Известия Гомельского государственного университета
имени Ф. Скорины, № 2 (131), 2022

УДК 314.14:616-036.21(100)

Study of the problems of collecting and analyzing data on the COVID-19 pandemic in different countries

O.V. Pugacheva

The article discusses important features of the statistics COVID-19 distribution around the world. The study is based on the well-known data of operational information on the countries. Particular attention is paid to countries with a large economy. The article shows how great the difference in the formation of data on a pandemic in different countries is. That makes it difficult to compare situations using standard statistical processing.

Keywords: pandemic, demographic statistics, operational information, tests redundancy, excess mortality rate.

В статье рассматриваются важные особенности статистики распространения COVID-19 по всему миру. Исследование основано на известных данных оперативной информации по странам. Особое внимание уделено странам с крупной экономикой. В статье показано, насколько велика разница в формировании данных о пандемии в разных странах. Это затрудняет сравнение ситуаций с использованием стандартной статистической обработки.

Ключевые слова: пандемия, демографическая статистика, избыточность тестов, оперативная информация, показатель избыточной смертности.

Introduction. According to Johns Hopkins University, the number of identified cases of coronavirus in the world exceeded 372,5 million as of 31.01.2022 [1]. Since the pandemic began, 5,658,016 people have died. The leader in the number of cases is the United States (74,235,709 people), followed by India (41,092,522 people) and Brazil (25,256,198 people).

The widely available and fairly detailed daily statistics on the spread of the COVID-19 pandemic across the countries of the world in 2020–2021 had a great impact on human behavior. In Russia, daily statistics came from the Information Center for Monitoring the Situation with Coronavirus (ICC), created in March 2020 to collect and analyze information on the development of the situation with the spread of COVID-19 infection. In Belarus, such data are generated by the Ministry of Health of the Republic of Belarus. Such information has enabled health systems in all countries to promote the necessary culture of behavior in the context of the global epidemic. With the rapid growth of those infected with a new and practically unexplored virus, doctors and all medical personnel made tremendous efforts to provide possible assistance to those who fell ill. In almost all countries, at the initial stage, there was a need for additional equipment and medicines.

Note that about a year before the pandemic was announced, the Global Health Security Index for 195 countries (GHSI) had been published. This Index was created by a solid group of international experts [2]. The Index is the result of a joint project by three organizations: the Johns Hopkins Health Security Center, the Nuclear Threat Reduction Initiative (NTI) and the Economist Intelligence Unit (EIU). The general conclusion of the invited panel of experts is: «The indicators show serious deficiencies in the ability of countries to prevent, detect and respond to disease outbreaks». And, unfortunately, this general conclusion was fully confirmed by the development of the situation with the COVID-19 pandemic.

A detailed analysis of the achievements and mistakes in the struggle for the health of citizens in extreme conditions is the field of activity of experimental scientists and medical practitioners. They gained invaluable experience, which will undoubtedly form the basis for subsequent practical recommendations. In this work only issues related to the statistical description of such processes are considered. There are several questions of the kind.

Registration of infected with COVID-19. The daily registration of the infected, the recovered and the dead, as well as the collection of this information from many medical units and the further publication of the resulting data on the Internet – all this turned out to be quite a difficult task. Technical information problems were largely resolved within a few months. But still, the harsh impossibility of correcting mistakes made in the past when collecting data from various sources sometimes led to the fact that negative values appeared in the reports, while by definition, these values could only be positive or equal to zero. We are talking about the daily number of people infected, recovered and died.

Recall that modern international statistical standards used during a pandemic include daily information on five indicators for countries (cities, regions). One of them – Inf (Infected with coronavirus) – characterizes the total number of people infected with coronavirus on a given day on a cumulative basis. And the other value – $dInf$ – characterizes the number of confirmed cases of infection on the given day, respectively. The relationship between the values of the indicated quantities for two consecutive dates $t-1$ and t is determined by the ratio:

$$dInf(t) = Inf(t) - Inf(t-1) \ .$$

Three other indicators determine the further detailing of the infected into three groups: patients with Ac (Active cases), deceased D (Deaths) and recovered R (Recovered). Of course with this method of keeping statistics, the D value does not mean the number of deaths due to the coronavirus. Obviously, on every fixed day, the identity

$$Inf(t) = Ac(t) + D(t) + R(t),$$

is realized.

