Methods

The findings in this two-part series are based on a mixed-methods analysis of U.S. news coverage from the first 100 days of 2020 as the coronavirus outbreak in China grew into a pandemic. Data were drawn from 125,696 news articles, published in English, from the websites of 66 mainstream and digital-only U.S. online news outlets available from Media Cloud, an open-source platform at the MIT Center for Civic Media[1] for large-scale media analyses of current events.

Three sets of questions guided this study:

1. How did the coronavirus story develop and spread during the first 100 days of 2020? Which news outlets published the most Covid-19 news stories?
2. How did photos from mainstream news outlets visually represent the story? What visual messaging and rhetorical strategies were employed in photos appearing in the rapidly changing news narrative?
3. How is news experienced in the 21st century? What does the Covid-19 coverage show about the massively distributed and interactive nature of the news ecosystem, and how news consumers are situated within it?

For Report 1, a three-part analysis was conducted between April 2, 2020 and July 31, 2020, that consisted of: (1) a computational analysis of articles from 66 news outlets to determine the "shape" of coronavirus coverage between January 1, 2020 and April 9, 2020 (N=125,696); (2) a follow-up analysis of this dataset to determine which news outlets produced the most news coverage (N=74,737); and (3) a qualitative analysis of selected news stories, published in English, from three waves of coverage.

For Report 2, a three-part analysis was conducted between June 15, 2020 through September 1, 2020, that consisted of: (1) a content analysis of news images (N=532) randomly selected from top 12 U.S. news outlets identified in Report 1 (N=74,737) to determine visual messaging themes; (2) a qualitative analysis of select news images (N=15) for their emotional themes; and (3) a related content analysis of the photos (N=15) and their visual messaging techniques and rhetorical strategies employed.

Defining "news"

Our premise in designing the research in this series was that "news" encompasses a wide spectrum.[2] Prior research PIL conducted about college students on news engagement practices (N = 5,844) confirmed this assumption, since we found the younger cohort defined news differently from the faculty and librarians teaching them.[3]

As researchers, our working definition of news was "events happening all around the world." We relied on broad meanings relating to the spectrum of information in this domain, from hard news to soft news to opinion: Hard news has been defined as "coverage of breaking events involving top leaders, major issues, or significant disruptions in the routines of daily life"; soft news is "typically more sensational, more personality-centered, less time-bound, more practical, and more incident-based than other news";[4] and opinion pieces reflect the author's subjective view about a topic and may have overt social or political implications.

All of these news categories were analyzed for this series, as were news briefs, updates, and opinion pieces from these widely read news outlets. The high volume of misinformation, memes, and false narratives was excluded from our analysis, as far as it could be determined, since they are beyond the scope of our inquiry about mainstream news coverage.[5]

Selecting the news outlet sample

In the first step of our research for Report 1, the analysis examined the peaks and valleys of Covid-19 coverage by using the Explorer tool from Media Cloud. A broad range of news outlets were chosen for our sample based on the following criteria: (1) The news outlet was available from Media Cloud's collection of "Top U.S. newspapers of 2018" or "Top U.S. digital native sources of 2018," as identified by Pew Research Center, and (2) the news outlet was the flagship publication for a media group (e.g., Wall Street Journal), rather than a specialized subsidiary news source (e.g., Market Watch).

Two news sources from Media Cloud's collection of "U.S. mainstream media" – Fox News and CNN – were added to the sample to incorporate news websites published by outlets otherwise known for their television broadcasts. These additions provided a more balanced representation of sources favored by news consumers with certain political orientations, i.e., right- vs. left-leaning.[6] Since coronavirus coverage permeated a broad range of news categories, such as business, health, politics, sports, and entertainment, our sample included publications that specialize in those topics as well as more comprehensive news outlets.

As a final criteria, a news outlet needed to have published more than 300 or more news articles about coronavirus in our search results to be included in our analysis. A complete list of the news outlets (*N*=66) used in the shape of news analysis appears in Table 1.

Table 1 (Methods): News outlet sample

Source	Type	N in sample
90min.com	Digital-only	348
Arkansas Democrat-Gazette	Metropolitan Daily	1,220
The Atlanta Journal Constitution	Metropolitan Daily	906
AZCentral	Metropolitan Daily	3,909
Baltimore Sun	Metropolitan Daily	1,368
bgr.com	Digital-only	567
bleacherreport.com	Digital-only	3,525
The Boston Globe	Metropolitan Daily	1,754
Business Insider	Digital-only	9,625
Bustle	Digital-only	541
Buzzfeed	Digital-only	1,186

