

# **Social Media as an Early Proxy for Social Distancing Indicated by the COVID-19 Reproduction Number**

Joseph Younis, Harvy Freitag, Jeremy S Ruthberg, Jonathan P Romanes, Craig Nielsen, Neil Mehta

Submitted to: JMIR Public Health and Surveillance  
on: June 11, 2020

**Disclaimer:** © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 18

0..... 18

Figures ..... 19

Figure 1..... 20

Figure 2..... 21

Figure 3..... 22

Figure 4..... 23

Figure 5..... 24



# Social Media as an Early Proxy for Social Distancing Indicated by the COVID-19 Reproduction Number

Joseph Younis<sup>1\*</sup> BSc; Harvy Freitag<sup>1\*</sup> BSc; Jeremy S Ruthberg<sup>1</sup> BA; Jonathan P Romanes<sup>2</sup> BSc; Craig Nielsen<sup>3</sup> MD; Neil Mehta<sup>3</sup> MBBS, MS, MD

<sup>1</sup>Case Western Reserve School of Medicine Cleveland US

<sup>2</sup>Kansas City University of Medicine and Biosciences Kansas US

<sup>3</sup>Department of Internal Medicine and Geriatrics Cleveland Clinic Foundation Cleveland US

\*these authors contributed equally

## Corresponding Author:

Joseph Younis BSc  
Case Western Reserve School of Medicine  
Health Education Campus  
9501 Euclid Ave  
Cleveland  
US

## Abstract

**Background:** The magnitude and time-course of the COVID-19 epidemic in the United States depends on early interventions to reduce  $R_0$  to below 1. It is imperative, then, to develop methods to actively assess where quarantine measures such as social distancing may be deficient and suppress those potential resurgence nodes as early as possible.

**Objective:** The objective of this study is to determine whether social media is reliable in the assessment of social behavior during an epidemic, using the United States as proof of concept. We ask if social media is an early indicator of public social distancing measures in the USA by investigating its correlation with the time-varying reproduction number ( $R_t$ ) as compared to social mobility estimates reported from Google and Apple Maps.

**Methods:** In this observational study, estimated  $R_t$  was obtained for the period between March 5th and April 5th, 2020 using the EpiEstim package. Social media activity was assessed using queries of “social distancing” or “#socialdistancing” on Google Trends, Instagram and Twitter, with social mobility assessed using Apple and Google Maps data. Cross-correlations were performed between  $R_t$  and social media activity or mobility for the USA. We used Pearson correlations and the coefficient of determination ( $\rho$ ) with significance was set to  $P < 0.05$ .

**Results:** Negative correlations were found between Google search interest for “social distancing” and  $R_t$  in the USA ( $P < .001$ ), and between search interest and state-specific  $R_t$  for 9 states with highest COVID-19 cases ( $P < .001$ ); most states experienced a delay varying between 3-8 days before reaching significance. A negative correlation was seen at a 4-day delay from the start of Instagram hashtag “#socialdistancing” and at 6 days for Twitter ( $P < .001$ ). Significant correlations between  $R_t$  and social media manifest earlier in time compared to social mobility measures from Google and Apple maps, with peaks at -6 and -4. Meanwhile, changes in social mobility correlated best with  $R_t$  at -2 and +1 days for workplace and grocery/pharmacy respectively.

**Conclusions:** Our study demonstrates potential use of Google Trends, Instagram, and Twitter as epidemiological tools in the assessment of social distancing measures in the United States during the COVID-19 pandemic. Their earlier correlation with future social behavior provides actionable insight into whether social distancing efforts will be enacted by the public and where. Social media can be used to help gauge the interest and potential success of this social distancing assisting in the creation of more accurate projections of the epidemic course. Limitations of this study are inherent to the use of Google Trends, Instagram, and Twitter that may represent a biased sample, including only those that are internet-literate and with access to internet, which may covary with socioeconomic status, education, geography and age.

(JMIR Preprints 11/06/2020:21340)

DOI: <https://doi.org/10.2196/preprints.21340>

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>



## Original Manuscript



## Social Media as an Early Proxy for Social Distancing Indicated by the COVID-19 Reproduction Number

Joseph Younis and Harvy Freitag<sup>1</sup>, Jeremy Ruthberg<sup>1</sup>, Jonathan P. Romanes<sup>2</sup>, Craig Nielsen<sup>3</sup>, Neil Mehta<sup>3</sup>

<sup>1</sup>Case Western Reserve University School of Medicine, OH, USA

<sup>2</sup>Kansas City University of Medicine and Biosciences, MO, USA

<sup>3</sup>Department of Internal Medicine and Geriatrics, Cleveland Clinic Foundation, OH, USA

### Corresponding Author:

Joseph Younis, B.S.

Case Western Reserve University School of Medicine

10001 Chester Avenue, 331

Cleveland, OH, 44106, USA

Tel: 630-407-4703

Email: jxy641@case.edu

**Running Title:** Can Social Media Predict Social Distancing Effort?

**Word count:** 2845

### Abstract

**Background:** The magnitude and time-course of the COVID-19 epidemic in the United States depends on early interventions to reduce  $R_0$  to below 1. It is imperative, then, to develop methods to actively assess where quarantine measures such as social distancing may be deficient and suppress those potential resurgence nodes as early as possible.

**Objective:** We ask if social media is an early indicator of public social distancing measures in the USA by investigating its correlation with the time-varying reproduction number ( $R_t$ ) as compared to social mobility estimates reported from Google and Apple Maps.

**Methods:** In this observational study, estimated  $R_t$  was obtained for the period between March 5<sup>th</sup> and April 5<sup>th</sup>, 2020 using the EpiEstim package. Social media activity was assessed using queries of “social distancing” or “#socialdistancing” on Google Trends, Instagram and Twitter, with social mobility assessed using Apple and Google Maps data. Cross-correlations were performed between  $R_t$  and social media activity or mobility for the USA. We used Pearson correlations and the coefficient of determination ( $\rho$ ) with significance set to  $P < 0.05$ .

**Results:** Negative correlations were found between Google search interest for “social distancing” and  $R_t$  in the USA ( $P < .001$ ), and between search interest and state-specific  $R_t$  for 9 states with highest COVID-19 cases ( $P < .001$ ); most states experienced a delay varying between 3-8 days before

reaching significance. A negative correlation was seen at a 4-day delay from the start of Instagram hashtag “#socialdistancing” and at 6 days for Twitter ( $P < .001$ ). Significant correlations between  $R_t$  and social media manifest earlier in time compared to social mobility measures from Google and Apple maps, with peaks at -6 and -4. Meanwhile, changes in social mobility correlated best with  $R_t$  at -2 and +1 days for workplace and grocery/pharmacy respectively.

**Conclusion:** Our study demonstrates potential use of Google Trends, Instagram, and Twitter as epidemiological tools in the assessment of social distancing measures in the United States during the early course of the COVID-19 pandemic. Their correlation, and earlier rise and peak in correlative strength with  $R_t$  when compared to social mobility may provide proactive insight into whether social distancing efforts are sufficiently enacted. Whether this proves valuable in the creation of more accurate assessments of the early epidemic course is uncertain due to limitations. These limitations include the use of a biased sample that are internet-literate with internet access, which may covary with socioeconomic status, education, geography and age, and the use of subtotal social media mentions of social distancing. Future studies should focus on investigating how social media reactions change during the course of the epidemic, as well as the conversion of social media behavior to actual physical behavior.

