A data capture model and its associate study on the public web published COVID-19 data

Zhiwei LIANG * ▲
Department of Biomedical Data Science
Stanford University School of Medicine
CA 94305-5464, USA
lzw@stanford.edu

Mathematical Engineering Academy of Chinese Medicine
Guangzhou University of Chinese Medicine
GD 510006,China
lzw@gzucm.edu.cn
(* Contact Author)

Pan ZHANG ▲
Dongguan & Guangzhou University of Chinese Medicine Cooperative Academy of Mathematical Engineering for Chinese Medicine
GD 523808,China
1224228964@qq.com
(▲ joint first author)

Baoyan LIU
Data Center of Traditional Chinese Medicine
China Academy of Chinese Medical Science
Beijing 100700, China
13601180524@139.com

Nenggui XU
Huanan Center of Acupuncture and Moxibustion
Guangzhou University of Chinese Medicine
Guangzhou, GD 510006, China
ngxu8018@gzucm.edu.cn

Lu TIAN
Department of Biomedical Data Science
Stanford University School of Medicine
CA 94305-5464, USA
lutian@stanford.edu

Ying LU
Department of Biomedical Data Science
Stanford University School of Medicine
CA 94305-5464, USA
ylu1@stanford.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313399

***Abstract*— Background and Objective**: The Coronavirus Disease 2019 pandemic situation is remaining severe worldwide. A single outbreak data source is not adequate for comprehensive analyses of the response to the pandemic. Such analyses need to seek proper integration of epidemic data for subsequent statistical analyses. **Methods**: 1) Considering reputations of publishers, activities, public users' accessibility, and retrievable historical data among several platforms, the World Health Organization (WHO), the US Centers for Disease Control (CDC), and Baidu's Real-time Epidemics (BRE) websites were selected as our data sources. 2) Data for 32 weeks until August 15th, 2020, were followed, including the US cumulative confirmed cases (CCCs), cumulative death cases (CDCs), cumulative discharged or cured cases (CD|CCs), daily new infective confirmed cases (DNCCs), and daily new death cases (DNDCs). 3) Estimators for the weekly current active infected confirm cases (CACs) and the weekly COVID-19 fatal rate in the US hospitals (WFRUSH) were derived. Graphic display modules demonstrated the risks associated with demographic data. **Results:** 1) CCCs reached 5,285,546 cases in the US on August 15th, 2020, which initially climbed from the 9th-11th week; the CDCs were 167,546. The fatality rate initially climbed from the 12th-13th week, but fast turned over to decrease from the 18th week, then gradually flattened out near 3.17% till the mid of August 2020. 2) The WFRUSH first rose sharply at the 10th-11th week and started to decline in the 12th week, although there was a repeated smaller fluctuation in the 13th-14th week, during the generally downward process. 3) The US demographic characteristics and CDCs showed that the proportion of fatal cases in the senior Americans (age group over 65) accounted was 78.8%, about 4 (3.83) times the proportions of the other age groups. Supposed the death cases of seniors, directly caused by the COVID-19 rather than caused by the fundamental diseases, the γ value of the seniors, a ratio between the senior CDCs proportion over the senior population proportion was 4.81. Such a γ value for seniors, indicated a much higher fatality risk than other age groups. **Conclusion:** Integrative capture data from the publicly web-published COVID-19 statistics helps extend analyzable data and estimate or derive new-useful indicators CACs, WFRUSH, and γ value for the demographic group. As of the including the working population age of over 45, would have a much higher fatality rate than younger ages. It seemed necessary to study further if these death were caused directly by the COVID-19. Additionally, the African Americans, and male Americans, had relatively higher fatality rates. These high risks require more attention to strengthening health prevention; including the working-age population, even although the WFRUSH as a more appropriate and vital indication becomes stable to a low level after July 2020, meaning the clinical interventions and treatments were improved, or the virus fatality power was declined.

