Update on estimates of numbers of COVID-19 deaths accounting for reporting delay

Shaun Seaman and Daniela De Angelis (17/06/20) on behalf of the MRC-BSU COVID-19 Working Group

1 Background

Death occurrences are subject to delay in reporting, and the analysis of deaths by date of death is inevitably distorted by such delay. The ability to now-cast and forecast the death burden necessitates appropriate correction of the observed data to estimate the number of deaths that have occurred but not yet been reported. This is the current situation with COVID-19 related mortality, where observed data do not provide enough information to identify trends in mortality and assess the impact of the social distancing measures.

This report provides updated estimates of the numbers of deaths that occurred on each day up to June 16th.

2 Data sources

Each day we receive a data extract with deaths from three sources:

- Demographics Batch Service (DBS): a mechanism that allows Public Health England (PHE) to submit a file of patient information to the NHS Spine for tracing against the personal demographics service (PDS). PHE submit a line list of patients diagnosed with COVID-19 to DBS daily. The file is returned with a death flag and date of death updated (started 20th March).

- NHS England: reports data from NHS trusts relating to patients who died after admission to hospital or within emergency department settings.

- Health Protection Teams (HPT): a select survey has been created by PHE to capture deaths occurring outside of hospital settings e.g. care homes (started 23rd March).

These data cover almost all the deaths with confirmed COVID-19 infection. Data contain information on day of death, date of report to each of the data sources and demographic information (e.g. age).

Also available are data on death registrations from the Office for National Statistics (ONS). However, we do not use these data, because the reporting delay distribution for these deaths is somewhat irregular and thus harder to model. These ONS data also became available later in calendar time than did data from the other three sources (NHS, DBS and HPT). A consequence of not using the ONS data is that we are excluding deaths that will ultimately only be reported by ONS and not by...
NHS, DBS or HPT. Thus, the numbers we report below will underestimate the total numbers of deaths. However, the degree of underestimation is not expected to change over calendar time, and so our estimates should provide a useful guide to the trends in numbers of deaths.

3 Data used in this update

We used data on deaths that occurred on or after March 22nd and were reported by June 16th. So, we were able to estimate the distribution of reporting delays up to 86 days, and to estimate the numbers of death occurring on each day from March 22nd up to June 16th. These numbers of deaths include only those deaths that are reported within 82 days of death. This is for two reasons. First, our data do not enable us to estimate the numbers of additional deaths that will be reported after more than 86 days. Second, we can reduce the variability in our estimates by restricting the maximum allowed delay (in this case, 82 days) to be slightly less than the maximum that we could allow (86 days).

We considered two types of data streams: 1) only those deaths reported by NHS; and 2) all deaths regardless of whether reported by NHS, DBS or HPT (when an individual was reported by more than one of these three sources, we used the earliest of these two or three reports).

4 Results

Figures 1 and 2 show, for each English region, the estimated numbers of all deaths (solid black line), with 95% point-wise confidence intervals (broken red lines). These numbers were obtained using a model for the reporting delay that allows the reporting pattern to be different on Sundays and Mondays from that on other days of the week (see Appendix). The estimate of the number of deaths on the most recent day (June 16th) is omitted, because there is a lot of uncertainty associated with this, due to few deaths being reported on the same day as they occur.

The blue circles show the numbers of deaths on each day that had been reported by 16th June.

Figures 3 and 4 shows the estimated number of NHS deaths. Note that in both Figures, the estimates for the most recent two days are very unstable and quite unreliable. The confidence intervals are narrow. This is because they are calculated under the assumption that, apart from the Sunday and Monday effects, the probability that a death is reported within a given number of days is the same regardless of which day that death occurred. In reality, there is more variability in this probability, but a more complex modelling approach would be required to allow for this.

