

# Truth or Dare? Detecting Systematic Manipulation of COVID-19 Statistics

Fatih Serkant Adiguzel\*, Asli Cansunar†, Gozde Corekcioglu ‡

September 4, 2020

*Forthcoming at the Journal of Political Institutions and Political Economy*

## Abstract

Which countries manipulate COVID-19 statistics? Does the party ideology of local governors affect the probability of data manipulation at subnational levels? How does democratic quality affect statistical transparency during the pandemic? In this article, we apply election fraud detection methods –various digit-based tests that exploit human biases in generating random numbers– to the daily announced official numbers of new and cumulative coronavirus infections. First, we use digit-based tests to identify countries that likely manipulated their pandemic statistics. We then move on to examine the empirical relationship between democratic quality and data transparency. We find suggestive evidence that data manipulation occurred in China, the United States, Russia, and Turkey. Second, we show that non-democracies, as well as countries without free and fair elections, are more likely to release data that display signs of statistical malpractice.

**Keywords**— COVID-19, digit-based tests, statistical malpractice, democracy

---

\*Department of Political Science, Duke University, 140 Science Drive, Durham, NC 27708, USA; serkant.adiguzel@duke.edu

†Nuffield College, University of Oxford, New Road, Oxford OX11NF, UK; asli.cansunar@politics.ox.ac.uk

‡Department of Economics, Kadir Has University, Cibali Mah. Kadir Has Cad. Fatih, Istanbul, 34083, Turkey; gozde.corekcioglu@khas.edu.tr

# 1 Introduction

The outbreak of COVID-19 has triggered global interest in all countries' daily new and cumulative coronavirus infections. The pandemic's prevalence in a highly globalized world means that government officials and international organizations make important decisions regarding travel restrictions, quarantine rules, and other precautions based on numbers produced by governments that have incentives to skew local statistics. Consequently, there has been a surge of interest in whether officially released COVID-19 statistics from some countries reflect the truth. Numerous journalists and academics have attempted to uncover the discrepancies between official statistics and other data types. An example of this is the high percentage of lost users from China Mobile Ltd., China's largest cellphone carrier. China Mobile reported that it had lost almost 8 million subscribers between January and February 2020 <sup>1</sup>. Although this loss could be attributed to the pandemic's adverse effects on the economy, many critics have expressed fears that the "lost" subscribers are evidence of a death toll that is much higher than the number formally announced by the Chinese government. Another issue is the irregular deviations in the number of deaths recorded in March and April 2020 compared to the average number of deaths recorded in the same period over the previous few years. It has been conclusively shown that using the excess mortality data rather than the official statistics on COVID-19-related deaths can lead to more accurate estimates of the severity of the outbreak in most countries (Aron and Muellbauer, 2020; Schellekens and Sourrouille, 2020).

Given that the COVID-19 crisis has been the ultimate litmus test for state capacity, and the countries that have failed to contain the virus's spread are facing severe political and economic consequences, politicians have good reasons to hide their governments' failures. It is well documented that adverse economic conditions, whether caused by the governments' inability to promote economic growth or unexpected catastrophic shocks, are constant civil conflict sources and a significant determinant of incumbent turnover in both democracies and autocracies (Miguel, Satyanath and Sergenti, 2004). Governments in authoritarian or democratically backsliding countries have more to fear: economic and social problems prompted by the pandemic can cause significant political disturbance. Governments either try to limit the dissemination of "bad news" or manipulate the self-reported economic indicators to exercise damage control in hard times, especially in non-democratic countries (Magee and Doces, 2015; Rozenas and Stukal, 2019).

Not only are the consequences of the governments' failure to address the adverse effects of the pandemic potentially more dangerous in non-democracies, but authoritarian governments are also better armed with technological and bureaucratic tools to manipulate the self-reported estimates (Guriev and Treisman, 2019).

---

<sup>1</sup>"China's Mobile Carriers Lose 21 Million Users as Virus Bites" Bloomberg Business, <https://www.bloomberg.com/news/articles/2020-03-23/china-s-mobile-carriers-lose-15-million-users-as-virus-bites>.

However, forging numbers is a dangerous game with significant trade-offs. This is because underreporting the number of active cases might induce a false optimism among citizens, resulting in non-compliance with the preventive measures and social distancing. Furthermore, as the crisis is a global pandemic, any attempt to cover the spread and the virus's actual damage might induce international diplomatic or economic retribution.

One major challenge when comparing the official statistics and other types of data to conclude that governments are dishonest when reporting COVID-19 statistics is the difficulty of isolating the impact of intentional fraud from unintentional underreporting due to low state capacity. Unsurprisingly, governments choose to spend their limited resources on containing the virus's spread rather than establishing communication networks to ensure data transparency. Additionally, most countries struggled to identify every patient who either had COVID-19 or died from the virus due to the scarcity of testing kits. Also, most statistics do not include individuals who have had mild cases of COVID-19 and were instructed to stay at home to save hospital beds for more severely ill patients.

However, not enough attention has been paid to identify the countries that intentionally underreport daily new and cumulative infections to signal that the government has successfully kept the public health crises under control. Furthermore, little is known about the relationship between democratic quality and truthful dissemination of COVID-19 statistics. This raises the questions of who forges COVID-19 data, who reports the truth to the best of their ability, and whether democratic quality predicts how governments release information related to the pandemic. Allegations of deliberate data tampering carry profound public health implications.

In this article, to empirically identify the statistical irregularities in reported data, as well as to investigate the relationship between democratic quality and statistical transparency, we apply election fraud detection methods to the daily announced official numbers of new and cumulative COVID-19 cases, including a variety of digit-based tests that exploit human biases in random number generation ([Mebane Jr, 2006](#); [Klimek et al., 2012](#); [Beber and Scacco, 2012](#)). We present a detailed account of the digit tests and focus on the following countries where the statistics were scrutinized by social media users and international media outlets: Brazil, China, India, Iran, Italy, Russia, Turkey, United Kingdom, and the United States. We then turn our attention to the inspection of state-level data from the United States to ascertain whether politicized responses to the pandemic have translated into different levels of truthful data dissemination between states with Democratic and Republican governors. Finally, we present a cross-country analysis using the daily reported new and cumulative COVID-19 cases from 132 countries that explores the correlation between democratic indicators and the likelihood of data manipulation.

