PREPRINT: SUBMITTED FOR PUBLICATION

Misinterpreting statistical anomalies and risk assessment when analysing Covid-19 deaths by ethnicity

Norman E Fenton[1,4], **Martin Neil**[1, 4] , **Scott McLachlan**[1,2], **Magda Osman**[3],

[1] Risk and Information Management, Queen Mary University of London, United Kingdom
[2] Health informatics and Knowledge Engineering Research (HiKER) Group
[3] Biological and Experimental Psychology Group, Queen Mary University of London, United Kingdom
[4] Agena Ltd, Cambridge, UK

31 July 2020

Abstract

When analysing Covid-19 death rates by ethnicity in the USA it has been shown that, although the aggregated death rate for whites is higher than for blacks, in each main age subcategory the death rate for blacks is higher than for whites. This apparent statistical anomaly is an example of Simpson's paradox. While the paradox reveals blacks are more at risk than whites, this is only because age is a much more important risk factor than ethnicity. Hence, any public policy with respect to Covid-19 that prioritises ethnicity over age is misguided. We have found similar evidence of Simpson's paradox in UK Covid-19 death rates but the widely publicised conclusions about the risk to the Black and Minority Ethnic (BAME) population are misleading. In particular, the conclusion in the recent UK Office of National Statistics (ONS) report, that blacks are more than four times more likely to die from Covid-19 than whites, may create an unjustified level of fear and anxiety among the BAME community. It is misleading for three reasons: 1) It appears to rely on old 2011 census data about the population proportions rather than on more recent estimates; 2) It is appears to be based on an 'age standardized' measure of risk that is very different from that used by the World Health Organisation (WHO); and 3) It focuses on relative rather than absolute measures of risk. Hence, we believe the ONS conclusions may be misleading from a risk assessment perspective and may serve as a poor guide to public policy.

Corresponding Author: Norman Fenton n.fenton@qmul.ac.uk

1. A hypothetical Example

Imagine there is a country called Bayesland that is divided into two distinct geographical areas - North Bayesland and South Bayesland with equal population sizes. The country has been struck with a new, novel, infectious disease called P-STAT. Statistics reveal that the death rate for this disease for Southerners is TWICE that of the death rate for Northerners in each different age category (Table 1).

Table 1

		deaths per million
age < 65	North	3
	South	6
age 65+	North	2001
	South	4000

This information suggests there is a consistent and worryingly much greater risk for Southerners than Northerners. From a public policy perspective, it would seem reasonable to focus resources on Southerners.

However, suppose the statistics also reveal that – when we aggregate the data over all age groups – the death rate for Northerners is FORTY times that of the death rate for Southerners (Table 2). Who do you now believe is at greater risk?

Table 2

		deaths per million
Total	North	1002
	South	26

The initial reaction by most people when presented with this kind of information is to assume that this is some kind of deception – that there surely cannot be such a complete reversal in the death rates simply by looking at the data in a different way. But there is no deception as the full date in Table 3 shows:

Table 3 Aggregated v disaggregated data

		deaths	***population***	***deaths per million***
Total	North	2004	2,000,000	1002
	South	52	2,000,000	26
age < 65	North	3	1,000,000	3
	South	12	1,990,000	6
age 65+	North	2001	1,000,000	2001
	South	40	10,000	4000

The situation whereby, drilling down into each sub-category gives results that are the 'opposite' of the aggregate result, is well known in statistics and is an example of *Simpson's paradox* (Simpson, 1951). Normally, statisticians look first at the aggregated data, and the paradox only becomes evident when they then look at the disaggregated data. It is widely assumed that, once Simpson's paradox has been revealed in cases like this, it is the disaggregated data that correctly identifies the group most 'at risk'. So, this might suggest that public policy should target Southerners and not Northerners as was initially assumed. However, things are not so clear-cut.

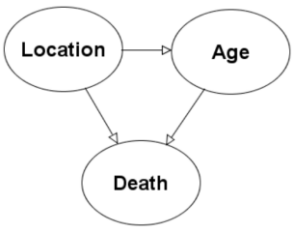

Figure 1 Causal explanation for observed data and Simpson's paradox

Simpson's paradox arises because there is an underlying causal explanation in the data that might not be initially obvious (Fenton, Neil, & Constantinou, 2019; Pearl & Mackenzie, 2018). In this case it is that Northerners are much more likely to be over 65 than Southerners, and the disease is far more deadly to those aged over 65 (see Figure 1).

