Challenges in estimating the distribution of delay from COVID-19 death to report of death

Shaun Seaman and Daniela De Angelis (17/04/20)

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 reported daily death 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 data do not provide enough information to identify trends in mortality and to 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).

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 set time zero to December 31st 2019. So, in particular, time 1 is January 1st 2020, time 82 is March 22nd and time 107 is April 16th. We used data on deaths that occurred on or after time 82 ($D \leq 82$) and were reported by time 107 ($D+T \leq 107$). So, we were able to estimate the probabilities that the delay equals t for each of $t = 0, \ldots, 25$. We considered two types of data streams: 1) just NHS deaths; and 2) all deaths regardless of whether reported by NHS, DBS or HPT

(when an individual was reported by more than one source, we used the earliest of these two or three reports).

To investigate how the estimates fluctuate as data accumulate, we also analysed just the reports that had been made by each of times 97, 98, ..., 106 and compared the estimates that were obtained.

In this report, the estimated delay distributions and the resulting estimates of numbers of deaths were calculating using a model that allows the reporting pattern to be different on Mondays from that on other days of the week (see Appendix for an explanation of how this is achieved through the inclusion of covariates).

4 Results

Figure 1 shows the estimated numbers of all deaths by day, with 95% pointwise confidence intervals. The circles indicate the number of deaths on each day that had been reported by 16th April.

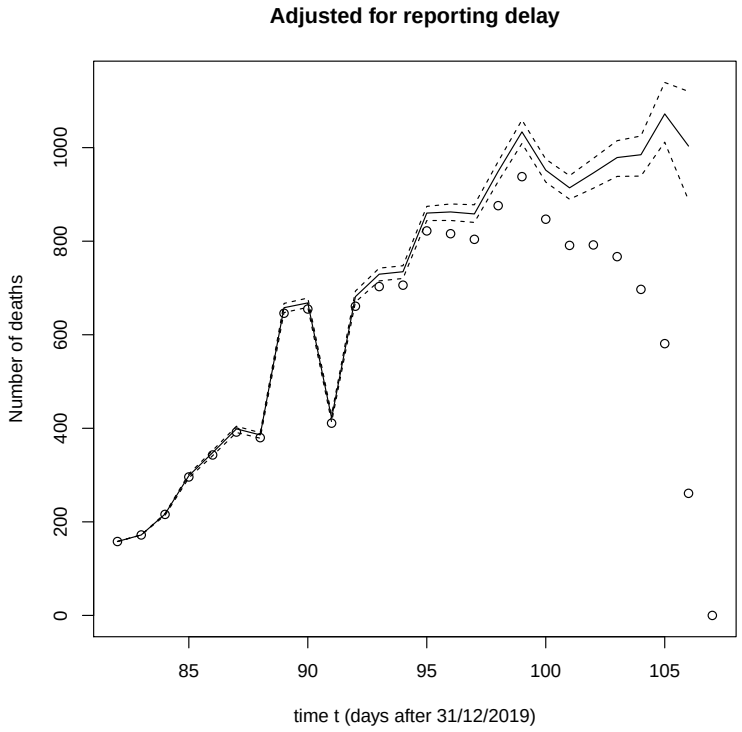


Figure 1: Estimates of the number of all new deaths over time

Figure 2 shows the corresponding numbers for NHS deaths only. Note that in both figures the estimates for the last two times are very unstable and quite unreliable. Also, due to statistical uncertainty and the use of different estimated delay distributions, the peak of the estimated daily deaths curve for NHS deaths is higher than the peak for all deaths.

