Risks of Data Inconsistency in Information Systems Used for Predicting the Pandemics Development

Volodymyr Bakhrushin[1[0000-0003-3771-5256]], Anna Bakurova[2[0000-0001-6986-3769]], Mariia Pasichnyk[3[0000-0002-5179-4272]] and Elina Tereschenko[4[0000-0001-6207-8071]]

[1,2,3,4]National University «Zaporizhzhia Polytechnic», Zaporizhzhia, Ukraine

[1]vladimir.bakhrushin@gmail.com,[2]abaka111060@gmail.com, [3]mary.pasechnik@gmail.com, [4]elina_vt@ukr.net

Abstract. Predicting the pandemics development is based on mathematical models and empirical data. Prediction errors can lead to ineffective decisions, both in terms of protecting human health and in terms of the economy. In this regard, it is important to prevent the risks associated with the irrelevance and inaccuracy of data contained in the information systems used for forecasting. Experience in predicting the development of COVID-19 pandemic shows that primary data are not always suitable for direct application in mathematical models. One of the problems is the reliability of data on cases and deaths. Different countries have different approaches to their detection and registration, which may also change over time. Another problem is the deviation of real dynamics from the assumptions of the basic models, in particular, due to spatial heterogeneity, changes in quarantine measures and different practices of their observance, and so on. This can result in significant errors in predicting the number of new cases, the number of deaths, the probability and expected parameters of the "second wave", and so on. In this regard, some indicators of pandemic development and possible approaches to eliminate the risks caused with the specifics of the relevant data contained in information systems were analyzed.
The proposed system of measures to identify and prevent the risks of data inconsistencies in information systems used to predict the development of pandemics that could be useful in the development of The Risk-Informed Systems Analysis (RISA).

Keywords. RISA, Information, COVID-19, prediction, risk, data, reliability, accuracy, sources of errors.

1 Introduction

The COVID-19 pandemic was one of the most critical events of 2020, resulting in numerous casualties, significant economic downturn in most countries and other negative consequences. The choice of effective solutions for the response of public authorities to the challenges of a pandemic requires reliable assessments of various risks, which requires relevant models and data. Therefore, to reduce the risks of ineffective health

care solutions, the economy needs to implement evidence-based policies. This requires reliable information on decision-making issues. For COVID-19, such data were almost non-existent at the initial stage. But researches are gradually emerging that could significantly change the prognosis of the pandemic and its aftermath in different scenarios and different strategies to prevent pandemic.

One key issue whose solution is needed to develop effective measures to counteract the spread of the COVID-19 pandemic is to make a substantial increase in the accuracy of forecasts for future morbidity, mortality, social and economic consequences. At the beginning of the pandemic, such predictions were based mainly on relatively simple mathematical models and limited data sets. But over time, new pandemic data were emerging in different countries that could be used to improve models, increase forecast accuracy, and make better decisions.

The difference between this year's pandemic and previous ones is large-scale research on the new SARS-CoV-2 coronavirus, testing of people who can be infected, studying the impact of SARS-CoV-2 on various systems of the human body, as well as social, economic and other consequences. One of the results of this research was the creation of large-scale data collection systems (which also include RISA), which are used to make strategic and operational decisions and recommendations at the level of governments and international organizations. However, over time it becomes clear that individual data are not reliable and relevant, similar data from different countries are not always comparable, available data are not always suitable for direct application in mathematical models used to build pandemic forecasts, and so on. This creates the risk of making big mistakes in forecasts and decisions based on them. Analysis and implementation of measures to prevent risks arising in the current conditions of collection, presentation and application of primary data on morbidity and mortality will significantly increase the effectiveness of strategic and operational decisions to limit the spread of the pandemic.

As part of the system risk analysis, The Risk-Informed Systems Analysis (RISA) helps support decision-making in a pandemic related to economics, reliability and security, provides the use of RISA-tools to quantify projected differences by region, reduce costs by reducing risks.

2 Related Works

Prognostic mathematical models are the basis for understanding the development of the pandemic and making effective decisions to prevent its spread [1]. The first solution to the development of the COVID-19 pandemic was robbed based on the results of predictions based on simple SIR, SEIR models and their modifications [2], which may have been prompted by the supervisors of the system and the extraordinary differential equations. Thus, SEIR models include such groups of people: S - Susceptible (number of people who has not been infected and has no immunity); E - Exposed (number of people who are currently infected, but are not contagious); I - Infected (number of people who are currently infected and are contagious); R - Recovered (number of recovered people who have immunity).