So, what relevance should this have on our policy decision? Established authorities already recognise the role that age plays in fatality statistics and recommend ways in which different age distributions can be considered when calculating aggregate risk. For example, the World Health Organisation (WHO)

calculate an "age standardized death rate" (WHO (World Health organization), 2006). If the age categories are as above, it is defined as:

> (<65 death rate)×(population proportion <65) + (65+ death rate)×(population proportion 65+)

But this definition is equivalent to the overall death rates provided in the final column of Table 3. So, according to the WHO, the age standardized death rate for North Bayesland is 1002, and for South Bayesland is 26. This conflicts with the assumption – based on identifying Simpson's paradox in the disaggregated data, that it is the Southerners who have the higher death rate. So, any public policy which, solely uses the disaggregated data, and prioritises Southerners in tackling P-STAT disease, would clearly be irrational.

The key point is that age is a much more significant risk factor than location. The death rate for people aged 65+ is nearly 270 times greater than those aged under 65 (2002 compared to 75). So, any public policy decision should take this into account. Imagine if the national Government of Bayesland had 1 million cryptos to invest in what it believes to be the group most 'at risk' of the disease, knowing that such an investment will save 10% of those who would otherwise die from it (assuming deaths would otherwise continue at the same rate in the next period). Imagine also if the only three options available were 1) target only Northerners (which might seem reasonable based on the aggregate death rates); 2) target only Southerners (which might seem reasonable based on the disaggregated death rates); 3) target only those aged 65+. Then, under each policy option they would save respectively:

1. About 200 lives (since there are a total of 2004 deaths among Northerners)
2. About 5 lives (since there are a total of 52 deaths among Southerners)
3. About 204 lives (since there is a combined total of 2041 deaths among 65+ in the North and South).

So, prioritising Southerners would be a severe misallocation of resources. Any rational allocation of resources should take account of both age and location based on the relative risk of both.

2. Replacing 'Location' with 'Ethnicity' in the real world for Covid-19

At this point you may be wondering what has this got to do with the Covid-19 death statistics and ethnicity? First, Simpson's paradox is highly prevalent in real world data, and it causes serious complications regarding how data might help assess risk and make policy decisions. Given this, we need to examine whether Simpson's paradox is present in demographic data to identify those at higher risk, and what implications it has when making decisions. ***It turns out that the real-world situation has similarities with the Bayesland example when we replace 'Location' with 'Ethnicity'***.

Based on data that has been gathered regarding specific demographics at higher risk of dying of Covid-19 in several countries, we are informed that older adults (65+) and Black, Asian and Minority Ethnic (BAME) individuals are in the high mortality risk group.

Using data collected by the USA's Center for Disease Control (CDC) Dana Mackenzie, a mathematician in the USA, has shown that, when comparing Covid-19 fatality statistics for whites and non-whites Simpson's paradox is evident: while the overall rate is higher for whites, in each main age sub-category it is higher for non-whites (Mackenzie, 2020). His causal 'explanation' is the same as that presented in our hypothetical example: whites (like Northerners) are more likely to be over 65 than blacks (like Southerners), and it is the 65+ group that is most at risk of Covid-19 death. But, irrespective of Simpson's paradox, the WHO age-adjusted death rate is higher for whites than blacks.

By repeating our analysis, we can examine whether Mackenzie's results are replicated in England and Wales death statistics on Covid-19 as collected by the Office of National Statistics (ONS) (Office for

National Statistics, 2020a). We looked at the most recent ONS report on Covid-19 deaths (registered up to 17 April 2020 from England and Wales) by ethnicity.

One of the main points the report made was the following:

> "When taking into account age in the analysis, Black males are 4.2 times more likely to die from a Covid-19-related death and Black females are 4.3 times more likely than White ethnicity males and females".

There was extensive national newspaper coverage of this worrying finding.

However, we found that the ONS report does not provide sufficient data or information to support these conclusions: it provides the total fatalities per ethnic group, including a breakdown of these for the <65 and 65+ age categories, but ***does not provide the critical information about the assumed proportion of whites and blacks in the population***. The conclusions that can be drawn are highly sensitive to small changes in the census year we might use for the population demographics and we present two different scenarios, one for the 2011 census and one from more recent estimates, as shown in Table 4.

