Critiquing a Dataset

In 2020, the COVID-19 pandemic threw the world into a state of unknown. Hospitals rapidly became overwhelmed and case numbers climbed at alarming rates. Quickly, the need for comprehensive data to understand the virus's transmission patterns and societal impacts became apparent. Recognizing this, the Delphi team at Carnegie Mellon University developed COVIDcast: a large-scale COVID tracking project that involved hundreds of thousands of individuals. By using traditional survey methods and online data sources, COVIDcast created a real-time analysis of the pandemic's effects across the United States. This project offered public health authorities’ valuable insights, leading to better informed decision-making, resource allocation, and communication strategies during a time of unprecedented change.

Specifically, the data used in the project comes from the Delphi group at Carnegie Mellon University. The Delphi group is a research organization specifically tasked with developing a data-driven guide to the COVID-19 pandemic. The purpose of the data collection is to track and analyze the spread of COVID-19 across the United States, providing information such as the rate of the pandemic’s change and its impact on different regions of the nation.

The data collection process used by the group is based on both survey questions and online data sources. These sources include social media platforms, internet search trends, and other publicly available data streams. One external source that the group used is the Facebook COVID-19 Trends and Impact Survey (CTIS) which collected over 100 million responses across over 200 countries. Another source of data was analyzing Google Trends to see how often people were searching for COVID related information, including common COVID symptoms, nearby clinics, and more information on vaccines. Moreover, the project included case report information from John Hopkins University’s Center for Systems Science and Engineering (CSSE) and data from USAFacts, a nonprofit initiative aimed to make government data easily accessible. The approach of using both survey data and online data trends allows for a comprehensive analysis of the pandemic's variables, capturing both official statistics and societal sentiments.

The people represented in the data were only individuals residing anywhere within the United States. The data completely excludes anyone living outside of the United States. However, certain parts of the United States population could be underrepresented due to poor access to technology or internet, language barriers, accessibility of healthcare, and other socioeconomic factors. This means that the data could hold potential biases. One hypothetical scenario that shows potential biases in the data concerns communities with lower socio-economic status. These marginalized areas may have limited access to high-quality personal protective equipment, potentially leading to higher infection rates. Furthermore, the compounded challenges of reduced healthcare services and digital technology, which are essential for reporting symptoms and cases, worsen this scenario. Thus, the data collected from these communities may underrepresent the true extent of COVID's impact, ultimately skewing overall statistics and undermining the pandemic's effects on these specific populations.

As mentioned previously, one key strength of the Delphi group's data is the combination of traditional survey questions and online data sources. This allows for a more informed view of the pandemic's impact, capturing both official statistics and societal sentiments. Another key strength is the use of online data sources themselves. This allows for real-time analysis, instead of having to wait for specific periods of time to analyze trends in survey data. The timeliness of the survey data can then be used in critical decision-making moments.

Though there are many strengths, the data collected by the project also has its limitations. One potential weakness is the reliance on self-reported data, which can be subject to biases and/or inaccuracies. For example, individuals may misremember or misrepresent their symptoms, leading to inaccuracies in the data. Furthermore, one possible symptom of the virus is “brain fog”, a phenomenon that explains difficulties in concentrating and remembering information. Statistically, the largest demographic of patients were the elderly, who already have elevated rates of ailing memory due to natural causes. These factors combined contribute to the idea that there is a plausible non-measurable percentage of survey data that is based on misremembered information. Another weakness that was touched upon previously was that the online data sources used by the project may not be representative of the entire population, as certain demographic groups could have limited access to digital platforms, skewing the data against these parts of the population.

The project itself assumes that the data collected accurately reflects the sentiments of the general population. However, this may not be true, and there are likely some biases in the data. For example, social media data may be biased towards younger age groups, as these are age groups that are more familiar with technology and the ability to report on social media platforms. Furthermore, from an ethical perspective, the data collection process itself has some possible concerns regarding user privacy. While the Delphi group itself has a list of privacy protocols for keeping personal data anonymous, the same may not hold true for the external data sources that the project used.

Another key point of ensuring the collected data is sound is testing the validity of the data against multiple factors. To assess the validity of online search trend data, researchers could correlate the search trends with multiple different variables and ensure that the trends in search data line up with the variable trends. Such variables could include the number of new confirmed cases or patients per day, or the number of patients reporting COVID-like symptoms via the survey. To evaluate the reliability of the survey data, researchers could conversely compare the survey responses with figures from official health records and ensure they are consistent.

This dataset could be useful to many different groups and organizations. Public health authorities can use this data to more accurately allocate resources according to where and when they are needed, as well as prepare such resources in advance. Researchers could use this data to study the spread of the virus and figure out which social trends are effective in reducing the spread. The data could also be used to see whether public awareness campaigns are effective in helping stop the spread of the virus, for example reminding people to use proper protective equipment, and reminding people to staying indoors if they are displaying symptoms.

The use of this dataset could however be potentially problematic as well. For example, going back to the scenario of marginalized communities reporting lower than actual rates of infection; health authorities looking at the data may conclude that these communities do not need any additional support for dealing with the virus. This would in turn worsen the situation, causing a loop of high infection rates and low report rates.

To improve future data collection methods, the Delphi group could consider expanding their outreach to marginalized communities. This could be done via public health campaigns or partnering with community organizations. Additionally, incorporating more diverse data sources, such as alternate methods of reporting cases (including non-electronic methods), could provide a more accurate view of the pandemic's impact.

In conclusion, the Delphi group's COVIDcast project is a strong attempt to analyze the COVID-19 pandemic in a data-oriented way. While the dataset has strengths in combining multiple data capture methods, it is also limited in failing to address potential biases. By attempting to improve on data collection measures and find weaknesses in the uses of the data, researchers can be more representative and ethical when addressing future public health crises.