First, we find that electoral fraud detection techniques reveal that data from the most prominent data manipulation suspects, China and the United States, exhibit irregularities that suggest the intentional manipulation of pandemic statistics. While we find suggestive evidence that Russia and Turkey’s data shows evidence of statistical malpractice, we find no significant evidence to conclude data from the United Kingdom, India, Italy, and Brazil show clear signs of manipulation. Second, through a detailed examination of state-level data from the United States, we find no partisan differences in data quality between states with Republican or Democrat governors. Finally, we find that non-democracies and countries without free and fair elections are more likely to release data that manifest signs of statistical malpractice. Because an effective global health response to the COVID-19 outbreak requires a constant and transparent flow of data across countries, our findings have profound policy implications.

## 2 Methodology and Data

With the rise of competitive authoritarianism, political scientists and statisticians have focused on developing methods to detect electoral fraud from subnational election data. One example of such “forensic” procedures is the last-digit and digit-pair tests developed by [Beber and Scacco \(2012\)](#) that exploit psychological biases in random number generation ([Beber and Scacco, 2012](#)). They show that humans do not select digits with equal frequency, avoid repetition, prefer serial sequences, and select pairs of distant numerals relatively infrequently. Therefore, they argue that manipulative data generation by humans can be detected in electoral data by identifying patterns that conform to these biases.

The last-digit test is based on the observation that humans fail to select digits with equal frequency. In the absence of data manipulation, the last digits of electoral statistics, such as vote counts or turnout numbers, should be distributed uniformly with an expected theoretical ratio of 0.1, provided two conditions are met: (i) counts do not cluster within a very narrow band of numbers, and (ii) counts do not include a large share of very small numbers ([Beber and Scacco, 2012](#)). Conversely, deviations from uniformly distributed last digits suggest that the data are unlikely to have emerged through a “natural” generation process. For example, an excess of zeroes vis-à-vis other numerals in the last digits of reported numbers signals potential manipulation. Absent fraud, however, we would not expect to observe a suspicious lack or excess of digit repetition.

Given the parallels between election data and pandemic statistics, we apply the electoral fraud detection tests to the numbers of daily reported new and cumulative COVID-19 cases disseminated by governments. Both daily new and cumulative cases are politically important numbers that the governments might have

reasons to keep under control fraudulently. While daily new cases are viewed as the signal of the seriousness of the outbreak by the general public, other governments often use cumulative cases per capita in determining international precautions such as travel restrictions or quarantine lists. Given the importance of tourism revenues and the continuation of international business activities for the economy, especially for developing countries, cumulative cases are also ought to be prone to political manipulation.

We use data from the COVID-19 Data Hub ([Guidotti and Ardia, 2020](#)) which compiles official coronavirus statistics from 183 countries. We complement this data set with information obtained from Brazil’s Ministry of Health on the local COVID-19 statistics ([Cota, 2020](#)). We limit the end date of data collection for two countries: Italy and China. In China, we exclude data after the last day of the Wuhan lock-down, April 9. In Italy, we collect data up until the last date of the national lock-down, May 19. For the last digit analysis, we pool daily COVID-19 statistics disseminated at the the most local level. Thus, our unit of observation is at the day-unit level, where unit refers to the most disaggregated administrative level at which data was provided. This allows us to examine the possibility of data manipulation at the level where the data generation happens.<sup>2</sup> We also limit our sample to days where the new and cumulative number of COVID-19 cases exceeds 100 in a local –or a national– unit, since last-digit analysis requires that the data do not contain a large proportion of single- and double-digit counts ([Beber and Scacco, 2012](#)). Finally, for all countries, we stop the data collection process (for a given local unit) if there are no new cases for seven consecutive days (in that unit) given that one of the assumptions of the digit-based tests requires that counts do not cluster within a very narrow band. Using these COVID-19 numbers, we compute a chi-squared test statistic and the corresponding p-value that measure whether the deviation of the distribution of last digits from the expected theoretical frequency of 0.1 is jointly significant ([Beber and Scacco, 2012](#)). We then use information from the last digit test to investigate the correlation between democratic institutions and data transparency during the pandemic.

### 3 Results

We begin our analysis by investigating the distribution of the daily cumulative coronavirus infections’ last digits.<sup>3</sup> Figure 1 shows the frequencies with which different numerals appear in the last digit of cumulative confirmed cases of COVID-19 for nine selected countries: the United States, Italy, Brazil, Russia, India,

---

<sup>2</sup>The number of observations varies across countries depending on the pandemic’s longevity and the administrative level at which the data was provided.

<sup>3</sup>In the Appendix, we complement this analysis with a digit-pair test which considers the distance between the last two digits of reported cumulative confirmed cases of COVID-19.

China, the United Kingdom, Iran, and Turkey. Horizontal lines indicate the lower and upper confidence bounds. We use 1,000 simulations to construct 95 percent confidence intervals around the expected values.

Since the beginning of the outbreak, there has been mounting international suspicion that China’s official COVID-19 statistics cannot be trusted. From GDP figures to investment data, scholars have carefully documented that the Chinese government and the local governments have a long history of manipulating data for political gain (Chen et al., 2019). China’s Premier, Li Keqiang, acknowledged that the COVID-19 numbers had been misreported by local governments when he warned officials that “being open and transparent means a new case must be reported once it is discovered. It is what it is. There must be no concealing or underreporting.”<sup>4</sup> Figure 1 reveals that the total confirmed cases reported by China display substantial evidence of manipulation. We observe significant deviations from the uniform distribution in the numbers of confirmed cases reported by Chinese provinces, particularly for the numeral 6. We present additional results from China in the Appendix.

Data from the United States, likewise, exhibit significant manipulation signs: the numeral 9 is underrepresented and lies outside the confidence intervals. Similarly, numeral 1 is overused in the cumulative daily COVID-19 cases of the US. Data from Brazil, Italy, Iran, Turkey, and the UK do not display fraudulent manipulation signs. All numerals from the last digits lie within the confidence intervals of an equifrequent distribution. Nevertheless, it should be noted that the confidence intervals are wider for Iran and Turkey due to reporting at a higher level of aggregation, which might plausibly mask unusual variation across digits.