According to these models, the dynamics of daily cases of the disease is described by a symmetrical or asymmetric peak, and the dynamics of the total number of cases has the shape of an S-shaped curve. Such models are still used for forecasting in many countries. It was on their basis that strict quarantine measures were introduced at the beginning of the pandemic [3]. However, they are oversimplified and are only suitable for qualitative analysis under certain conditions. In particular, they do not take into account the heterogeneity of the distribution of the active population, the impact of the demographic structure of the population on key indicators and so on. In addition, such models use a set of constants - basic reproduction number, effective contact rate, recovery delay, as well as empirical data on the number of people who can be infected, the number of infected and the number of those who recovered or died. However, these constants can be estimated only by indirect methods and are in fact quantities that change in time and space, and some empirical indicators are determined with large errors. From the point of view of strategic decision-making, an important disadvantage of these models is that they describe only the "first wave" condition, which is the only one according to such models. But minimizing morbidity and mortality through strict quarantine during the first wave does not answer the question of what will happen after the quarantine is relaxed. And whether the decisions made will remain optimal, given the longer period of time, as well as additional mortality due to stress, limited access to health care and diagnosis, deteriorating quality of life, and so on.

Recently, more complex models [1, 4-6] have been increasingly used to predict the development of a pandemic, which, in particular, can take into account a larger number of parameters and their temporal and spatial changes in computer implementation. However, the parameters of such models are determined by the quality of approximation of empirical data, which increases the impact of errors in these data on modeling results and forecasts. Therefore, even for relatively short-term forecasts, they can provide a scatter of results in 1-2 orders of magnitude higher [7].

One of the main problems is the significant underestimation of real data in information systems on the number of infected. Sample studies for the presence of IgG and memory T cells conducted in different countries, show that the total number of people who were infected and have antibodies to SARS-CoV-2 may be 1 - 2 orders of magnitude higher than the number officially registered cases: [10-13]. This problem is less critical in terms of forecasting the dynamics, unless there is a significant change in policy or scope of testing. But it is becoming very critical in choosing strategies to counter the pandemic and assess the likelihood and scale of new outbreaks. The latter significantly depend on the proportion of the population that is immune to infection. For its realistic assessment it is necessary to know the proportion of people who have already fallen ill and have immunity. It is also important for strategy selection to assess and compare risks, in particular expected mortality from COVID-19, mortality from other diseases, including side effects from pandemics and quarantine measures, and expected social and economic consequences from different solutions. From this point of view, the indicators of Case fatality rate (CFR) and Infection fatality rate (IFR) are important. The first indicator provides estimates of infection mortality based on primary data on the number of reported infections and deaths. Due to these problems, underes-

timation of the real number of infected CFR mortality estimates is significantly overestimated and for most countries is in the range of 1 - 15%. IFR estimates are more realistic. They are usually obtained on the basis of model parameters identification and sample surveys. According to the latest data, the most probable IFR values are in the range of 0.1 - 1%, and according to some data, this value may be less than 0.1% [14-16].

The Working Group on Mathematical Modeling of Problems Related to the SARS-CoV-2 Coronavirus Epidemic in Ukraine of the National Academy of Sciences of Ukraine, the National Academy of Medical Sciences of Ukraine and the Taras Shevchenko National University of Kyiv has developed its own mathematical model to the class of deterministic SEIR-models. It allows to take into account the presence of asymptomatic infected persons, takes into account three levels of complexity of the disease for patients with symptoms and allows for short-term prognosis [17]. However, as with most other similar models, attempts to forecast for a longer period of time (more than 1-2 weeks) lead to a significant increase in the range of uncertain-ty. The group of researchers from the Operations Research Center of the Massachusetts Institute of Technology [18] is also based on SEIR and takes into account the possibility of incomplete detection of infected people, the number of people in con-tact with the infected person during the day, and possible government's and societies actions. As in other similar models, the forecast is based on official data, which makes it sensitive to the relevance and reliability of this data.

Another area of research is the general risk analysis of information systems associated with the COVID-19 pandemic. In particular, [19] states that the main component of risk is uncertainty. The lack and unreliability of information, in particular, has led to inadequate risk assessment and uneffective solutions in China, resulting in the rapid spread of the COVID-19 worldwide [20]. The authors [19] believe that solutions for overcoming the pandemic are too complex and cannot be formalized in the form of certain algorithms. Therefore, they propose to apply adaptive management approaches. Many countries have started to develop their strategies "from scratch", not based on existing knowledge about the development of pandemics, relevant models and experience of countries that have encountered a pandemic in the past [21].

The INFORM collaboration [22] developed the INFORM COVID-19 Risk Index to support decision-making on the allocation of global and regional resources. It assesses the risks of the COVID-19's impact on health and the humanitarian situation that may lead to the need for international assistance. For a country-wide risk assessment, RIKA India has proposed a four-factor model (health, behavior, impact, social policy) [23].

The analysis of the recent studies results shows that to assess the risks of decision-making in a pandemic, the problem of the lack of common protocols for collecting primary information about the global pandemic remains. Different strategies and approaches of different countries to testing, collection, registration of data in information systems make it impossible to directly use primary information according to common models for all countries. There are problems in assessing the effectiveness of quarantine measures in different countries, including due to differences in approaches to collecting primary information.