Thus, in order to get into the published statistics, it is necessary to obtain the infected status. However, on the recommendation of the WHO, a person receives such a status if his situation refers to the so-called confirmed case, when «there is laboratory confirmation of the presence of COVID-19 infection, regardless of the clinical manifestations and symptoms of the disease». And this means that if a person for some reason did not pass the test, or if he received a false-negative test result, and, unfortunately, died, then he does not fall into the global statistics on coronavirus.

Table 1 contains data for 26 countries [3]: 24 countries belong to countries with a large economy [4], and two countries (Belarus and Israel), as representatives of small countries with approximately the same population. Countries are sorted in descending order by the fourth column – the amount of test redundancy as of 20.05.2021. Here, the redundancy of tests is understood as the value of the ratio of the total number of tests to the total number of registered infected. The last column shows data on vaccination with at least one dose of vaccine as of 25.06.21.

The second and third columns of data are provided to illustrate the economic and human potential of the countries in question. The second column shows the share of the country's GDP in total global GDP for 2019, according to the International Monetary Fund. This data is referred to as wGDP2019 in table 1.

The number of tests among the countries under consideration varies greatly. Of course, the recommendations on the need for laboratory tests to declare the number of infected in general are being followed: as of 20.05.2021, the redundancy of tests lies in the range of 2,9 (Mexico) – 1759,8 (China). The spread in the number of tests per million population is also large: from $5{,}26 \times 10^4$ (Mexico) to $2{,}52 \times 10^6$ (Great Britain).

Table 1 – Country data on the number of tests performed, the number of infected and the percentage of the population vaccinated

Country	wGDP 2019, %	Population, million people	Tests redundancy (Tests / Inf)	Number of tests per million people	Number of infected per million people	% of the population vaccinated on 25.06.21
China	16,29	1404,44	1759,79	$1{,}11*10^5$	64,7	43,21
Australia	1,62	25,97	589,95	$6{,}85*10^5$	1154,9	23,79
Korea	1,90	52,08	70,44	$1{,}84*10^5$	2575,2	29,82
Saudi Arabia	0,87	34,54	41,80	$5{,}16*10^5$	12668,5	no data
United Kingdom	3,24	67,25	38,69	$2{,}52*10^6$	66208,6	64,93
Russia	1,85	146,75	27,06	$9{,}20*10^5$	33900,6	14,82
Canada	1,99	37,82	25,14	$8{,}85*10^5$	35494,1	67,52
Japan	5,93	125,75	19,27	$1{,}05*10^5$	5552,7	20,21
Israel	0,44	8,71	17,30	$1{,}57*10^6$	88296,8	63,98
Belarus	0,07	9,51	16,13	$6{,}49*10^5$	43805,5	7,40
Germany	4,54	82,96	15,99	$6{,}91*10^5$	43788,0	52,95
Italy	2,32	60,72	15,27	$1{,}05*10^6$	68811,9	54,77
USA	24,46	331,81	13,87	$1{,}41*10^6$	101881,7	54,04
France	3,16	65,32	13,86	$1{,}25*10^6$	90590,9	50,47
Spain	1,64	46,86	13,47	$1{,}05*10^6$	77500,2	52,03

End of table 1

Belgium	0,61	11,53	12,92	1,15*10^{6}	90070,6	59,58
India	3,41	1369,56	12,39	2,30*10^{5}	19000,6	18,86
Switzerland	0,81	8,69	10,82	8,48*10^{5}	78958,8	48,53
Turkey	0,81	84,04	10,02	6,05*10^{5}	61404,4	38,22
Indonesia	1,26	269,86	9,00	5,70*10^{4}	6517,8	9,52
Sweden	0,63	10,42	8,99	9,34*10^{5}	101264,2	44,63
Netherlands	1,05	17,29	8,37	7,88*10^{5}	93435,5	54,39
Iran	0,56	84,15	6,61	2,17*10^{5}	33329,0	4,60
Argentina	0,55	45,55	3,73	2,77*10^{5}	74888,3	34,55
Brazil	2,25	211,21	2,98	2,20*10^{5}	74879,0	32,08
Mexico	1,42	127,09	2,88	5,26*10^{4}	18786,0	23,15

