Pandemic and official statistics; some comments on recent COVID-19 experiences

Olav Ljones
Retired former Deputy Director General Statistics Norway, Norway
E-mail: ola-lj@online.no

Abstract. The fight against the COVID-19 pandemic is dependent on quality information and statistics; various empirical data play an important role in the global fight against the virus. There are however some quality problems and lack of international comparability of the data in use; this comparability may be improved. This comment is inspired by the paper by Len Cook and hopefully will supplement his paper by adding some additional topics.

Keywords: Pandemic, official statistics

1. Preface

Nearly all countries have, during the winter and spring 2020 experienced the pandemic caused by a CORONA virus and the COVID-19 disease. In many countries a rather dramatic and sudden development with many people hospitalised, patients in respirators and rapid increases in number of deaths from COVID-19 have been observed. The social and economic consequences are also dramatic with a rapid increase in unemployment and reduction in national income.

In this time of a global crisis we notice that data and statistics are important for the description of what happens. Political decisions with wide consequences are taken based on statistics. The key figures for the prevalence and deaths from COVID-19 are usually sourced from epidemiological institutes. National Statistics Offices (NSO) may deliver background data. The NSO may also take responsibility for maintenance of important parts of the infrastructure that is needed for the overall data system.

In many countries it is a long tradition that NSOs cooperate with other governmental agencies in the collection of data and publishing official statistics. A well-organised system for civil registration and vital statistics, CRVS, that includes data on births, deaths and in and out migration is a crucial part of the infrastructure for official statistics. Such a CRVS is beneficial for both administrative purposes (e.g. health and epidemiology) and statistical purposes. For planning purposes and the operation of a health statistical system, basic facts of the population size and composition will be crucial. To secure a well maintenance of these systems, it is advantageous that the system includes a register of identities of all inhabitants, the ID system.

The statistics on persons tested, infected, and under care (including under intensive care) are in many countries undertaken by health authorities and based on medical and administrative criteria and in general little use is made of official statistics. In some countries, statistics published by the epidemiological institutes may also be labelled as official statistics.

2. Observing the epidemic, the first infected and deaths

The disease and diagnosis COVID-19 are new. This is a special challenge for official statistics. It is also perhaps fair to say that the outbreak came as a surprise. The pandemic in Wuhan China in January 2020 was observed by the World Health Organisation (WHO) and media informed the public.

Different strategies to fight the pandemic were presented to the politicians. Two main strategies came out. The differences in the words used for labelling the

two main strategies: may exaggerate the real difference since there are overlapping elements and in all strategies different actions are combined, The real differences are not so clear and distinct as the labelling may indicate.

– Knock down (of the virus) by a closing down of many social and economic activities in society supported by legal regulations. Strong regulations for 3–6 month and softer regulations for even some more than 12 months
– Restrain, reduce the spread of the disease by some closing down of activities and a more frequent use of recommendations than legal instruments. Unclear duration.

An important element in the choice of strategy is to avoid that the need for treatment in hospitals exceeds their capacity. It appears that there are different attitudes from governments as to whether government recommendations are sufficient or whether it is necessary to use legal instruments.

The director of the WHO, Dr Tedros Adhanom Ghebreyesus, recommended to perform testing and collecting data. He said in a TV appearance that to fight the pandemic without data is as stupid as to fight a fire blindfolded.

There are some cornerstones in a statistical description of the pandemic. If we lack these statistical cornerstones, the whole statistical construction may be unstable. One such cornerstone is the number of infections and another is the number of deaths caused by the COVID-19.

It is difficult, or perhaps impossible, to diagnose COVID-19 without a proper medical test. The test program (priority list for whom to test) will very often be based on epidemiological criteria. One strategy can be based on testing persons who feel sick, or to give priority to testing of vulnerable groups. Another group to focus on in such a strategy is key medical staff and other people on other key positions.

The statistician will however recommend that, since it is important to know and monitor the number of infected individuals in the total population, a recommended strategy should be to include in the test strategy a survey based on a random sample of the population. Since it may be important to test some priority groups by medical criteria it will be useful to combine such a random sample with a sample of tests taken from medical priority groups.

