African Journal of Teacher Education and Development
ISSN: (Online) 2958-0986

The importance of understanding large data, context, conventions and uncertainty in a pandemic

Authors:
Max Stephens[1]
Djordje M. Kadijevich[2]
Janelle C. Hill[1]
Mayamiko Malola[1]

Affiliations:
[1]Melbourne Graduate School of Education, Faculty of Education, University of Melbourne, Melbourne, Australia

[2]Institute for Educational Research, Belgrade, Serbia

Corresponding author:
Max Stephens,
m.stephens@unimelb.edu.au

Dates:
Received: 14 Aug. 2022
Accepted: 15 Sept. 2022
Published: 17 Oct. 2022

How to cite this article:
Stephens, M., Kadijevich, D.M., Hill, J.C. & Malola, M., 2022, 'The importance of understanding large data, context, conventions and uncertainty in a pandemic', *African Journal of Teacher Education and Development* 1(1), a7. https://doi.org/10.4102/ajoted.v1i1.7

Copyright:
© 2022. The Authors. Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

Read online:
Scan this QR code with your smart phone or mobile device to read online.

Background: The coronavirus disease 2019 (COVID-19) pandemic has provided rich data displays informing the public about the spread of infection, risks for certain population groups and the effectiveness of vaccines. These data sources offer opportunities for students and teachers to explore and discuss data of high relevance to their lives and their communities.

Aim: This article argues that in the teaching of statistics and probability, greater attention needs to be given to understand the three key elements of statistical literacy, namely context, conventions and uncertainty. The article also identifies several key areas linking theory and practice.

Setting: This article draws on different data displays using COVID-19-related websites internationally; nationally in Africa and South Africa; and locally in the State of Victoria in Australia.

Methods: By investigating and analysing different data displays, the article shows the importance of assisting students to understand context, data conventions, uncertainty and risk–benefit to understand COVID-19 data. The article examines pertinent 'frontier' areas in the teaching of probability and statistics.

Results: The article identifies important opportunities and challenges for the teaching of statistics in schools and for teacher education, including greater attention to frequentist expressions of probability, risk–benefit analysis, the importance of time series analyses and critical approaches to the evaluation of available data sets.

Conclusion: For schools, greater attention needs to be given to the different conventions by which data are expressed, including the use of dynamic dashboard representations.

Contribution: The article shows how available COVID-19 data can be used to enhance students' statistical literacy and enrich teacher education.

Keywords: probability; statistics; statistical literacy; data visualisations; dashboards; teacher education; COVID-19; pandemic.

The importance of COVID-19 from a teaching perspective

The ongoing pandemic of coronavirus disease 2019 (COVID-19) has had a huge impact on every country since the beginning of 2020. It has provided many opportunities for teachers and students to examine the progress of the course of COVID-19 and to look behind the daily reports. Teachers at all levels of school have necessarily been involved in explaining to students what the virus is about, how it spreads and its impact on school life. Many websites have been set up to record the data relating to daily infections, vaccinations, hospitalisation over time and the number of deaths because of complications arising from COVID-19. Several of these will be explored later in this article, showing how students and teachers can be assisted to make sense of the data and how it is presented.

This article takes two approaches to utilising current COVID-19 data and the teaching of probability and statistics in schools. The first approach, which we see as applicable across the compulsory years of school, is where teachers help students to use their available statistical knowledge and skills to make sense of and to explore the COVID-19 data. This exploration is intended to deepen students' mathematical knowledge and to provide them with the means, according to their age and stage of school, to understand what the data mean, to draw conclusions from the data and to distinguish fact from opinion. The second approach relates to the intersection of theory and practice and will contain recommendations for teacher education and teaching in the senior years of high school. Both approaches underline the importance of assisting students and

teachers to make sense of and to move confidently in a data-rich environment (Boaler, LaMar & Williams 2021).

Uncertainty and context

Although many components of the contemporary mathematics curriculum are intended to have applications across other school subjects and outside of the classroom (Australian Education Council 1991), two key aspects of statistics and probability, namely uncertainty and context, connect the application of these two curriculum areas to the outside world (Callingham, Watson & Oates 2021). Internationally and in every country, these two features, namely uncertainty and context, are experienced in how we think about the current COVID-19 pandemic (Watson & Callingham 2020). These authors remind us that uncertainty associated with chance events and the confidence associated with decisions in contexts where statistics have been collected differ from the rest of the Mathematics curriculum, which is based on facts and theorems. We are further reminded that context is essential to any meaningful data that are collected (Cobb & Moore 1997), and the entire statistical problem-solving process is based on anticipating, acknowledging, accounting for and allowing for variability in these data (Bargagliotti et al. 2020). In applying their statistical knowledge to pressing societal issues such as the spread and control of COVID-19, students need contextualised skills and understandings to apply to each step of their investigation.

Teaching and learning to make sense of the data

In many national curricula, teachers are encouraged to draw on real-life data to support the teaching of probability and statistics. These trends and the reasons for them have led to a richer definition of contemporary statistical literacy, as advocated by Boaler et al. (2021) and Callingham and Watson (2017). There are many advantages in having teachers work with their students on one major website. This allows students to become familiar with the ways in which data are presented and to track changes in the development of the pandemic over time. However, this practice also allows teachers and students to build their abilities to interpret and analyse data and to become critical of the kind of information that is displayed on a particular website in terms of its context, clarity, comprehensiveness and reliability. More importantly, they can identify where additional data are needed or how the presentation of data could be improved. These are important foundations for developing a scientific approach to the data being used and to examine the credentials of those who provide the data.

In many cases, the data will be provided by government or semigovernment agencies, such as universities, but they may also include other groups, such as media, including local newspapers which are familiar to students. Students and teachers should look at the scientific credentials of these sources and should expect to see a focus on presenting data in various forms with a minimum of opinion and speculation and to be alert to occasions when other people are presenting opinions that are not found in the data. After being accustomed to using one website, students and teachers may branch out with greater confidence to look at other websites locally and internationally. What will become apparent are the different conventions for describing and analysing the data.

