Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 1 January 2021

Research Article

COVID-19: A Data Analysis Using Regression Analysis

Cheryl Ann Alexander[1]* and Lidong Wang[2]

[1]*Institute for IT innovation and Smart Health, Mississippi, USA*

[2]*Institute for Systems Engineering Research, Mississippi state university, Vicksburg, USA*

***Corresponding Author:** Cheryl Ann Alexander, Institute for IT Innovation and Smart Health, Mississippi, USA.

Received: November 03, 2020

Published: December 09, 2020

Abstract

SARS-CoV-2 is a novel coronavirus that developed in Wuhan, China in the Hubei Province at the Huanan Wet Market, a fresh market that sells and prepares live and exotic animals. The SARS-CoV-2 virus was a moderately fatal, highly contagious disease like that of its sister SARS-CoV. SARS-CoV-2 was initially named the novel coronavirus 2019 by the World Health Organization (WHO); however, in February 2020, the WHO named the viral disease COVID-19. COVID-19 causes a potentially fatal pneumonia among other symptoms. The most common symptoms are fever greater than 38° Celsius, dry cough, shortness of breath, etc. The most interesting symptom of COVID-19 is the inflammatory response, often called the "cytokine storm", which can cause an acute respiratory distress syndrome (ARDS) and cascade failure which can be potentially fatal. Therefore, because COVID-19 is potentially fatal and highly contagious, with very few therapeutics available in the beginning and no vaccine as of today's publication in November 2020, the containment and limitation of the spread of the disease and mathematical projections of cases is critical to mitigating the spread. Big Data analytics can be used to project numbers of cases and using datasets of formerly published early papers, this paper uses a regression analysis to prove the projections of some experts were erroneous. This paper introduces COVID-19, data management systems, and conducts a regression analysis on a formerly published paper's dataset.

Keywords: SARS-CoV-2; COVID-19; Regression Analysis; Big Data; Data Management Systems

Introduction

In December 2019, a new virus causing an outbreak of fatal pneumonia occurred in Wuhan, China in the Hubei Province which is in the South of China. The virus was initially traced to an animal to human transmission from the Huanan Seafood Market of Wuhan, where they sold live and exotic animals. Afterward, China was faced with a virulent disease that was transmitted human to human [1-5]. Some of the symptoms of this pneumonia were fever greater than 38°, fatigue, chills, dypsnea, a dry cough, diarrhea, pulmonary inflammation leading to an interstitial fluid buildup causing Acute Respiratory Distress Syndrome (ARDS), hematologic complications such as microthrombi, inflammation and damage to the gastrointestinal tract and liver [3,5]. Renal impairment, myocardial inflammation leading to a heart attack or myocarditis (i.e., inflammation of the heart's lining), and some have even hypothesized a viral neurotropism, which may last far longer than other viral symptoms [3,6].

This new virus, like the Severe Acute Respiratory Syndrome (SARS) which also started in China in 2003 [4], is a β-coronavirus, which on the phylogenetic analysis and taxonomy prompted the Coronavirus Study Group of the International Committee on Taxonomy of Viruses recognized that this virus was caused by the SARS-C0V-2 virus [7] and was tentatively called the novel coronavirus by the World Health Organization (WHO) [5-7]. By February 12, 2020, the WHO had named this new viral disease COVID-19 [4,6,8]. Although the Chinese government took unprecedented action to control this novel COVID-19, including intensive surveillance, epidemiological investigations, and a total lockdown of Wuhan and

its surrounding cities, the authorities soon found that this was not enough to control the outbreak of the disease. Because the Spring Festival, better known as the Chinese New Year was approaching, and travel both international and within the nation of China would explode, Chinese officials sealed off national travel within China to and from Wuhan, however, international travel from Wuhan was not staunched [4].