Source: number of tests for coronavirus by country: URL: https://covid-stat.com/ru/kolichestvo-testov-koronavirus/; number of infected per 1 million people: URL: https://coronavirus-monitor.info/country/russia/; number of vaccinated from coronavirus in the world, there is no data on vaccination in Saudi Arabia: URL: https://gogov.ru/covid-v-stats/world.

The first three leaders in terms of test redundancy (China, Australia, the Republic of Korea) are also leaders in the lowest number of infected per 1 million population. Japan demonstrates moderate redundancy, however, in terms of the number of infected, it is closest to the aforementioned leaders. For the rest of the countries, there is no clear dependence of the number of infected people on the value of test redundancy. Although statistically, the correlation between the redundancy of tests and the number of infected is nonzero at the 5 % significance level and is -0,39.

Moreover, it is easy to see that the difference between countries in the data on the number of infected per 1 million of the population is large even with close values of test redundancy. These facts suggest that a clear effect on the number of tests is achieved with redundancies greater than 70. Perhaps the cause and effect can be reversed: namely, having achieved a small amount of infection by strict local quarantine measures, countries can afford to conduct a large number of tests.

The constructed mathematical model showed that people begin to be infectious, on average, 2–3 days before the onset of symptoms, and the peak of the ability to infect occurs in the period from 14 to 16 hours from the moment of the first manifestation of symptoms. Further, within a week, the ability to infect the virus decreases until it disappears completely.

These data cannot be considered accurate complete. For example, the moment of onset of primary symptoms was noted retrospectively: the patients themselves spoke about this after they were diagnosed. In addition, patients received treatment, which could affect the concentration of the virus in the mucous membranes, and, accordingly, the data on the rate of its decrease could also be underestimated relative to the drug-free variant of the development of events.

So, we can fix the following factors that make the daily published data on the number of infected, objectively not reliable enough:

– difficulties in organizing daily collecting data from numerous medical organizations;

– rather vague criteria for including citizens who have applied for help in the category of infected with the COVID-19 virus, taking into account the possibility of identifying the virus by tests;

– the existence of a fairly large number of asymptomatic patients who do not even go to doctors.

If the first factor presents mainly difficulties for building a mathematically adequate model according to these time series and ultimately is integrally corrected by correcting fictitious negative values4, then two other factors can lead to a significant decrease in the number of deaths from among those infected with the COVID-19 virus. In addition, these factors make it difficult to estimate the number of people who are already immune and do not yet need vaccination.

In many countries, there have been many such assessments from officials. But it is difficult to find out the true picture of the number of people infected with the COVID-19 virus, as the only official source of this information is the data provided by each country. And this data is practically unrelated to other indicators not included in the operational summary. The situation with mortality data is somewhat different.

Mortality statistics. At the initial stage of the pandemic, the lack of approved protocols for the treatment of a new disease required constant experimentation from doctors and the involvement of various specialists for consultation. The rapid spread of the pandemic to large segments of the population required the mobilization of the entire health system. That led, in many cases, to postponement of routine care for other chronic diseases, which also led to additional deaths that could have been avoided in the absence of a pandemic.

On the Internet, in some countries, you can find a lot of criticism about the published figures on the number of deaths among those infected. At the same time, as a rule, the authors had in mind precisely the fact that the deceased, who had symptoms of COVID-19, did not always fall into the «like» statistics, the openness and publicity of which had no previous analogue. But these new qualities of the «like» statistics made it possible to notice that the number of deaths in these statistics does not always correspond to the statistics of total mortality, which is also available in many (but, unfortunately, not in all) countries.

