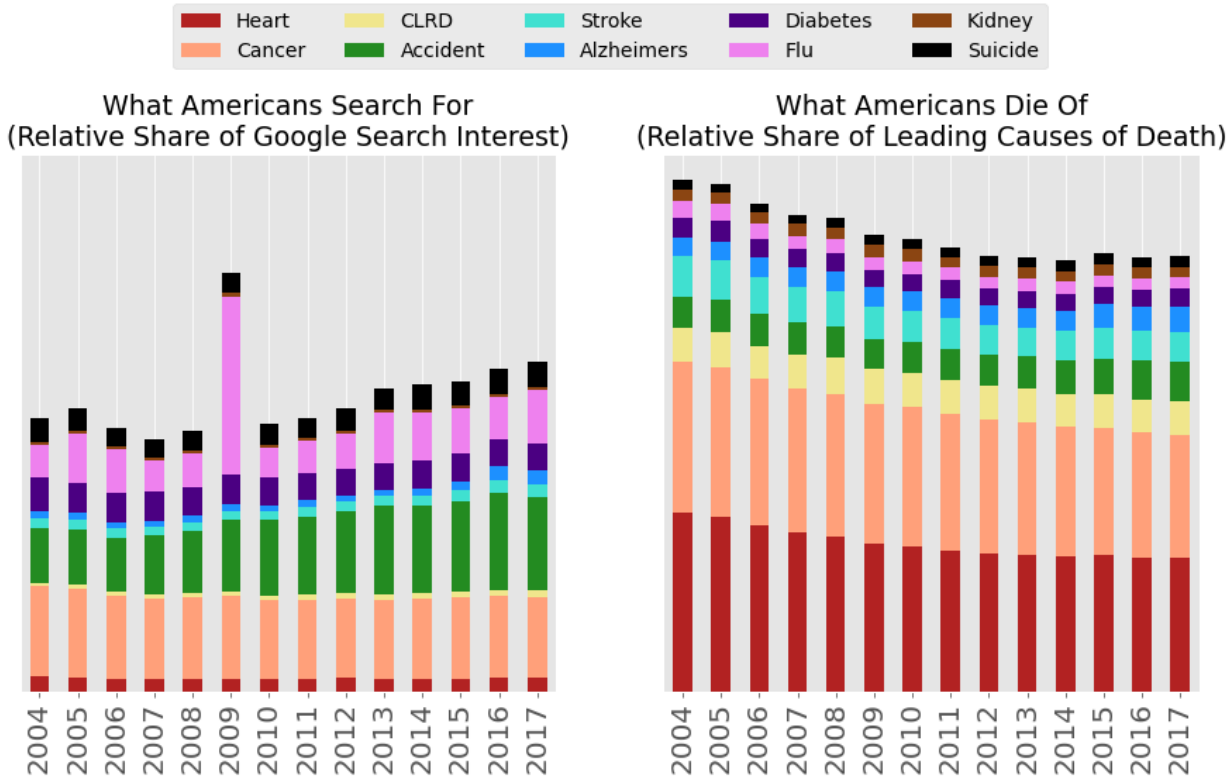
Comparative Analysis of Most-Searched Causes of Death VS Leading Causes of Death (United States)



Findings

This was a fascinating project to engage with due to the sheer number of ways that the data can be visualized and unraveled. My visuals for the purposes of this report however have been limited to the above stacked bar charts depicting the derived relative shares of the respective metrics, and some heatmaps of the United States with a calculated heat score corresponding to the gap that exists between interest and actual mortality for any given one of the leading causes of death.

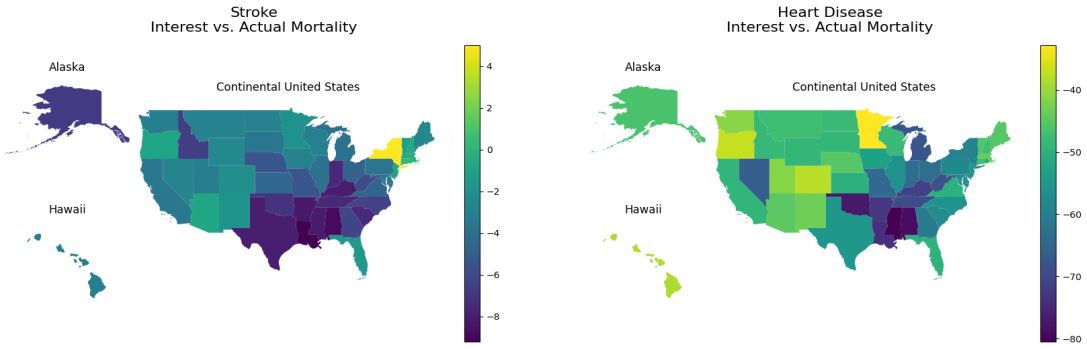
The most obvious and immediate findings visible above include the dramatic spike in flu interest in 2009 that has no correlation with a notable uptick in deaths by flu that year. The Swine Flu Pandemic of this year triggered a media frenzy, sparking concerns throughout the nation that something rivaling the Spanish Flu was on everyone's doorstep. The reality was just over 12,000 deaths - a fraction of the typical

