Time Analysis of an Emergent Infection Spread Among Healthcare Workers: Lessons Learned from Early Wave of SARS-CoV-2

Patricia AF Leme1,*, Mehrsa Jalalizadeh1,*, Cristiane Giacomelli da Costa1,*, Keini Buosi1, Luciana SB Dal Col1, Franciele AV Dionato1, Lucas M Gon2, Reza Yadollahvandmiandoab1, Leonardo O Reis1,2

1UroScience, School of Medical Sciences, University of Campinas, UNICAMP, Campinas, Sao Paulo, Brazil; 2Pontifical Catholic University of Campinas, PUC, Campinas, Sao Paulo, Brazil

*These authors contributed equally to this work

Correspondence: Leonardo O Reis, UroScience, School of Medical Sciences, University of Campinas, UNICAMP, Campinas, Sao Paulo, Brazil, Tel +55 19 35217481, Email Reisleo@unicamp.br

Aim: To find whether an emergent airborne infection is more likely to spread among healthcare workers (HCW) based on data of SARS-CoV-2 and whether the number of new cases of such airborne viral disease can be predicted using a method traditionally used in weather forecasting called Autoregressive Fractionally Integrated Moving Average (ARFIMA).

Methods: We analyzed SARS-CoV-2 spread among HCWs based on outpatient nasopharyngeal swabs for real-time polymerase chain reaction (RT-PCR) tests and compared it to non-HCW in the first and the second wave of the pandemic. We also generated an ARFIMA model based on weekly case numbers from February 2020 to April 2021 and tested it on data from May to July 2021.

Results: Our analysis of 8998 tests in the 15 months period showed a rapid rise in positive RT-PCR tests among HCWs during the first wave of pandemic. In the second wave, however, positive patients were more commonly non-HCWs. The ARFIMA model showed a long-memory pattern for SARS-CoV-2 (seven months) and predicted future new cases with an average error of ±1.9 cases per week.

Conclusion: Our data indicate that the virus rapidly spread among HCWs during the first wave of the pandemic. Review of published literature showed that this was the case in multiple other areas as well. We therefore suggest strict policies early in the emergence of a new infection to protect HCWs and prevent spreading to the general public. The ARFIMA model can be a valuable forecasting tool to predict the number of new cases in advance and assist in efficient planning.

Keywords: healthcare worker, COVID-19, SARS-COV-2, diagnosis, prediction, dispersion

Introduction

The World Health Organization defines an emerging infectious disease as “one that has appeared in the population for the first time, or that may have existed previously but is rapidly increasing incidence or geographic range”. Severe Acute Respiratory Syndrome-CoV (SARS), Middle East Respiratory Syndrome-CoV (MERS), hemorrhagic fever viruses (Lassa, Ebola), and new influenza A viruses (H5N1), (H7N9), and (H1N1) are all considered emerging infections of the 21st century up until the year 2019, when the new SARS variant, SARS-CoV-2, stunted the world for two years. Both MERS and the original SARS were mainly spread among HCWs and attempts at stopping their spreads succeeded in preventing worldwide pandemics. Monkeypox is the most recent emerging infection; fortunately, it is not airborne.4

Multiple studies have assessed the prevalence of SARS-CoV-2 infection among HCWs, most of which use serological evidence.5–7 Even though serological testing provides valuable epidemiological information, it does not show individuals that are currently infectious or have been infectious in the past few days, because the test might...
Remain positive for months after active infection.\textsuperscript{8} In this article, we aim to assess the temporal spread of infection between HCWs and non-HCWs based on nasopharyngeal RT-PCR testing to show which population actively spreads the disease during each pandemic wave. In Brazil, the first wave of COVID-19 reached its peak around the end of May 2020 until September 2020. The second wave started in early November 2020 and peaked by the end of March 2021.\textsuperscript{9} The secondary aim of this article is to assess the predictive value of ARFIMA forecasting using weekly case numbers to forecast future new cases. We further provide a short review of literature and discuss our results in the setting of previous findings, such as the airborne as route transmission for SARS-CoV-2.\textsuperscript{10}

**Materials and Methods**

**Study Subjects**

After local ethics committee approval number 4.173.069, all patients over 16 years old referred for SARS-CoV-2 nasopharyngeal RT-PCR testing were added to the database. Trained nurses recorded each patient’s data: age, gender, occupation, zip code, and symptoms. The data include information on all tests carried out at university medical centers, which receive university employees as their main patients; and there are no other SARS-CoV-2 testing locations in the area covered. Exclusion criteria were inconclusive test results, duplicates or follow-up tests.