Keywords—
health prevention or clinical treatment interventions; coronavirus disease 2019 (COVID-19) in the US; associate study on the public web data; electronics data capture (EDC)

Project supported from:
1. Seed Fund of Asian Health at Stanford University;
2. World Federation of Acupuncture-Moxibustion Societies;

I. INTRODUCTION

The Coronavirus Disease 2019 (COVID-19)[1] is spreading globally. From the World Health Organization (WHO) released epidemic data as of August 15th, 2020, the total number of confirmed cases worldwide reached 21,026,758, and the total number of deaths reached 755,786, which had made a significant impact on global economic and social development[2]. As an essential factor resource for health and medical management and disease control, sufficient data of the epidemic, and public medical status are particularly important in the prevention and control stage. It still lacks adequate or perfect data, e.g., without retrievable history data, inadequate indicators, or lack of uniform interface or reporting standards among the USA's states, if one just captured data from a single official website among the publicly web-published information. Therefore, we intend to build and utilize methods for epidemic data collection, analysis to make suggestions or auxiliary

© IEEE 2021. This article is free to access and download, along with rights for full text and data mining, re-use and analysis.

II. MATERIALS AND METHODS

A. *Data source and category*

We evaluated the official authorities, the retrievable history data, the comprehensiveness the rapidity, the reputations, the access-able-populations, and the interoperability for different websites and selected appropriate websites such as the World Health Organization (WHO), the US Centers for Disease Control (CDC), and Baidu's Real-time Epidemics (BRE). Then, we integrate their public published COVID-19 data about the US. (Table.1) into our data system.

The data spans 32 weeks due to August 15th, 2020. we captured the COVID-19 data containing Cumulative Confirmed Cases (CCCs), Cumulative Death Cases (CDCs), Cumulative Discharged or Cured Cases (CD|CCs), New Confirmed Cases (NCCs), and New Death Cases (NDCs). NDCs data came from the WHO official website, and the CD|CCs were retrieved from the Baidu Epidemics website. The NCCs, CDCs, as well as the CDCs' demographic characteristics, gathered from the US CDC[6].

Table.1 The comparison for the characteristics of the epidemic data source from web sites

Data plat form	Data characteristics					
	Academic or Administrative Authority [3]	Retrievable history and needed data	Comprehensiveness [4]	Rapidity	Reputations	Accessible-populations, and inter-operability[5]
WHO*	☑	☑	☐	☐	☑	☑
CDC*	☑	☑	☐	☑	☑	☑
BRE*	☐	☑	☑	☑	☑	☑
1M3F	☐	☑	☑	☑	☐	☑
JHU	☑	☐ Δ	☑	☑	☑	☑

Note: * It indicates the websites optioned in this study as the data sources of the COVID-19 situation about the US after integrative comparison.
Δ Compared with other sources, it lacks some retrievable history data which this study needs, such as the Discharged or Cured Cases (CD|CCs).

B. *Data processing and data visualization*

All data is saved to or read from one of the structured texts such as CSV, EXCEL, HTML5 via the data exchanging interface if needed. γ is defined as a matrix of ratios between proportions of COVID-19 death (*d*) in the US over the corresponding proportion in US population (*p*), i.e., γ=d/p. γ was calculated by age group, gender, race and ethnicity,.

Apply Python or Microsoft Excel to plot the figures with the data, and the integrative indicator or deducing indicators data. The indicators contain the total Cumulative Confirm Cases (CCCs, plot #D), the total Cumulative Discharged or Cured Cases (CD|CCs, plot #E), the total Cumulative Death Cases (CDCs, plot #F), the Current infected confirmed and Active cases (CACs, plot #G), the Weekly New Confirmed Cases (WNCFCs, plot #H), the Weekly New Cure Cases (WNCCs plot #I) the Weekly Death Cases (WDCs, plot #J), the Fatal Rate (FR, plot #M), the ratio made from the CD|CCs (plot #E) divided by CCCs (plot #N), the Weekly Hospital Output Flow (plot #O: #I plus #J), the Weekly Fatal Rate in the US Hospital (WFRUSH, plot #P: #J over the sum of #I plus #J), and the zooming in or zooming out the plots for some of the mentioned above, such as the plots #H, #I, #M, #O, #P.

III. RESULT

A. *The overall situation and trend of COVID-19 cases in the US*

As original or deducing indicators and contributed by the captured data, fifteen variables plotted with Python for data visualization. Among the indications and their line graphs, as of August 15th, 2020, the total Cumulative Confirm Cases (CCCs, plot #D) had climbed to 5,285,546, and the total Cumulative Death Cases (CDCs, plot #F) were 167,546[7]. The CCCs initially raised quickly in the 9th-11st week, and accelerated from the 25th week till the mid of August 2020. The current infected confirmed and active cases (CACs, plot #G) were 224,1920, which was the difference between CCCs – CDCs – CD|CCs.