---

<sup>4</sup>“China’s premier warns local officials not to hide new coronavirus infections” CNN, <https://edition.cnn.com/2020/03/25/asia/china-coronavirus-li-keqiang-intl-hnk/index.html>.

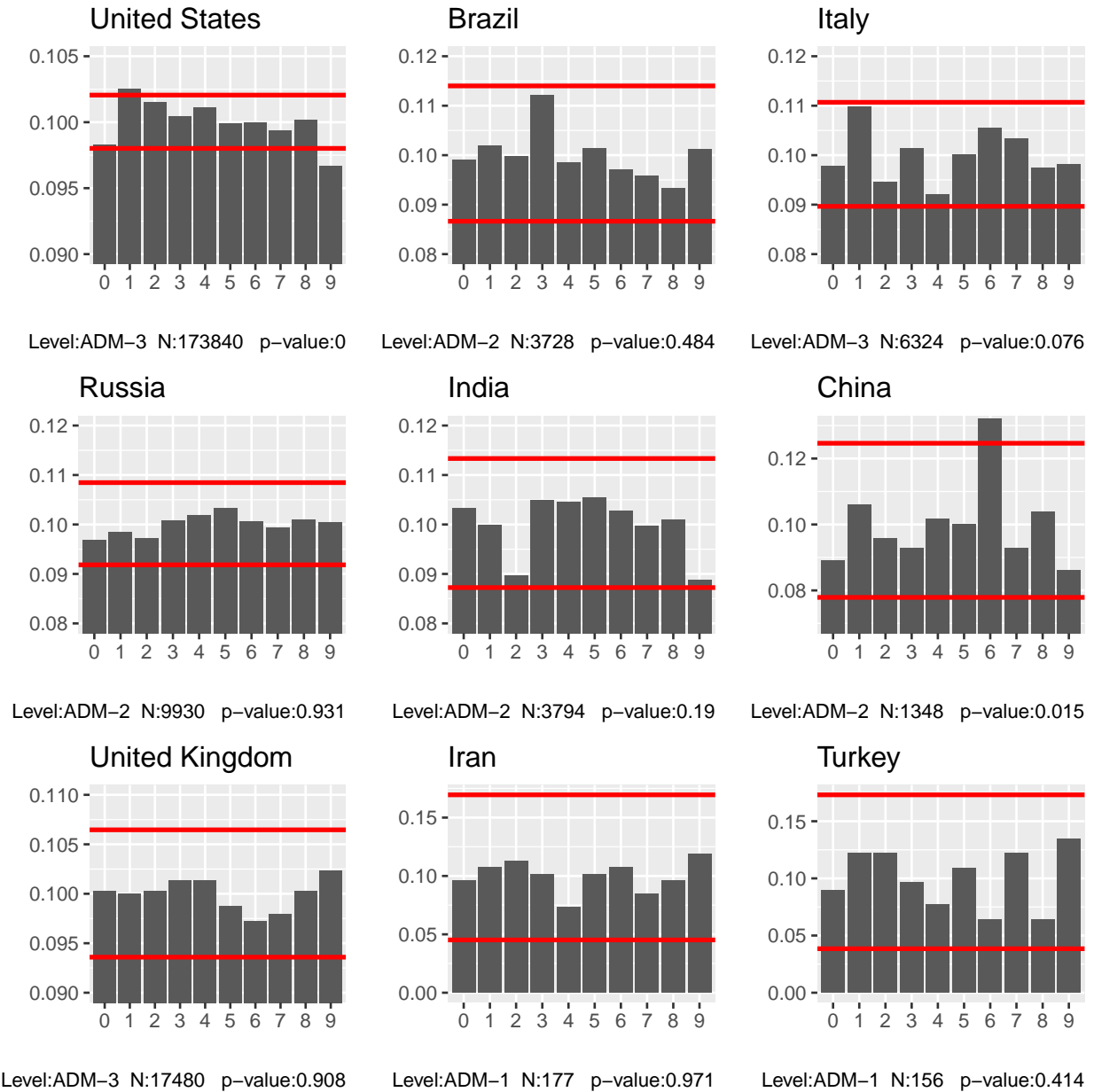


Figure 1: Frequency of last digits for cumulative COVID-19 cases. Administrative level at which data is used, number of observations, and p-values are reported for each country. Administrative levels correspond to 1= national, 2 = state/province, 3 = sub-state/sub-province.

We then move on to investigate the last digit distributions of daily reported new cases from the same set of countries excluding the United Kingdom and China. There are not enough observations that exceed 100 new cases per day due to the highly local level of data dissemination in both countries. Figure 2 confirms the finding that the data from the United States show significant signs of manipulation. Data from Brazil, Italy, India, and Iran do not display irregularities. In contrast to the analysis with the cumulative numbers, Turkey and Russia now show suggestive evidence of statistical malpractice.