## Introduction:

Public health measures are the epicenter of global efforts to combat the COVID-19 pandemic [1]. The premise of these measures converges on a central notion: decreasing the  $R_0$  of the novel coronavirus below 1 to suppress transmission. With a  $R_0$  value below 1, the virus can no longer sustainably propagate from one person to another, eventually halting its spread [2]. The most championed of these efforts is the idea of “social distancing”, or the practice of distancing yourself from others in order to reduce respiratory droplet transmission, the primary mode of transmission for COVID-19 [3]. However, social distancing has not been inconsequential, with primary concern to socioeconomic health. Several macroeconomic reports exploring the supply and demand shock of COVID-19 describe that its effects may rival that of the 1918 Spanish Flu and the Great Depression [4].

Transmission of COVID-19 was first detected in the United States on February 2020, and by mid-March, all 50 states and four U.S. territories had reported cases of COVID-19 [5]. The total number of confirmed cases continued to rise exponentially before this trend was broken in early April. In an effort to slow transmission, several states implemented strict lockdowns, curfews and business restrictions [6]. New York Governor Cuomo declared a state of emergency on March 7th, and New York City implemented one of the first large-scale lockdowns of schools, temples and other large gathering places in Rochelle. This further extended to include stay-at-home orders in other areas of New York, California and Illinois. Restrictions on businesses deemed non-essential were eventually implemented in more than 40 states [6].

In response, dramatic drops across several economic sectors have been witnessed leading to financial strain on American households. An unprecedented over 10 million unemployment claims were filed in the 2 weeks ending on March 28<sup>th</sup> [7]; for reference, the previous peak was at 695,000 claims in October 1982. Despite national-level interventions, such as mandated paid time off and a historic \$2 trillion stimulus package (CARES Act) to mitigate the broad impact of COVID-19 [8].

Early intervention is ideal for the mitigation of a pandemic’s socioeconomic and health costs, but such potential is often a post-hoc discovery. A more practical approach is active scrutiny and revision of the implemented measures, ideally in the early phases of the pandemic’s course [9]. Recent efforts have attempted to quantify social distancing efforts using Google or Apple maps’ user activity [10,11]. While these tools accurately reflect social behavior at a point in time, we hypothesize that Google Trends and social media yield earlier actionable insight that can help control the pandemic’s

trajectory.

Google Trends, and social media (Instagram and Twitter) are used extensively in the scientific literature and have been validated against external reference data sets in numerous public health and health surveillance studies [12-17]. With an estimated 35% and 27% of all US Citizens using Instagram and Twitter on a regular basis, and 89.7% of digital users searching on Google, these avenues remain the most practical tools for study [18-20]. Likewise, studies during this pandemic are investigating the utility of social media in the dissemination of preventive health information [21-23] while Twitter recently provided full access for prospective social media data tracking for COVID-19 research. Despite this, their use as epidemiological tools in the assessment of social behavior in early epidemic courses remains to be determined.

In this study, we investigate the use of Google Trends, Instagram, and Twitter as tools for the evaluation of social distancing measures by the public in the early epidemic phase. We first highlight a correlation between social distancing measures as captured by social media and national and state specific  $R_t$ , an epidemiological estimate of  $R_0$  throughout an epidemic. We then compare the correlation of these social media avenues with  $R_t$  to the correlation of Google and Apple maps' user activity with  $R_t$ . We focus on the top 9 affected states from the time of writing, April 10<sup>th</sup>, 2020. We collect the most recent social media data using Google Trends, Twitter and Instagram and use the updated confirmed cases compiled by the Centers of Disease Control COVID-19 Case Data and John Hopkins Coronavirus Resource Center [24,25].

## **Methods:**

### ***Database inclusion***

We utilized Google Trends, Instagram and Twitter. In addition to their established use in the scientific literature, we also focus on Instagram and Twitter because their demographic overlaps significantly with the public-facing jobs [18-20] most likely to be affected by social distancing. Furthermore, a poll conducted by Morning Consult between March 27<sup>th</sup> and 30<sup>th</sup> reported 88% of Americans between the ages of 30-54 are practicing social distancing to some extent [20,26], an age range closely resembling Instagram and Twitter's median ages of 34 and 40, respectively.

The choice to include only the top 9 states by COVID-19 incidence was made because lower incidence states yielded insufficient social media and incidence data. When the analysis was run on the bottom 9 states by COVID-19 incidence, the results displayed erratic patterns of social distancing search interest with no clear peak and days with no data suggesting low search volumes; additionally, the reproduction number ( $R_t$ ) displayed large error margins and could not be calculated continuously over the study period.

### ***Google Trends***

Google Trends records billions of data points from search terms entered by the general public. It then compares the summative search volume of each search query (defined as the exact term entered into Google's search bar) to the day of highest search volume to yield a search volume index (SVI) score of 1-100. SVI is assigned to each day and represents that day's relative search frequency. Google Trends contains a geo-filtering feature to that allows search data from within the United States or, to be more granular, from specific states.

Google Trends data for the search query "Social Distancing" was collected on April 10<sup>th</sup>, 2020, for March 1<sup>st</sup>, 2020, through April 10<sup>th</sup>, 2020.

### ***Instagram and Twitter***

Instagram and Twitter are social networking platforms that can be accessed on a phone application or internet website. As of 2018, there are 107 million Instagram users in the United States. Similarly, as of January 2020, 59 million Twitter users are American, comprising the largest percentage of Twitter's user base together, these social networking services capture a large percentage of the American population [20,26].

Unamo.com search algorithms were used to capture the historical frequency of mentions for the



hashtag “#socialdistancing” in the United States on Twitter and Instagram between March 1<sup>st</sup> and April 10<sup>th</sup> [27].

### **Calculation of $R_0$**

$R_0$  is the number of individuals infected by a single infected individual during his or her entire infectious period, in a population which is entirely susceptible.

$$R_0 = \frac{\kappa\beta\lambda}{(\mu+\kappa)(\mu+\gamma)\mu}$$

Where  $\kappa$  is the rate at which an exposed individual becomes infectious,  $\beta$  is the probability that a susceptible individual becomes infected upon interaction with an infected individual,  $\lambda$  is birth rate of susceptible individuals,  $\mu$  is the per capita natural death rate, and  $\gamma$  is the per capita recovery rate.

The  $R_0$  for COVID-19 has varied in value from 1.4 to as high as 11.1 reported from some communities in China and Singapore [28-29]

The time varying reproduction number ( $R_t$ ) is an epidemiological estimate of  $R_0$  calculated using two variables, (1) the daily incidence of onset of acute respiratory illness, and (2) the distribution of the serial interval (time interval between symptoms onset in a case and in their infector).

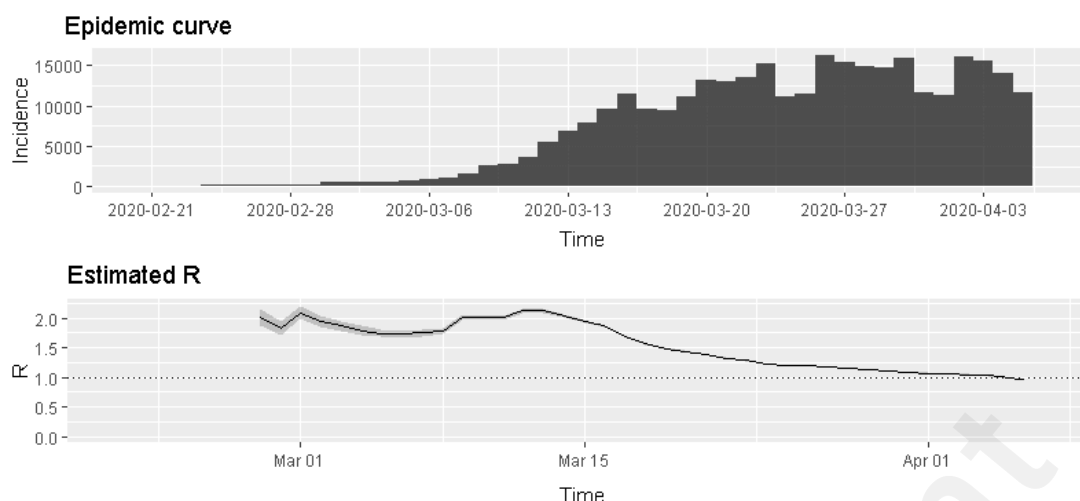
The daily incidence of COVID-19 in the USA was obtained from estimations of symptom onset provided by the CDC COVID-19 Case Data, which contains data up to April 5<sup>th</sup> [24]. The state-wide incidence rate is based on confirmed cases obtained from the John Hopkins Coronavirus Resource Center. The serial interval was obtained using available parametric data computed previously for the initial outbreak of COVID-19 [30].