As a result of the international travel and trade, the WHO declared SARS-CoV-2 a pandemic on March 11, 2020 [3,4,6,7,9]. However, the pandemic had already spread to approximately 200 countries by March 2020 [7]. More dangerous to the elderly and vulnerable patients with comorbidities, children do not seem to share the excessively virulent spread of COVID-19, although further research is necessary to verify this [6]. By late January, the SARS-CoV-2 virus had spread to Washington state, with the first case in Puget Sound, Washington; the initial epicenter of the US outbreak of COVID-19. By late February 2020, the first US death from COVID-19 had occurred in Washington [10]. In March 2020, the US joined most countries in Europe, Canada, Asia, Russia, Central America, and South America in shutting down everything but essential businesses [11]. In addition to the personal tragedies and global economic losses brought on by this global pandemic, the world has suffered an economic recession brought on by multiple small business closures, large corporate bankruptcies, and furloughs and layoffs of thousands of people starting in March 2020. This dire situation has caused multiple researchers to look for solutions in both medicine and engineering [9].

Database management systems

Most industrial systems use database management systems (DBMS) to control and organize data for the business. Critical to data storage and management, DBMS play an essential role in business globally, providing a series of transaction management mechanisms and ensure logical and consistent data management. The entire DBMS depends upon databases being correct in atomicity, isolation, and temporal correctness. Most conventional methods, however, lack the ability to analyze the interplay of these qualities, and they do not scale well for most systems with a large amount of transactions. Therefore, Big Data analytics is often used to handle complex data [12-15].

With the many technological revolutions ongoing in today's climate, the Internet of Things and connected devices play an important part in daily life for most individuals. Therefore, what amounts to an unimaginable amount of massive data is produced and conventional data methods are simply unable to handle noisy, unstructured, streamed data. The data is representative of a diverse number of events and controls a large amount of activities in daily life. This makes it essential that a capable, accurate, and secure DBMS be available for any computing platform. Data should also be securely and well-managed in situations where IoT is used as cybersecurity is always a consideration [13].

The deployment of additional innovative devices which then connected to a series of objects which are deployed at the edge of the network, can undoubtedly ease most people in to the use of these smart devices for such tasks as air quality monitoring, home fire detection, home door locks, etc. These small devices often produce huge amounts of data that must be analyzed. In addition, with large scale situations it is critical for all devices to work in a sense-process-attenuate control loop, which guarantees a more secure, much better performance. Most of these devices are deployed in the cloud for better security and much better management of massive amounts of data produced by these devices [13]. Computer databases have the capacity to significantly enhance the precision and effectiveness of information organization. This plays a critical role in information management in an organization; however, DBMS can often play a key role in other functions. While conventional DBMS can no longer efficiently handle the vast amounts of data, Big Data analytics is often used to manage data that conventional methods cannot. With an understanding of all the data within a database, which can be an assortment of tables, schemas, figures, text, etc., this differentiates it from a filing system. Therefore, data can be used by many users and shared in many applications also. While artificial intelligence (AI) and artificial neural networks (ANNs) have been used to extrapolate data and process data within a database, the data must be carefully collected and accurate. Using a DBMS fosters accuracy and care when processing collected data [15].

Big Data analytics has been proven a useful set of tools with which to conduct data management, particularly when the data are enormous and complex. Big Data analytics can accommodate structured, unstructured, and semi-structured data and give a real-time glimpse of high-streamed data [14]. In January 2020, when the WHO knew that China had a highly transmissible and contagious virus, SARS-CoV-2, on its hands, data management and analysis became key to predicting the virus movements, predicting mor-

bidity and mortality, and eventually, predicting measures to stop the virus [4]. However, reporting and monitoring systems cannot be cobbled by ineffectual DBMS due to a hurried or mistaken response when dealing with a pandemic. Data collection and more importantly, data analysis must be first and foremost purely done for tracking, management, and relevant decision-making [14]. The Crisis Management Cell was mandated in January of 2020, as part of a mandated COVID-19 response team. Data collection and data repository were to be done using Big Data analytics tools. Functioning non-stop, the world was on a war-like footing with the virus and data manipulation was its key tool for defense [16]. Therefore, the purpose of this article is to review the statistical data collection and statistical outcomes for the target article using the appropriate tools [4,12-16].