We defined HCW as any occupation that involves direct (HCW-D) or indirect contact with patients (healthcare associates or HCW-A). We compared these two groups to those without healthcare-related jobs (non-HCW). All statistical analyses were performed using R version 4.0.2 (2020.06.22) on RStudio platform version 1.3.1073 and using the following packages: tidyverse, lubridate, forecast, quantreg, splines, ggmap, pracma, fractaldim, and janitor. Normality for samples of n < 5000 was calculated using the Shapiro–Wilk test, and for n > 5000 the shape of the histogram was considered. The significance was considered when p < 0.05.

**Time Series Analysis and Forecast**

Time was the most important independent variable in our analysis; therefore, we allocated our data into weekly time fractals ie we set the unit of time in our analysis as one week. We calculated the proportions of female gender, positive results, and HCW per week as dependent variables and the Mann-Kendall trend test was used to detect change over time (trends). To better understand outcomes and detect underlying relationships between the occurrence of positive cases in time, we transformed the count of positive cases to fractal dimension via the R package fractaldim. This transformation reveals the spread of infections across time and shows its memory process.

After confirming a long memory process for COVID-19, we avoided traditionally used forecasting models (such as Autoregressive Integrated Moving Average [ARIMA] models or machine learning models) and instead, incorporated ARFIMA (Autoregressive Fractionally Integrated Moving Average) forecasting model, which accounts for long memory by fractional differencing the time series.

The ARFIMA model is like the Box-Jenkins ARIMA model, except that the integrated part of the ARIMA model can be a fractional number defined that is the inverse of the Hurst parameter.\textsuperscript{11} The fractional differencing allows a better representation of the data since it decays slower than other types of time series. We tested our model’s forecasting ability using data gathered three months later (May–July 2021) to estimate the reliability of our 12-week forecasting tool while keeping the investigators blind to actual values.

**Results**

From February 2020 to April 2021, after excluding duplicates or follow-up tests, a total of 8998 SARS-COV-2 RT-PCR tests were included; a total of 1914 (21.3%) of them returned positive, 70.6% returned negative, and 8% were either inconclusive or not available. Inconclusive or not available results were excluded for further analysis.

Fifty patients needed hospitalization, and three died. The mean age of participants was 38.0, minimum 16 and maximum 95 years old. SARS-COV-2 positive patients were on average 1.8 years older than the negatives (95% CI: 1.2–
2.4 years, p < 0.0001). There was a gender discrepancy, with 70% of the patients being female. This discrepancy was investigated by occupation analysis of our dataset.

**Occupation Based Analysis**

A total of 3462 (42%) patients were HCW-D, 933 (11%) HCW-A, and 3877 (47%) patients were non-HCW. The gender discrepancy was significant in healthcare-related groups but not in the non-HCW: 82% of HCW-D and 81% of HCW-A were female, while only 57% of non-HCW were female.

Overall, SARS-CoV-2 was confirmed in 27%, 25%, and 19% of HCW-D, HCW-A, and non-HCW, respectively. The odds ratio of positive result was slightly higher for females (OR = 1.12, 95% CI 1.00–1.25, p = 0.041). Mean age was significantly lower in the non-HCW group: 35.6 ± 13.6 years versus 41.0 ± 11.4 for HCW-D and 39.9 ± 10.2 for HCW-A (p < 0.0001, one-sided ANOVA).

We also combined both healthcare-related groups into one group (HCW) and compared them to non-HCW for interpretation of results. The odds ratio for positive SARS-CoV-2 testing was OR = 1.5, 95% CI 1.35–1.67 (p < 0.0001) for the HCW group versus non-HCW.

**Temporal Analysis**

Data were allocated into weekly time fractals. There were on average 137 tests performed per week with 32 of them returning positive (23.4%). The proportion of females to males significantly declined over the time period (tau = −0.48, p < 0.0001). Figure 1A shows all tests taken over time based on job types plus parallel timeline of public health policies. In the positive cases, we analyzed changes in the proportion of HCW (both direct and indirect contact) to non-HCW over time (Figure 1B).

At the beginning of the pandemic, HCW were the majority of the positive cases; however, their proportion declined significantly over the first 10 months (tau = −0.52, p < 0.0001). In the final 4 months of our data gathering, the proportion of HCW remained steady (tau = +0.10, p = 0.62) (Figure 1B). Mann-Kendall trend test showed a steady decline in this ratio over the entire period (tau = −0.58, p < 0.0001).