We used the R statistical software along with the EpiEstim package to calculate the  $R_t$  using the above parameters for the period of March 5<sup>th</sup> to April 5<sup>th</sup>.  $R_t$  for the USA and top 9 states by confirmed COVID-19 cases was derived for this time period. For subsequent calculations, we included data after the onset of at least 100 confirmed cases in each state, as the  $R_t$  prior to that had standard deviations in excess of 0.5.

Google Trends SVI for “social distancing” was then independently compared to  $R_t$  for the top 9 affected states (NY, CA, PA, MA, NJ, FL, LA, MI, IL) and the USA as a whole. Analyses were performed using Pearson correlations with significance set to  $\alpha < 0.05$  then plotted on logarithmic graphs. Correlations were obtained using raw data and after varying periods of time delay between Google SVI or social media mentions and change in  $R_t$ .

Cross-correlations for the relationship between  $R_t$  and measures of social distancing in the USA were performed, using available data based on Google maps tracking that measures changes in percent social mobility. This data was available for separate locations, including grocery and pharmacy stores, recreation and retail stores, and workplaces. In addition, the cross-correlations between  $R_t$  and “#socialdistancing” mentions on Instagram and Twitter were also performed. The coefficient of determination ( $\rho^2$ ) was calculated and graphed, which represents the strength of the correlation at different time delays between  $R_t$  and each of the social mobility and social media measures. The peak of the coefficient of determination for each of these measures are tabulated along with the delay for which the greatest strength of relationship was found.

### **Results:**



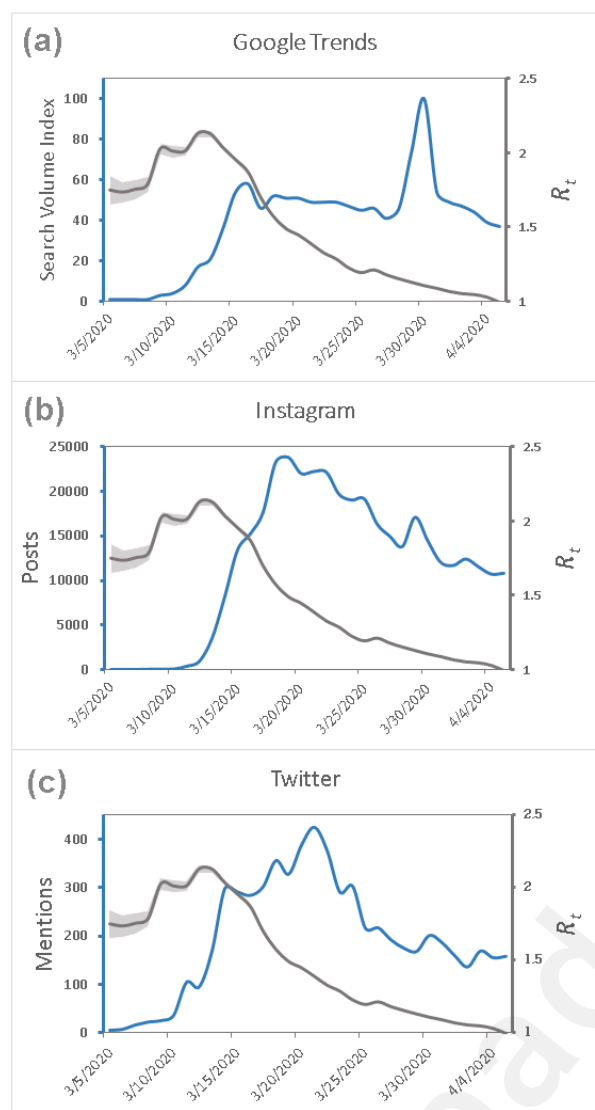
**Figure 1. Estimated  $R_t$  Based on Time of Symptom Onset for Cases of COVID-19 in the USA**

The time-varying reproduction number,  $R_t$ , represents the mean number of secondary cases generated by a primary case over a sliding weekly window.

The estimated  $R_t$  is shown for the period of February 28<sup>th</sup> to April 5<sup>th</sup> calculated from the number of COVID-19 cases by symptom onset, with a mean serial interval of 3.96 days and standard deviation of 4.75 days. The shaded error bands are equal to one standard deviation of the estimated  $R_t$  for each date.

Significant negative correlations were found between Google SVI for the search query “social distancing” and the time-varying reproduction number ( $R_t$ ) between the dates of March 5<sup>th</sup> and April 5<sup>th</sup> in the USA ( $P < .001$ ). The relationship between estimated  $R_t$  and Google SVI can be visualized graphically in Figure 2a. The strength of the correlation reached a peak at 4 days delay from the start of the searches when considering all cases in the USA, with a Pearson Correlation coefficient of 0.72 ( $P < .001$ ).

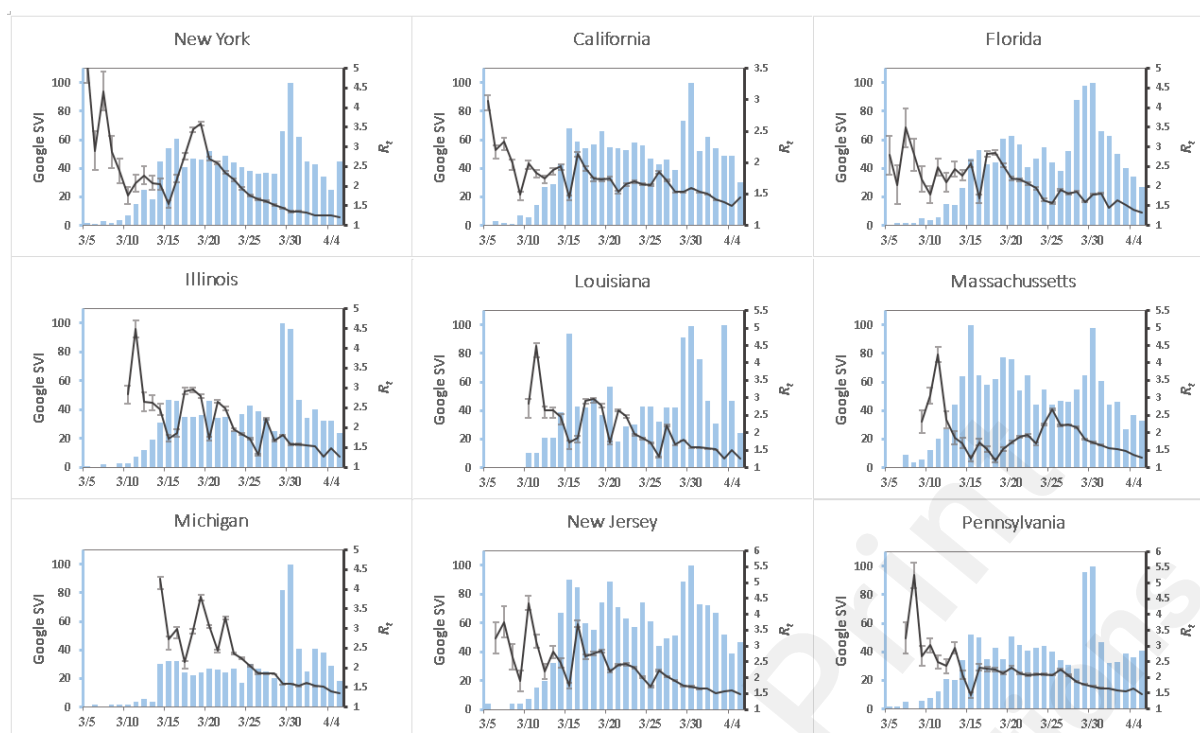
There was a total of 376,067 “#socialdistancing” mentions on Instagram and 6,470 on Twitter in the studied time period. The increase in “#socialdistancing” mentions on Twitter and Instagram predate the appearance of a decrease in  $R_t$  seen in Figure 2 (b) and (c). The relationship between  $R_t$  and Instagram mentions (Figure 2c) is significant, and strongest at a 4-day delay ( $P < .001$ ) from the start of Instagram hashtag “#socialdistancing” mentions. Significance for Twitter is seen only at a 6-day delay ( $P < .001$ ).



