hw3

Jason Pekos

27/09/2021

Plots

For the first plot, I'll just do a standard CI-type plot for each region separately. I think this keeps clutter to a minimum, and emphasizes the subcategories in the original table. There are some other options here that I decided against:

- Bar plots
- Connecting the points with `geom_line()`

I didn't like bar plots here because I feel that just using `geom_point()` emphasizes that this is a point estimate from a regression, and I feel that the bars add unnecessary clutter to the error bars. Also the presence of negative point estimates here doesn't look great on a bar plot (in my opinion).

I decided not to use `geom_line()` because I feel it helps with absolutely nothing, even though I saw it used online a lot for this type of task. If there was some change over time, I would probably consider a super light version of it (thin mid-opacity dashed lines), but that isn't a part of this table.

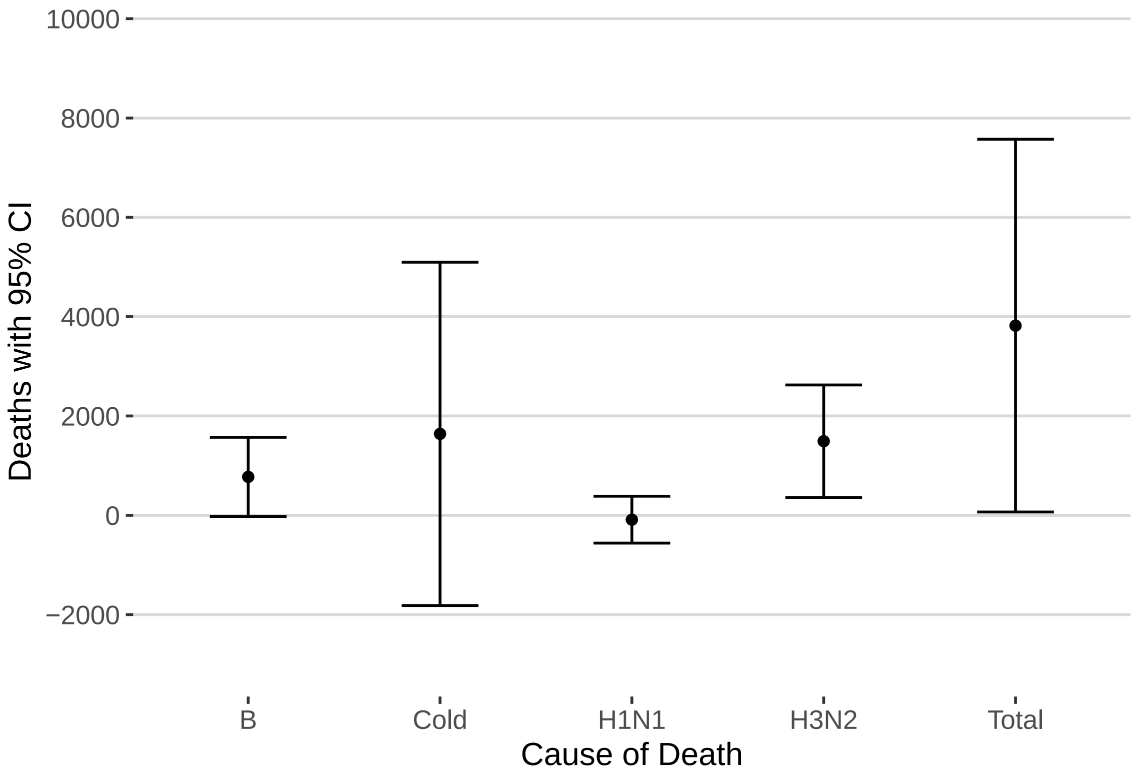
I made a few other minor additions — most importantly, more breaks to help with interpretability.

Overall though, there's only a few pieces of information we actually care about encoding here:

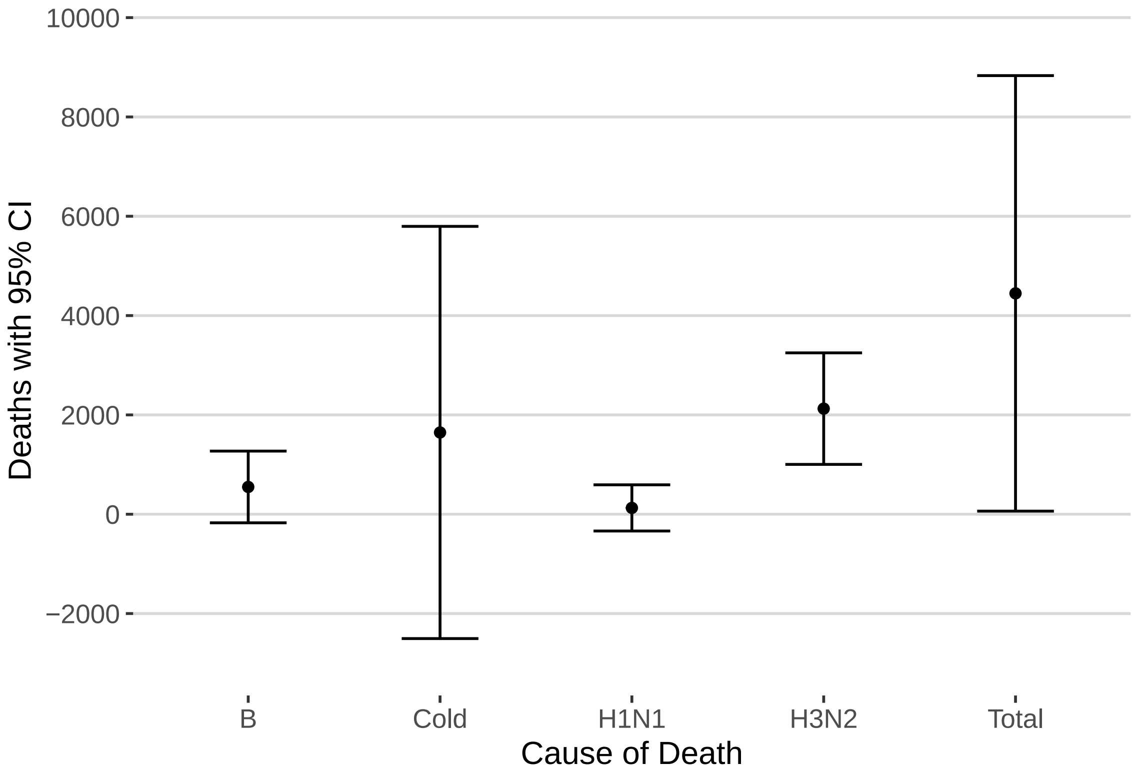
- Geographic area studied
- Point estimate of Deaths
- Disease name
- Error estimates