Thus, the inaccuracy and irrelevance of available pandemic data collected in information systems is a significant risk factor for pandemic fore-casting and decision-making. Therefore, the presented study is devoted to identifying the risks associated with the available data, as well as developing approaches to reduce their impact on forecasting results.

3 Proposed methodology

The research methodology provided that forecasting the dynamics of pandemic development in individual countries/regions can be based not only on the use of classical or modified models, but also on statistical analysis of data available in information systems (total and daily reported cases of infection and death, PCR and ELISA testing, etc.) in those countries/regions where the outbreak started earlier. Despite the differences in the dynamics of indicators due to differences in testing and registration of morbidity and mortality, demographic, social and other differences, etc., statistical analysis of indicators allows to determine the expected range of values at a certain stage of a country's pandemic data relating to countries or regions where these stages occurred earlier. Such analysis can also establish the link between the factors influencing the development of the pandemic and the dynamics of the studied indicators. For the COVID-19 pandemic, the maximum of the first pandemic wave in China was in January, in South Korea in early March, in Spain, Italy, Luxembourg, New Zealand, Norway and a number of other countries in the second half of March. In Ukraine, on the other hand, the maximum of the first wave was in early May, and in many countries of Asia, Africa and Latin America (Brazil, India, South Africa) it was reached in the first half of July or even not reached yet. The short-term forecast of the time and height of the maximum daily incidence for the first wave of the pandemic in Ukraine made in [24], despite the limited range of reference countries available at that time, agreed well with the actual data and forecast of NASU based on mathematical model [25].

To build such forecasts, it is necessary to use large arrays of data. Data on the COVID-19 pandemic are now available in many databases and information systems. In this study, the source data were taken from [26], where the European Center for Disease Prevention and Control (registered cases of infection and death), Our World in Data (official test reports), and the United Nations, World Bank, Global Burden of Disease, Blavatnik School of Government (other data). Official government resources of European countries and the United States, in particular www.cdc.gov, www.cebm.net, www.epicentro.iss.it and others, were used as additional sources.

Analysis of the daily morbidity and mortality dynamics shows that it usually belongs to several typical patterns: symmetrical or asymmetrical isolated peak, peak with a "wide flat top" (plateau), a mixture of several normal peaks or peaks from the plateau. Accordingly, the dynamics of total morbidity and mortality can usually be described as a single S-shaped curve, or the sum of such curves. Based on this, an isolated peak can be considered as the main element of the dynamics of daily cases (Fig. 1). This corresponds to the basic SIR and SEIR models. As the main characteristics for the peak

description can be taken as its date (t_m), height (h), half-width (t_2 - t_1) - the time interval between the dates when the daily incidence was h/2 and asymmetry (t_2 - t_m)/(t_m - t_1).

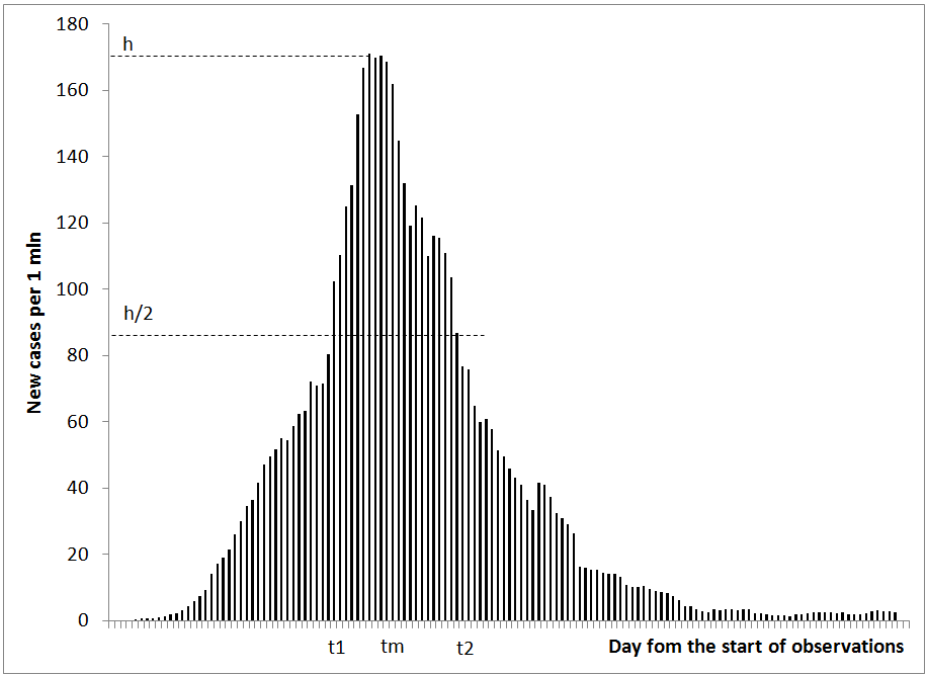

Fig. 1. Daily morbidity peak (Ireland)