Methodology and purpose of the study

COVID-19 has unfortunately brought healthcare all over the world to its knees, and scientists are struggling to find meaningful data in many areas. During the first part of the pandemic, from January 2020 to around May 2020, most countries were just starting on initiating testing, beginning lockdowns of everything but essential businesses, and looking for meaning in the scant amount of data that was coming from China [16-19]. Crisis resilient hospitals in the US, however, were working hard to make models using Big Data tools and GIS tools, plus other AI tools such as ANNs. Engineering and medicine were working hand in hand to find a solution for handling this virus [4]. Disease surveillance is the bedrock behind any pandemic response. The analysis of time, place, person, and distribution of disease is critical for finding existing real-time trends and future trends in multiple subpopulation groups. This led to many attempts by scientists for predicting the disease outcomes, disease mortality, and locations and time periods that were important to the viral disease. But thousands were dying, and scientists had to adjust to finding solutions quickly. Not only did they need to control their operations during the pandemic, they also needed to respond to deviations and learn from those deviations [4,16].

Predicting the mortality and morbidity of COVID-19 patients

Predicting the rate of infection and the morbidity and mortality of infected patients was the primary goal for Liu., *et al*. [4]. However, medicine has, for the most part, focused on several issues related to the understanding of a) symptoms; b) characterizations; c) estimating incubation periods; and d) accurately predicting the number of patients who may be infected with the virus [8]. These numbers are critical in creating and planning for Intensive Care Unit (ICU) bed use and how many infected patients may have to be quarantined within the community. It is more than essential, it is quite critical that the models be right, using the right data, the right formats, and the right DBMS tools. Unfortunately, within the first few months of COVID-19, scientists knew little about this novel coronavirus and many of the estimated models were quite wrong, or the data was wrongly plugged in, or either the wrong tools were used [1,8,10]. Unfortunately for Liu., *et al*. [4], the authors used a methodology that was wrong for the data they had, and it is possible that even the data was wrong at that time. Often during the beginning of COVID-19 many peer-reviewed journals offered quick review and acceptance of any article within the scope of COVID-19. To project the data for their study, Liu., *et al*. [4] created a modified four-stage Susceptible-Exposed-Infectious-Removed (SEIR) Model to capture the evolutionary trajectory of infection in Wuhan, China. However, the Model failed to take into consideration the asymptomatic spread of infection plus some other valuable pieces of data, thus their trajectory was skewed from the start [4,6,9].

Predicted versus real data

Although the predicted data versus the real data should have been close or on target, in the study by Liu., *et al*. [4], the data was off. Either the data was not collected real-time, which will skew the data, or the data was mishandled either in collection or transcription. The authors used t-tests to accurately reflect whether the data was significant. However, as the author of this paper recreates the experiment, it is easy to see the discrepancies within the data collection [4]. However, in medicine, most data can be characterized by its features. For example, most providers and experts want models that are truly accurate pictures of impending infection rates as medicine must be the frontline to care for these patients within an often-restricted number of ICU beds. Therefore, the most common methods of data analysis are mathematical modeling and data analytics [9].

Enhancing the study and limitations

In March, the WHO ordered countries globally to analyze their data, keep accurate records, and develop COVID-19 research centers to stop the spread of this virulent and often fatal disease. In the early months of March, April, and May of 2020, the WHO warned of letting the uncontrolled spread of the SARS-C0V-2 virus could lead to massive numbers of deaths and a global pandemic of a lifetime [20]. Government, however, was focusing on the unprecedented

disruption to the economy, but there was a need for tighter restrictions and more scientists to do COVID-19 research. Global advisories have been to social distance, wear a mask, and sanitize one's hands repeatedly throughout the day. However, it seemed that the WHO could only make these recommendations after the data in Wuhan, China became available, which was late in March. Chinese government officials had been tight-lipped about the pandemic that started in Wuhan and was not willing to share information with the rest of the world [20-22].

Other researchers conducted research into other facets of the viral pandemic once the lockdowns began. Other focus areas arose such as supply chain, how other diseases were affecting COVID-19 or not being treated, and how hospitals needed to quickly realize they may not have enough beds in the US to hold the thousands of critical patients flooding the Northeastern and Northwestern states [2]. For example, several academic organizations banded together to form a department in John Hopkins Medical Center School of Medicine to forecast patient infected numbers using GIS systems. The Center launched its dashboard in late January and attempted to give real-time data analysis to a model set up by scientists at the engineering school from John Hopkins. The "dashboard" was well-known as it was displayed quite frequently on CNN, Fox News, MSNBC, etc. However, it too had problems of its own. It was only updated certain times of the day and there was a delay until the local area data was reported [21].