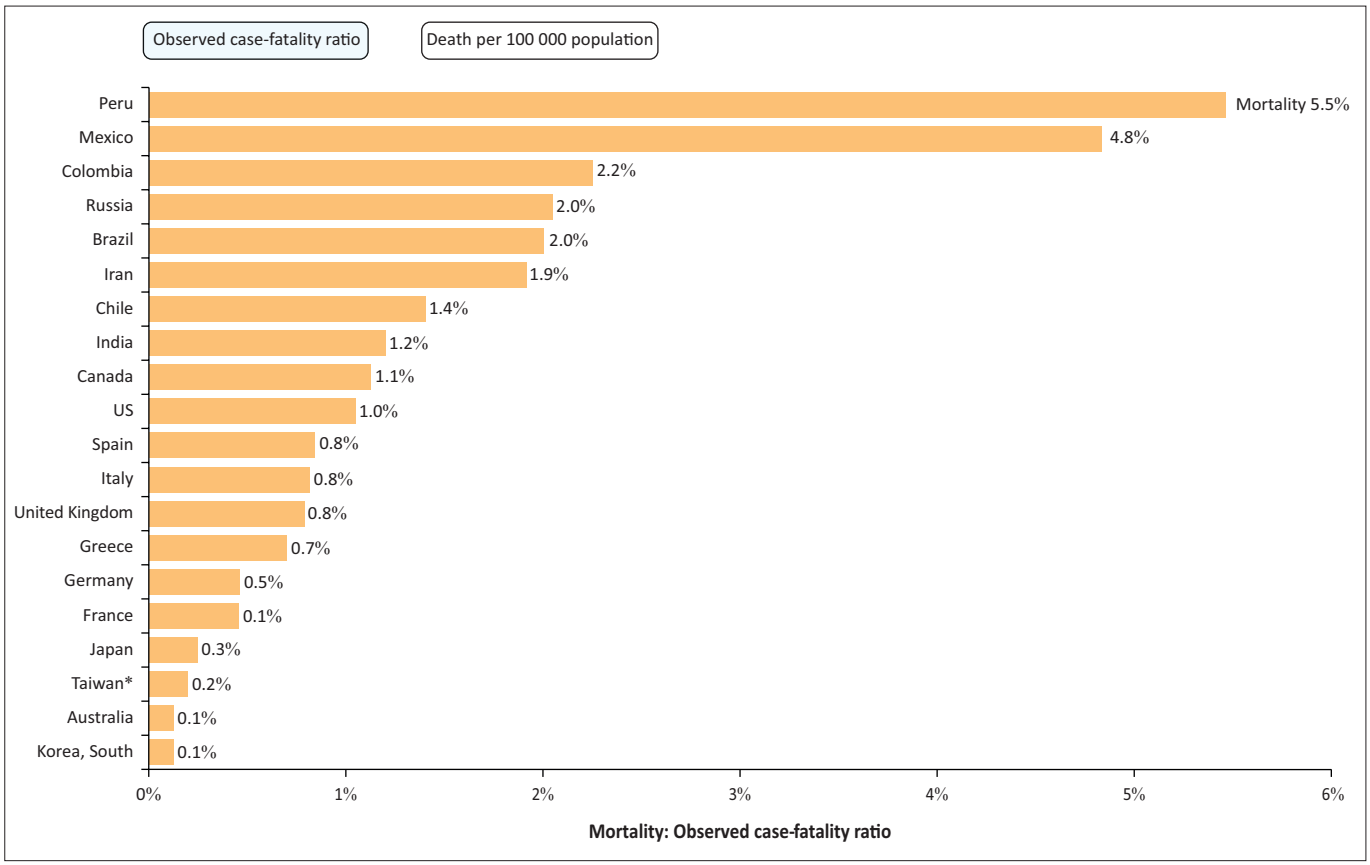
In the next section of the article, we look at four different presentations of COVID-19 data. The first involves international comparisons. The second involves an African-wide perspective. The third looks at COVID-19 data as reported from the Republic of South Africa. A fourth perspective is from the State of Victoria in Australia. Each of these presentations allows us to look at different contexts and to discuss different conventions used to report data. These provide expanded opportunities for students to develop statistical literacy and an understanding of uncertainty.

International comparisons

Different conventions were used in these international comparisons, and discussing them requires mathematical, as well as statistical, literacy. For example, the Johns Hopkins University Coronavirus Resource Center has a mortality analysis of COVID-19 for the top 20 affected countries (https://coronavirus.jhu.edu/data/mortality). This website uses two measures to record the impact of COVID-19. The first measure, shown in Figure 1, is called the observed case–fatality ratio, which is represented as the number of deaths per 100 confirmed cases. This ratio is looking at deaths among those who have been identified to be sick with COVID-19. The second ratio, shown in Figure 2, looks at deaths because of COVID-19 as a proportion of the population of confirmed cases and healthy people. To enable comparisons to be made across countries, this ratio is expressed in terms of deaths per 100 000 population.

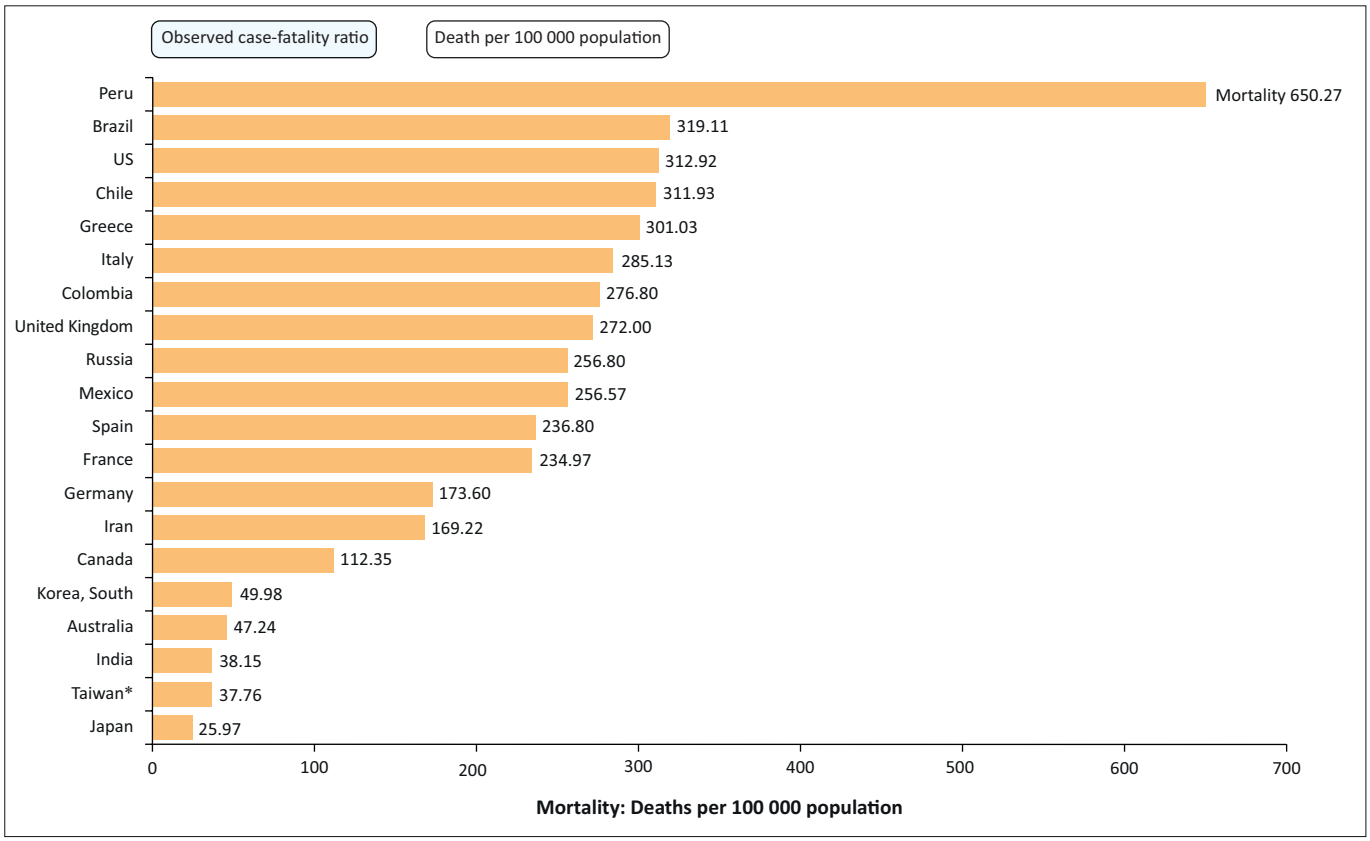
Teachers need to assist students to explore different data visualisations and the different measures of probability that they employ – a position which is strongly endorsed by Muniz (2022), Boaler et al. (2021) and Ridgway, Nicholson and McCusker (2013).