**Figure 2. Relationships Between Social Media Activity for “Social Distancing” and Estimated  $R_t$  for the USA and New York**

On the left shows the relationship between Google Trends Search Volume Index (SVI) for “social distancing” and estimated  $R_t$  for the (a) USA and (b) New York over the time period from March 5<sup>th</sup> to April 5<sup>th</sup>. Light blue line refers to Google Search Index (SVI) for the query “social distancing.” Dark grey line refers to the estimated  $R_t$ . On the right shows the relationship between estimated  $R_t$  and total social media mentions for “#socialdistancing” in the USA on (c) Instagram and (d) Twitter over the same time period. Error bands shown in light grey represent one standard deviation from the mean  $R_t$  calculated at each date.

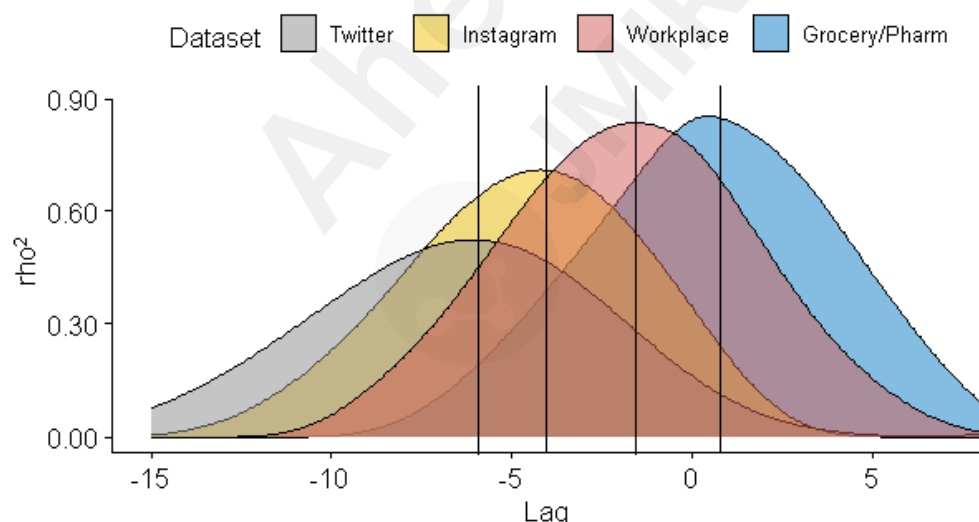
When evaluated by state, New York, New Jersey, Massachusetts, Michigan, Pennsylvania, California, Louisiana, Illinois, and Florida all showed significant negative correlations between “social distancing” SVI and the state-specific  $R_t$  ( $P < .005$ , refer to Figure 3). These correlations reached peak significance at different delay periods.  $R_t$  for some states, such as Massachusetts, experienced an early correlation with increasing searches for “social distancing”. Other states such as New York and Louisiana experienced a larger time delay from the start of Google searches to a decrease in  $R_t$ , at 6 and 8 days respectively. Most states experienced a delay varying between 3-8 days before reaching peak significance.



**Figure 3. State-wide Relationships Between Estimated  $R_t$  and Google SVI for “Social Distancing”**

The estimated  $R_t$  was calculated individually for the Top 9 states with the most confirmed COVID-19 cases on April 5<sup>th</sup>. They are graphed for the period of March 5<sup>th</sup> to April 5<sup>th</sup> for each state (dark line) along with the Google search index for “social distancing” in each state for the same time period (light blue bars). Error bands represent one standard deviation from the mean estimated  $R_t$  at each date. The data for New York can be found in Figure 2.

Significant correlations between  $R_t$  and social media appear to manifest themselves earlier in time when compared to social mobility measures, with peaks at -6 and -4 days for the relationship between  $R_t$  and Twitter and Instagram mentions respectively ( $P < .001$ , refer to Figure 4). Social mobility correlated best with  $R_t$  at -2 and +1 days for workplace and grocery/pharmacy, respectively.



**Figure 4. Cross-Correlations Between Changes in Social Mobility or Social Media Mentions and Estimated  $R_t$**

Cross-correlations (represented by  $\rho^2$ , the coefficient of determination) between estimated  $R_t$  and social mobility changes and social media mentions in the USA. The black lines represent the peak correlation. Social mobility changes include traffic at grocery, pharmacy, and workplace locations based on Google Maps tracking. Social media measures include mentions of “#socialdistancing” on Instagram and Twitter.

The relationship between  $R_t$  and social media or social mobility ( $P < .001$ ) reaches its strongest point

at different delay periods, tabulated below. The increase in social media mentions predate the decrease in  $R_t$  the earliest, with a lag-time of 4 to 6 days. Social mobility data also predate the decrease in  $R_t$  although at later times of 0 to 3 days. The table also shows the strength of correlation between each of the measures and  $R_t$ , represented by  $\rho^2$ . The strongest correlations are between social mobility data and  $R_t$  with comparatively lower correlations between social media and  $R_t$ . Google trends, however, shows a comparable  $\rho^2$  with data from Apple maps, but not Google maps, which exhibits the strongest correlations all domains except parks.

**Table 1. Peak Correlations Between Social Media and Social Mobility Measures and Associated Time-Delay**

	Dataset	$\rho^2$	Lag (days)
Social Media	Instagram	0.68	-4
	Twitter	0.47	-6
	Google Trends	0.72	-4
Apple Maps	Driving	0.75	-3
	Transit	0.80	-3
	Walking	0.73	-2
Google Maps	Grocery/Pharmacy	0.89	+1
	Transit	0.83	-2
	Workplace	0.86	-2
	Parks	0.66	-2
	Recreation	0.84	0
	Residential	0.85	-2

### Discussion:

In our study, we found that increased social distancing mentions on social media correlated with reduced USA  $R_t$ , with Google Trends correlating with reduced state-specific  $R_t$  as well. We also found that the correlation varied when social distancing mentions or search queries are lagged by a few days, this effect depended on the state and social media platform. The delay to reach peak strength discrepancy between Instagram and Twitter is interesting because the reach of Instagram in the United States is much greater, indicating possible time-sensitive influence on behavior imparted by user reach. Why the delay periods differed between states is unclear but may be partly explained by the unequal implementation of top-down public health interventions.