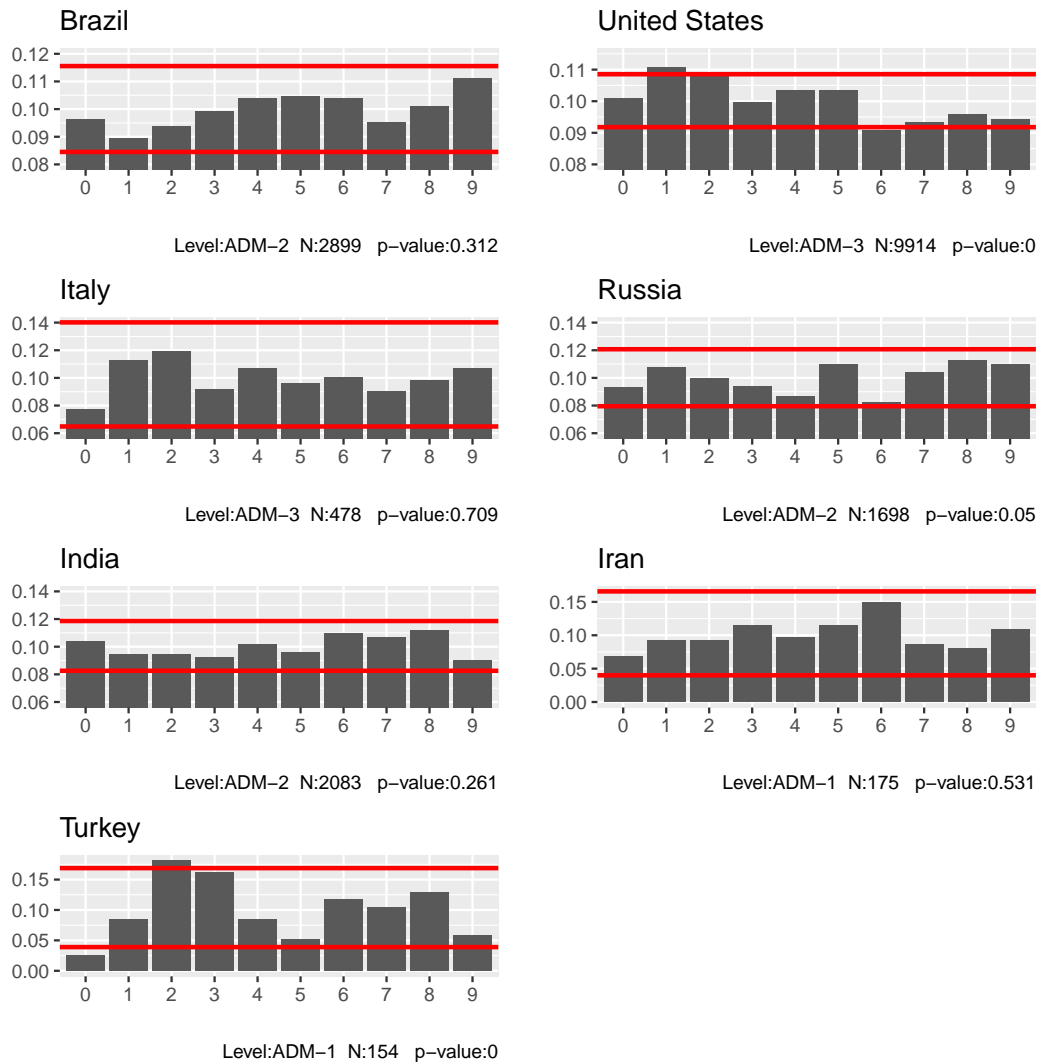


Figure 2: Frequency of last digits for daily new COVID-19 cases. The administrative level at which data is used, the number of observations, and p-values are reported for each country. We exclude China and the UK due to a low number of observations. In both cases, since the dissemination of statistics happens at very local levels, there are not enough observations that exceed 100 new cases per day for a meaningful analysis. Administrative levels correspond to 1= national, 2 = state/province, 3 = sub-state/sub-province.

How can we reconcile different findings pertaining to Turkey and Russia when we use different types of COVID-19 statistics? We contend that if the governments aspire to signal that they are successfully containing the disease, the optimal behavior is to craft the most popular statistics amongst the public and calculate the other numbers accordingly. For both Turkey and Russia, we find that domestic political attention is predominantly directed to daily new infections. In the Appendix, we include screenshots of the daily announcements of COVID-19 statistics by governmental agencies for these two countries. Announcements by health officials show that daily new coronavirus infections are significantly highlighted relative to the daily cumulative cases.



The well-documented politicization of the optimal public health response to COVID-19 raises the possibility that data manipulation could occur at the sub-national level, carried out by localities governed by different parties. For example, the response to COVID-19 in the United States has been deeply political, and elite messaging from the administration of President Donald Trump may have produced different levels of pressure on state governors with different party identifications. Republican governors, under pressure from the president to prematurely relax lockdown restrictions, might have sought to forge cumulative cases intentionally. To investigate the extent of partisan differences in data reporting transparency, we examine the COVID-19 statistics disseminated by counties in states with governors from the Democrat and Republican parties in the US. We implement the last-digit analysis separately for each state by pooling daily COVID-19 statistics from counties.

Figures 3 and 4 display the last-digit distributions of daily cumulative coronavirus infections from states with the highest and lowest p-values respectively. States with Democrat governors are shown in blue, while states with Republican governors are marked in red. Taken together, these figures confirm that states with both Republican and Democrat governors show displays of statistical malpractice. In contrast to widespread news and social-media discussion that accused Republican governors of fabricating data that suit their political agendas, a majority of states with the highest p-values are those with Republican governors.