The bars in the two charts (Figure 1 and Figure 2) show the number of deaths using two different measures. One measure given in Figure 1 shows the number of deaths per 100 confirmed cases (called the observed case–fatality ratio). The Johns Hopkins website allows students to move between these two measures. In Figure 1, Peru and Mexico are at the top with 5.5% and 4.8%, respectively. Alternatively, the number of deaths can be shown per 100 000 population, as shown in Figure 2. The latter measure represents a country's general population, including both confirmed cases and healthy people. In the second figure of the Johns Hopkins data display, Peru and Brazil are at the top with 650 and 319 deaths, respectively, per 100 000 population. Students should



Source: Johns Hopkins University Coronavirus Resource Center, 2022, *Mortality analyses*, CRC, USA.

FIGURE 1: Mortality in the most affected countries (observed case–fatality ratio).



Source: Johns Hopkins University Coronavirus Resource Center, 2022, *Mortality analyses*, CRC, USA.

FIGURE 2: Mortality in the most affected countries (deaths per 100 000 population).

notice that Mexico is now ranked number 10 with 257 deaths per 100 000 population. They should also note that in the second figure, the United States of America is now ranked third with 312 deaths per 100 000 population.

Students can see, for example, how the positioning of the top 10 affected countries changes according to which measure is used. Among the five countries listed at the bottom of each figure, countries such as Japan, Taiwan, Australia and South Korea are present in both representations.

These two measures allow students to compare countries. However, countries recording more deaths per 100 000 of the population are not necessarily recording the most deaths overall. Students could well discuss the reasons for different mortality numbers and why many data displays prefer to express probabilities in terms of frequencies. We return to this important issue later in the article. Students will also notice that no African country appears in the Johns Hopkins website.

These data need to be placed in context, as the Johns Hopkins website explains:

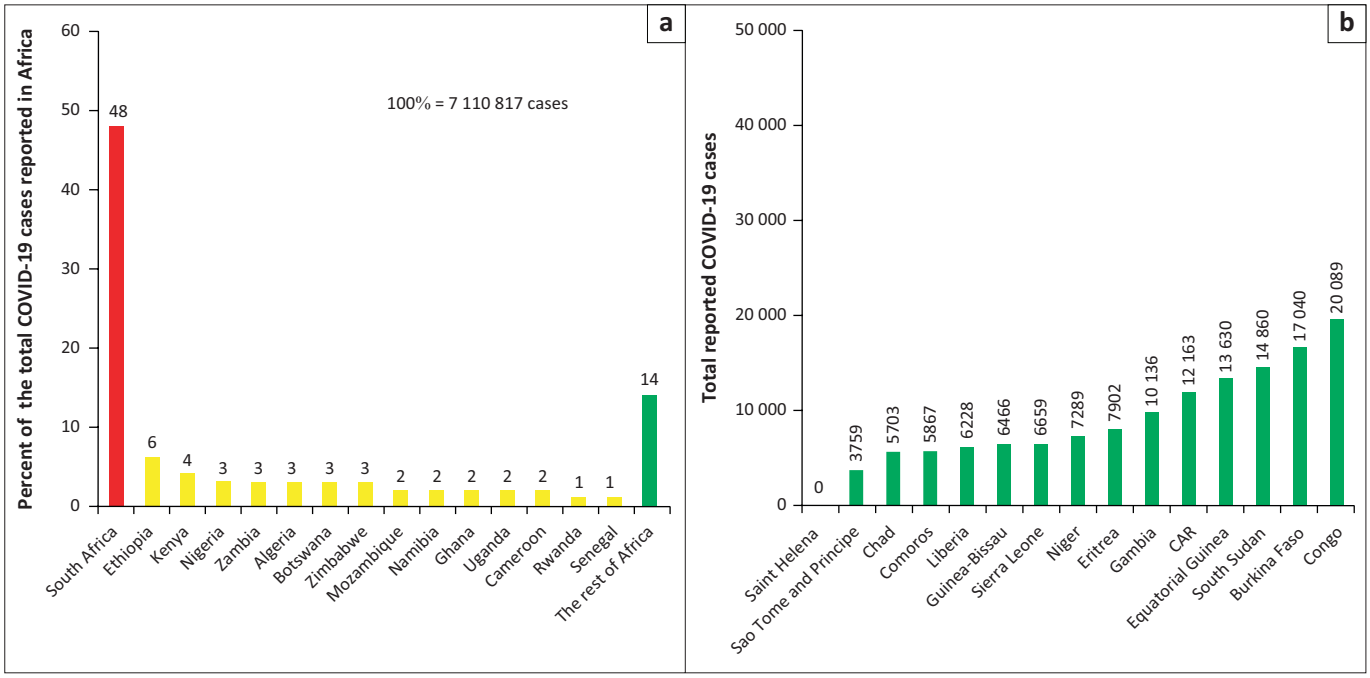
> Countries throughout the world have reported very different case fatality ratios – the number of deaths divided by the number of confirmed cases. Differences in mortality numbers can be caused by differences in the number of people tested: With more testing, more people with milder cases are identified. This lowers the case–fatality ratio. Demographics: For example, mortality tends to be higher in older populations. Characteristics of the healthcare system: For example, mortality may rise as hospitals become overwhelmed and have fewer resources. Other factors, many of which remain unknown. (https://coronavirus.jhu.edu/data/mortality)

The above quotation contains two important sentences which should be discussed by students: '*With more testing, more people with milder cases are identified. This lowers the case–fatality ratio*'. Consequently, there is uncertainty about the accuracy of data. The ability to carry out and record the results of widespread testing is a direct consequence of how well-funded a country's health system is. Therefore, students should understand that local and national contexts will impact on the accuracy of data gathered and recorded. In some lower-income countries, deaths because of COVID-19 may not be recorded as such. This helps students, even young students, to understand that data are what is recorded and are never a direct reflection of reality (Boaler et al. 2021). Appreciating and attending to data omissions and mistakes in recording apply to both small and large data sets.

An African perspective

Figure 3 is taken from a study by Bwire, Ario and Eyu (2022) and shows the top and least COVID-19-affected countries in Africa after excluding the island states (which had very few cases), except Comoros. This analysis is based on COVID-19 data reported to the World Health Organization (WHO) for the period 2020–2021. Section A of Figure 3 shows the top COVID-19-affected countries in descending order. Section B shows the least COVID-19-affected countries in ascending order after excluding the island states.

Figure 3 allows students to ask pertinent questions about how WHO data are acquired and presented. These questions would not be so obvious if students had not previously looked at and discussed websites such as the Johns Hopkins website presented in Figure 1 and Figure 2. This all-African representation is from an article by respected medical scientists writing on the COVID-19 pandemic in Africa



Source: Bwire, G., Ario, A.R. & Eyu, P., 2022, 'The COVID-19 pandemic in the African continent', *BMC Medicine* 20, 167. https://doi.org/10.1186/s12916-022-02367-4.

FIGURE 3: Top (a) and least (b) COVID-19-affected countries in Africa.