3. Pandemics – a shock or an expected event?

The definition of a pandemic is an epidemic that is observed in more or less all countries. With interna-

tional relations growing stronger, e.g. in trade and economic development, tourism and migration, the spreading of virus and antibiotic resistant bacteria has become an international issue and consequently needs also to be regarded as a global challenge. The global consequences from the COVID-19 pandemic are very visible (e.g. world wide dramatic increases in unemployment and poverty).

On the national level, the epidemic seemed to enter society as a surprise, even when many countries had several weeks to prepare themselves for the outbreak. When times comes to evaluate the way countries/governments reacted on the outbreak, a topic that needs to be included in such evaluation is how these weeks before the real outbreak came were used for preparations.

There are also several examples of more general warnings to the international community for future pandemics to the international community. An important report is "A world at Risk Annual report on global preparedness for health emergencies" prepared by the Global Preparedness Monitoring Board. The Board was convened in May 2018 with Ms. Gro Harlem Brundtland (former Norwegian Prime minister and Director General of the WHO) and Mr Elhadj As Sy (Secretary General Red Cross) as vice chairs.The founding for this report were provided by the WHO and the World Bank. The report of this Board was released in September 2019 and warned about future pandemics. A clear expression of this warning is the title of one of the chapters: "Preparing for the worst: A rapidly spreading, lethal respiratory pathogen pandemic." The report is important, however, it is not easy to understand it as a warning and to identify what actions that should be taken to prepare for a pandemic. Furthermore, the report is rather silent on the necessity of improved epidemiological surveillance and statistics. A document that helps to better understand the setting for the 2019 report is the WHO document: "International Health Regulations" (2005). This technical document contains in one chapter guidelines and recommendations about reporting and exchange of information about some diagnoses especially those that are classified as contagious and can end as epidemic or pandemic. This reporting of statistics on diagnostic information is important for early warning systems for pandemics. Based on the current experiences it is clear that a further discussion on how to improve the dissemination of such important statistics is a real necessity. Relevant questions in this context are if such reporting by epidemiologist could benefit from improved cooperation with official statistics? Surely, at

the moment this is not in the scope of this WHO report and the term *statistics* or *official statistics* is not at all used in this document.

4. How should we prepare for a pandemic?

Norway observed the first incidents of COVID-19 in March 2020 soon at national level the limited access to key medical equipment like respirators, equipment for testing and protection outfits, could be observed. The result of a nearly complete close-down of international aviation was that the supply of many types of such very needed equipment stopped.

It became also very quickly after the outbreak clear from interviews and other media messages that in many countries there were no strategic stockpiles of such equipment in private and public ownership or a reduced maintenance of such stocks. Considering this stockpiling to be of a too high economic cost. Just in time' delivery in a production chain seems nowadays to be the overarching principle in industry and production. It was easy to observe that when the close-down of international air transport appeared the direct and indirect effects on stocks of a variety of goods were enormous.

It is doubtful whether official statistics would have been able to reliably monitor the size and content of such international strategic stockpiles of medical equipment. It might be that such statistics are considered too sensitive to have official status.

5. Statistics and comparisons

Even when international comparability is a common objective for official statistics, perfect comparability is difficult to achieve. For example in statistics on the causes of deaths we observe that national adaptations give some differences in how causes of death are measured. Some countries only counted deaths that occurred in hospitals while others have performed a more complete count.

Regarding statistics on the numbers of people infected with COVID-19 it seems the only way will be to have reliable statistics will be based on medical testing. Even traditional medical consultations will give insecure diagnoses and unreliable statistics.

As argued in paragraph 2, it is difficult to estimate the total prevalence of the COVID-19 infection from most national test data, since these tests are not based on a representative (probabilistic) sample of the total population. The absence of this statistical cornerstone makes other empirical work e.g. to estimate the lethality from COVID-19, complicated.