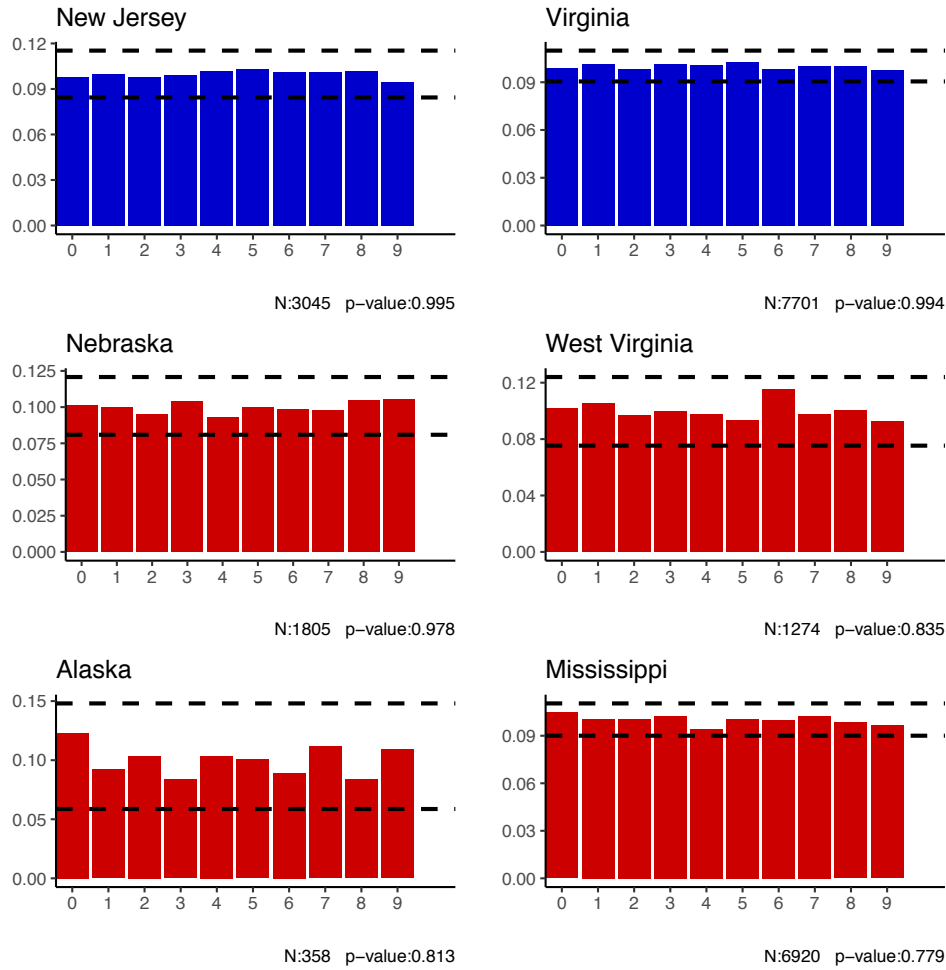


Figure 3: Last-digit analysis for daily cumulative cases. States with the highest p-values.

We now move on to explore the link between statistical manipulation, indicators of democracy, and economic development. We operationalize our dependent variable, *data manipulation*, in three different ways. First, we use the p-value from the last-digit tests of cumulative coronavirus cases as a continuous dependent variable. Recall that low p-values point to a higher probability of statistical malpractice. Second, we use a binary variable that categorizes countries as involved in data manipulation if the p-value from chi-square tests of the null hypothesis of equally frequent last-digit numerals is less than 0.05. Finally, we use numerals' variance from the last digits of the reported statistics as our dependent variable. Since the theoretical expectation is that every numeral in the last digits of announced COVID-19 cases should be 0.1

under a non-fraudulent data generation process and a sufficiently high number of observations, the higher variance between the numerals signal that it is more likely that manipulation occurred in the dissemination of pandemic statistics. We employ an ordinary least squares regression for our analysis. In these regressions, we restrict our sample to countries with at least 100 observations,<sup>5</sup> in addition to dropping units where we do not observe new cases for 7 consecutive days.

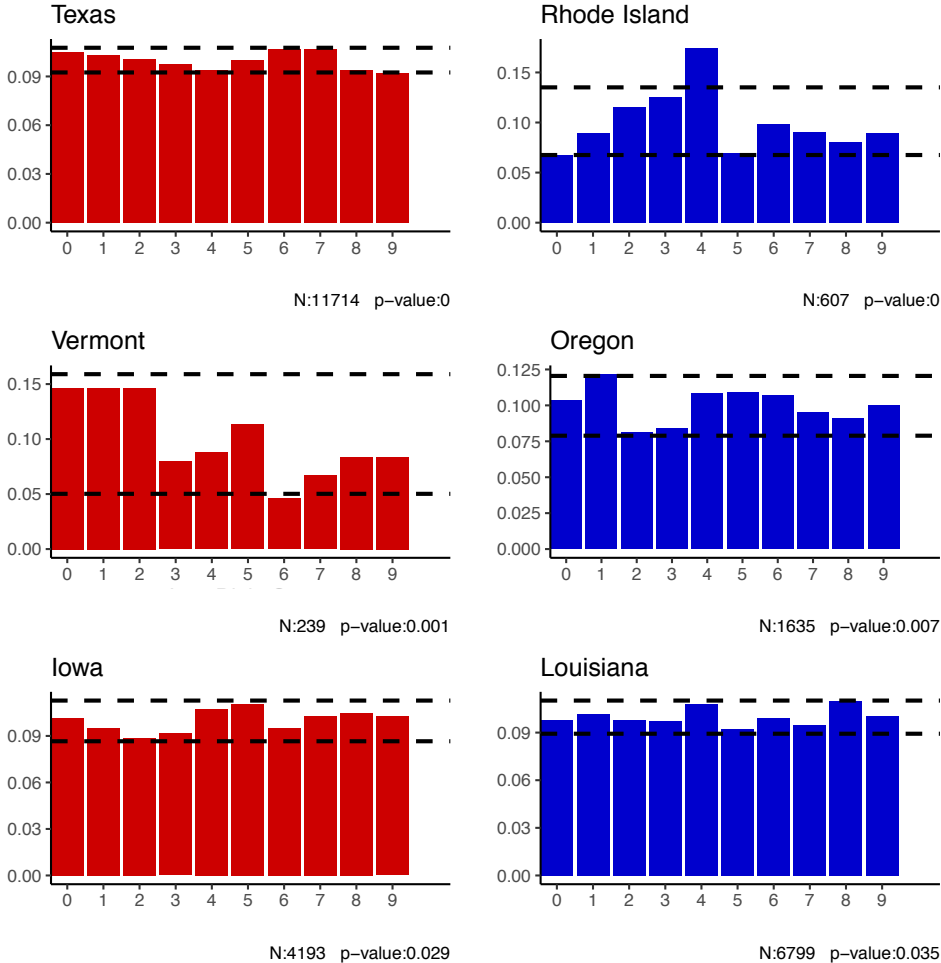


Figure 4: Last-digit analysis for daily cumulative cases. States with the lowest p-values.

Our key independent variables are based on indicators of democratic quality compiled by [Cheibub, Gandhi and Vreeland \(2010\)](#), which measure whether a country is a democracy (*democracy*=1 if a country

<sup>5</sup>We refrain from doing the same empirical exercise for p-values of the digit tests using daily new cases because the sample size drops to 44 as a result of this restriction.

