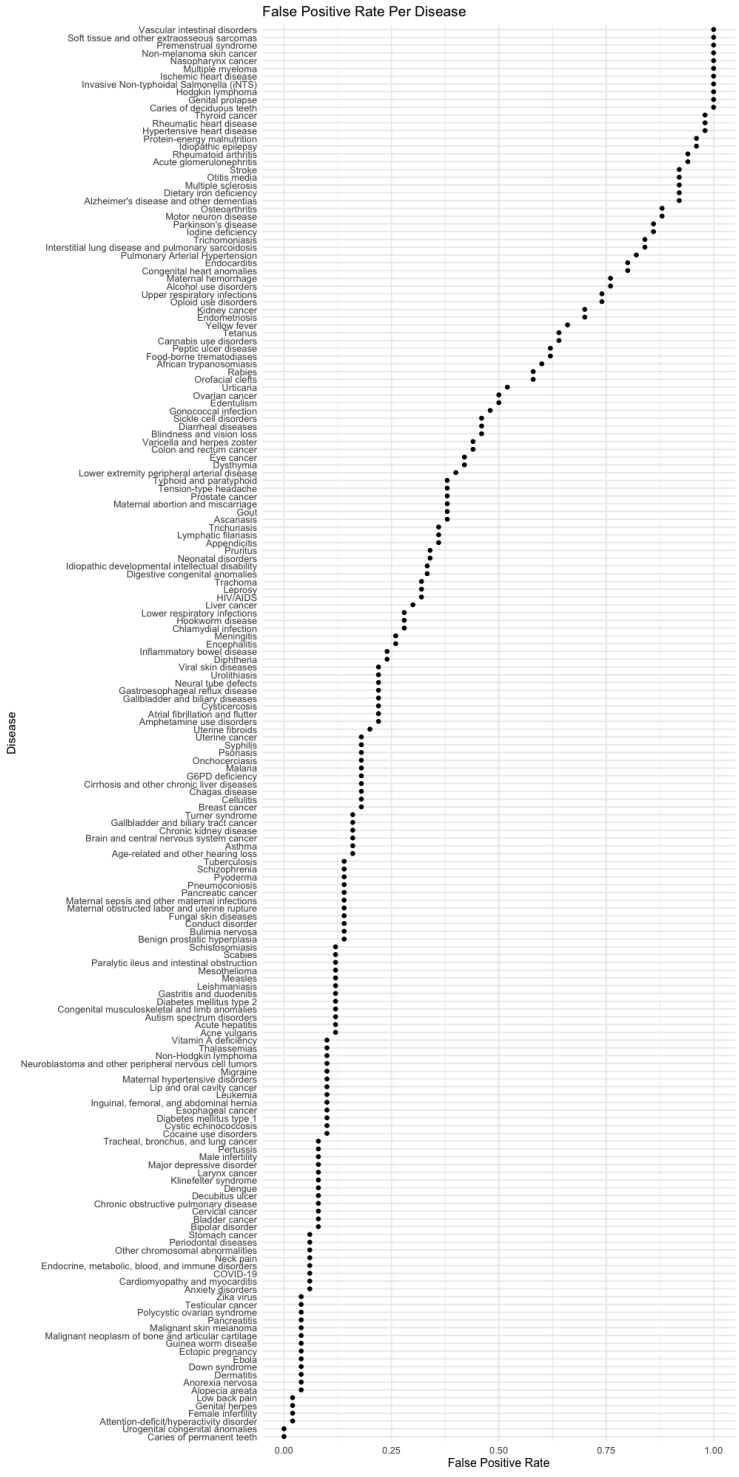

Figure 2

Regression of submissions on stigma

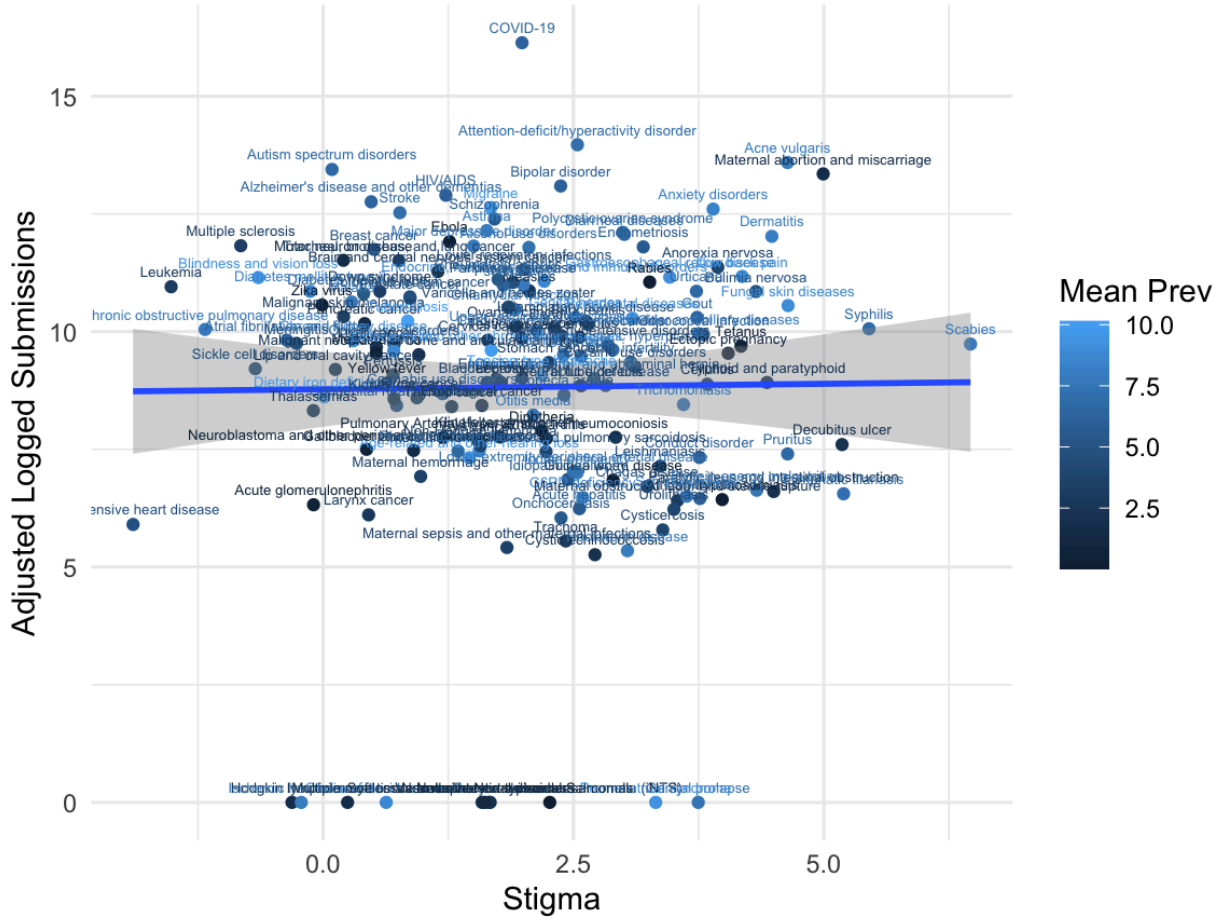