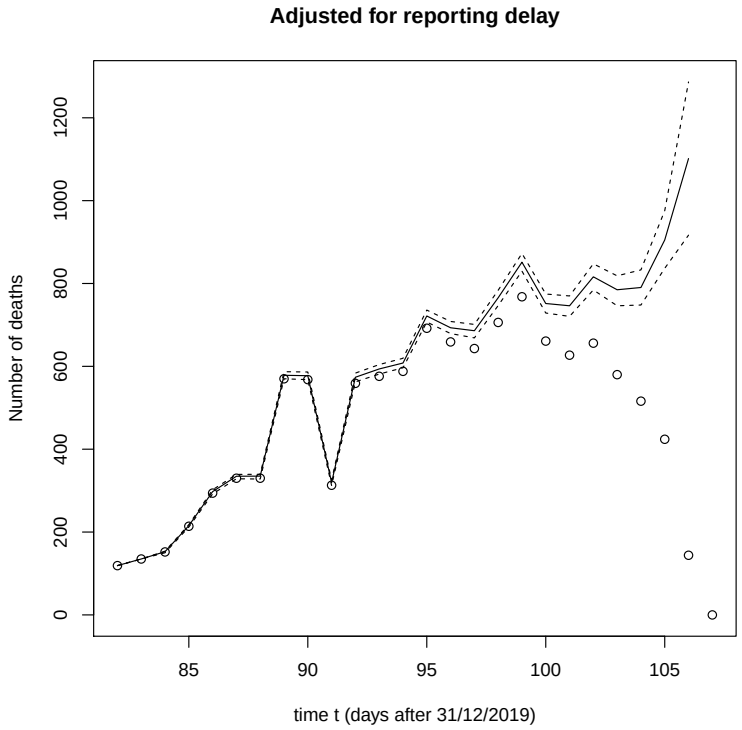


Figure 2: Estimates of the number of new NHS deaths over time

4.1 Sensitivity of the estimated reporting delay distribution and adjusted counts

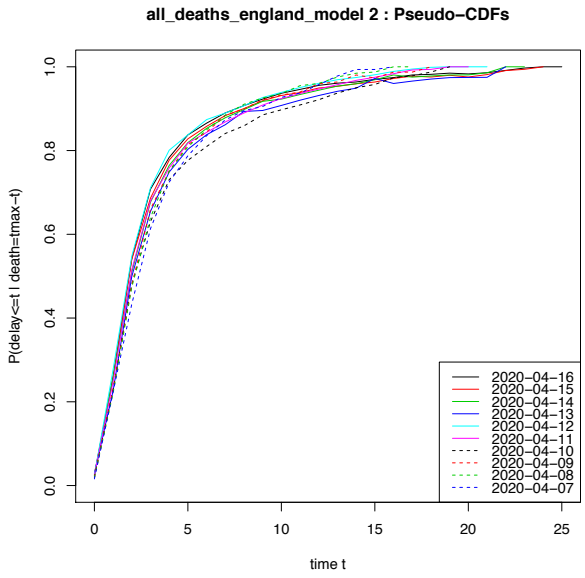
The estimated delay distributions and the resulting estimated numbers of deaths are extremely sensitive to the irregular reporting patterns. Figure 3 shows some of these sensitivities for all deaths in England.

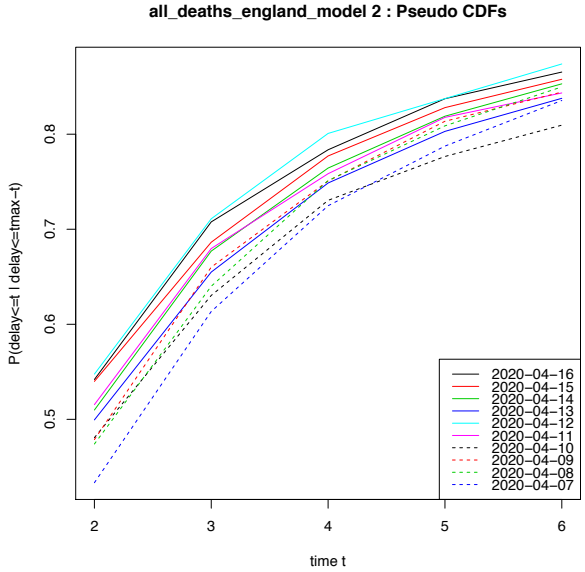
The first panel shows how the cumulative distribution function of the delay changes as data accumulate, from the 7th of April to April 16th. The delay distribution is different for different dates of death, because of the Monday reporting effect. For this reason, each cumulative distribution function shown here (the 'pseudo-CDF') is an amalgam of different cumumulative distribution functions (see Appendix for details), and so is not necessarily a monotone function of time.

The second panel of Figure 3 shows how the probability of a reporting delay of two days or less changes dramatically over the period considered, with an obvious consequence on the weighting of the incomplete number of deaths in the most recent times. The probabilities of a report with short delay are particular influential on the resulting estimated numbers of deaths.

The third panel in Figure 3 demonstrates the volatility of the resulting estimated numbers of deaths. The same phenomenon was observed for the NHS deaths.

In conclusion, it can be quite challenging to estimate the number of deaths that occurred on each day, and these estimates can be sensitive to the dataset used.