Instagram and Twitter mention of “#socialdistancing” correlated earlier with reduced COVID-19  $R_t$  in the USA than did social mobility measures from Google and Apple maps. Interestingly, Twitter showed the earliest correlation with  $R_t$  but also has the lowest coefficient of determination. This finding may be explained by the fact that Twitter reaches the smallest user base compared to Instagram or Google Maps. Social media in general exhibited a weaker correlation with  $R_t$ . This is expected since social mobility measures directly relate to the density of people congregating in an area, whereas social media is an indirect measure of social distancing and likely represent a smaller proportion of the population. Nonetheless, these findings confirm our hypothesis that social media may serve as earlier indicators of future social behavior.

The idea that lagging social distancing efforts as captured by social media produces significant reductions in  $R_t$  implicates a predictive role for social media. This is consistent with the interpretation that Google Trends, Instagram, and Twitter model the dissemination of information that may lead to individual decisions to undergo social distancing. While the strength of the correlation for social media was found to be weaker than that for social mobility, the value is in the relationship of the correlation to time. Furthermore, the strength of the correlation may improve with subsequent studies using more accurate measures of social distancing in the media to actual social distancing behavior in the public.

An additional interpretation for the significance found in lagging social media mentions is that the delayed drop in  $R_t$  is also consistent with the expectation that social distancing is a method of primary prevention, as early practice prevents a future increase in  $R_t$ . It is tempting to consider whether these effects depend on the incubation period for COVID-19. About 50% of infected individuals show symptoms by 5 days, and 97.5% by 12 days [2,31]. Our study shows that all nine states exhibit a significantly reduced  $R_t$  with an 8-day lag period for social distancing search interest, supporting a quarantine timeframe that confidently covers the upper limit of the incubation period. On the contrary, quarantine times closer to the median incubation period of COVID-19 may be insufficient as only 30% of the states showed significant reductions in  $R_t$  when the lag period was shorter than median incubation. A parallel can be drawn from these findings, albeit speculatively: there may also be a threshold in this pandemic's trajectory in the United States before which a termination of social distancing efforts may be too early.

### Limitations:

Whether this proves valuable in the creation of more accurate assessments of the early epidemic course is uncertain due to limitations. Limitations of this study are inherent to the use of Google Trends, Instagram, and Twitter because they are presumably indirect measures of public behavior. The data represents a subtotal amount of mentions on Instagram and Twitter, and the study period is short and during the early course of the epidemic where testing and reporting COVID-19 was imperfect. Additionally, we focused on only the top 9 states by incidence; while this was in effort to reduce false positive findings from unreliable low-incidence states, it does introduce barriers to generalizing results to other states. Furthermore, social media may represent a biased sample of those that are internet-literate and with access to internet, which may effectively covary with socioeconomic status, education, geography and age.

### Conclusion:

Our study demonstrates the utility of Google Trends, Instagram, and Twitter as epidemiological tools in the assessment of social distancing measures in the United States during the early course of the COVID-19 pandemic. Their correlation, and earlier rise and peak in correlative strength with  $R_t$  when compared to social mobility may provide proactive insight into whether social distancing efforts are sufficiently enacted. Whether these findings translate to the hypothesized clinical value is uncertain due to limitations. While social media remains a candidate to gauge the success of this containment measure in the early epidemic period, future studies should need to investigate how social media reactions change during the course of the epidemic and whether these correlation patterns with  $R_t$  persist.

### References:

1. Pan A, Liu L, Wang C, et al. Association of Public Health Interventions with the Epidemiology of the COVID-19 Outbreak in Wuhan, China. JAMA. 2020;323(19):1915–1923. DOI: 10.1001/jama.2020.6130
2. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus, Journal of Travel Medicine 2020;27(2). DOI: 10.1093/jtm/taaa021
3. Delen D, Eryarsoy E, Davazdahemami B. No Place Like Home: Cross-National Data Analysis of the Efficacy of Social Distancing During the COVID-19 Pandemic. JMIR Public

- Health Surveill 2020;6(2):e19862. DOI: 10.2196/19862
4. Athreya, KB, Mather, M, Mustre-del-Río J, Sanchez, JM. COVID-19 and Households' Financial Distress. Part 2: The Spread of COVID-19 and (Financial) Pre-Existing Conditions. [https://www.richmondfed.org/publications/research/coronavirus/economic\\_impact\\_covid-19\\_03-30-30](https://www.richmondfed.org/publications/research/coronavirus/economic_impact_covid-19_03-30-30). Published March 2020. Accessed April 5, 2020
  5. Bialek S, Bowen V, Chow N et al. Geographic Differences in COVID-19 Cases, Deaths, and Incidence — United States, February 12–April 7, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:465–471. DOI: 10.15585/mmwr.mm6915e4
  6. Lee, JC, Mervosh S, Avila Y, Harvey B, Matthews, AL. See How All 50 States Are Reopening (and Closing Again). <https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html>. Published June 2020. Accessed August 2020.
  7. Culter D. How Will COVID-19 Affect the Health Care Economy? JAMA Network. JAMA Health Forum: COVID-19. <https://jamanetwork.com/channels/health-forum/fullarticle/2764547>. Published April 9, 2020. Accessed April 27 2020
  8. HR 748—CARES Act. 116th Congress, 2020. <https://www.congress.gov/bill/116th-congress/house-bill/74>
  9. Walensky RP, del Rio C. From Mitigation to Containment of the COVID-19 Pandemic: Putting the SARS-CoV-2 Genie Back in the Bottle. *JAMA* DOI: 10.1001/jama.2020.6572
  10. COVID-19 Mobility Trends Reports. Apple. <https://www.apple.com/covid19/mobility>. Accessed April 15, 2020.
  11. COVID-19 Community Mobility Reports. Google. <https://www.google.com/covid19/mobility>. Accessed April 15, 2020.
  12. Park JH, Christman MP, Linos E, Rieder EA. Dermatology on Instagram: An Analysis of Hashtags. *J Drugs Dermatol* 2018;17(4):482-484. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6831082/>. Published April 2018. Accessed April 12, 2020
  13. Dorfman RG, Vaca EE, Mahmood E, Fine NA, Schierle CF. Plastic Surgery-Related Hashtag Utilization on Instagram: Implications for Education and Marketing. *Aesthet Surg J* 2018;38(3):332-338. DOI: 10.1093/asj/sjx120
  14. Sinnenberg L, Buttenhein AM, Merchant RM. Twitter as a Tool for Health Research: A Systematic Review. *Am J Public Health* 2017;107(1): e1–e8. DOI: 10.2105/AJPH.2016.303512
  15. Bloom R, Amber KT, Hu S, Kirsner R. Google search trends and skin cancer: Evaluating the US population's interest in skin cancer and its association with melanoma outcomes. *JAMA Dermatol*. 2015;151(8):903–905. DOI: 10.1001/jamadermatol.2015.1216
  16. Solano P, Ustulin M, Pizzorno E, et al. A Google-based approach for monitoring suicide risk. *Psychiatry Res* 2016;246:581–586. DOI: 10.1016/j.psychres.2016.10.030
  17. Moccia M, Palladino R, Falco A, Saccà F, Lanzillo R, Brescia Morra V. Google Trends: New evidence for seasonality of multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2016;87:1028–1029. DOI: 10.1136/jnnp-2016-313260
  18. Lu Y, Zhang L. Social Media WeChat infers the development trend of COVID-19. *Journal of Infection* 2020. DOI: 10.1016/j.jinf.2020.03.050
  19. Wojcik S, Hughes A. Sizing Up Twitter Users. Pew Research Center. Internet & Technology.