Figure 3

Regression of submissions on prevalence

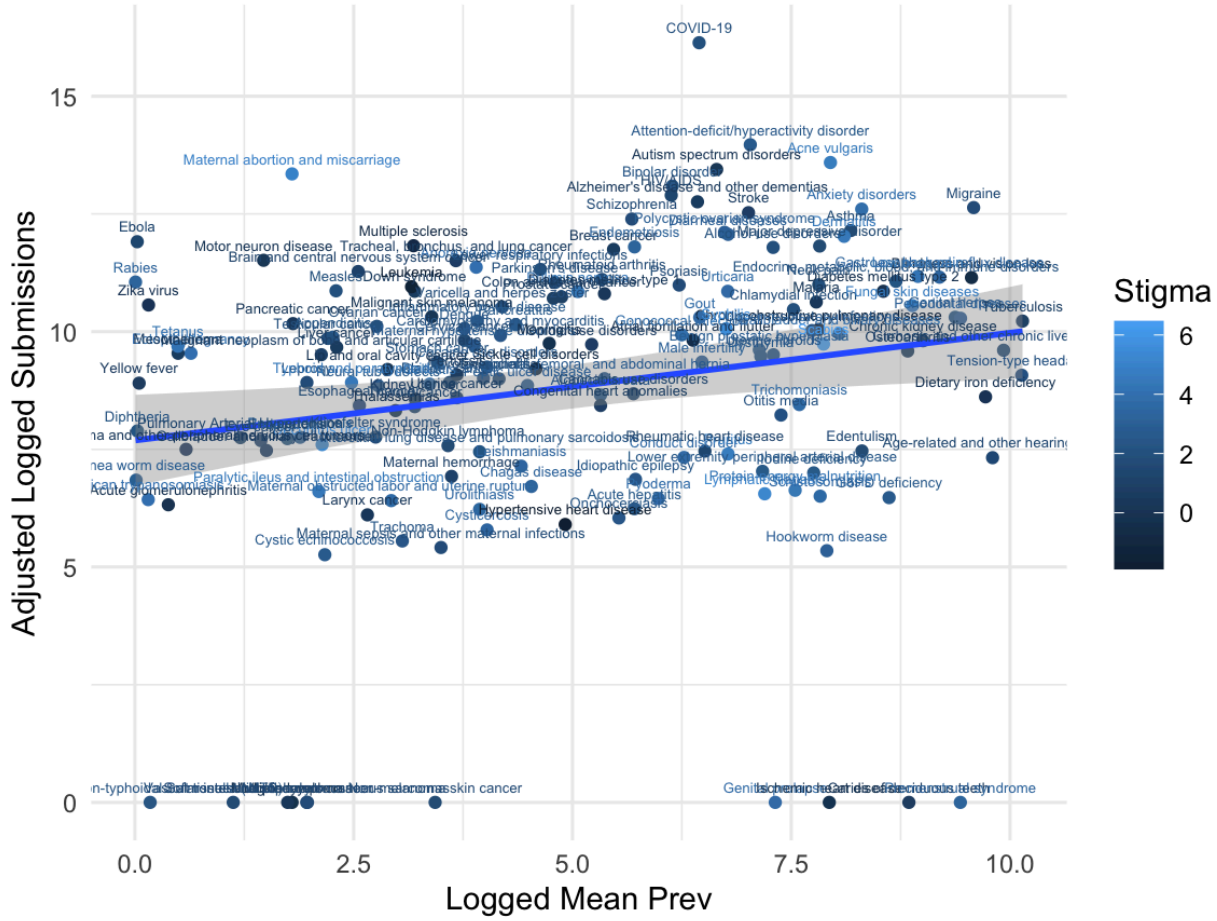