The fatal rate (FR, plot #T for #M) increased rapidly from the 12nd-13th week but converted to decline from the 18th weeks, and gradually flattened out, near 3.17% till the mid of August 2020. The Weekly Fatal Rate in the US Hospital (WFRUSH, plot #V for #P) rose sharply at the 10th-11th week and started to decline in the 12th week, although there were repeated fluctuations in the 13th-14th week, during the generally downward process until wandering at a low level less than 3.2% within the last 7-8 weeks. (Fig.2-3)

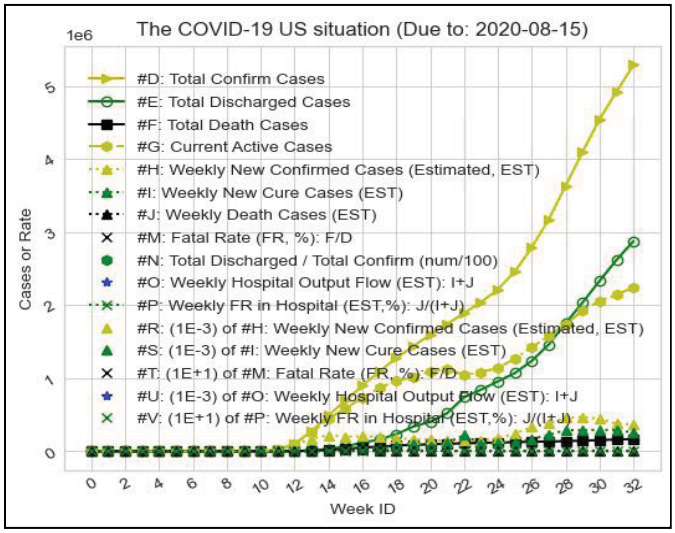

Fig.2 Changes COVID-2019 cases in the U.S. as of August 15th, 2020

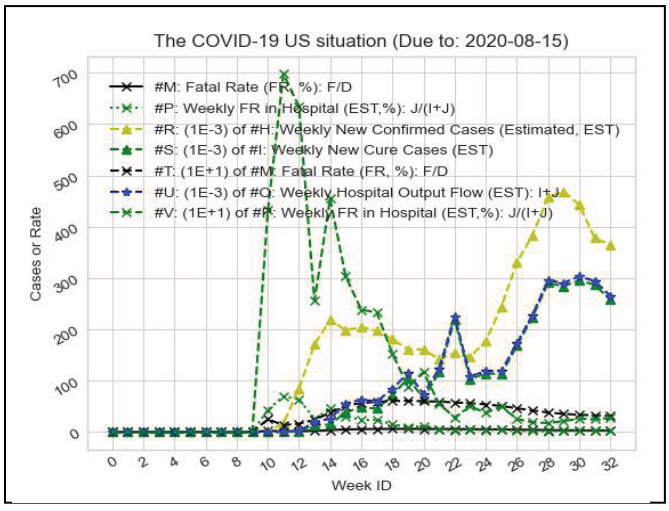

Fig.3 The weekly COVID-10 death cases and the estimated cure cases as well as the weekly fatal rate in the US hospital (WFRUSH, plot #V for #P) as of August 15th, 2020

B. The death ratio distribution in populations

1. The age group's death ratio distribution in populations

As of August 15th, 2020, among 125,686 Cumulative Death Cases (CDCs) in the US, showing in the bar graph, proportions under the age of 1 was 0.01%,1-4 was 0.01%, 5-14 was 0.02%, 15-24 was 0.18%, 25-34 was 0.78%, 35-44 was 2.03%, 45-54 was 5.34%, 55-64 was 12.79%, 65-74 was 21.63%, 75-84 was 26.49%, and the age over 85 was 30.71%. The proportion of seniors (age over 65) was 78.8%, exceeding 3/4th of the total CDCs. (Fig.4)

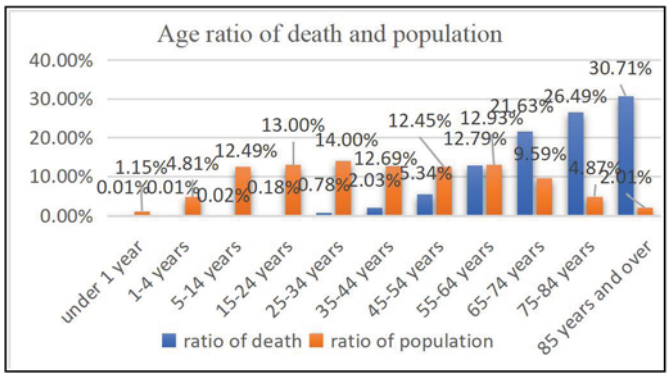

Fig.4 The COVID-19 cumulative death proportions among age-groups in the US compared with the population-ratio among their corresponding groups