**Address Heat Map**

A digital map of all patient zip codes was created using Leaflet JavaScript library maps, and it is available online at http://wdchealth.covid-map.com/shiny/covid-map/. Page one of this website shows a heat map of all positive cases. Page two shows a time series of individual cases based on their zip codes. In the time series, each dot represents a single new case and it stays on the map for 14 days. Cut scenes of the time series are shown on the Temporal Map Analysis page. The “Count of Cases by Zip Code” table on the website shows the top ten zip codes with the highest number of positive cases.

**ARFIMA Forecast**

We calculated the Hurst parameter for positive and negative cases, yielding 0.78 and 0.84, respectively. It revealed the self-similarity of this infection and a pattern that persists with a long-memory process over seven months, Figure 2A. The Autocorrelation Function (ACF - Figure 2B) plot indicates that the memory indeed persists over the range of data collection once exposed in a community. Negative cases also showed a long-memory process as the long memory process crosses the axis on the autocorrelation function with persistent memory.

We used our model on 1914 positive tests, aggregating them to the weekly counts to estimate future weekly new cases (Figure 2C). It showed a steady slow increase of future cases (purple region) of infection. The purple hues are the statistical bounds (80% and 95%) of our prediction, while the dark blue line indicates the prediction of future cases.
Testing the Model

The ARFIMA prediction model was created using data available until April 2021. In July 2021, we gathered data of the past 3 months ("actual value") and compared it to the "predicted values". Figure 2D compares predicted (red line) versus actual (blue lines) count of positive cases. The average error was ±1.9 cases per week.

Figure 1 Comparison of occupation type over time. (A) Count of all patients based on occupation type. Healthcare worker (HCW) predominance in the beginning months of the pandemic versus non-HCW in the final months of the data gathering. HCW-D = HCW with direct patient contact, HCW-A = HCW with indirect contact. The time of each public intervention is shown in parallel; lockdown was between March 23rd and May 30th 2020, start of vaccination of HCWs was in January 2021, and start of vaccination of non-HCWs was in April 2021. Face mask use was mandated for public with the start of lockdown. (B) Percent HCW (both direct and indirect) in all positive patients declined rapidly (τ = −0.52, p<0.0001) in the first months of the pandemic, then it stays relatively steady over the rest of the data gathering (blue line and dots, τ = +0.10, p=0.62).
Discussion

Our data come from university health records that cover university employees as well as the students’ population of the area, to whom the university provides health services. The area covered does not have any other testing center, and this data is collected before the introduction of the rapid SARS-CoV-2 testing which can currently be purchased in pharmacies. We can, therefore, have relatively good confidence in that observing the pattern of SARS-CoV-2 spread in this data can show valuable information on dispersion of this disease across the covered area.

However, our study is also limited by the population that it covered, the university setting of the area means that the study findings may not apply to all populations. Furthermore, our data did not have the variables to prove or disprove airborne nature of the disease or detect possible confounders. For the mode of transmission, we use previously published studies as reference to discuss our results with assumption of airborne spread.

Another limitation of our data is its inability to show effect of public support systems. The city of Campinas had lockdown quarantine starting March 23rd 2020 until end of May, when activities were reopened progressively and oscillated from 30% to 80%. During the periods of most restriction, people’s circulation ranged at about 40%. Most of the regular business restarted after August, including restaurants, bars, and malls, with restrictions on people and working hours. Therefore, no clear cut-off intervention can be applied to our data for interrupted time series analysis. Hence, we focus on healthcare facility policies and lessons that should be applied to the next emerging infection as soon as it is detected and before it starts a new pandemic based on our analysis, previous studies, and lessons learned from the original 2003 SARS.

Time analysis of our data proved to be crucial in understanding the spread of the disease, showing HCWs as the main bulk of the infected population at the beginning of the pandemic. The time analysis graph shows a rapid rise in the raw
count of infected HCWs in the beginning of the first wave, when the patients were almost exclusively HCWs. Then, the proportion of HCWs to non-HCWs steadily declined as time passed and the number of infected increased in the first and the second wave of the pandemic. Once the infection spread to the general population, the proportion of HCW patients remained almost steady.

Vaccination in the area started with HCWs in the early January 2021. The non-HCW population started receiving the vaccine in the late March 2021 (starting with the geriatrics). It is wise to assume that it may have helped reduce the spread in the HCW population and is somehow responsible for the decline of positive COVID-19 cases in this population, especially since vaccine hesitancy was reported to be low in Brazil compared to some other populations.

Moreover, the data reflect a period before the arrival of SARS-CoV-2 variants. Indeed, the first detected in Brazil was the P1 variant and it was very related to the healthcare system in the northern region of Brazil. Those variants were found near our city only on February 2021, so it did not largely impact our data.