Table 1: Signs of Statistical Malpractice and Political Indicators

<i>Last Digit Test</i>	<i>p-value</i>		<i>p-value &lt; 0.05</i>		<i>Variance</i>	
	(1)	(2)	(3)	(4)	(5)	(6)
Democracy=1	0.044 (0.054)		-0.115** (0.055)		-0.0002** (0.0001)	
Free and fair election=1		0.052 (0.054)		-0.142** (0.055)		-0.0002** (0.0001)
log(GDP)	-0.118** (0.057)	-0.119** (0.056)	0.063 (0.058)	0.066 (0.057)	0.00004 (0.0001)	0.00005 (0.0001)
log(Population)	-0.023 (0.017)	-0.023 (0.017)	-0.003 (0.017)	-0.005 (0.017)	0.00002 (0.00003)	0.00002 (0.00003)
log(Health exp)	0.065 (0.054)	0.064 (0.054)	0.012 (0.055)	0.015 (0.054)	0.00001 (0.0001)	0.00001 (0.0001)
log(Observations)	0.129** (0.053)	0.129** (0.053)	-0.084 (0.055)	-0.084 (0.054)	-0.0003*** (0.0001)	-0.0003*** (0.0001)
level: ADM-2	-0.401** (0.159)	-0.405** (0.158)	0.423** (0.162)	0.434*** (0.161)	-0.00002 (0.0002)	-0.00001 (0.0002)
level: ADM-3	-0.722*** (0.265)	-0.726*** (0.265)	1.094*** (0.271)	1.105*** (0.269)	0.001 (0.0004)	0.001 (0.0004)
Observations	132	132	132	132	132	132
R <sup>2</sup>	0.114	0.116	0.221	0.234	0.306	0.314

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

is classified as a democracy and *democracy*=0 if not) and whether a country holds free and fair elections (*free and fair election*=1 if elections were conducted, they were free and fair, and there was a peaceful turnover of legislative and executive offices following them; *free and fair election*=0 otherwise). We include logarithmic transformations of the GDP per capita and population to account for differences in the quality of data dissemination based on countries' size and economic development levels. We also control for the logarithmic transformations of health expenditures per capita in current US dollars to capture the effect of health capacity on the probability of statistical fraud. Finally, we control for the logged number of observations (days with more than 100 confirmed cases) and include dummy variables for the administrative level at which data is made available.

Table 1 documents the results from these cross-country regressions. In all specifications, the signs of the coefficients of the main independent variables indicate that democracy and the existence of free and fair elections are negatively related to the likelihood of statistical malpractice during the COVID-19 outbreak. The coefficients of democracy and holding free and fair elections are statistically significant in cases where we use *p-value < 0.5* (columns 3-4) and *variance* (columns 5-6) as dependent variables. These results are robust to the inclusion of economic and demographic controls.

## 4 Conclusion

Fraud allegations concerning reported COVID-19 statistics have been a hotly debated issue, almost as prominent as the disease’s outbreak. Nevertheless, it is challenging to empirically identify countries in which the statistics have been tampered with, due to the difficulty of distinguishing the effects of low state or health capacity from intentional data manipulation. In this article, we propose the application of election forensic tools to evaluate the fabrication of reported COVID-19 statistics.

While one cannot precisely determine the statistical distribution of the number of COVID-19 cases or deaths, it is possible to theorize that the last digits are expected to follow a uniform distribution under certain conditions given that psychologists have uncovered specific patterns that emerge when humans generate numbers. We have applied a last digit-based test to governments’ daily pandemic statistics. We have tested whether the reported last digits of publicly announced COVID-19 cases deviate from theoretical distributions and demonstrate traces of human biases in random number generation. Among controversial countries, we have detected atypical distributions in China, Russia, Turkey, and the US. We have also identified a statistically significant negative relationship between democratic institutions and the probability of statistical fraud.

The forensic tool used in this paper has several caveats. First, the identification of suspicious statistics is dependent on the availability of data: It is possible that manipulation is not identified in certain countries precisely because they do not disseminate sufficient information to allow forensic procedures. A few notable examples are Turkmenistan and North Korea, where the oppressive governments’ leaders have repeatedly argued their countries have no COVID-19 cases (Ibbotson, 2020). There is no useful way of accurately assessing the coronavirus outbreak’s gravity in these countries with no data to analyze. Given that we find a negative correlation between democratic institutions and data transparency even when many authoritarian countries drop from our cross-national sample by not disseminating enough information to be analyzed, our empirical results indicate that data from non-democratic countries should be taken with a grain of salt when giving important decisions about traveling or adherence to social distancing measures. Second, these techniques are not intended to identify the exact method through which data manipulation takes place—as is intended in other types of fraud detection methods—but are instead used to decide whether abnormal numbers contaminate a whole data set.

What new information is revealed about the COVID-19 statistics from this exercise? Most importantly, digit-based tests allow us to detect abnormalities even in cases where data is scarce. For example, many scholars have relied on comparing excess deaths to expected deaths to predict the degree to which the

coronavirus outbreak has affected their countries. First, although these exercises are useful to identify the real magnitude of the death toll, it is a poor predictor of the probability of intentional fraud, especially given that almost all countries' expected deaths have diverged from the reported deaths during the pandemic. Secondly, while the excess death data is available for most advanced industrialized countries, these numbers are often not collected or transparently disseminated by most governments in the developing world. The scarcity of excess death data makes it extremely hard to conduct an empirical global analysis that includes more than 130 countries.

Another question that emerges through our analysis is that, why would governments choose to manipulate numbers rather than not disseminating data at all? Although this is a plausible approach for countries like Turkmenistan and North Korea, for any country that is fully embedded in international trade and political networks, such a strategy would cause considerable international resentment.

Although our results are purely correlational and digit-based tests do not provide conclusive evidence, they are undoubtedly useful in inspecting statistical malpractice during the COVID-19 crisis. The conclusions from digit-based procedures should encourage citizens to continue to comply with the precautionary measures such as social distancing, washing hands, or wearing masks rather than blindly trusting questionable statistics produced with political or economic concerns.