31 countries were taken for analysis, where as of July 20, 2020, clear peaks in daily morbidity and mortality were identified. In many countries, these figures are significantly weekly. Therefore, to determine the characteristics of the maxima, smoothing by the moving average method with a 7-day smoothing interval was used. Data on the dynamics of weekly indicators obtained by grouping daily primary data were also used to clarify the maximum position. To ensure data's comparability, all morbidity and mortality rates were used per 1 million inhabitants. However, even with such an adjustment, the use of data for forecasting needs further analysis, as the available indicators relate to the country as a whole and do not take into account the regional distribution. The importance of taking this into account is illustrated by the results for the US states. Here, after the first maximum, which was reached in early April, and the two-month plateau, the second peak of morbidity began. According to the analysis, this trend is due to the non-simultaneous spread of infection in different states. In March-April, the main contribution to the overall incidence was made by New York, New Jersey, Massachusetts and several other states, but in June-July the number of new cases here decreased by 5 - 15 times. Instead, the main contributors are California, Texas, and Florida, where the daily number of new cases has increased by more than an order of magnitude since March. Data on the regional distribution of key indicators were analyzed

in Ukraine in the context of risk analysis associated with the further development of the pandemic.

4 Results and Discussions

As noted, one of the key problems in forecasting the development of pandemics is the incorrect data on the total number of infected and lethal.

The case fatality rate (CFR) commonly used for decision making is obtained by dividing the number of mortality by the number of registered patients, or by dividing the number of mortality by the sum of mortality and the number of patients who have recovered (respectively, lower and upper grades). For countries where the number of active cases is a small percentage of the total number of reported cases, these CFR estimates are close to each other. For example, for China, where the share of active cases on 25.07.2020 is 0.31% of the total, they are equal to 5.53% and 5.55%, respectively. As of February 15, when the share of active cases was 83.8%, and the daily number of new cases was close to the maximum, they differed significantly and were equal to 2.43% and 13.1%, respectively. For the United States, where the share of active cases on April 26, 2020 was 82.1%, these CFR estimates were 5.65% and 31.5%, respectively, and as of July 25, 2020, when the share of active cases decreased to 48.8%, they are, respectively, 3.50% and 6.82%. Both estimates are significantly different for different countries due to differences in testing policies and different stages of development of the COVID-19. Therefore, these CFR estimates can be used to short-term predict the development of a pandemic in a particular country in the absence of changes in testing policies, or to compare countries with the same testing policies. But they are unsuitable for decision-making based on estimates of the true proportion of fatal and severe cases.

Table 1 shows the data on the share of infected people from the total population, obtained from sample surveys of the population.

Table 1. The share of infected people from the total population according to sample surveys

Country	The part of infected by the results of sample surveys [27], %	The share of those informed according to official data at the end of the respective period, the calculation according to github.com, %	The share of infected according to official data on 18.07.2020, calculated according to github.com, %
Austria	4,7 (18 week)	0,18	0,22
Belgian	2,9 – 6% (mid-April)	0,29	0,55
Bulgaria	4,8 (13 – 17 weeks)	0,019	0,12
Spain	5,0 – 5,47 (17 – 19 weeks)	0,57	0,66
Luxembourg	1,97 (17 – 19 weeks)	0,62	0,88
Finland	1,0 – 4,3 (16 – 23 weeks)	0,13	0,13
The Czech Republic	0,0 – 4,0 (18 week)	0,073	0,13

As can be seen from the above data, the number of people with antibodies to SARS-CoV-2 coronavirus is 3-54 times higher than the number of officially registered cases of infection. According to the latest data on the study of memory T cells [8-10], the actual number of infected may be 2-3 times higher. However, even with such an adjustment, only in some regions the share of the population with immunity to COVID-19 today is approaching 50%. In most cases, it does not exceed 5-10%, which makes probable new waves of disease. This assumption is confirmed by a significant increase in morbidity in Bulgaria, Luxembourg, the Czech Republic and a number of other European countries in June-February.

Another approach to estimating the actual number of infected is based on IFR estimates. According to the above data, taking the range of the most probable values of 0.3 - 0.6%, you can get lower and upper estimates of the actual number of cases of infection on 20.07.2020, which are shown in Table 2.