Figure 4

Regression of submissions on DALYs

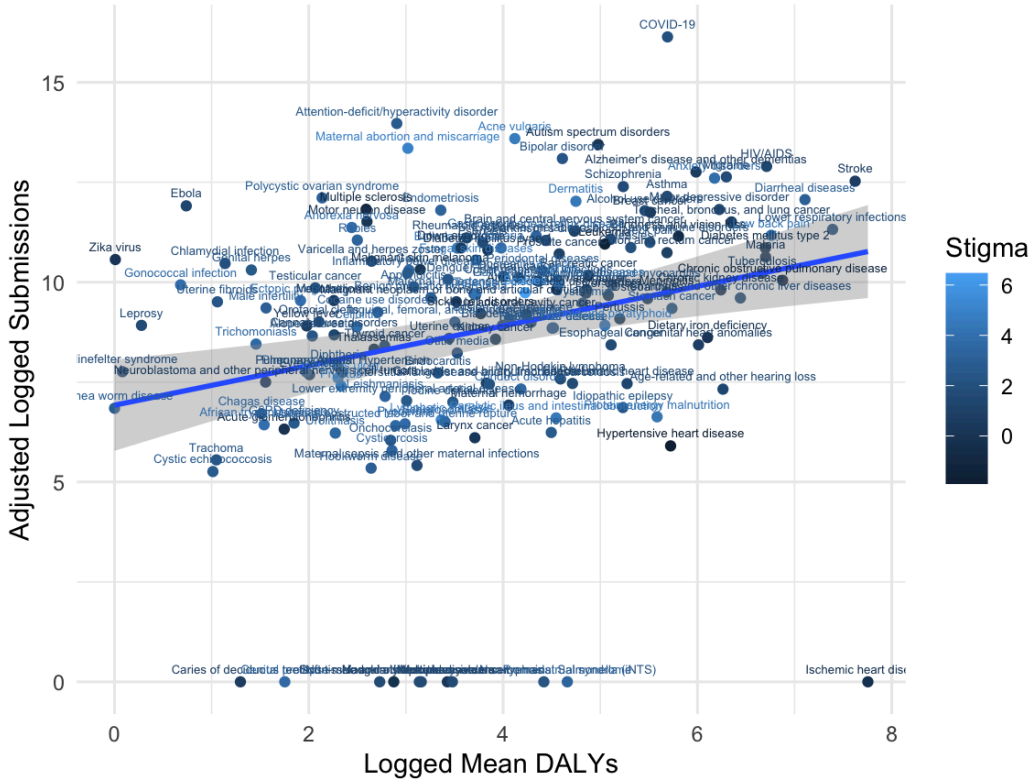

Figure 5

Top 10 Diseases