- <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>. Published April 2019. Accessed April 12, 2020
20. Clement J. Percentage of U.S. adults who use Instagram as of February 2019, by age group. Internet: Social Media & User-Generated Content. Statista.com. <https://www.statista.com/statistics/246199/share-of-us-internet-users-who-use-instagram-by-age-group/>. Published August 2019. Accessed April 15 2020
  21. Simonsen L, Gog JR, Olson D, Viboud C. Infectious disease surveillance in the big data era: Towards faster and locally relevant systems. *J Infect Dis* 2016;214(Suppl 4):S380–S385. DOI: 10.1093/infdis/jiw376
  22. Merchant RM, Lurie N. Social Media and Emergency Preparedness in Response to Novel Coronavirus. *JAMA*. DOI: 10.1001/jama.2020.4469
  23. Al-Dmour H, Masa'deh R, Salman A, Abuhashesh M, Al-Dmour R. Influence of Social Media Platforms on Public Health Protection Against the COVID-19 Pandemic via the Mediating Effects of Public Health Awareness and Behavioral Changes: Integrated Model. *J Med Internet Res* 2020;22(8):e19996. DOI: 10.2196/19996
  24. Previous U.S. COVID-19 Case Data to April 15 2020. National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases. CDC.gov. <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/previouscases.html>. Page last reviewed: April 1, 2020. Accessed April 22 2020
  25. Gardner L. Public Health: Mapping 2019-nCoV. Johns Hopkins Center for Systems Science and Engineering. <https://systems.jhu.edu/research/public-health/ncov/>. Published January 23, 2020. Accessed April 26 2020
  26. Clement J. Worldwide desktop market share of leading search engines from January 2010 to January 2020. Internet: Search Engine and SEO. Statista.com. <https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/>. Published March 2020. Accessed April 15 2020
  27. SEO and Social Media analytical tools for online marketing professionals. Unamo. <https://unamo.com/>. Accessed April 15 2020
  28. Fox S, Rainie L. The Web at 25 in the U.S. Pew Research Center. Internet & Technology. <http://www.pewinternet.org/2014/02/27/the-web-at-25-in-the-u-s/>. Published Feb 2014. Accessed April 12, 2020
  29. Stepleton K,. Unemployed Insurance Weekly Claims. Release Number: USDL 20-691-NAT. U.S. Department of Labor. <https://www.dol.gov/ui/data.pdf>. Published April 23 2020. Accessed April 28 2020
  30. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Ancel Meyers L. Serial interval of COVID-19 among publicly reported confirmed cases. *Emerg Infect Dis* 2020;26(6). DOI: 10.3201/eid2606.200357
  31. Lauer SA, Grantz KH, Bi Q, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application [published online ahead of print, 2020 Mar 10]. *Ann Intern Med* 2020;M20-0504. DOI:10.7326/M20-0504