We deal with the first by simply making two plots. We encode the second with position on a common scale, and we add the error estimates — which are in the same units — to the same scale. The only thing I'm unsure about is (not) encoding disease name with colour. I didn't end up doing this because it's already 'encoded' in some sense by position on the x-axis. It's a very simple table, and there isn't an interpretability issue with a graph like this. Especially with only five categories, I feel this would add needless clutter to a nice and simple graph.

Estimated Deaths in New York Metropolitan Area



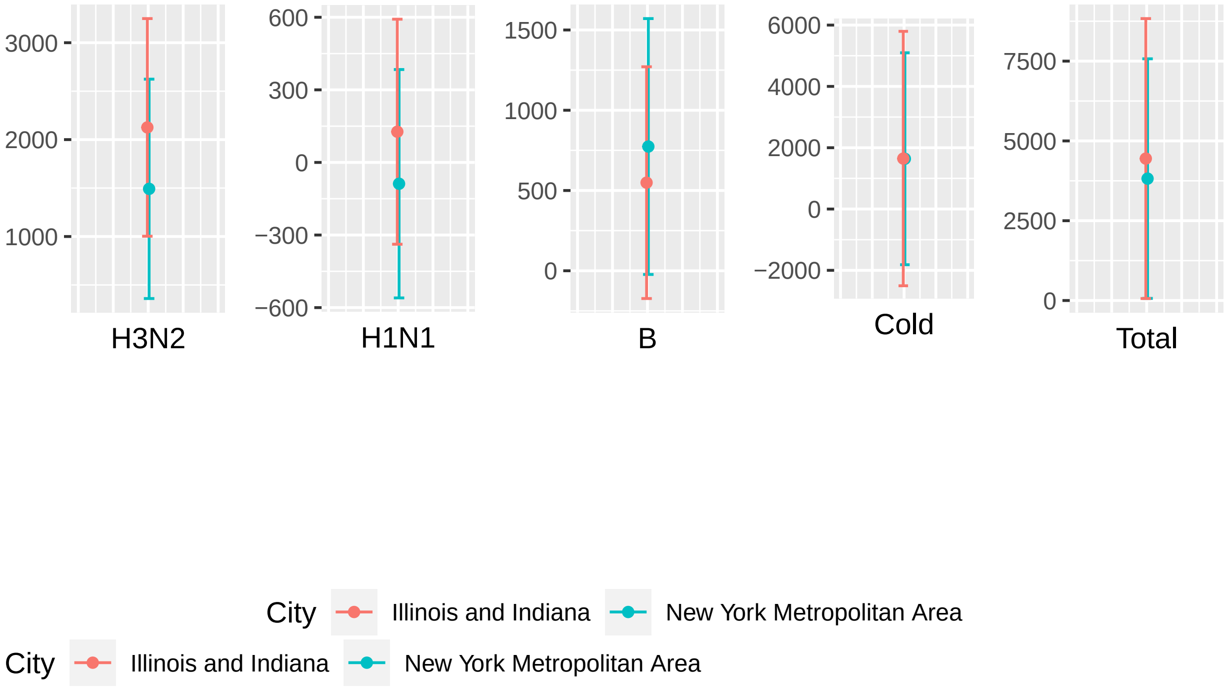
Estimated Deaths in Illinois and Indiana



Alternate Graph:

Simply separating out the categories seems a little like cheating, so an alternate graph where we plot the same disease for both cities on the same graph, but splitting out each disease — we do this because the wide range of estimates across diseases and the tiny variation within diseases makes differentiation geographically super tough.

Graph of estimated deaths + 95% CIs for each respiratory illness:



Purpose and Discussion

The purpose of these graphs is to show the point estimates + 95% error bars from a regression model intended to measure seasonal respiratory virus mortality.

I've already talked about the considerations for graph one. In graph two, we now encode geographic location by color, split the illness classifications into different sub-graphs, and encode deaths via position on a common scale. The purpose of this graph is to compare deaths due to the same disease in different municipalities — I'd argue that the most important metric here is still deaths, and that even though we encode location primarily by colour, we also implicitly encode it into position by keeping one point estimate from each location in each box. This makes it easier to compare the metric we care about (deaths) — keeping the points of interest next to each other.

This respects Cleveland's hierarchy for the most part, and I find the graph very readable.