Table 2. Lower and upper estimates of the share of infected in the total population, calculated by IFR,%

Country	Lower estimate	Upper estimate	Relation to the share of infected, calculated by the number of registered cases	
			Lower estimate	Upper estimate
USA	5,4	21,7	4,6	18,4
Brazil	4,7	18,7	4,7	18,9
India	0,3	1,0	3,1	12,3
Spain	7,6	30,4	11,6	46,2
UK	8,3	33,4	19,2	76,8
Italy	7,3	29,0	17,9	71,7
Germany	1,4	5,5	5,6	22,5
France	5,8	23,1	21,6	86,3
Sweden	7,0	27,8	9,1	36,3
Belgium	10,6	42,3	19,2	76,7
Ukraine	0,4	1,7	3,2	12,6
Netherlands	4,5	17,9	14,8	59,3
Poland	0,5	2,2	5,1	20,3
Armenia	2,7	11,0	2,3	9,3
Switzerland	2,8	11,4	7,3	29,3
Moldova	2,1	8,5	4,1	16,3
Serbia	0,7	2,7	2,8	11,3
Austria	1,0	4,0	4,5	18,1
Czechia	0,4	1,7	3,3	13,1
Denmark	1,3	5,3	5,8	23,1
Bulgaria	0,5	2,2	4,3	17,1
Finland	0,7	3,0	5,6	22,3
Luxembourg	2,2	8,9	2,5	9,9
Hungary	0,8	3,1	17,3	69,0

The given data generally correspond to the estimates given in Table 1 according to the data of sample surveys. They also confirm the above conclusions about the significant

underestimation of official data on cases of infection, even during applying the lower estimates. At the same time, even using the upper estimates, it can be concluded that there is a high risk of new outbreaks in most of these countries. However, such a conclusion can be significantly adjusted in view of the following circumstances.

Firstly, according to [8], immunity to COVID-19 can have not only individuals who have relapsed into the COVID-19 infection. It can also be found in people who have previously had SARS or other coronavirus infections.

Secondly, from the data [28-29], it follows that collective immunity can be formed at significantly lower than 50-70% of the infected population due to the heterogeneity of the system.

Thirdly, the analysis of the available data in the information systems shows that in recent months the CFR for the COVID-19 has decreased significantly. This may have various explanations, some of which are related to the decrease in IFR. Also, as noted above, available IFR estimates may be significantly overestimated. Then, even the above estimates obtained using IFR, in some cases may be significantly lower than the actual level of morbidity.

The IFR estimates, that mentioned above, are based on sample studies or identification of pandemic spread patterns. Another approach to estimating IFR can be based on the analysis of the distribution of CFR values. It can be assumed that the lowest CFR values will be observed in the countries with the highest proportion of infected persons. Therefore, they will be closest to the real IFR values. Analysis of the available data shows that for different countries, the CFR can vary from a few hundredths of a percent to more than 10%. Data for some countries are shown in Table 3.

Table 3. Indicators characterizing the development of the COVID-19 pandemic for some countries (as of 25.07.2020)

Country	Number of cases per 1 million inhabitants	Number of deaths per 1 million inhabitants	CFR	Number of tests per 1 million inhabitants
USA	12830	448	3,50	158610
Great Britain	4387	673	15,3	210452
Italy	4062	581	14,3	106994
Qatar	38691	58	0,15	165494
Sweden	7819	564	7,21	74353
Oman	14430	70	0,49	56851
Ukraine	1437	36	2,51	21437
Singapore	8435	5	0,06	199896
Iceland	5399	29	0,54	353657

As can be seen, there is a large variation in CFR values, due primarily to different approaches to identifying infected individuals. In countries whose strategies provide for the most complete identification of such individuals (Iceland, Singapore, etc.), the CFR is in the range of 0.06 - 0.5%, which can be taken as an empirical upper estimate of IFR. This estimate is also obviously somewhat inflated, as in no country does the test

cover all patients, but it confirms the conclusion of other studies that the IFR does not exceed a few tenths of a percent. It should be noted that actual IFR values may change over time due to improvements in treatment protocols, and may vary significantly between countries due to different demographics and different capabilities of health systems. These factors must also be taken into account when forecasting the development of a pandemic, in particular, when applying data from one country to another.

In fig. 2 presents data on the distribution by Ukraine and United States regions of officially registered cases per 1 million people.

