Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of Covid-19 in England

Lay summary

The reproduction number (R) is the average number of infections caused by each infected person. This represents how quickly an epidemic spreads, which both policy-makers and the public have used to better understand the Covid-19 pandemic. Researchers usually estimate R over time (Rt) using public health data, including positive Covid-19 tests, hospital admissions, and deaths. However, these data may not fully represent the whole population, because each is influenced differently by testing rates and the average age of cases. We explored how this impacts estimates of Covid-19 transmission in England. For example, in spring 2020 Rt estimated from hospital admissions was higher than Rt estimated from all positive tests, probably reflecting outbreaks among vulnerable care home residents. This means we should consider the data used to estimate Rt before making conclusions about population-wide transmission.

Formal abstract

The time-varying reproduction number (R<sub>t</sub>: the average number of secondary infections caused by each infected person) may be used to assess changes in transmission potential during an epidemic. While new infections are not usually observed directly, they can be estimated from data. However, data may be delayed and potentially biased. We investigated the sensitivity of R<sub>t</sub> estimates to different data sources representing Covid-19 in England, and we explored how this sensitivity could track epidemic dynamics in population sub-groups.

We sourced public data on test-positive cases, hospital admissions, and deaths with confirmed Covid-19 in seven regions of England over March through August 2020. We estimated R<sub>t</sub> using a model that mapped unobserved infections to each data source. We then compared differences in R<sub>t</sub> with the demographic and social context of surveillance data over time.

Our estimates of transmission potential varied for each data source, with the relative inconsistency of estimates varying across regions and over time. R<sub>t</sub> estimates based on hospital admissions and deaths were more spatio-temporally synchronous than when compared to estimates from all test-positives. We found these differences may be linked to biased representations of subpopulations in each data source. These included spatially clustered testing, and where outbreaks in hospitals, care homes, and young age groups reflected the link between age and severity of disease.

We highlight that policy makers could better target interventions by considering the source populations of R<sub>t</sub> estimates. Further work should clarify the best way to combine and interpret R<sub>t</sub> estimates from different data sources based on the desired use.