COVID-19 patients were tested using a PCR technique where the swab goes deep inside the nasal passage until there is an adequate specimen; the drawback to this test is that is sometimes takes as long as 10-14 days to come back in the height of the surge of case numbers as labs became overrun [5]. However, newer tests have proven more effective at identifying those infected quickly [23]. So, patients with symptoms could be counted although a false positive was obtained as data was consistent with test and test method as well as whether the patient was asymptomatic [5]. The Case Fatality Rate (CFR) was a key component of testing, analyzing the data, and the John Hopkins Dashboard. It was important early on to know the CFR so that health officials and government officials could know when they were flattening the curve [24].

Data analytics based on regression and deep learning

Using the database in the Liu., *et al*. [4] paper, the number of newly confirmed cases of COVID-19 before and after adjustment are shown in their copy of table 1 [4]. Data were used after an adjustment. A regression method and algorithm based on the data (i.e., Table 1 of Liu., *et al*. 2020) using *R* Language which is a very useful tool for data analytics. Figure 1 shows the result of the regression. Some data from Figure 1 of Liu., *et al*. [4] are very close to the data in the published paper, but other data are not close at all.

Date	Input	Output
31/1/2020	0	710
1/2/2020	1	1161
2/2/2020	2	1434
3/2/2020	3	1777
4/2/2020	4	2635
5/2/2020	5	2568
6/2/2020	6	2436
7/2/2020	7	3054
8/2/2020	8	2582
9/2/2020	9	3256
10/2/2020	10	3022
11/2/2020	11	2708
12/2/2020	12	2810
13/2/2020	13	4113

Table 1: The new confirmed cases after adjustment.

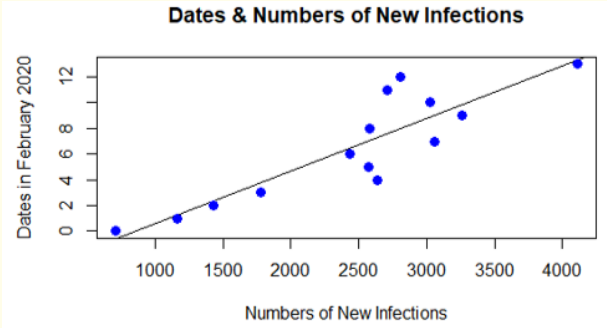

Figure 1: Dates in February 2020 and the numbers of new infection cases.

According to the Liu., *et al*. [4], the forecasted peak period ranges from February 18 to 23, 2020, which is the same as the actual data. However, during that period, there were between 37,000 and 39,000 active infective patients in Wuhan, China.

A regression model was then used to perform the prediction of the number of COVID-19 infection cases. The independent results

are shown in table 2. Considering the newly confirmed cases after adjustment in table 1 (e.g., the total number is 34,266) and the new case numbers (e.g., the total number is around 16,620) from February 14 to 17, 2020, the number of active infective patients in Wuhan from February 18 to 23, 2020 is between 55,516 and 81,508 which is much larger than the numbers (i.e., 37,000 and 39,000) in the published paper.

Date	Input	Predicted Case Numbers
14/2/2020	14	3870.495
15/2/2020	15	4060.218
16/2/2020	16	4249.941
17/2/2020	17	4439.664
18/2/2020	18	4629.387
19/2/2020	19	4819.110
20/2/2020	20	5008.833
21/2/2020	21	5198.556
22/2/2020	22	5388.279
23/2/2020	23	5578.002

Table 2: The new number of confirmed cases based on prediction.

Other data analytics methods were also tried so that a prediction of the newly confirmed case numbers of COVID-19 could be made. Deep learning, specifically, deep neural networks (DNN) based on R Language and the published data in table 1 (i.e., the newly confirmed cases after adjustment) was used at first. A prediction for the new case numbers from February 18 to 23, 2020 was then made. That predicted number for each day during this period was 2447.571, however, this result was no good. Perhaps the reasons lie in the following: a) the data quality in table 1 [4] is no good; 2) the sample size in table 1 is too small, not enough for deep learning; therefore, the result is not good. Figure 2 shows the final situation after running the deep learning algorithm [25-31].