In many countries, statistics on total mortality are in a weekly or monthly format and are published with some delay. The level of detail of these data varies. Eurostat has adopted a weekly format disaggregated by sex and age. At the same time, official national mortality statistics are provided weekly from 29 European countries or subnational regions within the framework of the joint EuroMOMO (European mortality monitoring) network. This network is supported by the European Center for Disease Prevention and Control (ECDC) and the World Health Organization (WHO) and is hosted by the Statens Serum Institut, (Denmark).

Regular information on total mortality from African countries, as well as from India, Indonesia, China, Saudi Arabia, Turkey, is generally unavailable. In Russia, Rosstat publishes monthly statistics showing reported deaths by cause of death, but with a delay of about 1,5 months. The statistics are presented by regions and three cities (Moscow, St. Petersburg, Sevastopol), but there is no breakdown by sex and age. For 2020, the information on causes of death not clearly related to the pandemic is not provided, but there is monthly information on deaths associated with COVID-19. The information is presented for the country as a whole, for regions and three cities [5].

This statistics include cases from four causes of death:

1) COVID-19, the virus has been identified;

2) COVID-19 is possible, but the virus has not been identified;

3) COVID-19 is not the main cause of death, but had a significant impact on the development of fatal complications of the disease;

4) COVID-19 is not the main cause of death, and did not have a significant impact on the development of fatal complications of the disease.

In the previous section, it was shown that the data on those infected with the COVID-19 virus is difficult to verify. But it turned out that the data on the number of deaths from COVID-19 is not a very adequate indicator, because in different countries the deaths from COVID, and presumably from COVID are taken into account in different ways, even the WHO recommendations on this are somewhat contradictory. In the article by I. Danilova back to the spring of 2020, it was shown that the published data on the number of infected and deaths may not be comparable between countries, as countries use different criteria for both testing for the virus and determining deaths from COVID-19 [6].

There were first assertions that a reasonable indicator in this case could be the amount of excess mortality (the so-called Excess mortality). And now many see it as the only adequate overall indicator. Sometimes it is even called the «gold standard». Various options for finding this indicator have been proposed.

The simplest way to calculate excess mortality is the difference between monthly data on total mortality of the studied year and the corresponding data of the previous year. We denote such a model calculating excess mortality as M1. Consider the situation in Russia for 2015–2021, using data from Rosstat [5]. Figure 1 shows graphs of total mortality in Russia for 2018–2021.

From the data in figure 1 it follows that the total mortality in Russia from April 2020 to March 2021 significantly exceeds the total mortality for previous 2015–2019, and total mortality for 2015–2019 differs little over the years, but has a general downward trend. Let's calculate excess mortality in 2020 in the simplest models as the difference between the values of the total mortality in 2020 and 2019 (model M1) and compare it with daily published data on the coronavirus of the Information Center for Monitoring the Situation with Coronavirus (ICC) for 2020 and detailed data from Rosstat, indicating the impact of the COVID-19 virus on some of the total reported deaths.

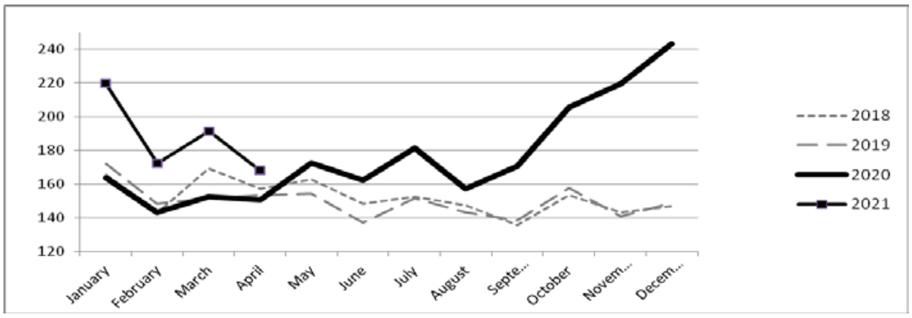

Figure 1 – The total number of deaths in Russia (thousand people) by months for 4 years