A key statistic for the pandemic is also the number of deaths caused by COVID-19. For this rather basic count there are various practices between countries and confusing presentations of the figures. For example, in some countries the national figures only cover deaths at hospitals, in other countries people that decease in their private home or at a nursery home, are included. In short, there is an urgent need to improve the implementation of international classifications and standards on diseases and causes of deaths

Since it is difficult to improve the system for coding and statistics – and errors will always appear – another method to compare is by calculating the total number of people dying during the COVID-19 season with the number of deaths in a normal season. Such comparison will of course be valuable, but, because of the rather wide seasonal intervals, it is not to expect that this comparison may replace statistics based on cause of deaths statistics.

Lack of comparability is a result of low international consistency and differing practices in the coding of causes of death. This coding is complex and requires to be done by competent and experienced staff. An often occurring challenge for an accurate diagnosis is the fact that there are combinations of diagnoses on the death certificate. Causes of death statistics do not only serve the COVID-19 analysis, From the perspective of multiple diagnosis's it may be advisable to follow standard procedures, e.g. to mark the immediate cause of death as well as the underlying cause. For the monitoring of epidemics and especially COVID-19, it could be relevant to do counts of all people that die with the diagnosis COVID-19 even if it is not the underlying cause of death. It seems as some countries follow this path.

Official statistics and government epidemic surveillance will be produced by distinct institutes. Patient data that are collected for epidemic surveillance, will be administrated for other objectives and with other principles than those governing statistical registers. The division of work and cooperation between statistical offices and the epidemic administration may vary between countries. In Norway registers with individual patient data are under the responsibility of the health sector and the data are legally owned by health authorities. This authority produces the official statistics from these data sources. The traditional role for the NSO is for what concerns health information limited to produce health surveys based on personal interviews of the

population and statistics on resources that are spent in hospital sector

What is important for arriving at high quality health statistics, is to find an overall organisational model for health registers and other registers – including statistical registers – that combines respect for protecting individual data and the merging of data sources as far as possible. In this exercise, it is clear that the main objective for epidemic authorities will differ from those from official statistics. For epidemic authorities it is important in some cases also to identify infected (individual) people and isolate them to avoid further spreading of the epidemic. The objective for official statistics is to describe the overall picture of the pandemic. How many are sick, the social and demographic composition of the infected and so on.

One international observation is the role of nursing homes in the outbreak of this specific pandemic. In the first phase the main concern in media was towards hospitals, especially focussing on the patients in intensive care including those who receive respiratory treatment. The main reason for recognising this subgroup was the observation of the limitations in strategic stockpiles of respiratory equipment. At a later phase it became clear from the figures that over 60 percent of the deceases from COVID-19 occurred in nursery homes. These patients seemed to have no or only little access to intensive care. There is little or no statistics to describe the medical treatment the elderly at nursery homes receive.

A relevant concept linked to pandemic is immunity: the proportion of individuals in a population that are *immune*. Immunity will normally be gained in two different ways, after vaccination or after infection and having recovered from the illness. For clarification on immunity specific medical tests are developed. Immunity is important for the dynamics of an epidemic and the level and development of immunity after illness may play an important role for the development of the epidemic.

6. Analytical concepts

The identification of vulnerable groups, Comorbidity. For epidemics it is of great interest to identify groups with a relatively high probability to be infected and to identify those groups that have a high probability for fatal outcome. For a statistician this identification seems to be an interesting task. The success in identifying vulnerable groups will benefit from support from medical expertise, and moreover, the other way around, the identification of the vulnerable group is important to support the design of optimal epidemiological strategies. Vulnerable groups may be asked to follow specific quarantine rules. Population groups with a relatively high rate of infection may also be asked to keep distance from the members of vulnerable groups. The vulnerable may also be a separate priority group as regards access to testing for COVID-19.

For COVID-19 the most important characteristic to identify vulnerability is age. The more complex question is if age makes COVID-19 a very serious disease or if it is the combination with suffering from other diseases Medical research indicates that diagnosis as lung problems, heart disease, diabetes, high blood pressure etc. will make a patient more vulnerable for the disease. Such results of additional diagnoses are important for how the public should interpret the advice on distancing etc. Comorbidity is in general defined as a situation where two or more diseases are observed simultaneously, and it is of course of interest to include COVID-19 in analysis of comorbidity. Several empirical analyses of single risk factors have been performed but it seems as a good strategy to do multivariate analysis of age and other diseases to see properly the importance of the combination of various diseases and age.