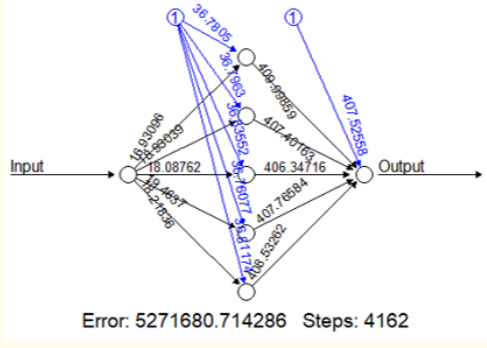

Figure 2: The final situation after running the deep learning algorithm.

Discussion and Conclusion

The published paper shows that "the predicted data are almost the same as the actual data at the first 45 time-series data points. The forecasted peak period ranges from February 18 to 23, 2020 during which the calculations were the same as the actual data. This illustrates that there is no significant difference among the predicted data, the approximate real data, and the actual outbreak data." COVID-19 is very complicated and many things regarding this disease are still unknown. Incubation, asymptomatic cases, possible mutations, etc. have made predicting case numbers a great challenge. The model in the published paper did not include major factors such as asymptomatic situations and effective actions or measures, the modeling and predicted results were exactly the same as the actual data or almost the same as the actual data, which is somewhat unbelievable, especially in February 2020 and the collected data (called actual data in the paper) are not very accurate.

Even if it was in October 2020, predicting the confirmed cases of COVID-19 is still not easy. The following table (Table 3) shows the number of new infection cases in four states in the USA. The data are from TV news or the websites of the four states. This COVID-19 data is tracked daily for personal research purposes every day. Table 3 indicated that it is not easy to predict the number of new infection cases. Also, many people do not go for a test at all although they may have been infected.

Dates in October 2020	MS	LA*	NC	GA
17 (Saturday)	751	N/A	2102	1554
16	1116	863	2684	1701
15	1322	823	2532	1686
14	876	331	1926	1331
13	713	653	1734	1021
12	296	63	1276	937
11	294	1168	1719	1162
10 (Saturday)	957	N/A	2321	1279

Table 3: New infection cases in four states in the USA.

* No data announced on Saturday in Louisiana (LA).

We think there is too much wrong information in some published papers. Many professional journals implemented the fast review and fast publication policy for manuscripts regarding COVID-19. The review and publication process of some manuscripts took a short time (e.g., one or two weeks) before May 2020 and their methods and results are somewhat weak.

Bibliography

1. Capuzzi E., *et al*. "Psychiatric emergency care during Coronavirus 2019 (COVID 19) pandemic lockdown: Results from a Department of Mental Health and Addiction of northern Italy". *Psychiatry Research* (2020): 113463.

2. Das T and Das D. "COVID-19 and economic loss of first phase of (21-Day) lockdown in India". *Space and Culture, India* 8.1 (2020): 21-26.

3. Lansiaux É., *et al*. "Covid-19 and Vit-d: Disease mortality negatively correlates with sunlight exposure". *Spatial and Spatio-temporal Epidemiology* 35 (2020): 100362.

4. Liu M., *et al*. "Modelling the evolution trajectory of COVID-19 in Wuhan, China: Experience and suggestions". *Public health* 183 (2020): 76-80.

5. Yun H., *et al*. "Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 patients". *Clinica Chimica Acta* 507 (2020): 94-97.

6. De Luca D., *et al*. "The EPICENTRE (ESPNIC Covid pEdiatric Neonatal Registry) initiative: Background and protocol for the international SARS-CoV-2 infections registry". *European Journal of Pediatrics* 179.8 (2020): 1271.

7. Guraya SY. "Transforming laparoendoscopic surgical protocols during COVID-19 pandemic; big data analytics, resource allocation and operational considerations; a review article*". International Journal of Surgery* 80 (2020): 21-25.

8. Agley J. "Assessing changes in US public trust in science amid the Covid-19 pandemic". *Public Health* 183 (2020): 122-125.