# References

- Aron, Janine and John Muellbauer. 2020. "Measuring Excess Mortality: the case of England During the Covid-19 Pandemic." *INET Oxford Working Paper* (2020-11).
- Beber, Bernd and Alexandra Scacco. 2012. "What the numbers say: A digit-based test for election fraud." *Political Analysis* 20(2):211–234.
- Cheibub, José Antonio, Jennifer Gandhi and James Raymond Vreeland. 2010. "Democracy and dictatorship revisited." *Public choice* 143(1-2):67–101.
- Chen, Wei, Xilu Chen, Chang-Tai Hsieh and Zheng Song. 2019. A forensic examination of China's national accounts. Technical report National Bureau of Economic Research.
- Cota, Wesley. 2020. "Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level." *SciELOPreprints* .  
**URL:** <https://doi.org/10.1590/scielopreprints.362>
- Guidotti, Emanuele and David Ardia. 2020. "COVID-19 Data Hub." *Journal of Open Source Software* 5(51):2376.  
**URL:** <https://doi.org/10.21105/joss.02376>
- Guriev, Sergei and Daniel Treisman. 2019. "Informational autocrats." *Journal of Economic Perspectives* 33(4):100–127.
- Ibbotson, Sophie. 2020. "COVID-19: Approaches, Outlooks, and Power Dynamics in Central Asia." *Asian Affairs* pp. 1–14.
- Klimek, Peter, Yuri Yegorov, Rudolf Hanel and Stefan Thurner. 2012. "Statistical detection of systematic election irregularities." *Proceedings of the National Academy of Sciences* 109(41):16469–16473.
- Magee, Christopher SP and John A Doces. 2015. "Reconsidering regime type and growth: lies, dictatorships, and statistics." *International Studies Quarterly* 59(2):223–237.
- Mebane Jr, Walter R. 2006. Election forensics: Vote counts and Benford's law. In *Summer Meeting of the Political Methodology Society, UC-Davis, July*. pp. 20–22.
- Miguel, Edward, Shanker Satyanath and Ernest Sergenti. 2004. "Economic shocks and civil conflict: An instrumental variables approach." *Journal of political Economy* 112(4):725–753.

- Rozenas, Arturas and Denis Stukal. 2019. "How autocrats manipulate economic news: Evidence from Russia's state-controlled television." *The Journal of Politics* 81(3):982–996.
- Schellekens, Philip and Diego Sourrouille. 2020. "Tracking COVID-19 as Cause of Death: Global Estimates of Relative Severity." [https://www.brookings.edu/wp-content/uploads/2020/05/Tracking\\_COVID-19\\_as\\_Cause\\_of\\_Death-Global\\_Estimates\\_of\\_Severity.pdf](https://www.brookings.edu/wp-content/uploads/2020/05/Tracking_COVID-19_as_Cause_of_Death-Global_Estimates_of_Severity.pdf).



# Truth or Dare? Detecting Systematic Manipulation of COVID-19 Statistics Online Appendix

Fatih Serkant Adiguzel\*, Asli Cansunar†, Gozde Corekcioglu ‡

September 4, 2020

## Contents

<b>1</b>	<b>Digit-Pair Tests</b>	<b>2</b>
<b>2</b>	<b>Cumulative Cases in China</b>	<b>3</b>
<b>3</b>	<b>Additional Figures</b>	<b>4</b>

---

\*Email: serkant.adiguzel@duke.edu. Department of Political Science, Duke University

†Email: asli.cansunar@nuffield.ox.ac.uk. Department of Political Science and International Relations and Nuffield College, University of Oxford.

‡Email: gozde.corekcioglu@khas.edu.tr. Department of Economics, Kadir Has University.

# 1 Digit-Pair Tests

Digit-pair tests are based on a similar intuition to the last-digit test. Under a natural number generation process, the last two digits of any number are identical with a probability of 0.1; similarly, the distance between the last two digits is 1 with probability 0.2; finally, the distance between the last two digits is greater than 1 with a theoretical frequency of 0.7. Given that people avoid repetition, prefer serial sequences, and select pairs of distant numerals relatively infrequently when they are asked to produce numbers, any statistically significant deviation of the last two digits from these parameters should be taken as an indicator that there has likely been human manipulation of the data.

Figure A1 examines the patterns of the last two digits of the daily reported total of COVID-19 cases. Recall that the last two digits of a number are identical with a theoretical probability of 0.1, the distance between the last two digits is 1 with probability 0.2, and the distance between the last two digits is greater than 1 with a theoretical frequency of 0.7. We investigate whether the observed occurrences diverge from the theoretically expected frequencies. Ticks represent the 95% confidence intervals for each tested pattern, and the dots report the observed frequency of each pattern in the data from the nine countries under analysis.

Overall, we can plausibly state that the last-digit and last-two-digits analyses complement each other and provide consistent results. In particular, numbers from the US follow patterns perfectly consistent with human biases in number generation: lack of digit repetition, an excess of adjacent digits, and a dearth of distance numerals.