yearly losses to flu, and a tragedy that likely primarily impacted the same vulnerable population for whom the seasonal flu is life-threatening.

More noteworthy in my eyes, however, is the dramatic difference between American concern and research into cardiac disease when compared to the tragic over-representation of that ailment in our obituaries. A lack of education and awareness, coupled with a decades-strong epidemic of obesity and poor cardiovascular health come together to ensure that these numbers will likely continue to dominate the depressing landscape of US fatalities for years to come.

There is a brighter note, however, in the overall downward trends that can be observed. Less people are dying - not just of these leading causes of death, but in general - and whether that is reflective of our entrance into a generational divide that will be corrected with the aging of the Baby Boomer generation or not, it is noteworthy. While the COVID-19 epidemic has claimed over a million lives, it is no secret that healthcare quality is on an upward trajectory overall.

Finally, the heatmaps of the United States are included in this project's repository and convey interesting stories from a geographical perspective. The South, specifically the humid arc of it, sees increased prevalence of deaths by stroke (probably heatstroke) compared to interest, and this is just one of many little stories visible throughout the maps. It would be enlightening to animate these over time, perhaps with a running record of important news events that spread awareness regionally and nationally.



Methodology

Data was acquired from two sources and processed in myriad manners. First, I collected the Center for Disease Control's exhaustive but infrequently-updated dataset on Americans' Leading Causes of Death through the National Center for Healthcare Statistics. Processing included steps to simplify and compress what was, thankfully, an already pretty tidy dataset.

The Google Trends data was slightly more interesting to acquire: Google Trends does *not* provide concrete numbers on the numbers of search engine queries occurring for any given terms/phrases. Rather, it serves up a "score" that comprises the relative

popularity of a term over a length of time between its minimum and maximum. The saving grace lies in the fact that multiple queries can have their results fetched at any given time, so I could simply include the literal control term “Control” in each downloaded dataset. Combining this with terms relevant to each cause of death (i.e. “nephritic disease”, “nephrosis”, and “kidney disease”, or “accidental death”, “homicide”, “unintentional injury”) allowed me to seize a wide swathe of tables containing truly comparable data despite the individually arbitrary nature of Google Trends downloads.

Both of these datasets needed to have their date ranges trimmed, with the Google Trends data only reaching so far back as 2004, and the NCHS table only documenting as recently as 2017. Each dataset contained numbers for the 50 US states and the District of Columbia, and the most significant further processing consisted of several steps to bring the data into a comparable state.

What I settled on was an approach that went about determining the minimum and maximum for each set, marking the leading cause of death on its most pronounced year 100, and doing the same for the most searched term on its best-performing year. From there, numbers cascaded downwards towards relative obscurity, allowing such graphics as those above to take shape.

To tell the story I found most interesting, I specifically decided to exclude the Y-axis from the primary plots, as it served only to confuse with numbers what I see as a compelling example of comparative scales in action. I had wanted to strive for meaningfulness and accessibility in my visualization, and so researched the colors associated with charities and awareness organizations aligned to each illness or incident; This proved not to be viable, though, as everyone seems to love their red and purple ribbons. However, I was able to devise a color palette that remains discernible in grayscale or with the various modes of color-blindness.

Significance

There exists a significant divide between the concerns that contribute to the mental landscape of the American psyche and those that actively threaten their wellbeing. Understanding this divide and working to address it serves many purposes, chief among them being the lessening of anxiety over relatively unlikely occurrences and increasing due diligence with regards to the physical menaces that tragically trim collective millenia every year from our time with our families, passions, and lives.

Project GitHub Repository: https://github.com/QuinnKWolter/InfoViz_CausesOfDeathUS