Even though HCWs are logically the most at risk in a pandemic, they are also the most educated and by-default the most readily armed against pandemics. However, in practice, we see many pitfalls in applying lessons that we have learned (or not learned) from previous pandemics and we end up repeating the same mistakes. In developing countries, HCWs reported low preparedness, improperly fitting personal protective equipment (PPE), and low confidence due to insufficient education in PPE usage. In developing countries, we saw clinics filled with suspicious patients sitting next to other patients in the same waiting area, hospital staff loosely using PPE, and HCWs coming to work even after they were tested positive for SARS-CoV-2 despite the facility’s “official” policy of home-quarantine for infected HCWs. Importantly, at the beginning of the pandemic, the mode of transmission for this new infection was uncertain which leads to delay in stringent use of masks in unprepared healthcare settings. This is the first important lesson that can be learned from our study: as soon as a new mutated infection emerges, “the worst-case scenario” (airborne spread) must always be assumed, and until enough research proves other modes of transmission, stringent PPE use must apply to the entire facility to prevent rapid spread among HCWs.

Our study is not the first of its kind, and many studies have compared infection rate between HCWs and non-HCWs during the early phases of the pandemic. Some confirm our results and some claim that HCWs had the same infection rate as the population during that time. These discrepancies can be attributed to multiple reasons. The first reason is the type of test used to confirm infection; our study is slightly superior as it uses RT-PCR testing instead of serology testing. Serology testing can remain positive for months and does not show actively infectious individuals, therefore it cannot be used for detailed time-analysis of the spread of the disease to compare different waves and discover patterns of spread among groups in short time periods.

Our study shows that the infection rate of an emergent airborne disease is higher among HCWs only during certain time early in the spread of the disease, so timeframe might also be a confounder in other studies. A South Korean study that used RT-PCR showed that “nosocomial infection” clusters were more common in the first wave, similar to the Italian and Norwegian studies that used RT-PCR. Studies that pooled data, used serology, and did not perform detailed time analysis are less likely to discover this pattern.

Third reason is differences among healthcare facilities in policies, routine protocols, preparedness for airborne disease control, as well as work-culture. A hospital in Italy, dedicated to referred COVID-19 patients reported low seroprevalence of the disease in their staff during the first wave of the epidemic. The hospital attributed this low prevalence to rapid and adequate adoption of infection control measures. Similar findings were reported from hospitals in Ontario, Canada and Austria, who adopted rapid effective infection control measures. Whereas partial use of PPE while working in departments with high patient load and turnover (emergency setting) is reported as a significant risk factor for infection. The 2003 SARS outbreak at Prince of Wales Hospital is particularly edifying in this matter: the index case (the first case) was hospitalized but was not isolated for 15 days. In these 15 days, he infected 50 HCWs, 17 medical students, 30 patients, 42 visitors, and four relatives.

Another interesting characteristic of SARS-CoV-2 spread is that according to some studies, HCWs do not need to work directly with COVID-19 patients to be at risk of infection. This could mean that SARS-CoV-2 does not necessarily need to infect an individual to travel but can spread on objects and improper handling of human secretions. A study suggests a higher chance of infection spread among co-workers in healthcare setting compared to patient-HCW
This is the second lesson that can be learned from our study. If HCWs are the first population that spread an emergent infection, stricter work rules need to be applied to them as soon as a new infection is discovered in the area. Relationship between HCWs, hand shaking, socializing while eating, and romance between coworkers must immediately stop until the new infections are well studied and under control.

According to the WHO booklet, the 2003 SARS was beaten by “contact tracing, quarantine, and isolation.” In the case of SARS-CoV-2, however, these simple steps would only work in the first wave of the pandemic or even before the wave, when the cases were few and traceable, and as shown in our data and many published articles, mainly limited to HCWs. A 2022 systematic review on contact tracing found evidence for effectiveness of this policy but with low strength, possibly indicating that this policy should be implemented as early as possible to be effective. Furthermore, WHO “no longer” recommends contact tracing for COVID-19 and focuses on population vaccination instead, as this disease is shifting from pandemic to endemic.

One reason that 2003 SARS should not be compared with the new SARS-CoV-2 is the higher case fatality rate and the lower possibility of asymptomatic cases for 2003 SARS. Low possibility of asymptomatic cases increases the chance of excluding the virus to the hospital settings and decreases the transmission of the virus outside hospitals. The long-memory and increased doubling time in our ARFIMA model can be attributed to asymptomatic cases who keep the virus in the population but have less chance of being identified and are less likely to transmit the disease according to one study.