The age γ ratios was highest in the age-group 85 and over as 15.26, then the 75-84 age-group as 5.45, the 65-74 age-group as2.256, and less than 1 for age group under 64 years old. The age 45 and over accounted for 96.97% of deaths, whereas this age group accounted for only 41.85% of US population. The γ value for age above 45 was 2.32. Noticeable, because of the unsure present COVID-19 cases accompanied with death due to underlying diseases, more appropriate model and algorithm should compare the relative risk of 2020 COVI-19 mortality rate relative to the historical mortality rate in previous years.

2. The gender group's death ratio distribution in populations

By the mid of August 15th, 2020, 124,840 cases with gender information show that the male CDCs accounted for 54% of CDCs, which is 8% higher than the female proportion of 46%. The proportion of women was 50.8%, and men was 49.2%. The counting values of the former proportion over the latter one was 0.91 (female), also lower than 1.10 (male). (Fig.5)

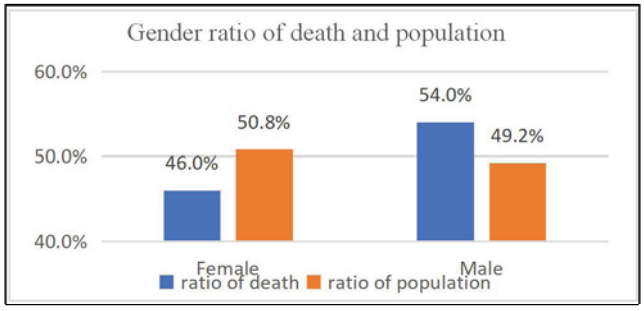

Fig.5 The COVID-19 cumulative death proportions among gender-groups in the US, compared with the population-ratio among their corresponding groups

3. The race or ethnic group's death ratio distribution in populations

The race or ethnic group distributions of CDCs in the US by August 15th, 2020 were Native Hawaiian (0.2%), American Indian (0.8%), Multiple (4.6%), Asian (5.1%), Hispanic or Latino (16.9%), African Americans (22.3%), and White (50.1%); whereas the population distribution for the corresponding groups were Native Hawaiian (0.2%), American Indian (1.3%), Multiple (2.8%), Asian (5.9%), Hispanic or Latino (18.5%), African Americans (13.4%) and White (57.8%). (Fig.6)

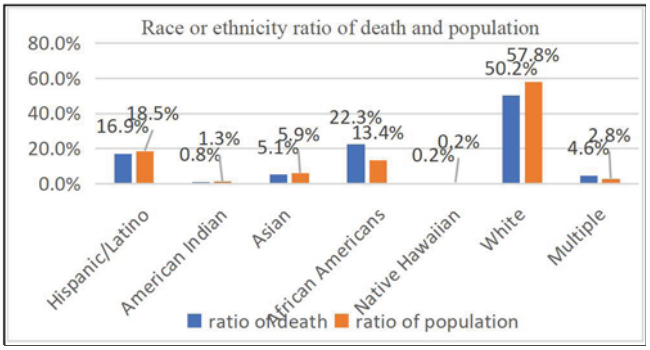

Fig.6 The COVID-19 cumulative death proportions among race or ethnic groups in the US, comparing with the population-distribution among their corresponding groups.

The race and ethnic γ values, from small to large, were 0.62 in American Indian, 0.86 in Asian, 0.87 in White, 0.91 in Hispanic or Latino, 1 in Native Hawaiian, 1.64 in Multiple, and 1.66 in African Americans. The front four groups' values are relatively smaller than 1.00; thus, the left bar shows lower than the right in their group. The Native Hawaiian group's value is 1, and both of the left and right bars are the same at a low level. The rear group's value is larger than one; therefore, the left bar is higher than the right.