Figure 5 shows the estimated numbers of all deaths by age group. Figure 6 shows the corresponding numbers of NHS deaths.
Figure 1: Estimates of the number of all new deaths by date of death
Figure 2: Estimates of the number of all new deaths by date of death
Figure 3: Estimates of the number of new NHS deaths by date of death
Figure 4: Estimates of the number of new NHS deaths by date of death
Figure 5: Estimates of the number of new (all) deaths by age group and date of death
Figure 6: Estimates of the number of new (NHS) deaths by age group and date of death
5 Appendix - Method

The method used to estimate the numbers of deaths was that described by Brookmeyer and Damiano (1989) ‘Statistical Methods For Short-Term Projections of AIDS Incidence’, Stat Med, vol 8, pp 23–34. Let $Y_{td}$ denote the number of deaths occurring on day $t$ and reported on day $t + d$ ($t = 0, \ldots, 86; d = 0, \ldots, 28$), where $t = 0$ means March 22nd, $t = 1$ means March 23rd, $\ldots$, $t = 86$ means June 16th. Also, let $Y_{t,82}$ be the number of deaths occurring on day $t$ and reported between days $t + 29$ and $t + 82$ (inclusive). We observe $Y_{td}$ when $t + d \leq 86$; it is missing otherwise. Our reason for grouping together all deaths with delays greater than 28 days is to reduce the computation load.

Fit the Poisson regression model with

$$
\log E(Y_{td}) = \alpha_t + \beta_d + \gamma_{\text{sun}} I(t + d = \text{Sunday}) + \gamma_{\text{mon}} I(t + d = \text{Monday})
$$

where $\alpha_t$ ($t = 0, \ldots, 86$), $\beta_d$ ($d = 0, 1, \ldots, 27, 28, 82$) and $\gamma_{\text{sun}}$ and $\gamma_{\text{mon}}$ are unknown parameters. The inclusion of the two indicator functions $I(t + d = \text{Sunday})$ and $I(t + d = \text{Monday})$ as covariates means that this model allows for Sunday and Monday reporting effects. Let $\hat{\beta}_d$, $\hat{\gamma}_{\text{sun}}$ and $\hat{\gamma}_{\text{mon}}$ be the estimates of, respectively, $\beta_d$, $\gamma_{\text{sun}}$ and $\gamma_{\text{mon}}$ thus obtained.

For an individual who dies on day $t$, the estimated probability that the reporting delay equals $d$ days given that it is less than or equal to $d$ days is

$$
\hat{g}_{td} = \frac{\exp\{\hat{\beta}_d + \hat{\gamma}_{\text{sun}} I(t + d = \text{Sunday}) + \hat{\gamma}_{\text{mon}} I(t + d = \text{Monday})\}}{\sum_{k=0}^{d} \exp\{\hat{\beta}_k + \hat{\gamma}_{\text{sun}} I(t + k = \text{Sunday}) + \hat{\gamma}_{\text{mon}} I(t + k = \text{Monday})\}}
$$

(this can be interpreted as a hazard in reverse time). So, the estimated probability that, for an individual who dies on day $t$ and is reported within 82 days, the reporting delay is less than or equal to $d$ days is

$$
\hat{F}_{td} = \prod_{j=d+1}^{28} (1 - \hat{g}_{tj}) \times (1 - \hat{g}_{t,82})
$$

if $0 \leq d \leq 28$, and is $\hat{F}_{t,82} = 1$ if $d = 82$.

The estimated number of deaths occurring on day $t$ ($t = 0, \ldots, 86$) is then

$$
\sum_{d=0}^{28} Y_{td} + Y_{t,82}
$$

if $0 \leq t < 5$,

$$
\frac{\sum_{d=0}^{28} Y_{td}}{\hat{F}_{t,28}}
$$

Note: A better strategy than grouping all delays of greater than 28 days into one group would have been to group these longer delays into weeks.
if $5 \leq t \leq 58$, and

$$\frac{\sum_{d=0}^{86-t} Y_{td}}{F_{t,86-t}}$$

if $59 \leq t \leq 86$.