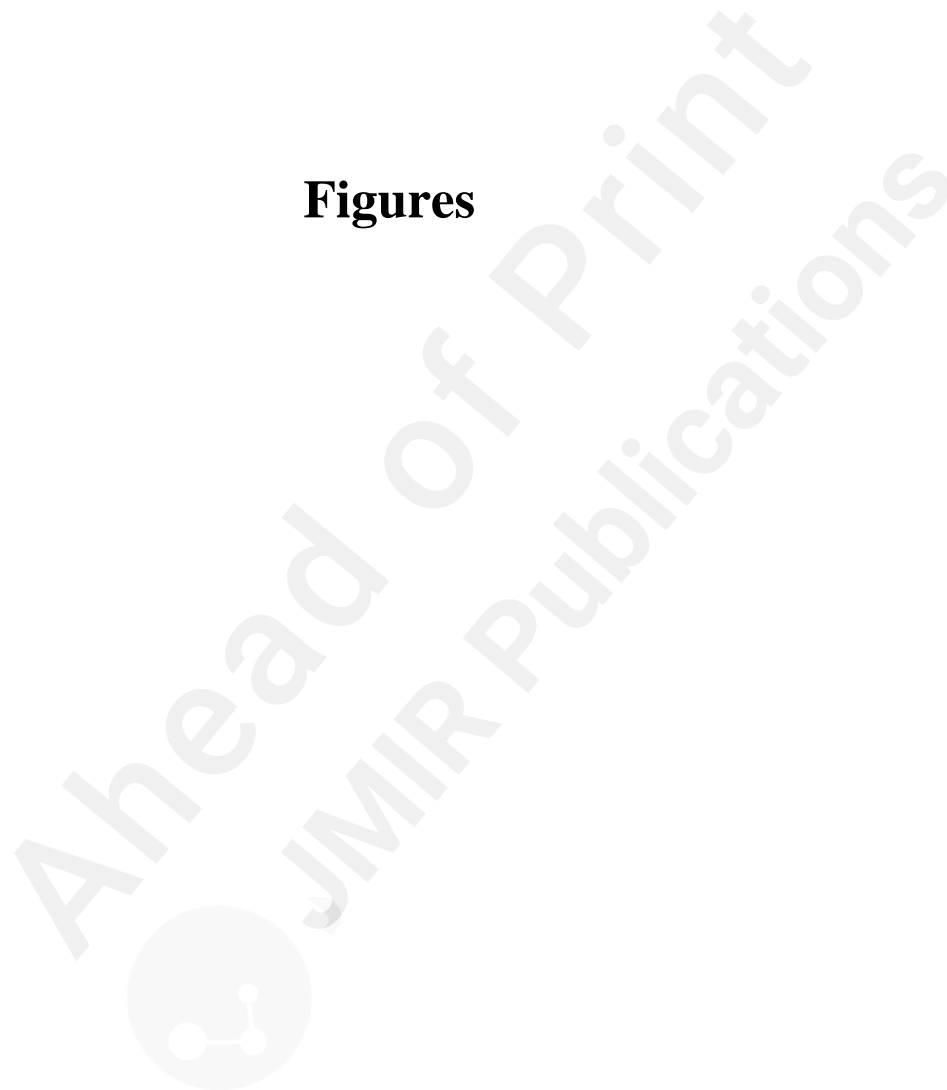


## Supplementary Files

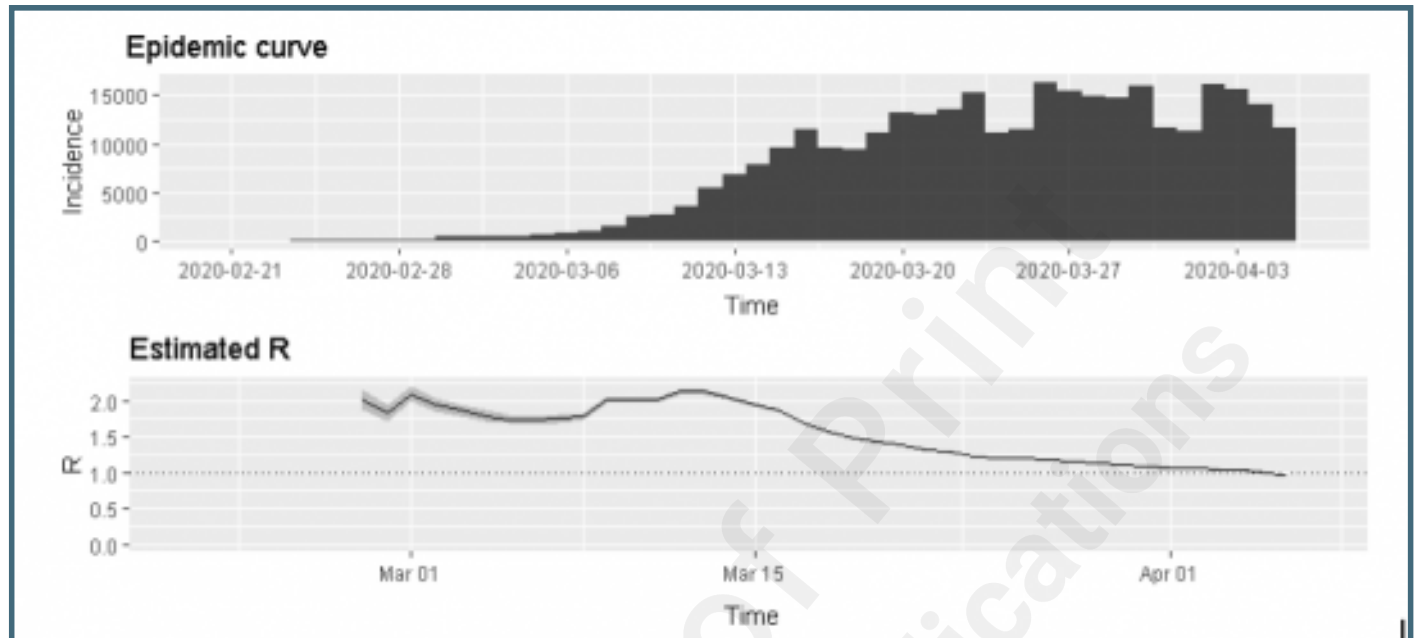
Untitled.