The existence of a long memory process was first explored by Hurst (1951) as a solution for regulating the flow of the Nile River. The Nile River example has since become the most famous example of the existence of a long memory process: long periods of high flow levels were followed by long periods of low flow levels as Hurst observed. Mandelbrot and van Ness (1968) introduced the Hurst parameter (H) to describe the long-term memory of a time-series process. As H gets closer to 1, the more persistent the time series is considered and at values less than or equal to 0.5, the long-memory process does not exist. It is important to note that the Hurst parameter can only be approximated. We used Whittle’s approximation using the maximum likelihood estimation (MLE) approach to estimate the Hurst parameter.

In our study, the “fractal dimension plot” (Figure 2B) shows that while quarantines might help with reducing the spread of the virus, it does not stop the spread due to its long memory process. Figure 2A shows that once the infection has begun, it is not possible to slow it down until the virus has run through its infection period but also puts other members of a community at risk, given its fractal nature. Comparing predicted values to actual values three months later (May–July 2021) showed that our model is a reliable tool to forecast new cases with an average error of ±1.9 cases, making it one of the most reliable prediction tools in this case, compared to other tools. Our prediction only deviates from actual in the ‘far future’ (late June), which is expected from any forecast. The unexpected rise in the number of new cases can also be explained by the rise of the new delta variant.

As we transition from pandemic to endemic, we recommend studies and doctors to consider two things: first consider co-infections; based on what we learned from sexually transmitted diseases (STD), the possibility of having one STD, increases the chance of having other STDs. As SARS-CoV-2 becomes endemic, it is possible for at-risk people to get co-contaminated, especially after massive gatherings (eg festivals). Secondly, consider changing the cut-off and sensitivity of the tests as we transition from epidemic to endemic, especially with serology testing.

**Conclusion**

Studying epidemiology of an emergent infection is more informative early in the spread of the disease. The index case needs to be diligently observed and reported. The pattern of spread would be more revealing early in the course of dispersion when the number of cases is limited and tracing is possible. Healthcare facilities should not wait for a pandemic to start isolating patients with suspicious airborne infections, appropriately apply PPE, and separate the HCWs that service those patients.

**Summary**

Our SARS-CoV-2 dispersion data indicate that an airborne disease rapidly spread among health care workers during the first wave of the pandemic. The ARFIMA model can be used as a valuable forecasting tool to predict the number of new cases a few weeks in advance and assist in efficient planning. Available literature data on the issue must be analyzed with
caution, once results are impacted by the type of test used to confirm infection, timeframe analysis due to its inherent time-related dynamic behavior and worldwide differences among work-culture policies, routine protocols and preparedness for airborne disease control.

Data Sharing Statement
The datasets during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Approval
University of Campinas ethics committee. Number: 4.173.069, consent or parental informed consent obtained per protocol from the parents/legal guardians of study participants under the age of 18 prior to study commencement. The current study complies with the Declaration of Helsinki.

Acknowledgments
To the involved institution(s), the patients, and those that provided and cared for study patients. Patricia AF Leme, Mehrsa Jalalizadeh, and Cristiane Giacomelli da Costa are co-first authors for this study.

Author Contributions
All authors contributed to data analysis, drafting or revising the article, have agreed on the journal to which the article will be submitted, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Funding
Coordination for the Improvement of Higher Education Personnel – CAPES: 88887.506617/2020-00 and National Council for Scientific and Technological Development – CNPq. Research Productivity: 304747/2018-1. The funder had no involvement in study design, data collection, data analysis, manuscript preparation, and/or publication decisions.

Disclosure
The authors declare that they have no competing interests in this work.

References
3. Weber DJ, Rutala WA, Fischer WA, Kanamori H, Sickbert-Bennett EE. Emerging infectious diseases: focus on infection control issues for novel coronaviruses (Severe Acute Respiratory Syndrome-CoV and Middle East Respiratory Syndrome-CoV), hemorrhagic fever viruses (Lassa and Ebola), and highly pathogenic avian influenza viruses, A(H5N1) and A(H7N9). Am J Infect Control. 2016;44(5):e91–e100. doi:10.1016/j.ajic.2015.11.018
4. Berta Suárez Rodríguez B ón DF. Epidemiologic features and control measures during monkeypox outbreak, Spain; 2022.
43. World Health Organization. Strategy considerations for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and other respiratory viruses in the WHO European Region during autumn and winter 2022/23: protecting the vulnerable with agility, efficiency, and trust; 2022.