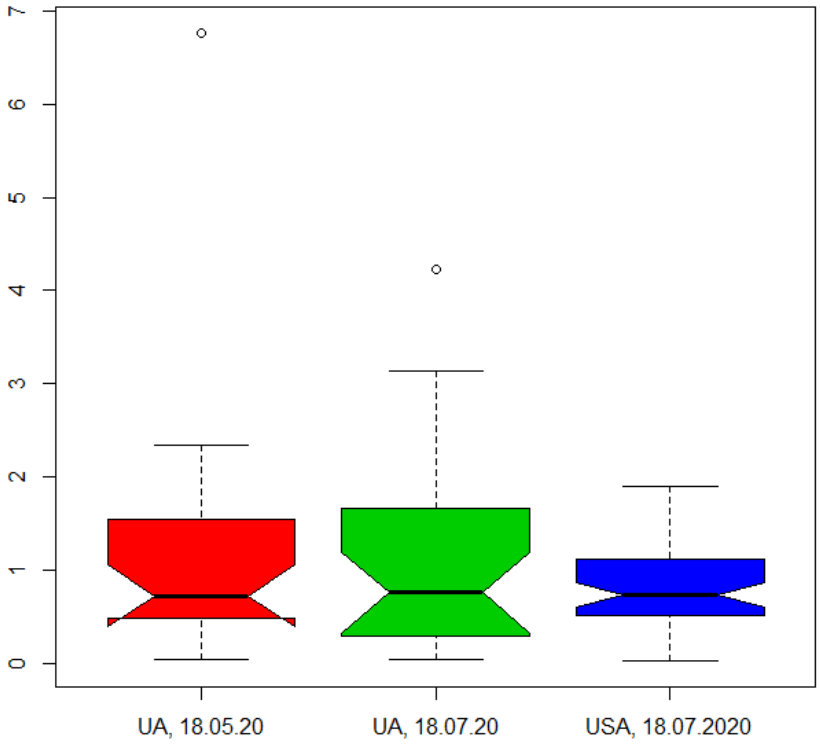

Fig. 2. Distribution by Ukraine and USA regions of the number of officially registered cases per 1 million people in relation to the average levels

These data indicate a significant heterogeneity in the regional distribution of morbidity in both countries. For example, on July 18, 2020, the maximum value for Ukraine exceeds the minimum value by 119 times, and the third quartile exceeds the first by 6.2 times. For the United States, similar ratios are 94 and 2.2. This indicates the incorrect use of averages to predict the further development of the pandemic based on models such as SIR, SEIR. In addition, the average level of officially registered morbidity in Ukraine as of July 18, 2020 is about 1.4, and in the United States - about 11.6 people per 1 million inhabitants. Even taking into account the fact that the number of tests per 1 million inhabitants is 7.2 times higher than in Ukraine, this gives grounds to conclude

that there is a risk of a significant increase in morbidity in Ukraine, which may mainly occur at the expense of regions where today the incidence rate is the lowest.

As noted above, the main element that can be used to describe an outbreak of a pandemic are the peaks in daily morbidity and mortality. Analysis of github.com data allowed us to identify 31 countries where clear peaks can be identified. This work did not take into account countries, in particular France, where data were repeatedly corrected by formally attributing large numbers of unaccounted cases (or subtracting erroneously credited cases) to certain dates, as well as some other countries for which there are doubts about the reliability or the reliability of statistical data, in particular, due to the low incidence rate as of 20.07.2020. Of these 31 countries, only cases were considered for Liechtenstein, as all data on mortality in information systems are formally assigned to one date.

The analysis shows that the data on peak heights and the total number of infected and dead at the dates tm and t2 per 1 million inhabitants have a significant (within 3 orders of magnitude) variance. This is due to different testing policies and sometimes the introduction of quarantine measures. In some countries, such as South Korea and Thailand, small domestic sources of infection have been rapidly tracked and isolated. Therefore, even on 26.07.2020 in these countries the number of officially registered cases is about 0.028% and 0.0047%. In Qatar, on the other hand, the number of reported cases exceeded 2% of the total population during the first outbreak.

However, the available data analysis in the information systems indicates a significant correlation between the individual parameters of the peaks. For example, for registered cases, the coefficients of determination for the linear model are equal to: 0.99 for the relationship between the total number of cases on the date t2 and tm, 0.76 for the relationship between the daily and the total number of cases on the date t2, 0,44 for the relationship between the time of outbreak (the number of days between the date when the total number of cases was 30 people per 1 million inhabitants, and the date tm) and the half-width of the peak. For fatalities, the first two figures are 0.86 and 0.73, respectively.

Instead, the link between similar peaks in morbidity and mortality is much weaker. In particular, for the heights of the corresponding maxima, it is equal to 0.16, for the total number of cases on the dates of the corresponding maxima - <0.01. This may be due to the fact that the absolute data on the number of registered cases deviate significantly from the actual number of infected. However, their understatement is significantly different in different countries due to different approaches to testing and case registration. Therefore, mortality data are more reliable for predicting the dynamics of pandemic outbreaks than data from reported cases. To estimate the total number of infected, which is important for estimating the likelihood and extent of new outbreaks, more reliable estimates can be obtained using mortality data and IFR estimates based on sample data than on the number of reported cases.

Table 4 shows the statistical characteristics of some more stable indicators of morbidity and mortality.