IV. Discussion

A. A key indicator to reflect the COVID-19 death in the US

Accurate mortality relies on accurate diagnosis and adequate testing for diagnostics confirmations. Otherwise, a suitable alternative indication needs to be considered as a reference. In the US, the weekly fatal rate in the US hospital (WFRUSH, Fig.4 plot #V for #P) can be an appropriate, valuable and vital indication to reflect the COVID-19 situation or interventions, other than the usual consideration on morbidity and mortality, because of the severe-type COVID-19 individual admitted to the hospital; in contrast, the mild-type (possibly including suspect type) stayed at-home quarantine. The WFRUSH initially rose quickly to a high peak within three weeks and then trended to a low level after the next ten weeks, i.e., after July 2020. Here are some of our considerations for the leading causes but not limit: (a) the clinical interventions or treatments became efficient. (b) the health knowledge, awareness, and culture improved. (c) the virus fatality power declined. Apart from the clinical treatment measurement, the next step study needs to combine the population, region, climate, and environment, with popular trends and potential impacts, to conduct extensively related in-depth mining and analysis; and to explore data statistics, machine learning, and calculation methods with effective models for accurately predicting popular trends[8].

B. Demographics fatality rate distribution and some Asian American health behaviors

(1) The COVID-19 age group death ratio over 45 (97%), over 55 (91.6%,) or over 65 (78.8%) is relatively much higher than those of the other groups. The higher age has, the higher the death-ratio has. Moreover, the death ratio in the age group of over 65 (78.8%) is approximately 4 (3.72) times to the sum (21.2%) of the other groups, and 4.79 (i.e., the age group γ value) times to the corresponding groups' proportion in population (16.5%). Besides, an implication for the age group γ value in the working population age-range 50-65 (even 40-65) is also relatively high, these age-range working population would have a potentially relatively high fatal risk if the death cases directly caused by the COVID-19. The elderly, especially those with

serious underlying diseases, such as heart disease, lung disease, or diabetes, seems cannot be excluded more likely to die due to severe complications of COVID-19; the current data also does not rule out factors such as the resumption of work, resumption of school, family gathering, and other factors[9]. More in-depth studies need to separate the COVID-19 age death distribution being caused directly from itself or indirectly from underlying diseases and compare with the historical population mortality ratio. On the other hand, children's incidence and death need to draw attention, too, although children's mortality proportion(0-4year) was rare (less than 0.1%). In any case, the early death means longer loss of potential life expectancy. Many premature deaths of the laboring population would also negatively contribute to the general domestic production (GDP) index. Thus, to take accurate, timely, and targeted daily protective measures and methods for different risk groups and susceptible groups in response to the epidemic, to a certain extent, will help reduce the fatality rate of COVID-19[10]; thereby relieve negative influence on the average life expectancy of the people.

(2) Females account for more in population than males, but the proportions of COVID-19 cases and mortality were less than that of males, indicating that males have a higher risk of contracting new coronary pneumonia and death than females. Similar trends have been reported in Italy, China, South Korea, etc.[11]. It means that the death rate of men is higher than that of women. Other considerations of possible reasons for the COVID-19 gender differences are factors such as hormone levels, immune function, lifestyle, and socioeconomic status. The mechanism leading to gender differences is not yet exactly understood and needs further research and exploration to provide opportunities for patients' follow-up care or disease prevention.

(3) In the race or ethnic group, the race or ethnicity group γ value implies the higher COVID-19 fatality in the African American and the multiple groups compared to the population census for their distribution. The values are also noticed that the Asians group γ value is 0.86, suggesting that the fatality is relatively low in this population than the other groups' in their census except the American Indian. Such low γ value does not entirely exclude the influence that the traditional health culture or habits, including acupuncture, moxibustion, herbals, or some health exercises such as Tai Chi or Qigong, perhaps more or less influence Asian American individual behaviors. Moreover, some traditional medicine websites show American traditional medicine organizations, such as the American Association of Chinese Medicine & Acupuncture (AACMA), established a particular epidemic advisory group during the pandemics[12]. The organizations took active health culture spread, individual health protection advice on their customers or members, and health managing measurements, even writing to the US CDC to suggest the potential benefits of Chinese medicine and acupuncture to the prevention of COVID-19. Conduct public welfare activities such as Chinese medicine acupuncture and moxibustion free consultation and expand humanistic care. These measures have promoted the spread of Chinese medicine involving acupuncture in health culture, thereby enhancing residents' awareness of behaviors containing daily habits (such as diet, daily exercise, wearing a mask) and social activities (such as avoiding gatherings). On the other hand, the social software application for the health culture and knowledge spread to coping with the pandemic diseases among Asians has increased the prevention awareness on the covid-19. Although traditional Chinese medicine's participation in the prevention and treatment of COVID-19 in China suggests positive effects, traditional Chinese medicine's participation and benefits on preventing, managing, or recovering COVID-19 in the US still needs further practical evidence and studies.

