HOW MANY PEOPLE DIED FROM INFLUENZA IN 1918?

David J.D. Earn
Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada, L8S 4K1

The "Spanish flu" pandemic that began in 1918 was responsible for an extraordinarily large number of deaths. Until the 1990s, the most commonly quoted figure was "20 million people worldwide". Today it is more common to hear much larger numbers, such as 50 million (more than the current population of Canada). **Wikipedia** says "50 to 100 million" died in the Spanish flu pandemic, consistent with one of the most highly cited research articles on the subject[1].

What is the truth?

WHY IS THE NUMBER OF PANDEMIC FLU DEATHS HARD TO ESTIMATE?

Suppose we were interested simply in the number of people who died *during* the pandemic, without worrying about whether it was actually influenza that killed them. In countries such as Canada where the official registration of deaths began long before 1918, it is relatively straightforward to obtain counts of the total number of people who died that year. But in 1918, many countries around the world were not in the practice of systematic record-keeping in relation to deaths, so inevitably any estimate of the number of deaths worldwide will be subject to potentially substantial error. Remember that this is true even if we're just trying to estimate the total number of people who died, not only those who died specifically from flu.

Let's focus on what happened during the Spanish flu pandemic in Canada, and in particular the province of Ontario, which my research group has been studying for a number of years.

There were three epidemic waves that are usually considered part of the pandemic. The first wave was in the spring of 1918, the second in the fall, and the third in the winter and early spring of 1919[2]. In Ontario, these three distinct waves are not easy to identify. It isn't clear when the first wave started and ended, and the third wave appeared to start before the second wave ended. To be sure we don't miss the start and that we catch all three waves, let's consider the time period from January 1, 1918, to April 30, 1919, shown in the graph at right (Figure 1).

Figure 1: *Total mortality in Ontario during the 'Spanish Flu' pandemic.*

---
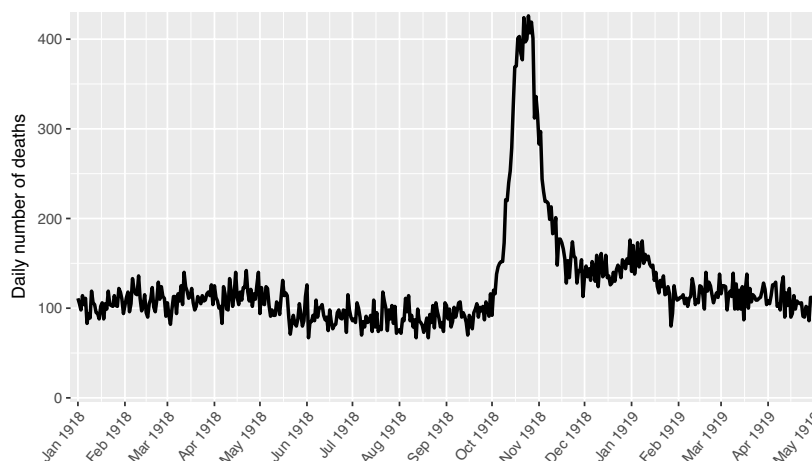
1   Johnson and Mueller 2002.

2   He et al. 2013.

The total number of deaths registered in Ontario in those 16 months was 60,547. It is possible that a very small number of deaths may have escaped registration or that some were registered incorrectly. But the figure of 60,547 is probably not wrong by much. To reflect the fact that we can't rely on the reliability of this number down to the last individual, we'll round it up to 61,000.

Determining how many of those deaths were caused by "Spanish flu" is much more difficult.

It is reasonable to assert with confidence that fewer than 61,000 Ontarians died of *influenza* during the pandemic. (It would be absurd to suggest that nobody died of anything else for 16 months!) The population of Ontario in 1918 was approximately 2,750,000 (estimated by **linearly interpolating** between the 1911 and 1921 census populations). Consequently, the percentage of the population that perished from the flu cannot have been greater than

$$\frac{61,000}{2,750,000} \times 100 = 2.2\%.$$

This *upper bound* is helpful. Can we also calculate a *lower bound* on the percentage of the population that *died of flu* in 1918-1919 in Ontario?