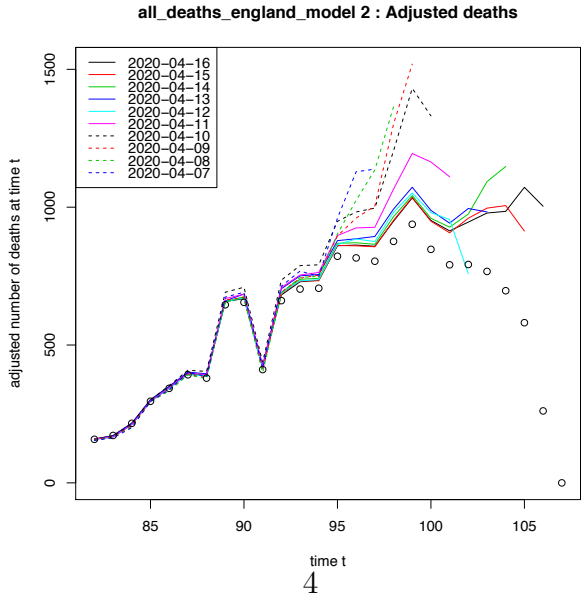


Figure 3: Sensitivity of results to the choice of dataset

Appendix: Statistical methods

Consider a (finite) population of individuals who will ultimately die from Covid-19. Let D and T denote an individual's time of death and 'delay' (i.e. time from death to report), respectively. Both of these are both measured in discrete time (days). Let X denote a vector of 'baseline' covariates, i.e. covariates whose values are determined at or prior to time D.

Our sample consists of all the individuals in the population for whom $D \geq 0$ and $D + T \leq t_{\max}$ for some time $t_{\max} > 0$, i.e. all the individuals who died at time zero or later and whose deaths had been reported by time $t_{\max}$. We use i ($i = 1, \ldots, n$) to index the sampled individuals. Our objective is to estimate m_j, the number of individuals in the population who died at time j ($j = 0, \ldots t_{\max}$).

If we knew $F(t, x) = P(T \leq t \mid X = x)$, we could estimate m_j as

$$\sum_{i=1}^{n} \frac{I(D_i = j)}{F(t_{\max} - j, x_i)}.$$

However, $F(t, x)$ is unknown, and so we have to estimate it.

As the sample only includes individuals who are reported within $t_{\max}$ days of their death, we cannot (non-parametrically) estimate $F(t, x)$, but we can estimate $F^*(t, x) = F(t, x)/F(t_{\max}, x)$ for $t \leq t_{\max}$. If we plug this estimate into the formula

$$\sum_{i=1}^{n} \frac{I(D_i = j)}{F^*(t_{\max} - j, x_i)}, \tag{1}$$

we get an estimate of the number of individuals in the population who died at time j ($j = 0, \ldots t_{\max}$) and who have been or will be reported within $t_{\max}$ days of their death.

To estimate $F^*(t, x)$, we use the semi-parametric estimator of (among others) Brookmeyer and Damiano (1989 Stat Med). Note that, when there are no covariates X, this estimator reduces to the non-parametric maximum likelihood estimator, which is available in closed form.[1] Define

$$g_j(x) = P(T = j \mid T \leq j, X = x),$$

which can be interpreted as the hazard at time j in reverse time. Clearly,

$$F^*(t, x) = \prod_{j=t+1}^{t_{\max}} \{1 - g_j(x)\}$$

for $t = 0, \ldots, t_{\max} - 1$, and $F^*(t_{\max}, x) = 1$. We therefore proceed to estimate $g_j(x)$, which we do as follows.

Let Y_{dj} denote the number of individuals in the sample with $D = d$ and $T = j$ (for $0 \leq d \leq d + j \leq t_{\max}$). Fit the Poisson generalised linear model with

$$\log E(Y_{dj}) = \alpha_d + \beta_j + \gamma^\top h(X, d, j)$$

[1] It is a Kaplan-Meier estimator in reverse time, with late entry to the risk set.

where α_d $(d = 0, \ldots, t_{\max})$, β_j $(j = 0, \ldots, t_{\max})$ and γ are unknown parameters, and $h(X, d, j)$ is some specified vector function of d, j and the covariates X. In particular, we use $h(X, d, j) = \phi(d + j) \times \big(I(j = 0), \ldots, I(j \leq 7), I(j > 7)\big)^\top$, where $\phi(z)$ equals one if day z is a Monday and equals zero otherwise. This Poisson model therefore allows the pattern of reporting to be different on Mondays from the pattern on other days. The probability $g_j(x)$ can now be estimated as

$$g_j(x) = \frac{\exp\{\beta_j + \gamma^\top h(X, d, j)\}}{\sum_{k=0}^{j} \exp\{\beta_k + \gamma^\top h(X, d, k)\}}$$

In this report, we use the term 'pseudo-CDF' to refer to the function $F^\dagger(t) = F^*(t, t_{\max} - t) = P(T \leq t \mid D = t_{\max} - t, T \leq t_{\max})$.