chicago.suntimes.com	Metropolitan Daily	1,039
Chicago Tribune	Metropolitan Daily	1,976
cincinnati.com	Metropolitan Daily	1,088
cleveland.com	Metropolitan Daily	836
CNET	Digital-only	1,142
CNN	National	6,870
Columbus Dispatch	Metropolitan Daily	1,748
comicbook.com	Digital-only	1,140
Daily Beast	Digital-only	1,332
Daily News	Digital-only	2,378
Dallas Morning News	Metropolitan Daily	535
The Denver Post	Metropolitan Daily	788
digitaltrends.com	Digital-only	389
Fox News	National	7,861
freep.com	Digital-only	908
Gizmodo	Digital-only	501
hollywoodlife.com	Digital-only	661
HonululuAdvertiser	Metropolitan Daily	1,305
houstonchronicle	Metropolitan Daily	1,742
HuffPost	Digital-only	2,388
Indystar.com	Metropolitan Daily	437
Los Angeles Times	Metropolitan Daily	3,333
Marketwatch	Digital-only	3,980
Mashable!	Digital-only	788
Milwaukee Journal Sentinel	Metropolitan Daily	1,024
mlive.com	Digital-only	1,180
Newsday	Metropolitan Daily	1,724
New York Post	Metropolitan Daily	6,035
New York Times	National	4,791
nj.com	Metropolitan Daily	1,089
Orange County Register	Metropolitan Daily	1,585
Orlando Sentinel	Metropolitan Daily	464
Pittsburgh Post-Gazette	Metropolitan Daily	540
Politico	Digital-only	2,214

Refinery29	Digital-only	365
San Antonio Express News	Metropolitan Daily	349
San Francisco Chronicle	Metropolitan Daily	4,199
San Jose Mercury News	Metropolitan Daily	2,630
Seattle Times	Metropolitan Daily	3,649
Slate.com	Digital-only	491
South Florida Sun-Sentinel	Metropolitan Daily	947
St. Louis Post Dispatch	Metropolitan Daily	1,627
St. Paul Pioneer-Press	Metropolitan Daily	881
Star Tribune	Metropolitan Daily	1,997
Tampa Bay Times	Metropolitan Daily	972
techradar.com	Digital-only	472
theroot.com	Digital-only	315
thisisinsider.com	Digital-only	1,174
TMZ	Digital-only	605
uproxx.com	Digital-only	854
USA Today	National	2,795
Verge	Digital-only	765
Vox	Digital-only	837
Wall Street Journal	National	2,709
Washington Post	National	6,303

Data analysis

For Report 1, the date parameters were January 1, 2020 through April 9, 2020. An iterative process helped to develop the most effective search using the available operators to allow for changes in terminology: "coronavirus" OR "covid" OR "covid19" OR "covid-19" OR (chin* AND pneumonia). Of note, in the first weeks of 2020, news coverage often described the virus as a pneumonia in China or affecting Chinese people, so our search terms allowed for these variations. Search results were cleaned as much as possible to remove multiple records for the same story appearing in the same publication from the net sample.

The Explorer tool on Media Cloud was used to determine the 12 news outlets that produced the most news from the sample of 66 news outlets. Results showed that three-fifths of the news articles came from these publications, thus validating the results that these publications were key players in producing and circulating Covid-19 news.

Another Media Cloud tool, Topic Mapper, was used to examine the structure of the coronavirus news coverage and provide a more in-depth analysis of that coverage than Explorer offers. Based on these results, a heat map was generated of daily Covid-19 coverage from our subsample of top 12 U.S. news outlets. Individual articles from the Topic Mapper results were used for developing the narrative timeline of the coronavirus coverage.

For Report 2, a systematic content analysis of news images was conducted to explore how the coronavirus story was visually represented. A random sample of 532 news images was selected from the 74,737 news articles in the 12 news outlets identified in Report 1. Nearly all of these images were photos (95%); video clips were excluded. This sample was proportionally weighted per wave: Wave 1 (N=21), Wave 2 (N=127), and Wave 3 (N=384).

Content analysis and coding methods were used for analytic reduction and a systematic interpretation of underlying patterns in the news photo sample. Sixteen coding properties were used to analyze all 532 images. These properties were intended to capture the subject, composition, activity, location, ethnicity, age and gender of subjects, and affective messaging evoked by the images. Coders used manifest coding to count the instances of concrete properties in a photo, e.g., a street sign. Latent coding was used for thematic coding, which required coders to make a qualitative and critical interpretation of photos, e.g., *the emotion of fear was evoked by the image.*

Krippendorff's alpha (KALPHA), considered the most rigorous means of testing intercoder reliability, was calculated on results from two pilot test rounds coded by two PIL researchers. KALPHA takes into account chance agreement among content analysis coders. While there is no universally accepted standard for intercoder reliability using Krippendorff's alpha, communications researchers have suggested that a coefficient between 0.81 and 0.99 is "almost perfect," between 0.61 and 0.80 is "substantial," and 0.41 to 0.60 is "moderate." During the second pilot round, the coding practices reached the acceptable reliability level of 0.82.

