

Using Text Data Mining to Analyze the NIH's Response to the Coronavirus Pandemic

¹Megan Donegan*, ¹Nancy Praskievicz*, ¹Jacob Scholl, ¹Kirk Baker, ¹Judy, Riggie, ²Sheryl Brining

¹Division of Scientific Categorization and Analysis, Office of the Director

²Office of Research Information Systems, Office of Extramural Research, Office of the Director

As a requirement of the NIH Reform Act in 2006, the National Institutes of Health (NIH) implemented the Research, Condition, and Disease Categorization (RCDC) System to report NIH-funded project listings for various research areas to the public with the inaugural data sets published on the RePORT Categorical Spending page in 2008. The RCDC process involves trans-NIH subject matter experts defining research areas while using a sophisticated thesaurus-based, text mining process to develop the publicly available categorical spending reports each fiscal year (FY). As of FY2019, there are currently 292 publicly available reports that range in topics from Valley Fever to broad research areas like Biotechnology.

Retrospective analysis of RCDC categorical data through FY2008 reveals trends in government-funded research in response to public health areas of interest. In addition to analyzing NIH research projects, the RCDC system can also be applied to data sets other than NIH research projects, like patents and publications. Subsequent analyses of these data reveal the natural progression of research from health crisis onset, response in government funded research, subsequent publications, and then patent awards. Amid the opioid crisis in 2018, RCDC analyzed categorical data retrospectively illustrating an increase in the number of opioid-related awards since FY2008 and demonstrated a notable rise in the number of total awards and funding in FY2018. NIH's response to rising rates of opioid overdoses was evident in the increases to the funding in this research area and is expected to be similar in the center of the COVID-19 pandemic. We intend to apply the same methods and observe a similar rise in the number of coronavirus-related applications in FY2020. We also aim to compare this data to the number of applications and awards for similar outbreaks that have occurred since 2008 including Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in 2013 and Ebola virus (EBOV) in 2014. Furthermore, we will extend the RCDC process to examine research outputs in publication data following these outbreaks to illustrate the link between COVID-19, MERS-CoV, and EBOV and the responsiveness of the NIH in funding scientific research during public health crises.