Table 4 UK Covid-19 death rates (based on assumed England and Wales population of 62,411,850[1]) Simpson's paradox is evident in the 2020 estimate but not the 2011 census as shown by highlighted highest death rates in each category)

Data from ONS Report			2011 Census Population: 85% white, 3.4% black 14% of whites 65+ 5% blacks 65+		2020 Estimate Population: 78% white, 5.8% black 14% of whites 65+ 5% blacks 65+	
		fatalities	*population*	*fatalities per 100K*	*population*	*fatalities per 100K*
Total	white	10726	53050073	20.2	48681243	22.0
	black	766	2122003	36.1	3619887	21.2
age < 65	white	1036	45623062	2.3	41865869	2.5
	black	185	2015903	9.2	3438893	5.4
age 65+	white	9690	7427010	130.5	6815374	142.2
	black	581	106100	547.6	180994	321.0

Although 9 years out of date, the 2011 UK census is the most recent UK census with reported population proportions (85% white, 3.4% black). The 2020 estimated proportions (78% white, 5.8% black) are based on more recent estimates including extrapolations from the 2011 census about population changes in the previous 10 years. In both scenarios, we use the most recent ONS report on population age by ethnicity, which confirms that, indeed, whites are a much older population than blacks: 14% of whites are 65+ compared to only 5% of blacks.

Simpson's paradox is again evident in the 2020 estimate scenario where we see that, despite the death rate for whites being slightly higher overall, the death rates for blacks are more than twice that of whites in each disaggregated age subcategory. However, these results are not consistent with the ONS report conclusions. But, those of the 2011 census scenario are: although Simpson's paradox is not

[1] The most recent ONS report on UK population (Office for National Statistics, 2019) estimates the mid-2019 population of England as 56,286,961 and Wales as 3,152,897. It also estimates annual growth of 0.5%. Hence, we estimate current England and Wales population of 62,411,850.

evident here, in each age category the death rate for blacks is about four times that of whites as stated in the ONS report conclusions. This strongly suggests that the ONS analysis was based on the out of date 2011 census data.

Irrespective of whether we use the 2011 or 2020 estimates, in each age subcategory the death rate for blacks is over twice that as for whites. So, for any GIVEN age range a black person is over twice as likely to die of Covid-19 than a white person of the same age. Based on this presentation of the statistics we again ask the question: Which of the two groups (whites or blacks) is at higher risk of dying from Covid-19? The problem is that, the answer again is not as clear-cut as it seems - even when using the 2011 census data. What the ONS call the 'age adjusted figure' (and which appears to be some kind of averaging of the individual age category rates) is very different from the WHO definition of the age standardized death rate. As we noted, the WHO definition is equivalent to the overall death rate. So, in the 2011 census scenario it is higher for blacks (36.1) than whites (20.2) but nowhere near the four times 'age adjusted' figure claimed by the ONS, while in the 2020 estimate scenario it is higher for whites (22) than blacks (21.2).

It is vital to be as accurate as possible in analysis of data that carry such enormous consequences for public health policy making, yet, the analysis reveals two worrying insights.

1. The differences between the analysis here based on ONS data, and what the ONS claimed (due to using both the 2011 census data and the non-standard age adjusted death rate calculation applied).
2. As in the case of our hypothetical Bayesland example, finding that the death rate for blacks is higher than whites in the disaggregated data does not mean that the focus for public policy should be ethnicity. In either of the scenarios we considered, while both ethnicity and age are clearly shown as 'risk factors', **of the two, it is age which is the dominating factor**. In fact, whereas the death rate for blacks in each age group is between 2 to 4 times greater than that of whites**, the death rate for those aged over 65 is 54 times that of those aged under 65** (146.8 per 100K compared to 2.69 per 100K based on the 2020 estimates). There are, of course, many other factors that we have not considered here, but returning to the policy making perspective the conclusions stated in the ONS report are potentially highly misleading - much as they were for the recent ONS report on Covid-19 deaths by religion (Office for National Statistics, 2020b) as reported in (Fenton, 2020).