Table 4. Distribution quarters of separate parameters of peaks characterizing daily numbers of new cases and dead

Quarter	Growth time	Half-width	Asymmetry	The ratio of the total number of cases on the date t_2 and t_m
New cases				
min	-3	13	0,5	1,37
0,25	14	20	1,0	1,65
0,50	18	24	1,42	1,82
0,75	26	31	2,0	2,23
max	82	62	4,8	3,20
Dead				
	Number of days between peaks of mortality and infection	Half-width	Asymmetry	The ratio of the total number of cases on the date t_2 and t_m
min	-6	7	0,27	1,23
0,25	4	18	1,06	1,70
0,50	9	28,5	1,5	2,00
0,75	13	35	2,3	2,34
max	28	52	6,3	5,83

As can be seen from these data, the variance between the values of the given indicators of peak morbidity and mortality is much (1-2 orders of magnitude) smaller than the variance between the absolute values of the indicators. This makes it possible to use them to more accurately predict the development of a pandemic.

5 Conclusions

The results indicate that there are significant risks associated with data contained in information systems used to predict the development and decision-making of the spread of the COVID-19 pandemic. This risk reasons are:

1. Systematic errors in the primary data concerning the registration of cases of infection and mortality from corona-viral infection. Data on cases of infection are significantly underestimated, which affects the risk assessments of new outbreaks of the pandemic. Mortality data can be both underestimated and overestimated. This affects IFR estimates, but in any case they are more reliable than infection data. Because of this, estimates of general morbidity obtained by indirect methods may be more relevant.

2. Significant regional heterogeneity of cases of infection, which affects the possibility of their direct application as parameters of mathematical models of pandemic development, which leads to an increase in the risk of significant modeling errors. To reduce this risk, it is necessary to use models that take into account the available heterogeneities and empirical data, for individual regions, rather than the whole country.
3. Wrong strategic and operational decisions that can either increase mortality from coronavirus infection due to insufficient countermeasures, or increase the risk of negative social and economic consequences, including increased mortality due to pandemic and quarantine stress, complications of chronic diseases, limited access to medical care, etc. To reduce such risks, it is necessary to develop special optimization models that use more powerful information systems that contain verified data not only on epidemiological indicators, but also other data needed to correctly assess the socio-economic consequences.

To reduce these risks, it is necessary to adjust the data used in predictive models, in particular through the use of more reliable data on lethal and severe cases to estimate the number of infected, as well as estimates of the number of infected on the basis of sample studies of immunity in the population. The second method to improve forecasts and improve the efficiency of decisions made on their basis is the using of statistical estimates based on the use of information systems data on similar indicators of countries where the pandemic is similar, but significantly ahead of the country for which it is made forecast. It is important to increase the accuracy of forecasts to take into account in mathematical models the heterogeneity of pandemic development by region and the use of regional data in modified models.

The work was made by non-governmental organization "system research".

References

1. Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo & Marta Colaneri (2020) Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. https://www.nature.com/articles/s41591-020-0883-7 [Accessed 16 August 2020].
2. José M. Carcione, Juan E. Santos, Claudio Bagaini and Jing Ba (2020) A simulation of a COVID-19 epidemic based on a deterministic SEIR model. https://www.frontiersin.org/articles/10.3389/fpubh.2020.00230/full [Accessed 16 August 2020].
3. Christina Atchison, Leigh Bowman, Jeffrey Weaton, Natsuko Imai, Rozlyn Redd, Philippa Pristera, Charlotte Vrinten, Helen Ward (2020) Report 10: Public Response to UK Government Recommendations on COVID-19: Population Survey, 17-18 March 2020 https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-03-20-COVID19-Report-10.pdf [Accessed 16 August 2020].
4. Kaihao Liang (2020) Mathematical model of infection kinetics and its analysis for COVID-19, SARS and MERS, Infection, Genetics and Evolution, 82, doi: 10.1016/j.meegid.2020.104306.