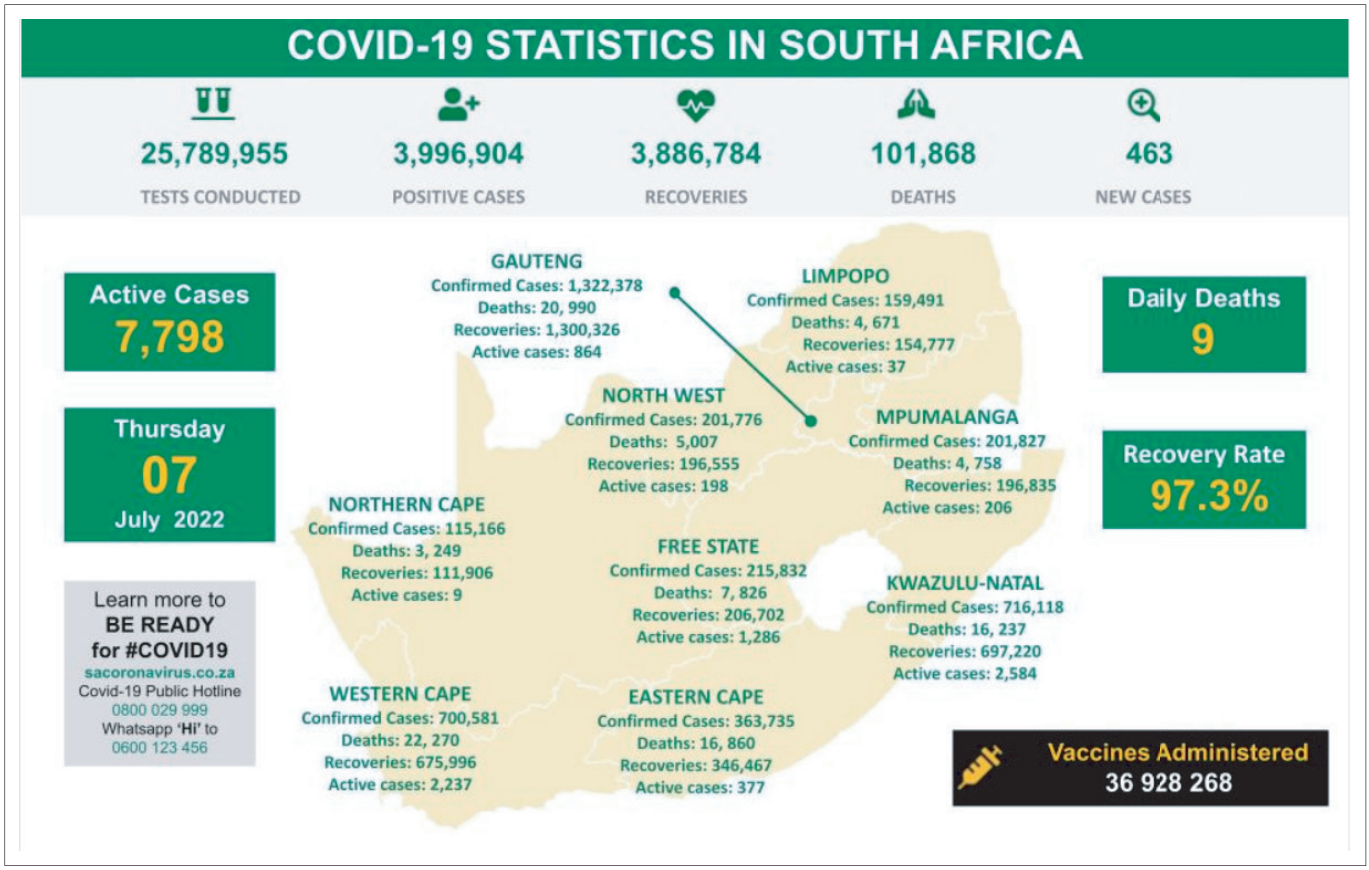
during 2020 and 2021. Yet students will be struck by the graph in Section A which shows South Africa as the 'top most-affected country' with nearly 50% of the 7 110 817 recorded cases. The next most-affected country is Ethiopia, with only 6% of reported cases. This discrepancy should prompt students to ask how uniform testing practices were across the African continent. Section B uses a different metric on its vertical axis, showing the total number of cases reported to the WHO. The cumulative number of positive cases for South Africa is shown in Figure 4 as 3 996 904 which includes 2022 data and is reasonably close to the number implied above. The unanswered question is how positive cases were identified in other African countries. This question is about context, especially about the efficiency and extent of testing regimes employed across Africa. We should not assume that these were uniformly the same.

National data from the Republic of South Africa

Figure 4 displays a collection of COVID-19 data for the Republic of South Africa for Tuesday, 07 July 2022. The website itself, https://sacoronavirus.co.za/, is in fact a dashboard allowing real-time data to be displayed according to the date inserted. Dashboards are data visualisations that provide access to real-time data and allow students to measure changes in data over time. Students can access other COVID-19 websites embodying dashboard capacities, such as the WHO (https://covid19.who.int/region/afro/country/za). Dashboards are increasingly used in business and industry allowing users to interact with data displays and look at visual representations of data over time. Mathematical literacy and statistical literacy require students to become familiar with these tools of the 21st century.

Different kinds of data are represented on the above website. The assumptions behind these need to be discussed. Across the top, cumulative data show the number of tests conducted starting from the beginning of the COVID-19 pandemic when data began to be recorded, the number of positive cases, the number of recoveries and the number of deaths. The last figure denoting new cases is not cumulative, showing 463 new cases on the date 07 July 2022. The cumulative data also show that the number of recoveries plus the number of deaths approximates but is not equal to the total number of positive cases. What might be the reasons for this? Similar questions can be asked of the data for each of the provinces. *Confirmed cases*, *deaths* and *recoveries* refer to cumulative data, whereas *new cases* refers to new cases identified for the given date.

Students can also be encouraged to apply the two different conventions as used by the Johns Hopkins website for analysing the data for each province, namely a case–fatality ratio showing deaths per 100 confirmed cases or a ratio showing deaths per 100 000 population. For example, taking two provinces, namely Gauteng and the Eastern Cape, the cumulative case–fatalities ratios are, respectively,



Source: Department of Health, 2022, *COVID-19 online resources and news portal*, Republic of South Africa

FIGURE 4: COVID-19 statistics in South Africa and its nine provinces (https://sacoronavirus.co.za/).