Another dimension that has been introduced in statistical analysis is the difference in prevalence between various *immigration groups* (Previous immigrants by country of origin). Questions about to what group of population a person belongs are sensitive and need qualified statistical methods to avoid misinterpretations for example on the relation between the migrant status and being infected with the Covid 19 virus.

The magic R – reproduction coefficient. To understand the dynamics of an epidemic, mathematical epidemiological models and the reproduction coefficient R are introduced. "R" is the number of individuals in a population infected through one infected individual from the same population. R can only be measured directly based on a complete chain/cluster of infected individuals. It is reasonable to assume that R depends on the level of contagiousness of the COVID-19 virus including the length of the contagious period of the virus and the start of the symptoms of contagiousness. How many persons one infected person will infect in average will also depend on characteristics like number of persons in the household with whom the person is in close contact , type of work and e.g. commuting practice

For epidemiology it is important to understand the dynamics of an epidemic. How does it spread, the increase in number of infected people, how immunity may develop, how many become severely ill and will

need treatment including respiratory treatment and how many will die? Intuitively the coefficient R is important. When $R > 1$ there will be a period with exponential growth in the number of infected. If $R < 1$ we will observe a decline in the number of infected people.

The mathematical model of the epidemic that is used in projections, should in an ideal world have parameters, such as R, that are based on real and actual data.

In Norway it seemed that the first model that was referred to in the media was the international model published in a paper by the Imperial college Covid-19 Response Team (2020). The modelling exercise included results from Denmark, Italy, Germany, Spain, UK, France, Norway, Belgium, Austria, Sweden and Switzerland. The time series data of the estimated R with confidence intervals shows differences in the initial level of R between the countries. For all countries it was observed that the R declined after various epidemiological policy measures had been implemented

At the beginning of the outbreak the use of the R values from this Imperial College paper had a significant influence on the international debate and public understanding of the dilemmas in the war against COVID-19. Later the focus became more based on national R estimates. In Norway these were based on a national model from the Norwegian Institute of Public Health (Folkehelseinsituttet, 2020). This model is now used for regular updating of the estimates of R. When it was observed that R was at the level of 0.63 for several weeks the Norwegian Government choose to loosen some of the strict anti-corona regulations.

The Folkehelseinsituttet model is complex. It is not easy presently to find a complete documentation of the model. One challenge for the calibration of the model will be that we don't have access to reliable data for the number of infected people. Data for performed test and results, hospitalised patients and deaths are however available. The techniques for estimating the model parameter is said to be a mechanical Bayesian technique. It seems as the number of infected is estimated indirectly via data on hospitalised people and also the number of deaths. Some model parameters are included without an empirical base. The estimated parameters, like R, will probably be sensitive to the choice of these parameters, It is obvious that the quality of estimates has to be examined. The model results are presented with estimates on the variance on the estimated parameters. These estimates on variance are very clearly explained in the presentations of the model.

The present situation on the understanding of COVID-19 by the public at large can be described by a high degree of common understanding of the relevance of the R. At the same time we observe in several countries that model documentation is limited and the professional debate about the quality of the model estimates is growing. It seems however that even insecure estimates of R play an important role for decision makers and the public debate.

Another important concept is immunity. This is a medical situation where people will not get sick even when they are exposed to the virus. It is a complex medical concept. It is common for such virus caused disease like COVID-19 that after recovery the individual will have gained immunity. Before a vaccine is available this is the only way of becoming immune. Details about immunity and COVID-19 seem not yet fully clear.

"Herd-immunity" is another important concept. It involves the notion that an epidemic will stop after a while and retreat. This will happen when the number of immune people is so high that it reduces the probability for an infected individual to infect another person. One strategic question will be whether is realistic to gain herd immunity without entering a situation with so many infected and sick that it exceeds the capacity in the health system.

