Adjusting COVID-19 deaths to account for reporting delay

Shaun Seaman and Daniela De Angelis (26/04/20) on behalf of the MRC-BSU COVID-19 Working Group and the PHE Modelling Cell

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.

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).

3 Data used in this update

We used data on deaths that occurred on or after March 22nd and were reported by April 25th. So, we were able to estimate the distribution of reporting delays up to 34 days, and to estimate the numbers of death occurring on each day from March 22nd up to 25th April. These numbers of deaths include only those deaths
that are reported within 34 days of death: our data do not enable us to estimate the numbers of additional deaths that will be reported after more than 34 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

Figure 1 shows, 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 blue circles show the numbers of deaths on each day that had been reported by 25th April. Figure 2 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.

Figure 3 shows the estimated numbers of all deaths by age group. Figure 4 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 new NHS deaths by date of death
Figure 3: Estimates of the number of new (all) deaths by age group and date of death
Figure 4: 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 $t + d$ ($t = 22$nd March, ..., 25th April; $d = 0, \ldots, 34$). We observe $Y_{td}$ when $t + d \leq 25$th April; it is missing otherwise.

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 $\alpha_t$ ($t = 22$nd March, ..., 25th April), $\beta_d$ ($d = 0, \ldots, 34$) 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 34 days, the reporting delay is less than or equal to $d$ days is

$$\hat{F}_{td} = \prod_{j=d+1}^{34} (1 - \hat{g}_{tj})$$

if $0 \leq d \leq 33$, and is $\hat{F}_{t,34} = 1$ if $d = 34$.

The adjusted number of deaths occurring on day $t$ is then

$$\frac{\sum_{d=0}^{34-1} Y_{td}}{\hat{F}_{t,34-t}}.$$