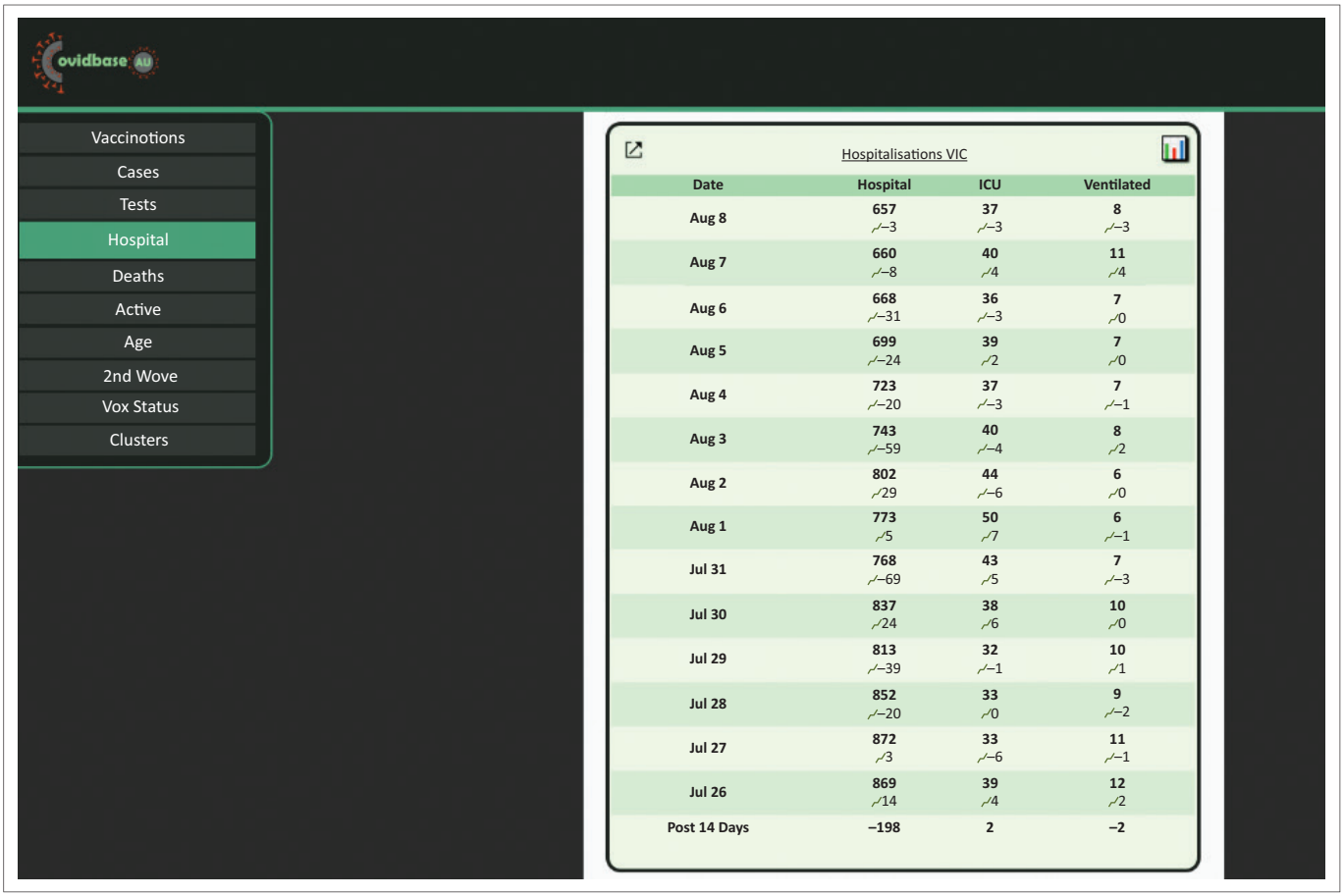
1.58% and 4.64%, given by dividing the number of deaths by the number of confirmed cases. (In the case of Gauteng province, after dividing 20 990 by 1 332 378, students need to interpret a calculator-derived result showing 0.0157537876 as a percentage [i.e. per 100 confirmed cases] to a given number of decimal places.) The second measure requires students to have an estimate of the population of each province. The population of Gauteng province, in which South Africa's largest city, Johannesburg, is situated, has been estimated in 2022 as 16.1 million (equal to 100 000 × 161), giving the number of deaths per 1 000 000 for Gauteng province as 20 990/161, which is approximately 130 per 100 000. By comparison, the Eastern Cape province, with a population in 2022 of approximately 7 582 566 (100 000 × 75.8), would have a ratio of approximately 222 deaths per 100 000, which is lower than that of France in Figure 2. Using these ratios to compare South African data to international data as shown on the Johns Hopkins website allows students to appreciate the importance of being able to make international comparisons. Students can be assisted to use dashboard data like these in many ways.

Data from a single state or province: Victoria, Australia

Figure 5 is taken from the Victoria Department of Health and Human Services (DHHS) website (https://covidbaseau.com/vic/) that provides data updated daily on 10 measures. This dashboard introduces a category relating to hospitalisations, along with other data that students will have met before relating to the number of new infections, new cases and so on. Data relating to hospitalisations are not typically provided in international and national data. Students need to ask why the number of hospitalisations becomes important in this local context. Moreover, we can see that the data are broken down further into those who are in hospital and in intensive care (ICU) and those who are also being ventilated while in ICU.

A focus on *hospitalisations* as shown in Figure 5 becomes important in a local context because it lets people know how their immediate health system is coping under the pandemic. These data are also used to convey to citizens the urgency of the situation in the sense that some people who are infected are seriously ill and those in ICU more seriously. Data of this kind are not only shaped by local contexts, but they also require students to think about how these data are recorded and the conventions used.

For a more accurate statistical analysis in estimating the hospitalisation rate, that is, number of hospitalisations per number of infected, it is important to consider the time delay between onset of symptoms and admission to hospital. In examining the proportion of hospitalisations, the number of



Date	Hospital	ICU	Ventilated
Aug 8	657 ↗–3	37 ↗–3	8 ↗–3
Aug 7	660 ↗–8	40 ↗4	11 ↗4
Aug 6	668 ↗–31	36 ↗–3	7 ↗0
Aug 5	699 ↗–24	39 ↗2	7 ↗0
Aug 4	723 ↗–20	37 ↗–3	7 ↗–1
Aug 3	743 ↗–59	40 ↗–4	8 ↗2
Aug 2	802 ↗29	44 ↗–6	6 ↗0
Aug 1	773 ↗5	50 ↗7	6 ↗–1
Jul 31	768 ↗–69	43 ↗5	7 ↗–3
Jul 30	837 ↗24	38 ↗6	10 ↗0
Jul 29	813 ↗–39	32 ↗–1	10 ↗1
Jul 28	852 ↗–20	33 ↗0	9 ↗–2
Jul 27	872 ↗3	33 ↗–6	11 ↗–1
Jul 26	869 ↗14	39 ↗4	12 ↗2
Post 14 Days	–198	2	–2

Source: Victorian Department of Health and Human Services, 2022, *COVID Base AU/VIC*, Victoria.

FIGURE 5: Victoria COVID-19 breakdown (hospitalisations in August and July 2022).

hospitalisations should ideally be offset by several days before forming a proportion with the number of infections and the consequent number of hospitalisations. The time delay itself is approximate and may vary for different age groups (see the Belgium study by Faes et al. (2020) for a more advanced treatment of this topic).

Teachers need to help students to clearly appreciate the definitions that are attached to different populations referred to in the media, such as the number of new daily infections and those who have been hospitalised. As mentioned on the Victorian DHHS website (https://covidbaseau.com/vic/), the number of new infections refers to new infections reported in an immediately preceding time period. However, 'hospitalisations' does not refer to new hospitalisations in the same time period. These data are collected differently and show the total number of people in the hospital because of COVID-19 on a given day.

Other key data refer to the number of people who have been tested on a given day and the proportion of those who have tested positive. This probability is rarely given as a percentage of the total population. On the contrary, the percentage showing the proportion of those vaccinated is a cumulative percentage over time, showing those vaccinated as a proportion of the total population *over twelve years of age*. (In many countries, this age range has changed several times over the past two years.) The different subsets of a population can be shown diagrammatically by a Venn diagram (see Figure 6), indicating, for example, that those who are in hospital will be a subset of those who are infected but not necessarily vaccinated. Diagrams such as this assist students to identify and discuss the various subsets of their population – another key feature of statistical literacy.