9. Eltoukhy A E., *et al*. "Data Analytics for predicting COVID-19 cases in top affected countries: Observations and recommendations". *International Journal of Environmental Research and Public Health* 17.19 (2020): 7080.

10. Deeds S A., *et al*. "Leveraging an electronic health record note template to standardize screening and testing for COVID-19". *In Healthcare* 8.3 (2020): 100454.

11. Grote L., *et al*. "Sleep apnoea management in Europe during the COVID-19 pandemic: Data from the European Sleep Apnoea Database (ESADA)". *European Respiratory Journal* 55.6 (2020).

12. Cai S., *et al*. "Statistical model checking for Real-Time Database Management Systems: A case study". 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA): 2019 24th IEEE International Conference On, 306-313. IEEE (2019).

13. Sengupta S., *et al*. "SFDDM: A secure distributed database management in combined Fog-to-Cloud systems. 2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD): 2019 IEEE 24th International Workshop On (2019): 1-7. IEEE.

14. Simsek Z., *et al*. "New ways of seeing big data". *Academy of Management Journal* 62.4 (2019): 971-978.

15. Wannalai N and Mekruksavanich S. "The application of intelligent database for modern information management". 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON): 2019 Joint International Conference On (2019): 105-108. IEEE.

16. Bakht N., *et al*. "Crisis management cell for monitoring COVID-19 situation in Pakistan Armed Forces—a case study". *Pakistan Armed Forces Medical Journal* 70.2 (2020): S620-S628.

17. Centers for Disease Control and Prevention. "CDC launches national viral genomics consortium to better map SARS-CoV-2 transmission". MLO: Medical Laboratory Observer 52.6 (2020): 7.

18. NIAID strategic plan details COVID-19 research priorities. MLO: Medical Laboratory Observer 52.6 (2020): 7.

19. NIH-supported research survey to examine impact of COVID-19 on rare disease community. MLO: Medical Laboratory Observer 52.6 (2020): 7.

20. Proença Caetano A., *et al*. "Development of a Portuguese COVID-19 Imaging Repository and Database: Learning and Sharing Knowledge During a Pandemic". *Acta Médica Portuguesa* 33 (2020): 447-448.

21. Santhanavanich T., *et al*. "Integration of heterogeneous coronavirus disease COVID-19 data sources using OGC Sensorthings API". *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* 6 (2020): 135.

22. Singh J A., *et al*. "The impact of the COVID-19 pandemic response on other health research. World Health Organization". *Bulletin of the World Health Organization* 98.9 (2020): 625-631.

23. Parnham J C., *et al*. "Half of children entitled to free school meals did not have access to the scheme during COVID-19 lockdown in the UK". *Public Health* 187 (2020): 161-164.

24. Verma A., *et al*. "Impact of lockdown in COVID 19 on glycemic control in patients with type 1 diabetes mellitus". *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* 14.5 (2020): 1213-1216.

25. Ayaz C M., *et al*. "Out-patient management of patients with COVID-19 on home isolation". *Le infezioni in medicina* 28.3 (2020): 351-356.

26. Dana Barlow R. "Satisfying the COVID-19 effect". *Healthcare Purchasing News* 44.6 (2020): 12-15.

27. Flaczyk A., *et al*. "Comparison of published guidelines for management of coagulopathy and thrombosis in critically ill patients with COVID 19: implications for clinical practice and future investigations". *Critical Care* 24.1 (2020): 1-13.

28. Fuwape I A., *et al*. "Impact of COVID-19 pandemic lockdown on distribution of inorganic pollutants in selected cities of Nigeria". *Air Quality, Atmosphere and Health* (2020): 1-7.

29. Harrison G. "How will COVID-19 affect emerging technologies?" *Database Trends and Applications* 34.3 (2020): 1-5.

30. Khan H S., *et al*. "Impact of COVID-19 Pandemic associated lockdown on admissions secondary to cardiac ailments in a tertiary cardiac centre of Pakistan". *Pakistan Armed Forces Medical Journal* 70.1 (2020): S342-436.

31. Nachimuthu S., *et al*. "Coping with diabetes during the COVID–19 lockdown in India: Results of an online pilot survey". *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* 14.4 (2020): 579-582.