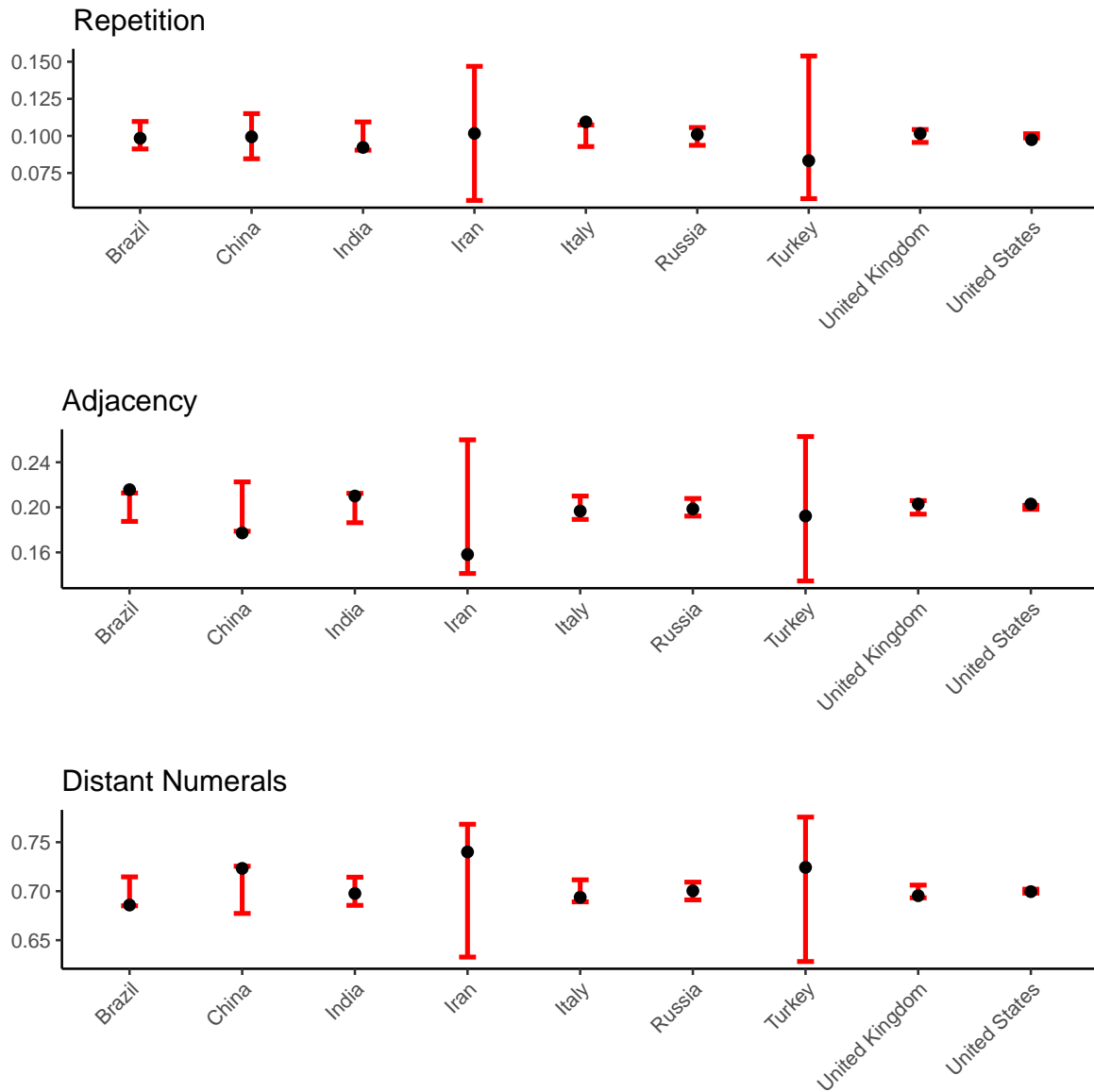


Figure A1: Frequency with which the last two digits are identical, adjacent, and have distance greater than one.

## 2 Cumulative Cases in China

In this section we provide additional complementary evidence on Chinese statistics. We refrain from imposing any rules on the data (such as dropping observations upon 7 consecutive days of no new cases) and take a closer look at the raw data as disseminated by the Chinese authorities. Figure A2 exhibits the last digit frequencies of daily announced cumulative confirmed cases before April 9. We detect several interesting patterns that might have remarkable implications

in the Chinese culture.

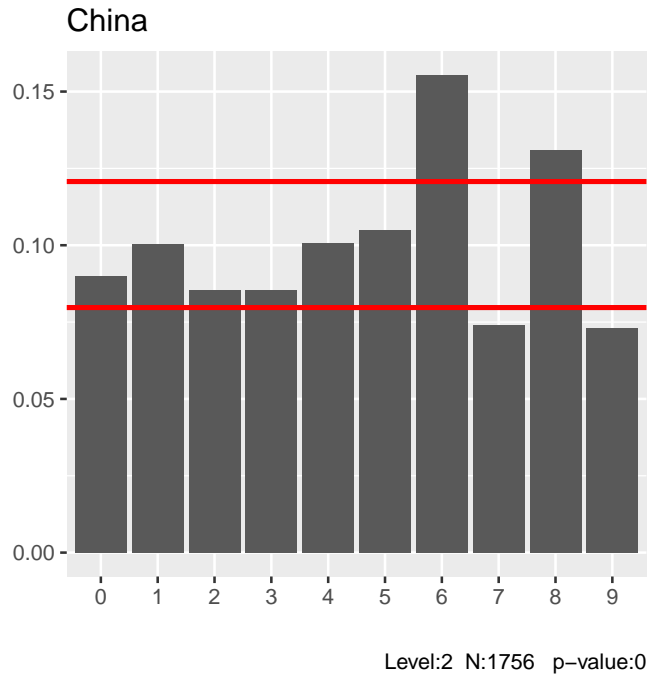


Figure A2: Frequency of last digits for cumulative COVID-19 cases in China. Administrative level at which data is used, number of observations, and p-values are reported. Administrative levels correspond to 1= national, 2 = state/province, 3 = sub-state/sub-province.

Given that numbers that are considered to be ‘lucky’ in Chinese culture often affect economic, and business transactions, an overrepresentation of 6 and 8 in the last digits of the confirmed COVID-19 cases suggests that the data were manipulated in a meaningful direction (Shum, Sun and Ye, 2014; Yang, 2011). Consistent with this is that the two underused digits, 7 and 9, also have interesting connotations. In Chinese, the number nine sounds like “long-lasting”, which is an undesirable characteristic for a pandemic, and number seven, which sounds like “to deceive”, is considered an unlucky number.

### 3 Additional Figures

In this section we include screenshots of daily public announcements of COVID-19 statistics by the primary official sources in China, Russia and Turkey. While the Chinese government reports both daily cumulative and new COVID-19 cases, Russian and Turkish statistics emphasize daily

new COVID-19 cases, as illustrated in Figures A3-A6 below.



Figure A3: Screenshot taken from the official Twitter account of the Chinese Center for Disease Control and Prevention. The data are provided by the Chinese National Health Commission. Source: <http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm>



Figure A4: The image is taken from the Instagram account of stopcoronavirus.rf, the official Russian Health Ministry website dedicated to the coronavirus pandemic.



Figure A5: Screenshot of a tweet by Turkey’s Health minister, who releases daily official coronavirus statistics from his Twitter account.



Figure A6: Anadolu Agency, a state-run news agency, tweets the daily official coronavirus statistics in both Turkish and English.

## References

- Shum, Matthew, Wei Sun and Guangliang Ye. 2014. “Superstition and “lucky” apartments: Evidence from transaction-level data.” *Journal of Comparative Economics* 42(1):109–117.
- Yang, Zili. 2011. ““Lucky” numbers, unlucky consumers.” *The Journal of Socio-Economics* 40(5):692–699.