5. Stelios Bekiros, Dimitra Kouloumpou (2020) SBDiEM: A new mathematical model of infectious disease dynamics, Chaos, Solitons & Fractals, 136, doi: 10.1016/j.chaos.2020.109828.
6. Muhammad Altaf Khan, Abdon Atangana, Modeling the dynamics of novel coronavirus (2019-nCov) with fractional derivative, Alexandria Engineering Journal, 2020, doi: 10.1016/j.aej.2020.02.033.
7. Forecast of the COVID-19 epidemic in Ukraine in the period April 20-27, 2020 (in Ukr) http://files.nas.gov.ua/PublicMessages/ Documents/0/2020/04/200422180841445-9406.pdf Accessed 16 August 2020.
8. Le Bert, N., Tan, A.T., Kunasegaran, K. et al. SARS-CoV-2-specific T-cell immunity in cases of COVID-19 and SARS, and uninfected controls. Nature (2020). doi: 10.1038/s41586-020-2550-z
9. Li, J., Wang, J., Kang, A.S. et al. Mapping the T cell response to COVID-19. Sig Transduct Target Ther 5, 112 (2020). doi: 10.1038/s41392-020-00228-1
10. Yang, L. T., Peng, H., Zhu, Z. L., Li, G., Huang, Z. T., Zhao, Z. X., Koup, R. A., Bailer, R. T., & Wu, C. Y. (2006). Long-lived effector/central memory T-cell responses to severe acute respira-tory syndrome coronavirus (SARS-CoV) S antigen in recovered SARS patients. Clinical immunology (Orlando, Fla.), 120(2), 171–178. doi: 10.1016/j.clim.2006.05.002
11. COVID-19 Antibody Seroprevalence in Santa Clara County, https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v1.full.pdf [Accessed 16 August 2020].
12. Up to 2.7 million in New York may have been infected, antibody study finds, https://www.nbcnewyork.com/news/local/new-york-virus-deaths-top-15k-cuomo-expected-to-detail-plan-to-fight-nursing-home-outbreaks/2386556/?fbclid=IwAR0i-J3TQ3idewt47akwVCQWkQU-AE4SOH0AExtM2koOYh3iLjS3W199MPg [Accessed 16 August 2020].
13. Spread of SARS-CoV-2 in Austria, https://www.sora.at/uploads/media/Austria_COVID-19_Prevalence_BMBWF_SORA_20200410_EN_Version.pdf [Accessed 16 August 2020].
14. How deadly is the coronavirus? Scientists are close to an answer, https://www.nature.com/articles/d41586-020-01738-2 [Accessed 16 August 2020].
15. Gideon Meyerowitz-Katz, Lea Merone (2020) A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates, doi: 10.1101/2020.05.03.20089854
16. Global Covid-19 Case Fatality Rates, https://www.cebm.net/covid-19/global-covid-19-case-fatality-rates [Accessed 16 August 2020].
17. Counteraction COVID-19, http://www.nas.gov.ua/EN/Activity/covid/Pages/wg.aspx [Accessed 16 August 2020].
18. Michael Lingzhi Li, Hamza Tazi Bouardi and other (2020) Overview of DELPHI Model V3 – COVIDAnalytics https://www.covidanalytics.io/DELPHI_documentation_pdf [Accessed 16 August 2020].
19. David Adam, Special report: The simulations driving the world's response to COVID-19. Nature 580, 316-318 (2020), doi: 10.1038/d41586-020-01003-6
20. Noah C Peeri , Nistha Shrestha, Md Siddikur Rahman, Rafdzah Zaki, Zhengqi Tan, Saana Bibi, Mahdi Baghbanzadeh, Nasrin Aghamohammadi, Wenyi Zhang and Ubydul Haque. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? International Journal of Epidemiology, 2020, 1–10 doi: 10.1093/ije/dyaa033

21. Yuri Bruinen de Bruina, Anne-Sophie Lequarrea, Josephine McCourta, Peter Clevestigb, Filippo Pigazzanic, Maryam Zare Jeddid, Claudio Colosioe, Margarida Goularta (2020) Initial impacts of global risk mitigation measures taken during the combatting of the COVID-19 pandemic, doi: 10.1016/j.ssci.2020.104773
22. Inform COVID-19 Risk Index Version 0.1.2 - Results and Analysis (17 April 2020). https://reliefweb.int/report/world/inform-covid-19-risk-index-version-012-results-and-analysis-17-april-2020 [Accessed 16 August 2020].
23. Ranit Chatterjee, Sukhreet Bajwa, Disha Dwivedi, Repaul Kanji, Moniruddin Ahammed, Rajib Shaw (2020) COVID-19 Risk Assessment Tool: Dual application of risk communication and risk governance, doi: 10.1016/j.pdisas.2020.100109
24. Bakurova, A., Pasichnyk, M., Tereschenko, E. & Bakhrushin, V. (2020) Data Analysis and Predicting of COVID-19 in Ukraine, doi: 10.13140/RG.2.2.35305.11369.
25. Forecast of the COVID-19 epidemic in Ukraine in the period April 20-27, 2020, (in Ukr), http://files.nas.gov.ua/PublicMessages/Documents/0/2020/04/200422180841445-9406.pdf [Accessed 16 August 2020].
26. COVID-19. Public data. https://github.com/owid/covid-19-data/tree/master/public/data [Accessed 16 August 2020].
27. Immune responses and immunity to SARS-CoV-2, https://www.ecdc.europa.eu/en/covid-19/latest-evidence/immune-responses [Accessed 16 August 2020].
28. Weitz, J.S., Beckett, S.J., Coenen, A.R. et al. Modeling shield immunity to reduce COVID-19 epidemic spread. Nat Med 26, 849–854 (2020). doi: 10.1038/s41591-020-0895-3
29. Tom Britton, Frank Ball, Pieter Trapman. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. Science 369, 846-849 (2020). doi: 10.1126/science.abc6810