V. CONCLUSION

Integrative capture data from the publicly web-published COVID-19 statistics, helps extend analyzable data and estimate or derive new-useful indicators CACs, WFRUSH, and γ value for the demographic group. As of the 32nd on August 15th, 2020, data shows that the older Americans, including the working population age of over 45, would have a much higher fatality rate than others if the death cases, seemed necessary to study further and make sure, were caused directly by COVID-19. Additionally, the Multiple and the African Americans, also male Americans, have relatively higher fatality rates. These high risks still need to alert more attention to strengthening health prevention; and deciding or arranging appropriately on the working-age population, although the weekly COVID-19 fatality rate in US hospitals, as a more appropriate and vital indication, becomes stable to a low level after July 2020, meaning the clinical interventions and treatments controllable, or the virus leading to the fatalities power declined.

ACKNOWLEDGMENT

Thanks to Professor Cao Shuo from the Department of Physics of Liaoning University for his data analysis suggestions. Thanks to acupuncturists Jun Hu, Xiansheng Huang, Yingqu Wang and members of the American Academy of Chinese Medicine and Acupuncture, who advocated and suggested appropriate information and activities to cope to the epidemics with traditional medicine and acupuncture culture and knowledge particularly in the Asian American group. Thanks, Ph.D. Qian Zhao of the laboratory for their preliminary data discussion, and thanks to Mr. Longdi Li, Ms. Zhu Yinong, and Ph.D. Geng Li from Dongguan & Guangzhou University of Chinese Medicine Cooperative Academy of Mathematical Engineering Academy for Chinese Medicine, for their data collection assistance.

REFERENCES

[1] WHO announced that the disease caused by the novel coronavirus would be named COVID-19. .2020-2-11. https://www.who.int/zh/dg/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020.

[2] Singhal T. A Review of Coronavirus Disease-2019 (COVID-19). Indian J Pediatr. 2020;87(4):281-286. doi:10.1007/s12098-020-03263-6.

[3] Pan D.New data sources and the right to speak: the enlightenment of global new crown pneumonia epidemic reports[J].Education Media Research,2020(04):69-70.

[4] Lu Qiang. Analysis of the application, deficiencies and countermeasures of data journalism in the new crown epidemic report-taking Caixin.com and Dingxiangyuan as examples [J]. News Communication, 2020(02): 15-16.

[5] Liu T,Liu YX ,Zhai C .Application of Data Visualization in Epidemic Reporting — Taking the New Coronary Pneumonia Incident Report as an Example[J].China Media Technology,2020(03):22-27.

[6] Demographic Trends of COVID-19 cases and deaths in the US reported to CDC.2020-08-15.

[7] CDC. Coronavirus disease 2019 (COVID-19) : Cases in the U.S.2020-8-15. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html.

[8] Luo H ,Yang ZY. Improving data governance capabilities in major public health emergencies: Taking the prevention and control of COVID-19 as an example[J].Journal of Xihua University (Philosophy and Social Sciences Edition), 2020,39(03): 45-54.

[9] American Association of Chinese Medicine Holds the Fifth California Government "Acupuncture Day" Free Clinic. https://aacmaonline.com/news.

[10] Pathak EB, Salemi JL, Sobers N, Menard J, Hambleton IR. COVID-19 in Children in the United States: Intensive Care Admissions, Estimated Total Infected, and Projected Numbers of Severe Pediatric Cases in 2020. J Public Health Manag Pract. 2020;26(4):325-333. doi:10.1097/PHH.0000000000001190.

[11] Spagnolo PA, Manson JE, Joffe H. Sex and Gender Differences in Health: What the COVID-19 Pandemic Can Teach Us. Ann Intern Med. 2020 Sep 1;173(5):385-386. doi: 10.7326/M20-1941. Epub 2020 May 8. PMID: 32384135; PMCID: PMC7249504.

[12] Operational Recommendations of Acupuncturists for Prevention and Control of New Coronary Pneumonia by the American Association of Chinese Medicine (Fourth Edition).2020-03-27. https://aacmaonline.com/news.