To explain some elements of herd immunity some model concepts are introduced.

> N = the total population N is divided in four subgroups. S = susceptible, I = infected, H = recovered and immune, D = dead

When the COVID-19 starts, group H is zero. The spread of the COVID-19 is measured by the R – how many individuals one infected individual will infect. Since in most countries individual histories on how people infect others are not available it is difficult to estimate the R. Instead use is made of data for I, data for D or (or the number of hospitalised), and then the model is fitted to the data and trends in data.

With the increase of the number of recovered and immune people the probability that an infected individual will *not* meet a susceptible individual is growing. The effective R will be lowered. By simple arithmetics the proportion of immunity can be calculated that will bring the efficient R (R(eff)) down to 1. R (0) is the initial R. The growth of infections will stop when R(eff) is down to 1. The arithmetic task is to calculate the H/N – the proportion that are immune when we gain $R = 1$. We include as an example the initial R (0) = 3. Then the equation is 1 = 3(1–H/N) and the solution is that when initial R is 3 the proportion of immune that is needed to stop the growth of the epidemic is 2/3.

One warning from epidemiologist will be that it may be difficult to gain 2/3 of the population with immunity without vaccine and without a possible conflict between the number of patients and the capacity of the health system.

7. The economic and social consequences

The immediate shocks caused by the COVID-19 pandemic in society with increased unemployment and reduced income are enormous. In all countries that are affected there is a dramatic economic recession foreseen. An important task for official statistics will be to monitor these consequences. A key indicator for the economic consequences will be the level and change in employment and unemployment including the temporarily laid off from the workforce. It will be beneficial for a country to have access to register data for employment and unemployment, Such data are often published with little time lag and may measure changes rather precise. Such statistics should be followed by calculations of the effect on the economy by estimation of the effect on national income. When monthly national account are prepared on a regular basis, it is close to a normal procedure to estimate the loss in national economy. A difficult task is however to project how long the recession will last and when the recovery will start.

8. Conclusions

For those who are interested in statistics, and particularly official statistics, there are many interesting examples of how important figures and statistics are relevant for the public debate and also for the creation of policies to fight the pandemic. During the last couple of weeks many epidemiological estimates and figures have grown to be familiar and commonly used concepts. The most important actor in all these discussions on the pandemic will be Health Authorities and their epidemiological expertise. The national statistical offices play a rather modest role.

We also observe differences between countries when key decisions are taken by the politicians at the top level or if key decisions are delegated to experts in epidemiological analysis in health administration. In Sweden the key decisions are taken by the epidemiological top experts while in Norway the key decisions are taken by the Cabinet. A lot of comparison exercise is done between neighbouring countries and surely at a later stage one topic will be assessed – who are the best for epidemiological decisions – politicians or experts.

It is obvious that in the aftermath of this crises it will be important to analyse what went wrong in the reaction by society and what needs to improved. For statisticians, including official statisticians, a point to start to study the lack of quality estimates of the number of infected individuals. There seem to be weak points in the systems for statistics on cause of deaths. Also improving the cooperation between health authorities and national statistical offices, at national and international level will need a priority treatment.

The consequences of the lock-downs are hard to meet for common people. Confidence in governments is a necessity for all democracies. The statistical description of the pandemic is an effective measure for improved common information. An important objective for international statistical activities in the next years must also be improving the international comparability of many figures.

References

[1] Folkehelseinstituttet (The Norwegian Institute of Public Health) (2020): Situational awareness and forecasting, Appendix A in a report (The title and main chapter of this report is in Norwegian only): COVID-19 epidemien: Kunnskap, situasjon, Risiko og response I Norge etter uke 14). Published by The Norwegian institute of Public Health April 5, 2020.

[2] Imperial College UK, the Imperial college COVID-19 Response Team (2020): Estimating the number of infectious and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. DOI: 10.2556177731.

[3] The Global Preparedness Monitoring Board.: A world at Risk Annual report on global preparedness for health emergencies. Geneva World Health Organization 2019,

[4] World Health Organisation: International Health Regulations (2005) Third edition.