There is also a third concern about the ONS analysis. We believe that, on reflection neither the age-specific death rates nor the age standardized death rates are especially useful as they exaggerate the real risk to people. Leading statistician and risk expert David Spiegelhalter (Spiegelhalter, 2019) has convincingly argued why it is better – when discussing risk – to use absolute, not relative, risk differences and to express these as expected frequencies. With this approach, based on the ONS data (under the 2020 population estimates) we can conclude:

- For every 100,000 black people under 65 we expect about 3 more to die of Covid-19 than for every 100,000 white people (5.4 compared to 2.5 respectively in total). Equivalently a black person under 65 has a 0.0029% increased probability (about 1 in 35,000) of dying compared to a white person under 65.
- For every 100,000 black people aged 65+ we expect 179 more to die of Covid-19 than for every 100,000 white people (321 compared to 142 respectively in total). Equivalently a black person aged 65+ has a 0.179% increased probability (about 1 in 600) of dying compared to a white person aged 65+.

3. Conclusion

There has been great concern about the increased risk of Covid-19 to the Black and Minority Ethnic (BAME) community. Our analysis suggests that, while there is an increased risk to BAME over whites, the conclusions stated in the recent ONS report may create an unjustified level of fear and anxiety among the BAME community. The conclusion, which asserts that blacks are more than four times more likely to die from Covid-19 than whites, is misleading for three reasons: 1) It appears to rely on old 2011 census data about the population proportions rather than on more recent estimates; 2) It is based on an 'age standardized' measure that is very different from that used by the World Health Organisation (WHO); and 3) It focuses on relative rather than absolute measures of risk. Hence, we believe that the ONS conclusion may be misleading when assessing risk and making decisions about public health policy. As in our previous related studies (Fenton, 2020; Fenton, Neil, Osman, & McLachlan, 2020; Neil, Fenton, Osman, & McLachlan, 2020), our analysis has shown the need for causal models and explanations to supplement traditional statistical analysis. Here we have shown that age is a much more significant risk factor for Covid-19 death than BAME, and the Simpson paradox observed for black versus white death rates is explained by the fact that there are far more elderly whites than blacks in the population. Moreover, the increased risk to the BAME population may be partly explained by environmental risk factors such as poverty, social distancing practices, and diet rather than genetics.

4. References

Fenton, N. E. (2020). *A Note on UK Covid-19 death rates by religion: which groups are most at risk?* Retrieved from http://arxiv.org/abs/2007.07083

Fenton, N. E., Neil, M., & Constantinou, A. (2019). *Simpson's Paradox and the implications for medical trials*. Retrieved from http://arxiv.org/abs/1912.01422

Fenton, N. E., Neil, M., Osman, M., & McLachlan, S. (2020). Covid-19 infection and death rates: the need to incorporate causal explanations for the data and avoid bias in testing. *Journal of Risk Research*, 1–4. https://doi.org/10.1080/13669877.2020.1756381

Mackenzie, D. (2020). Causal Analysis in Theory and Practice » Race, COVID Mortality, and Simpson's Paradox. Retrieved July 10, 2020, from http://causality.cs.ucla.edu/blog/index.php/2020/07/06/race-covid-mortality-and-simpsons-paradox-by-dana-mackenzie/

Neil, M., Fenton, N. E., Osman, M., & McLachlan, S. (2020). Bayesian Network Analysis of Covid-19 data reveals higher Infection Prevalence Rates and lower Fatality Rates than widely reported. *Journal of Risk Research*. https://doi.org/10.1080/13669877.2020.1778771

Office for National Statistics. (2019). *Population estimates for the UK, England and Wales, Scotland and Northern Ireland, provisional - Office for National Statistics*. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2019

Office for National Statistics. (2020a). *Coronavirus (Covid-19) related deaths by ethnic group, England and Wales - 2 March 2020 to 10 April 2020*. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronavirusrelateddeathsbyethnicgroupenglandandwales/2march2020to10april2020

Office for National Statistics. (2020b). *Coronavirus (Covid-19) related deaths by religious group, England and Wales: 2 March - 15 May 2020*. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/coronavirusCovid-19relateddeathsbyreligiousgroupenglandandwales/2marchto15may2020

Pearl, J., & Mackenzie, D. (2018). *The book of why : the new science of cause and effect*. New York: Basic Books.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Retrieved from https://www.bookdepository.com/Art-Statistics-David-Spiegelhalter/9780241398630

WHO (World Health organization). (2006). *Age-standardized death rates per 100,000 by cause*. Retrieved from https://www.who.int/whosis/whostat2006AgeStandardizedDeathRates.pdf