Because COVID-19 is an urgent public health phenomenon, many COVID-19 terms will be used for national and international reporting and therefore may be predefined, that is, standardised nationally and internationally, to ensure that data are collected in mutually agreed-upon and convenient ways. Unlike textbook problems, where definitions may not be contentious, teachers can discuss with students the reasons why a particular definition has been preferred and how this

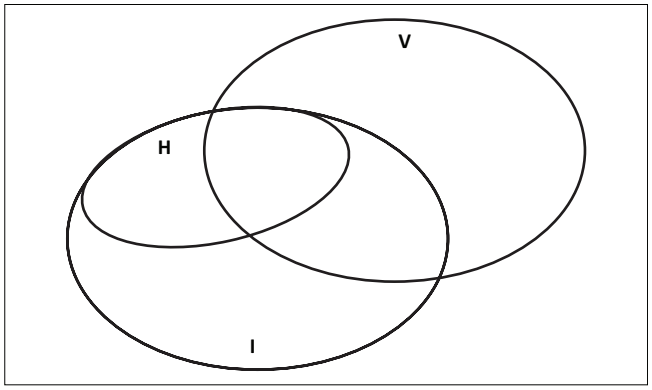


FIGURE 6: Venn diagram: vaccinated (V), infected (I) and hospitalised (H) in a population.

is related to the collection of important data relating to COVID-19. For example, the term 'hospitalisation' is not generally used as a measure of the number of *new* COVID-19 patients on a given day who have been admitted to the hospital after recording a positive test. Usually, it is based on hospital data recording the number of patients who are currently occupying COVID-19-designated hospital beds. Some COVID-19 patients may be sent home after a few days with appropriate medication, and others with more severe symptoms may be admitted to ICU, with a smaller number requiring ventilation or other life-support systems. It appears easier for hospitals to report on the number of COVID-19 beds that are occupied on a given day than recording the number of new admissions. Changes in the number of hospitalisations can be represented using a 7-day moving average.

Implications for teachers from these different representations

In line with the theoretical lenses of Boaler et al. (2021) and Ridgway et al. (2013), making sense of the data and helping students to understand and to create a narrative of the unfolding pandemic are essential tasks for teachers starting in the upper elementary years of school. Utilising websites such as the Johns Hopkins University Coronavirus Resource Center (https://coronavirus.jhu.edu/data/mortality), it is important to assist students to make sense of data displays. This interactive graphic showing a mortality analysis of COVID-19 for the top 20 affected countries demonstrates that the probability of death because of COVID-19, using two different measures, is not fixed but varies across the 20 countries. Students need help to discuss possible reasons for these variations. These may include the availability of vaccines, different capacities of health systems in the countries displayed, the extent to which different age groups have been infected and the impact of widespread testing where the inclusion of those who may have tested positive but with milder symptoms will lower the mortality ratios. Countries that employ compulsory testing in affected regions will have qualitatively different data from countries where testing is voluntary and undertaken by people who either feel sick or know that they have been exposed to someone who is. Context is vitally important to interpreting international, national and local representations of COVID-19 data.

Another important feature of the data for students to appreciate is that data, and hence the ratios and associated probabilities themselves, change over time, helping students to understand how their local situation and its wider national and international context are changing. These perspectives can be supported by tracking the available data over time, are important features of teaching in the compulsory years of schooling and provide a foundation for a more theoretically based focus on time series in the later years of school.

During COVID-19, teachers have been faced with the ethical challenges of guiding students' investigations responsibly

and scientifically. When real life and mathematics intersect so directly in times of COVID-19, teachers know that these challenges cannot be avoided. Especially in the middle and senior years of high school, it is essential for teachers to help students to distinguish between currently available data and future predictions which extrapolate from the current data based on various models.

The intersection of theory and practice

In this section, we consider how the experience of COVID-19 has provided opportunities to address and expand some highly pertinent 'frontier' areas in the teaching of probability and statistics. From this perspective, the goal is to ensure that theory and practice interact in productive and meaningful ways for teachers and students. Here we consider several examples briefly, with the first four discussing and evaluating some key issues for teaching probability. These are in the following order: the pros and cons of adopting a probabilistic versus a frequentist approach, the importance of distinguishing between absolute and relative risks, relative risk versus odds ratio and scientific evidence evaluation. These four areas are directly pertinent to teacher education, teacher professional learning and aspects of the senior high school curriculum. The first two areas are also pertinent to the curriculum in the compulsory years as well.

Probabilistic versus frequentist approaches

The Bayesian algorithm, which is used in the calculation of conditional probability, makes little sense for many students. It is not clear to them why the product of some probabilities must be divided by the sum of products of certain probabilities. Researchers such as Sedlmeier and Gigerenzer (2001) advocate simplifying matters and using a frequentist approach to replace the traditional, probabilistic approach. In so doing, a tree structure may be used as a graphical aid, as illustrated in the following example of a frequentist approach.

Suppose that 20% of the population carry a virus and that the test for the virus yields 10% false positives (meaning that as much as 90% of those who don't carry the virus test negative) and 20% false negatives (meaning that only 80% of those who carry virus test positive). Consider 100 people. Then, 20 (20%) of them carry the virus, where 16 (80%) test positive and 4 (20%) test negative. Of the 80 people without the virus, 8 (10%) test positive and 72 (90%) test negative. The frequencies of these four cases are represented in Figure 7. In the case of positive test, the probability of carrying the virus is 16/(16 + 8), which is exactly two-thirds.

Instead of sketching a decision tree, some teachers may prefer to create a two-way table, such as the one given in Table 1. No matter which graphical aid is used, translating percentages into the so-called natural frequencies would increase students' understanding about the situation at hand and

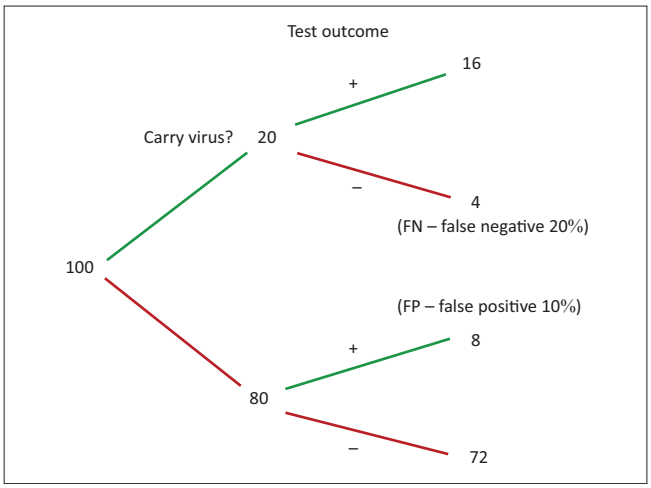


FIGURE 7: Bayesian reasoning: A tree-based frequentist approach.

TABLE 1: Bayesian reasoning: A table-based frequentist approach.