Methodological limitations

Several issues are associated with secondary analysis of an existing dataset. We took steps to avoid or minimize them. To enhance the reliability of our results, we used average numbers of news stories published each day per news outlet in our shape of news analysis. This sample of 66 sources represented a sampling distribution of all stories published in this time frame.

A confidence interval (95%) was calculated for these results to show a range of values into which the true value fell. An independent samples t-test was conducted on the results used to define organic breaks in the volume of reporting, and provide a basis for the three waves of coverage identified. A time-series analysis was used to extract the moving average in the figure.

A major limitation of our findings are the different objectives between the researchers who conducted the analysis and the researchers who originally built the dataset. As is common in secondary dataset usage, extra processing steps were needed to be taken to reduce duplications, handle updated and missing data, and adjust to changes in naming conventions.

A related issue arose when we discovered that Media Cloud continues to add past stories to its dataset through an RSS feed from news outlets over time. While this procedure helps Media Cloud create a more complete dataset, it affected the total count per publication in our analysis. Accordingly, the dataset a potential researcher may use at a later date will not be identical to the one used in the analysis for this Covid-19 series.[7] On a related note,

news websites often update their breaking news stories and the dates and URLs for all of the stories cited may be subject to change. Every effort to verify and update these addresses was made throughout our project but some URLs may have changed since the series was published.

Another limitation of our series is related to the representation of the 532 news photos analyzed in Report 2. Though this sample of news photos was randomly selected, it only constituted 0.007% of the 74,373 news articles in the dataset. The coding sample was not representative of news photos from the entire study sample, nor should it be regarded as such for this exploratory analysis of visual messaging.

We fully acknowledge the limits in our analysis and problems with the generalizability of the Media Cloud dataset to the larger news ecosystem. Instead of drawing conclusions about the output of *all* news outlets at large, the results are best viewed as being part of an analytical study about how U.S. pandemic mainstream news coverage developed and grew, and how media outlets defined certain narratives through words and visuals.

While further research is required to confirm our findings, the shape of news reported in our series has been validated in research conducted by the Media Cloud Research Team and Taboola's Newsroom Network.[8] As such, the data and results from our computational analyses used in this series are both informed as well as validated by these other research efforts. Together, these results provide a detailed snapshot of U.S. news coverage of Covid-19 and serve as a basis for further inquiry from a variety of disciplines.

Notes

1. See: https://mediacloud.org/.

2. The sample used in these analyses is made up of traditional news outlets, such as *The New York Times*, *Washington Post*, and the *Los Angeles Times* as well as new media "digital-only" sites such as *Business Insider* and *Hu Post.*

3. Alison J. Head, John Wihbey, P. Takis Metaxas, Margy MacMillan, and Dan Cohen (October 16, 2018), "How students engage with news: Five takeaways for educators, journalists, and librarians," Project Information Literacy Research Institute, http://www.projectinfolit.org/uploads/2/7/5/4/27541717/newsreport.pdf pp. 13-17; Alison J. Head, Barbara Fister, and Margy MacMillan, *Information literacy in the age of algorithms: Student experiences with news and information, and the need for change* (15 January 2020), Project Information Research Institute, https://www.projectinfolit.org/uploads/2/7/5/4/27541717/algoreport.pdf

4. Thomas E. Patterson (2000), "Doing well and doing good," Shorenstein Center, Harvard University, https://pdfslide.net/documents/doing-well-and-doing-good-shorenstein-cen-doing-well-and-doing-good-figure-1.html

5. While the subject of Covid-19 misinformation is beyond the scope of our report series, we acknowledge its importance in promoting news literacy and have provided an additional document, "PIL's Covid-19 misinformation resource list," as a supplement to our series.

6. "What news sources are left-leaning, centrist, or right-leaning?" (2014), University of Michigan Library Guides, https://guides.lib.umich.edu/c.php?g=637508&p=4462444 (includes data from Pew Research Center on political polarization)

7. The data Media Cloud provide at any one time about news coverage is subject to minor changes, based on myriad RSS feeds Media Cloud receives from any given news outlet. Accordingly, researchers interested in conducting their own data analysis about Covid news coverage using Media Cloud should pull their own unique dataset, noting the

date of their data download. Media Cloud is an open platform at MIT Media Lab, and written documentation and help from their team of experts is available, see: https://mediacloud.org/getting-started-guide.

8. For related studies, see Fernando Bermejo (March 22, 2020), "Information pandemic: Initial explorations of COVID-19 coverage — Media Cloud," https://mediacloud.org/news/2020/3/22/information-pandemic-initial-explorations-of-covid-19-coverage; Joshua Benton (April 14, 2020), "The coronavirus traffic bump to news sites is pretty much over already," Nieman Lab, Harvard University, https://www.niemanlab.org/2020/04/the-coronavirus-traffic-bump-to-news-sites-is-pretty-much-over-already