Fortunately, Ontario death registrations indicate a cause of death. Of the 60,547 deaths registered in the 16 months of interest, 10,476 list influenza as a cause. So the percentage of the total population that died of flu (at least as a contributing factor) was probably no less than

$$\frac{10,000}{2,750,000} \times 100 = 0.36\%.$$

This is definitely just a lower bound, since many deaths would have been attributed to the "last straw", e.g., pneumonia, a heart attack or another illness that would not have occurred (or would not have been fatal) if the person had not first contracted the flu. Incidentally, *some* of the deaths attributed to influenza in the death registrations were probably caused by other infections with similar symptoms. But during the 1918-1919 pandemic, the vast majority of "influenza-like illness" was probably truly flu.

With these calculations, we can now conclude that that number of people who died of influenza in Ontario between January 1, 1918, and April 30, 1919, was probably between 10,000 and 61,000. Put another way, it is likely that between 0.36% and 2.2% of the total population of Ontario died of flu during the pandemic.

That range, however, is not very precise. Can we do better?

People suffering from a bad case of influenza often contract pneumonia as a secondary infection. If they

die, then it is common for pneumonia to be listed as the primary (if not only) cause of death. The number of deaths in Ontario attributed to *either* influenza *or* pneumonia (or both) was 18,804 during the 16 months in question. If all these people really had the flu and died *because* of it then our percentage estimate becomes

$$\frac{19,000}{2,750,000} \times 100 = 0.69\%.$$

We could also count the number of people whose deaths were attributed to heart attacks, or various other causes that *might* be linked to an influenza infection. But it is difficult to determine if doing so would improve our estimate. Based on the simple calculations above, if somebody were to ask me how many people died of influenza during the Spanish flu pandemic in Ontario, I'd say "If you want one specific number then *around* **20,000** people is a reasonable estimate, but all I can say with confidence is that I think it is extremely likely that the true number is **between 10,000 and 60,000**". (Note that since I'm so uncertain, I'm giving only one digit of precision.)

Researchers work hard to improve on these types of estimates, because they help frame discussions about planning and preparedness for future influenza epidemics. Quantifying the "mortality burden" of influenza each year (not just during pandemics) is a never-ending challenge[3]. The most common approach is to look at mortality patterns over several years, attempt to subtract "normal" deaths that have nothing to do with influenza, and then attribute the remaining "excess deaths" to influenza. The same idea can be applied during pandemic years, but this is a messy business, to say the least. The graph below (Figure 2) compares mortality from all causes with mortality attributed to pneumonia and influenza (or just influenza). Does examining this graph give you more or less confidence in the crude estimates above?

Figure 2: *Mortality in Ontario during the "Spanish Flu" pandemic in three categories: All causes, Pneumonia & Influenza, and Influenza*
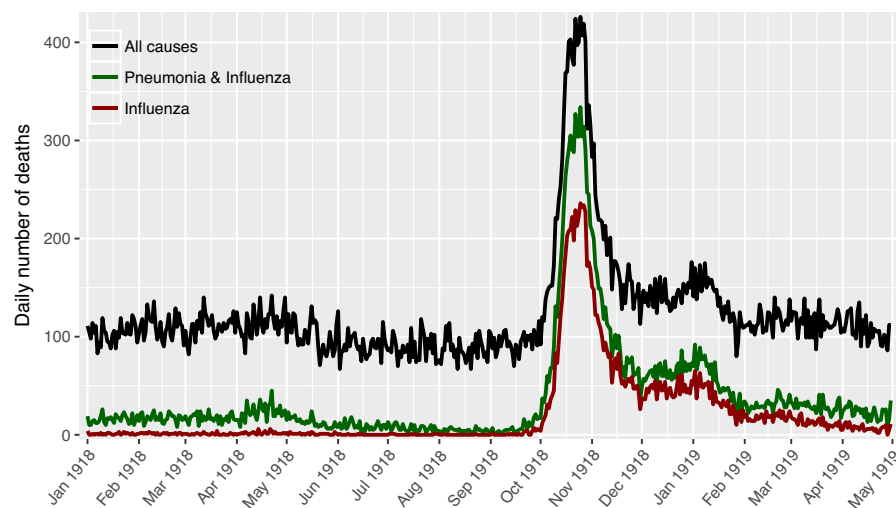
3   Iuliano et al. 2017.