Without going into details of the data by month, we will focus only on the resulting data for 2020, given in the line «Total for 2020» in table 2. Excess mortality for the year according to the M1 model is 323,8 thousand people. At the same time, according to the operational statistics of the ICC (the last column in table 2), only 57 thousand people are shown. And this natural comparison gives rise to many unpleasant questions. Of course, the excess mortality data includes all sorts of causes of death, but Rosstat reports on the number of deaths related to COVID-19 (columns 1–4 in table 2 in the line «Total for 2020») somewhat clarified the situation. It became clear that of the 323,8 thousand people included in the excess mortality, only 162,4 thousand can be directly related to COVID-19, as the sum of columns 1 + 2 + 3 + 4.

Table 2 – Excess mortality, mortality with the presence of covid according to Rosstat data and operational statistics of the ICC for 2020 and three months of 2021 (thousand people)

Month	Excess mortality	Cause of death				Sum 1	Sum 2	Operational statistics
2020	M1	1	2	3	4	1+2+3+4	1+3+4	Internet
January	-8,4	0	0	0	0	0	0	0
February	-5,2	0	0	0	0	0	0	0
March	0,7	0	0	0	0	0	0	0,017
April	-2,9	1,35	0,398	0,435	0,642	2,825	2,427	1,056
May	18,3	5,926	1,677	1,609	3,457	12,669	10,992	3,62
June	25,5	5,825	1,492	1,484	3,534	12,335	10,843	4,627
July	29,9	5,063	1,021	1,237	3,05	10,371	9,35	4,643
August	13,8	3,436	0,582	1,184	2,471	7,673	7,091	3,213
September	31,7	4,579	0,859	1,428	3,313	10,179	9,32	3,546
October	47,8	13,077	2,026	1,794	7,436	24,333	22,307	7,268
November	78,5	21,262	3,845	2,288	10,214	37,609	33,764	11,905
December	94,1	25,98	5,57	2,065	10,82	44,435	38,865	17,124
Total for 2020	323,8	86,498	17,47	13,52	44,937	162,429	144,96	57,019
2021								
January	47,319	22,747	4,708	2,337	8,086	37,878	33,17	16,163
February	24,181	14,791	2,329	1,956	5,708	24,784	22,455	12,94
March	39,4	15,003	2,454	1,401	4,857	23,715	21,261	12,728
Total for 04.20–03.21	448,3	139,04	26,961	19,22	63,588	248,806	221,85	98,85

Legend in table 2: M1-model of excess mortality as the difference between the data for 2020 (2021) and 2019 from the data in Figure 1; column 1 – COVID-19, virus identified; column 2 – possibly COVID-19, but no virus identified; column 3 – COVID-19 is not the main cause of death, but had a significant impact on the development of fatal complications of the disease; column 4 – COVID-19 is not the main cause of death, and did not have a significant impact on the development of fatal complications of the disease; column Sum1 – the sum of columns 1–4; column Sum2 – the sum of columns 1, 3, 4; the Internet column is from daily ICC data. All data are in thousand people.

This represents approximately 50 % of excess mortality. However, in the deaths listed in column 2, COVID-19 was not identified and therefore, formally, only the sum of the columns (1 + 3 + + 4) should have been included in the operational statistics, and this is only 144 thousand. But only 57 thousand people are reflected in the operational information.

Two conclusions follow from these comparisons. First, the ICC data were 2,5 times underestimated. And secondly, the prediction of Academician V.M. Polterovich about the possible impact of restrictive measures, expressed by him back in April 2020, unfortunately, was realized: about half of the excess mortality was not directly associated with the virus. And this discrepancy began in June 2020.