URL: <https://asset.jmir.pub/assets/905d19bc7d1031a9b228cb011d8488d9.docx>

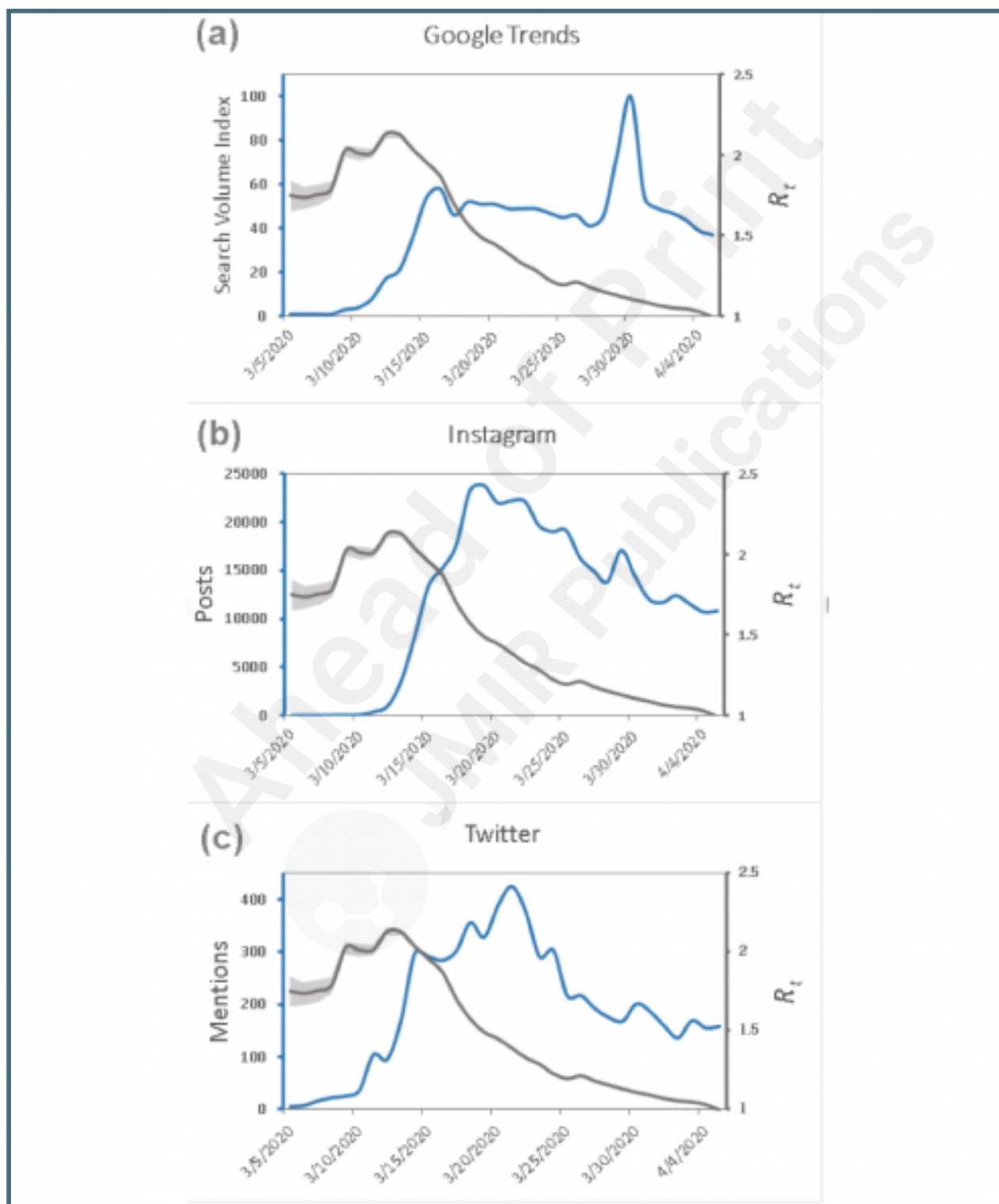
## Figures



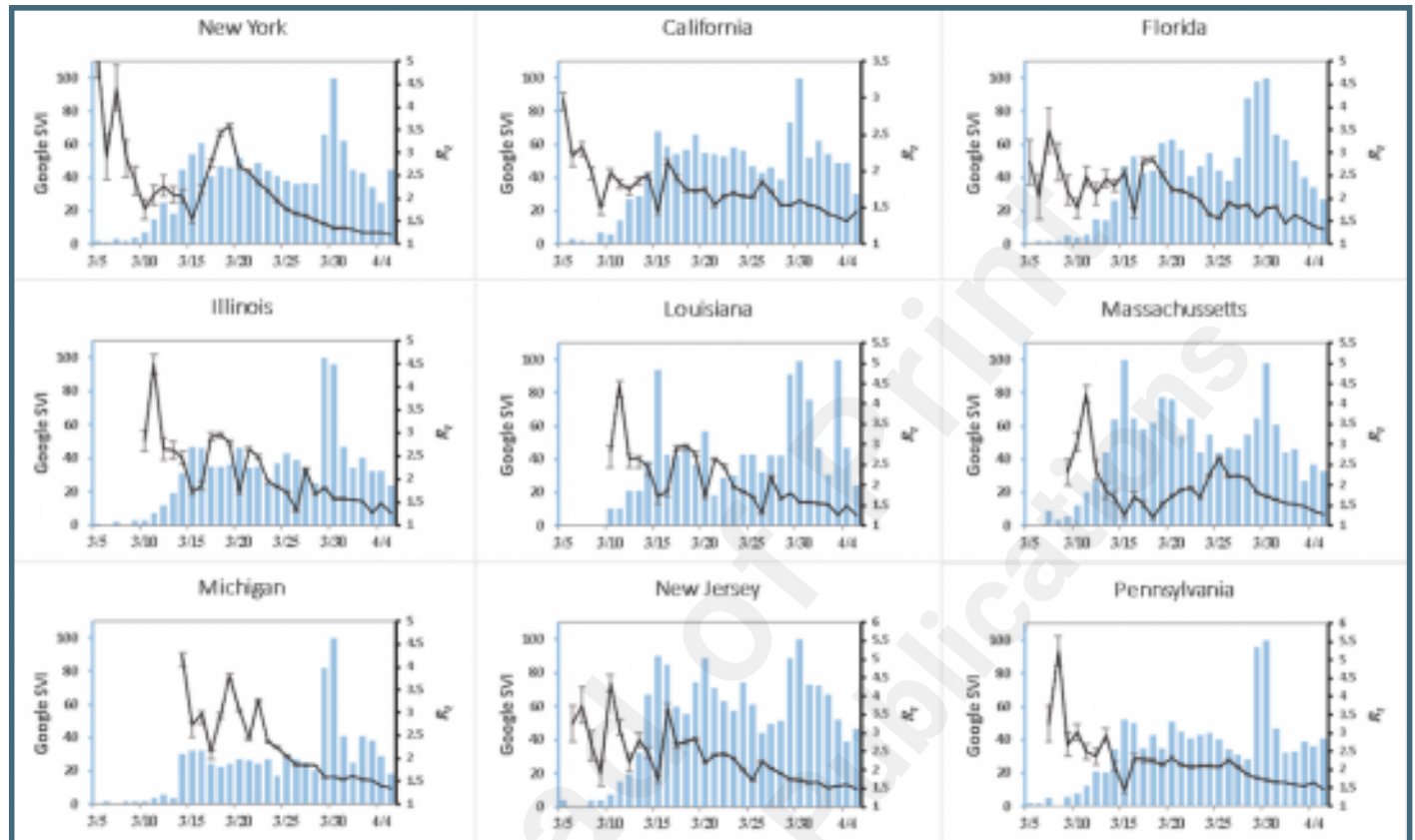
The time-variant or effective reproduction number,  $R_t$ , represents the mean number of secondary cases generated by a primary case over a sliding weekly window. The estimated  $R_t$  is shown for the period of February 28th to April 5th calculated from the number of COVID-19 cases by symptom onset, with a mean serial interval of 3.96 days and standard deviation of 4.75 days. The shaded error bands are equal to one standard deviation of the estimated  $R_t$  for each date.



On the left shows the relationship between Google Trends Search Volume Index (SVI) for “social distancing” and estimated  $R_t$  for the (a) USA and (b) New York over the time period from March 5th to April 5th. Light blue line refers to Google Search Index (SVI) for the query “social distancing.” Dark grey line refers to the estimated  $R_t$ . On the right shows the relationship between estimated  $R_t$  and total social media mentions for “#socialdistancing” in the USA on (c) Instagram and (d) Twitter over the same time period. Error bands shown in light grey represent one standard deviation from the mean  $R_t$  calculated at each date.



The estimated  $R_t$  was calculated individually for the Top 9 states with the most confirmed COVID-19 cases on April 5th. They are graphed for the period of March 5th to April 5th for each state (dark line) along with the Google search index for “social distancing” in each state for the same time period (light blue bars). Error bands represent one standard deviation from the mean estimated  $R_t$  at each date. The data for New York can be found in Figure 2.



Cross-correlations (represented by  $\rho^2$ , the coefficient of determination) between estimated  $R_t$  and social mobility changes and social media mentions in the USA. The black lines represent the peak correlation. Social mobility changes include traffic at grocery, pharmacy, and workplace locations based on Google Maps tracking. Social media measures include mentions of “#socialdistancing” on Instagram and Twitter.

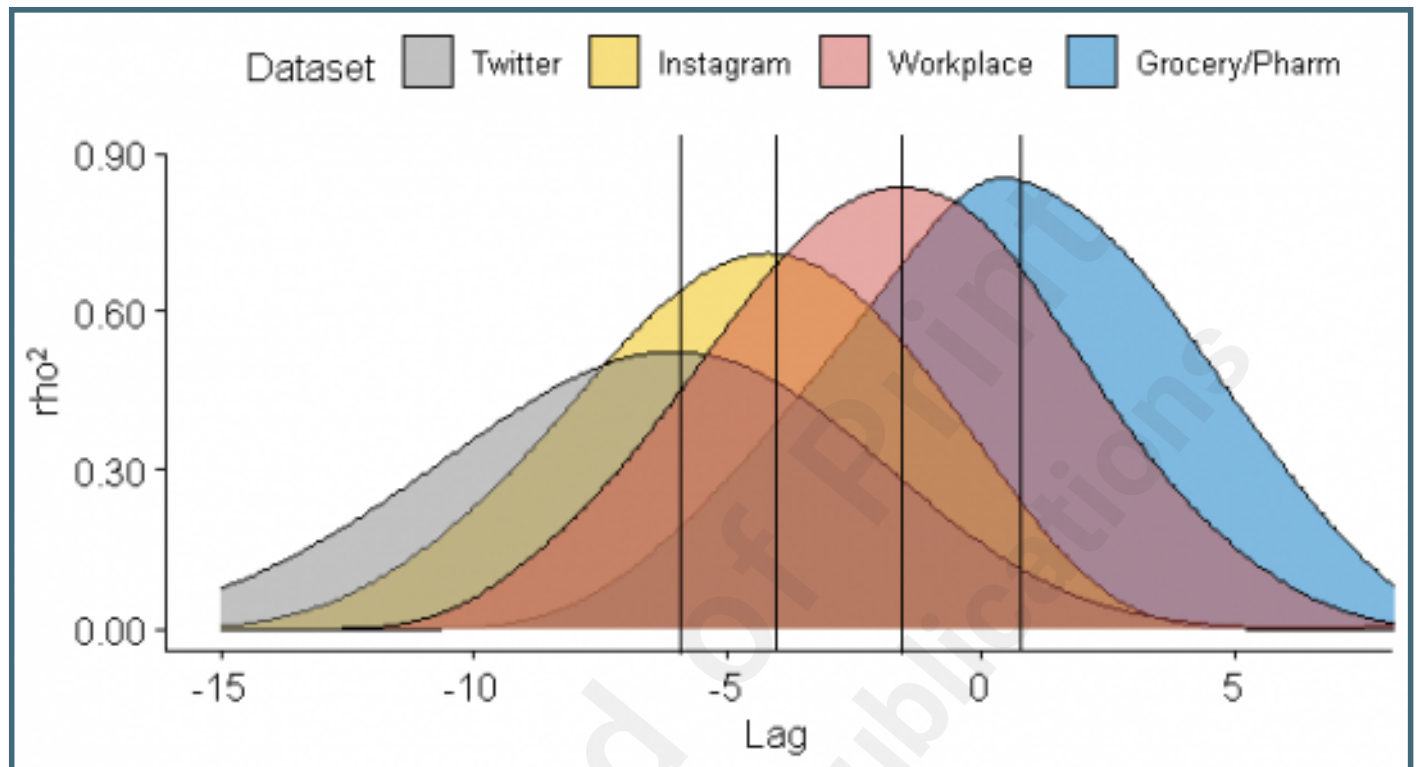


Table 1.

**Table 1. Peak Correlations Between Social Media and Social Mobility Measures and Associated Time-Delay**

	Dataset	$\rho^2$	Lag (days)
Social Media	Instagram	0.68	-4
	Twitter	0.47	-6
	Google Trends	0.72	-4
Apple Maps	Driving	0.75	-3
	Transit	0.80	-3
	Walking	0.73	-2
Google Maps	Grocery/Pharmacy	0.89	+1
	Transit	0.83	-2
	Workplace	0.86	-2
	Parks	0.66	-2
	Recreation	0.84	0
	Residential	0.85	-2