Another reason pandemic influenza mortality estimates vary is that there is no general agreement about exactly what time period should be included when counting the number of deaths "during the pandemic". Above, I considered the 16 month period from the beginning of 1918 to the spring of 1919. But there was another large influenza epidemic in 1920, which was associated with the same influenza virus. Some estimates include 1920, while others don't. You could take a more extreme position and argue that the flu epidemics in 1921, 1922 and later should also be included in the count, since the same virus was responsible for those outbreaks. At the other extreme, you might feel that it is better to focus only on the "main" wave that occurred in the fall of 1918. These issues make it more difficult to compare mortality estimates made by different researchers.

Summarizing, I think there are two main reasons why the number of people who perished in the Spanish flu pandemic is hard to estimate. First, it is very difficult to determine which deaths resulted from influenza infections. Second, the time period that is considered to be "the pandemic" is hard to define. Hopefully these comments help you understand why the number of people who died from influenza during "the Spanish flu pandemic" is so uncertain.

HOW SHOULD WE QUANTIFY UNCERTAINTY?

Research articles in the scientific literature normally report quantities together with some sort of estimate of uncertainty, often shown visually with an **"error bar"**. An error bar based on the range of uncertainty I gave above would represent a worst-case scenario (i.e., it is essentially impossible that the number of Spanish flu deaths during the pandemic in Ontario was less than 10,000 or greater than 60,000).

It is more common to report a **"most likely"** value and a **"95% confidence interval"** (CI) within which the most likely value lies. For instance, a conscientious author might report that in some particular city the number of people who died from flu in 1918 was "4960 (95% CI [1729,8128])". This means that the author's analysis indicates that it is most likely that 4960 people died, but that s/he is only 95% sure that the true number of deaths lies somewhere between 1729 and 8128. In contrast, I feel nearly 100% sure that the number of Spanish flu deaths in Ontario was between 10,000 and 60,000.

Defining what we really mean by **"likelihood"**, and how we can actually compute the likelihoods of various estimates, are fascinating challenges that are explored in university courses in statistics.

WHY IS UNCERTAINTY OFTEN NOT MENTIONED?

Measures of uncertainty (error bars) are usually ignored in articles about science or medicine in the popular press and other media. I think one reason for this is that, with the best of intentions, journalists tend to try to simplify complex messages so that readers can grasp them quickly. Of course, scientists who are interviewed can be guilty of the same tendency to simplify, for the same reason.

This is not a problem that is easy to overcome. Scientists are often given the impression that they should avoid emphasizing uncertainty, because attention to uncertainty could dilute the important messages and even be interpreted as a lack of expertise.

WHAT SHOULD WE DO ABOUT IT?

Be aware that the sources of numbers quoted in the media probably provided error bars, and the uncertainty might be extremely large. If a number is quoted as "1000" but the original author's 95% confidence interval was from 100 to 10,000 then just quoting "1000" is misleading. If the numbers you are reading about are important to you, then it is best to dig into the primary scientific literature where the original studies were published. If you feel that the media misrepresented the original studies, then politely point that out to the journalists (bearing in mind that most journalists are doing their best with the time they have available).

Of course, you won't always have time to pursue the sources. But it is a good idea in general to be skeptical about numbers that are quoted in the media. If no uncertainty is mentioned, ask about it. The degree of uncertainty associated with a number can be as important as the number itself.

ACKNOWLEDGEMENTS

REFERENCES

He, D., J. Dushoff, T. Day, J. L. Ma, and D. J. D. Earn. 2013. "Inferring the Causes of the Three Waves of the 1918 Influenza Pandemic in England and Wales." *Proceedings of the Royal Society of London, Series B, Biological Sciences* 280 (1766): 20131345. doi:**10.1098/rspb.2013.1345**.

Iuliano, A. D., K. M. Roguski, H. H. Chang, D. J. Muscatello, R. Palekar, S. Tempia, C. Cohen, et al. 2017. "Estimates of Global Seasonal Influenza-Associated Respiratory Mortality: A Modelling Study." *The Lancet* 391 (10127): 1285–1300. doi:**10.1016/S0140-6736(17)33293-2**.

Johnson, N. P. A. S., and J. Mueller. 2002. "Updating the Accounts: Global Mortality of the 1918-1920 'Spanish' Influenza Pandemic." *Bulletin of the History of Medicine* 76 (1): 105–15. doi:**10.1353/bhm.2002.0022**.