If we correlate the numbers M1 and ICC in the line «Total for 2020» in table 2, it turns out that the excess mortality in Russia as a result of the pandemic for 2020 is 5,7 times exceeds the ICC data. It is this fact that determined the title of D. Kobak's article «Excess mortality shows the true losses of Russia from Covid-19», published in the journal «Royal Statistical Society» [7]. The last line in table 2 characterizes the situation with data on deaths for the year from April 2020 to March 2021. During that time excess mortality according to the M1 model is 448 thousand (rounded off). To cases associated with COVID-19, column data can be referred to 1 + 2 + 3 + 4, which is 248 thousand (55 % of the excess mortality). However, only 221 thousand of them had the virus confirmed. But again, according to the data on deaths from the ICC, only 99 thousand cases were confirmed, that is, 2,24 times less. Thus, we can state that the ratio of excess mortality to the operational data of the ICC began to change rapidly after January 2021 and for the annual period from April 2020 to March 2021 was already only 4,5 instead of former 5,7.

There are different methods for calculating the value of excess mortality. The general calculation scheme is based on comparing the mortality rate of the studied year with the predictive mortality model – the so-called baseline. The latter is based on the analysis of mortality for several previous years. Model M1, which was used here for Russia and Germany, assumed that the predictive model of mortality for 2020 is mortality in 2019.

A more sophisticated method for calculating excess mortality in 2020 is based on an analysis of mortality over several previous years with the availability of weekly statistics and a smooth baseline curve for 2020. This method was used by the Institute for Health Metrics and Evaluation based at Washington State University in Seattle [8]. Obviously, the obtained values of excess mortality will depend on the applied calculation method, but most often the results of different methods differ within the calculated error.

Conclusion. Comparing data on today's pandemic in different countries is significantly different from comparing data from various laboratory experiments on which all knowledge in the scientific and technical field of knowledge is based. And this difference is due to the absence of uniform standards for recording data in different countries. And as a consequence of this, great difficulties arise in the analysis of these data by conventional statistical methods, which have proven themselves in the natural and scientific field. Therefore, the use of such methods, with such a wide variety of methods for generating data on a pandemic, can lead to doubtful conclusions.

In this sense, an indicative example is the limited data on the spread of the pandemic and mortality from it in the Republic of Belarus, which is associated with particular difficulties in obtaining reliable statistics on the spread of COVID-19 and the fact that more than a year after the end of 2020, when the coronavirus pandemic began, the Belarusian statistical office did not publish mortality data in 2020.

References

1. COVID-19 map. Johns Hopkins Coronavirus Resource Center [Electronic resource]. – Mode of access : https://origin-coronavirus.jhu.edu/map.html. – Date of access : 31.01.2022.

2. 2021 Global Health Security Index [Electronic resource]. – Mode of access : https://www.ghsindex.org/about/. – Date of access : 01.02.2022.

3. Chetverikov, V. M. Challenges to generating reliable COVID statistics : domestic and international experience / V. M. Chetverikov, O. V. Pugacheva, T. D. Vorontsova // Voprosy Statistiki. – 2021. – № 28 (4). – P. 45–66.

4. Chetverikov, V. M. Unique features and intensity of COVID-19 spread in large economies / V. M. Chetverikov // Voprosy Statistiki. – 2020. – № 27 (6). – P. 86–104.

5. Natural movement of the population in the context of the constituent entities of the Russian Federation [Electronic resource]. – Mode of access : https://rosstat.gov.ru/storage/mediabank/D4uPozCj/edn05-2021.htm. – Date of access : 13.09.2021.

6. Danilova, I. Morbidity and mortality from COVID-19. The problem of data comparability / I. Danilova // Demographic Review. – 2020. –№ 7 (1). – P. 6–26. (In russ.).

7. Kobak, D. Excess mortality reveals covid's true toll in Russia [Electronic resource] / D. Kobak // Significance. – 2021. – № 18 (1). – Mode of access : https://rss.onlinelibrary.wiley.com/doi/10.1111/1740-9713.01486. – Date of access : 03.02.2021.

8. Karlinsky, A. The World mortality dataset: tracking excess mortality across countries during the COVID-19 pandemic [Electronic resource] / A. Karlinsky, D. Kobak // medRxiv. – Mode of access : https://www.medrxiv.org/content/10.1101/2021.01.27.21250604v3. – Date of access : 01.02.2022.

Гомельский государственный
университет имени Франциска Скорины

Поступила в редакцию 01.02.2022