with the Highest DALYs

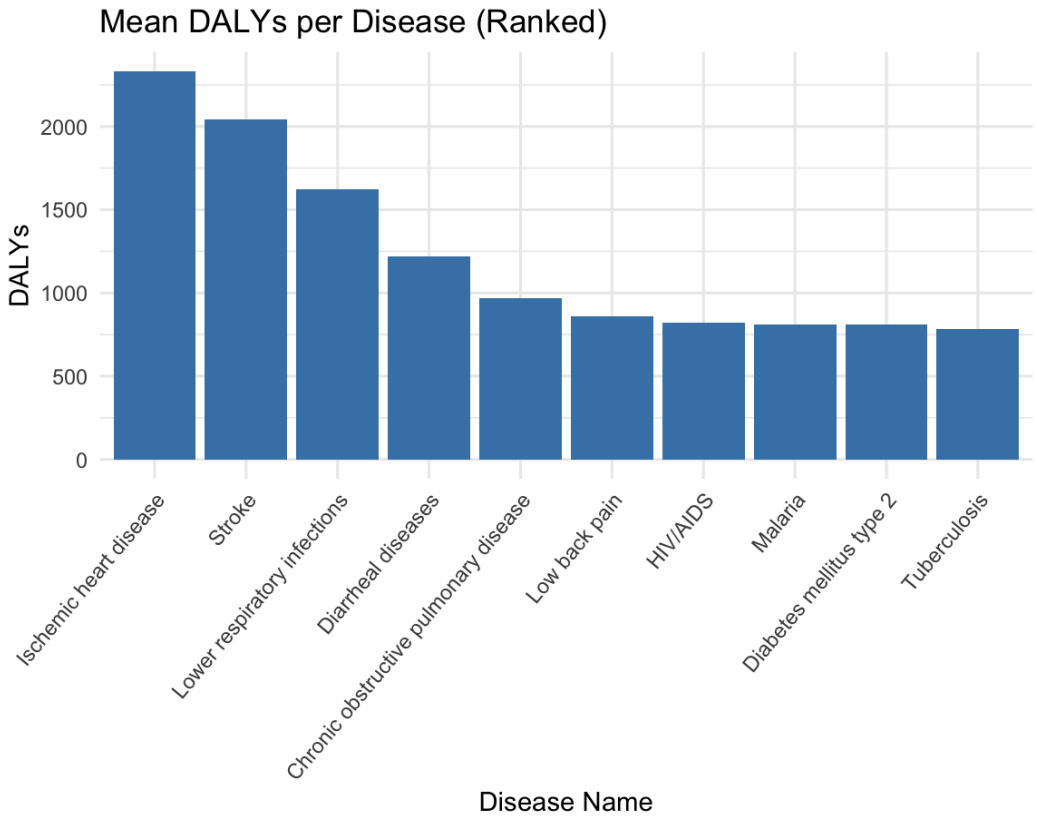

Figure 6

Top 10 Diseases with the Highest Stigma