Has virus?	Test outcome		Total
	+	–	
Yes	16	4 (20% FN)	20% of 100
No	8 (10% FP)	72	80% of 100
Total	**24**	**76**	**100**

FN, false negative; FP, false positive.

make the numbers in question more meaningful to them (e.g. Martignon 2016). Because of that, the teaching of probability should familiarise students with the frequentist approach, which may be used as a scaffolding tool for the traditional probabilistic approach, especially for students whose interest, motivation and understanding of complex probability calculations are at lower levels than needed.

Relative risks versus absolute risks

Another important area to cultivate is enabling students to distinguish between absolute and relative risk. It has been common, for example, to report on adverse reactions among those receiving vaccines. As data came in from across the world in early 2021, there was evidence that a few people who had been vaccinated with AstraZeneca – mostly in the under-50 age group – developed a blood clotting condition, thrombosis with thrombocytopenia syndrome (TTS), that required hospitalisation. This created a scare in Australia and other countries using this vaccine, with many people under 50 refusing to take AstraZeneca, exacerbating vaccine hesitancy. Were these concerns justified, and how were they addressed by the health professionals?

The statistics, however, are important to consider – for those under 50 years, the incidence of developing the TTS blood clotting illness because of AstraZeneca was 3.1 per 100 000, and for those between 50 and 59 years, the incidence was 2.7 per 100 000. The danger here is in focusing on absolute risk. To make an informed decision, people needed to look at the balance of risks comparing the relatively small risk of developing a blood clotting illness because of taking AstraZeneca compared with a higher risk of hospitalisation and death if *while being unvaccinated* one contracted

COVID-19. From a public health perspective, the key is to understand that risk has always to be measured against benefit. What are the implications of this issue and its importance for teacher education and for students in senior high school?

News periodicals, for example, often gain and maintain a large number of readers by communicating stunning stories, including health-related percentages, such as 'Smoking increases lung cancer by 30%!' During the COVID-19 pandemic, governments typically advocate vaccination as the best way to reduce the risk of getting the disease or encountering life-threatening health conditions if infected. Although the call to vaccinate is not problematic, the 'numbers' that usually accompany this call need to be questioned. What does it mean to claim that vaccination reduces the risk of having the disease by 20%? This percentage is far from negligible, but it does not mean that only 20 of 100 nonvaccinated people would have the disease. Not at all! A subtle issue that is very important for senior students to understand is that the 20% is the result of the comparison of two percentages, expressing the portion of nonvaccinated persons who get the disease in relation to the portion of vaccinated persons who experience the same. If these portions (in terms of percentages) are, for example, 10% and 8% (and 2% = 10% – 8% is clearly 20% of 10%), the call for vaccination has little empirical support, contrary to the situation when these percentages are higher, such as 25% and 20% (if we agree that a difference of 5% or more provides convincing evidence in favour of vaccination). Thus, to have relative risks increase or decrease meaningfully and to support proper decisions if needed, relative risks need to be accompanied by corresponding absolute risks. Without them, the situation at hand may only introduce confusion and promote misunderstanding. Through an examination of the two examples given below (or other similar reports often found in media), students would realise that it is indeed impossible to infer meaningful information when both relative risks and absolute risks are not given. However, their attempts to make sense of relative risks through what-if speculations may be productive if some questionable absolute risks are generated (e.g. a 20 times reduction from 20% to 1%), because they may suggest that some measures might be poorly defined or measured (see 'Scientific Evidence Evaluation' section).

To trace the place(s) where shortcomings concerning incomplete information were generated, the following steps may be examined: (1) selecting news from news or statistical agencies; (2) copying, adapting or elaborating the news selected; and (3) presenting the news in media. However, because of the content selected in (1), an elaboration needed to undertake under (2) may be limited or simply not possible (Kadijevich 2016). Studies in medical journals usually do not provide transparent information including absolute risks. A possible reason may be found in competing interests. Because major studies are often financed by the pharmaceutical industry, the results need to be presented in a way that would impress doctors, patients and policymakers. Undoubtedly, focusing on relative risk reductions is an effective way to do so (e.g. Gigerenzer et al. 2007) and usually exaggerates effect size (e.g. a 50% reduction from 10% to 5%; Shah et al. 2017). Media articles rarely themselves overstate the findings and implications reported, although they may limit viewpoints communicated to have news presented in a simplified way without conflicting information (e.g. Shah et al. 2017).

Scientific evidence evaluation

Understanding public policies and defending personal decisions call for adequate scientific evidence evaluation skills. Without them, news consumers cannot recognise various errors in scientific reasoning. These errors, usually called threats to validity, could be of various kinds: threats to internal validity, threats to construct validity, threats to statistical validity and threats to external validity. What follows is a summary of these threats examined in Shah et al. (2017).

Threats to internal validity are present when it is not clear what affects the dependent variable under study. Is it the applied treatment, one or more independent variables or another factor? One of such threats is causality bias: correlated variables need not be related causally. Another threat is the control of variables: have experimental variables been controlled appropriately (e.g. using a partial correlation or an analysis of covariance)? A third threat, related to the previous one, deals with the assignment of individuals to condition; if done properly, the so-called selection bias is not present. There are other threats to internal validity, such as when repeated measurement might mainly contribute to change.

Threats to construct validity are related to the measurement of the applied constructs. It is important that all constructs are not only clearly defined but also measured appropriately. If this does not apply, drawing conclusions from such compromised research might be worthless.

Threats to statistical validity are generated by, among other things, inadequate sampling, small sample size, unreliable measures and inappropriate statistical tests. It is not enough that a sample is random and of appropriate size. Measures should be reliable, and statistical tests should be appropriate.

Threats to external validity deal with problems in the applicability of findings to other subjects and situations. Such threats also present when effect sizes are small, even though large samples are used. External validity may also be threatened by relying on a specific number when reporting findings (e.g. smoking more than 10 cigarettes per day increases lung cancer risk by 40%).

Research shows that detecting threats to scientific validity is probably out of reach of most readers of scientific evidence. People rarely notice these threats, especially when the

presented evidence is congruent with their beliefs or behaviours. Possible reasons and factors affecting this undesirable state are given by Shah et al. (2017). Among these factors are numeracy and statistical reasoning skills. Domain knowledge is also a critical factor, but as one cannot be familiar with every scientific topic, he or she needs to learn and apply some general skills to evaluate the design of research and reliability of its findings. To develop these skills, teaching scientific evidence evaluation, as a part of a course on probability and statistics, may benefit from discussing the threats examined above. This discussion should be practised with students in the senior years of high schools. As there are many evaluation concepts, teaching should begin with the examination of one of the frequently misused concepts (e.g. causality and the bias it could generate) and addition of other concepts in a piecemeal way as required.

How should the teaching of statistics and probability change?

In this final section, we examine some consequences for teacher education and development in working with visualisations based on large data sets, often displaying real-time representations. New ways of visualising data, as we have shown in this article, provide challenges and opportunities. Ridgway et al. (2013) agree that the most obvious challenge is introducing teachers and students to the exploration and analysis of large data sets, either directly or, as we have illustrated, from visualisations derived from these data. Muniz (2022) reminds teachers that while data visualisations may appear to be objective descriptions of the world, they are only narratives produced by data science with an assumed objectivity to justify their consumption. Like Ridgway et al. (2013), Muniz (2022) underscores the importance of 'understanding how data visualisations are created, what they can represent, and how data, in general, are analysed and interpreted' (p.1).

The data we have discussed in this article are related to the pandemic. However, teachers and students may be interested in exploring the take-up of pop music or other social phenomena where large data sets are accessible. New digital and analytical tools are needed to investigate these issues. Tools such as dashboards, which were once accessible only to university students or corporate users, are now publicly accessible, and we can expect to see their use and impact grow in coming years.

The widespread availability of data relating to COVID-19 presents both opportunities and challenges for the study of probability and statistics. These challenges cannot be avoided as the phenomenon has a direct impact on our lives, and the opportunities to bring real-world data into the teaching of probability are too important to ignore. In times of uncertainty, individuals may rely on sources that provide simplistic answers to complex questions. The critical lenses provided by mathematical literacy and statistical literacy, as Boaler et al. (2021) rightly emphasise, provide an important foundation for the informed participation of all citizens.

In this article, we have made several recommendations for the improvement of students' mathematical literacy and statistical literacy. Some of these can be summarised as follows: students need to understand new conventions and representations, including frequentist approaches to the data that seem to be preferred by many websites. Assisting students at all levels to become familiar with and to evaluate available websites is an indispensable task for all teachers.

Distinguishing between absolute and relative risk is important, as is the ability to quantify risk using available data. Other ways of measuring risk, for example, through odds ratios, may need to be introduced. Students need to compare different measures and discuss possible reasons for variation across different contexts – starting with their local context, national and international contexts.

Among the challenges are the different definitions and conventions attached to contexts and websites. Another challenge that is relevant to the compulsory years of schooling as well is to focus on the adequacy of the data to answer questions over time. Variation over time is a feature embedded in almost all large data sets that are intended to capture and represent real-life phenomena. Simple ways of depicting trends using moving averages become more important. There is an equally strong case for the wider use of dashboards to support time series analysis.

From the earliest stages, students need to develop a scientific or critical approach to the available data. Learning to ask questions about data sources and their reliability and identifying when new forms of data are needed have been endorsed by Boaler et al. (2021) and Ridgway et al. (2013) as important elements of contemporary statistical literacy. Later, students need to develop more explicit evaluation skills to critique evidence and to search for errors in scientific reasoning, being aware of the main threats that usually compromise conclusions. Shah et al. (2017) agree that this is a challenge. In this respect, an important task is to train students to be sceptical about imputing causation simply because of probabilistic evidence.

The greatest opportunity is to better integrate theory and practice by utilising available data and their visualisations to support students' study of theoretical probability. Utilising these opportunities with large data sets cannot be achieved without introducing students, especially in the senior years of high school, to increasingly accessible forms of data processing software. That is a topic for another study. In this article, however, we have taken the position that COVID-19 has accelerated the use of data visualisations by government and nongovernment agencies in almost every country, creating a need for teachers and students to be better equipped to understand the conventions and assumptions of data visualisations and to build their capacity to understand, interpret and critically analyse data. These are critical 21st-century skills.

Acknowledgements

Mark Stephens provided the Venn diagram shown in Figure 6 and advice on the senior high school years.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Authors' contributions

M.S. contributed to the conceptualisation, methodology, formal analysis, investigation, writing of the original draft, review and editing. D.M.K. contributed to the conceptualisation, formal analysis, writing of the original draft and resources. J.C.H. contributed to the review and editing. M.M. contributed to the visualisation and resources.

Ethical considerations

This article followed all ethical standards for research without direct contact with human or animal subjects.

Funding information

The research done by D.M.K. was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (contract no. 451-03-68/2022-14/200018).

Data availability

All figures and tables are publicly available on the Internet and are cited appropriately. The JHU Coronavirus Resource Center stopped collecting data on 21 September 2022.

Disclaimer

The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of any affiliated agency of the authors.

References

Australian Education Council, 1991, *A national statement on mathematics for Australian schools*, Australian Education Council, Canberra.

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L. et al., 2020, *Pre–K–12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II)*, American Statistical Association, Alexandria, VA.

Boaler, J., LaMar, T. & Williams, C., 2021, 'Making sense of a data filled world', *Mathematics Teacher: Learning and Teaching PK-12* 114(7), 508–517. https://doi.org/10.5951/MTLT.2021.0026

Bwire, G., Ario, A.R. & Eyu, P., 2022, 'The COVID-19 pandemic in the African continent', *BMC Medicine* 20, 167. https://doi.org/10.1186/s12916-022-02367-4

Callingham, R. & Watson, J., 2017, 'Appropriate goals for statistical literacy at school', *Statistics Education Research Journal* 16(1), 181–201. https://doi.org/10.52041/serj.v16i1.223

Callingham, R., Watson, J. & Oates, G., 2021, 'Learning progressions and the Australian curriculum mathematics: The case of statistics and probability', *Australian Journal of Education* 65(3), 329–342. https://doi.org/10.1177/00049441211036521

Cobb, G.W. & Moore, D.S., 1997, 'Mathematics, statistics, and teaching', *American Mathematical Monthly* 104(9), 801–823. https://doi.org/10.2307/2975286

Department of Health, 2022, *COVID-19 online resources and news portal*, Republic of South Africa.

Faes, C., Abrams S., Van Beckhoven, D., Meyfroidt, G., Vlieghe, E. & Hens, N. et al., 2020, 'Time between symptom onset, hospitalisation and recovery or death: Statistical analysis of Belgian COVID-19 patients', *International Journal of Environmental Research and Public Health* 17(20), 7560. https://doi.org/10.3390/ijerph17207560

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L.M. & Woloshin, S., 2007, 'Helping doctors and patients make sense of health statistics', *Psychological Science in the Public Interest* 8(2), 53–96. https://doi:10.1111/j.1539–6053.2008.00033.x

Johns Hopkins University Coronavirus Resource Center, 2022, *Mortality analyses*, CRC, USA.

Kadijevich, D.M., 2016, 'Recognizing the shortcomings of statistics in media: What can novice do?', in J. Engel (ed.), *Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE), July 2016, Berlin, Germany*, ISI/IASE, The Haag.

Martignon, L., 2016, 'Empowering citizens against typical misuse of data concerning risks', in J. Engel (ed.), *Promoting understanding of statistics about society. Proceedings of the Roundtable Conference of the International Association of Statistics Education (IASE), July 2016, Berlin, Germany*, ISI/IASE, The Haag.

Muniz, B., 2022, 'Data science for a critical society', *Seminario Interuniversitario de Investigación en Ciencias Matemáticas*, La Universidad de Puerto Rico an Humancao, February 25, 2022.

Ridgway, J., Nicholson, J. & McCusker, S., 2013, '"Open data" and the semantic web require a rethink on statistics teaching', *Technology Innovations in Statistics Education* 7(2), https://doi.org/10.5070/T572013907.

Sedlmeier, P. & Gigerenzer, G., 2001, 'Teaching Bayesian reasoning in less than two hours', *Journal of Experimental Psychology: General* 130(3), 380–400. https://doi.org/10.1037//0096–3445.130.3.380

Shah, P., Michal, A., Ibrahim, A., Rhodes, R. & Rodriguez, F., 2017, 'What makes everyday scientific reasoning so challenging?', *The Psychology of Learning and Motivation* 66, 251–299. https://doi.org/10.1016/bs.plm.2016.11.006

Victorian Department of Health and Human Services, 2022, *COVID Base AU/VIC*, Victoria.

Watson, J. & Callingham, R., 2020, 'COVID-19 and the need for statistical literacy', *Australian Mathematics Education Journal* 2(2), 16–21.