Improving timely analysis of UK COVID-19 cases by Specimen Date

By David Hindley and John Roberts

The authors would like to thank Richard Kelsey for his review of this article.

9 October 2020

Summary

This article explains how data for UK COVID-19 can be structured and analysed to gain insights into the development of positive cases by Specimen Date. However, this analysis is not currently possible with the data that is made available on the UK Government's coronavirus dashboard[1]. There is clearly a need for the most up to date and accurate information possible regarding the trend in infections, as has recently been highlighted due to the significant amendment to the England cases date, as reported on 3rd and 4th October.

We show how standard actuarial techniques, widely used in insurance reporting, enable us to estimate the outcome of recent days' specimen results, where we have yet to receive the full data, to gain faster insight into trends.

We recommend Public Health England publishes the data in the way described to enable users of the data to gain additional insights into the development of case numbers by Specimen Date.

Introduction

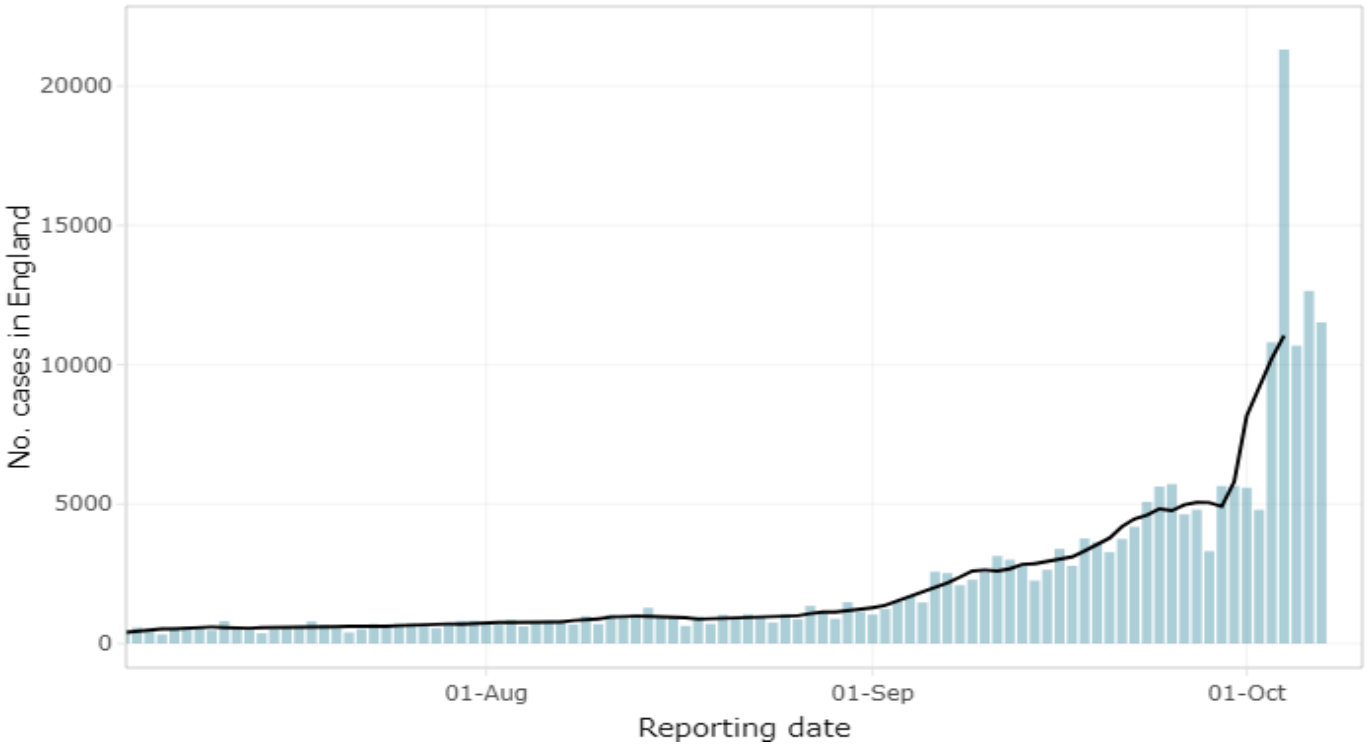
COVID-19 case data in the UK is typically available on two different date bases – Reporting Date and Specimen Date. On any individual day, the number of cases reported will relate to tests carried out on a number of different days (the "Specimen Dates"). The focus of much of the UK media's monitoring of case numbers is on the number reported. However, an analysis by Specimen Date can give a better and potentially more timely insight into the trend in case numbers over time, as it represents the number of persons testing positive on each date. The date a test is reported is influenced by several factors, including the way in which the test sample was collected, and the extent of any delays between the specimen being taken, its analysis and subsequent reporting. The recent data issue has highlighted this point, but the benefit to using Specimen Dates is wider than just that aspect.

Recent Data Comparisons

Below we show the difference in cases by Reported Date and Specimen Date, to demonstrate the differences seen, and some of the problems with the former measure. It is worth emphasizing that whilst the recent data error has magnified some of the problems with using Reported Date, we might expect timing variations to be a regular feature of the data. For the rest of the report we focus on the data for England alone, as data from the devolved administrations may exhibit different behaviours over time as they are relevant to the report. Figure 1 shows the number of cases reported each day in England from July to early October 2020, with cases reaching a peak in this period of over 20,000. Figure 2 shows the same data, but by Specimen Date, where the peak number of cases (so far) is approximately 11,000.

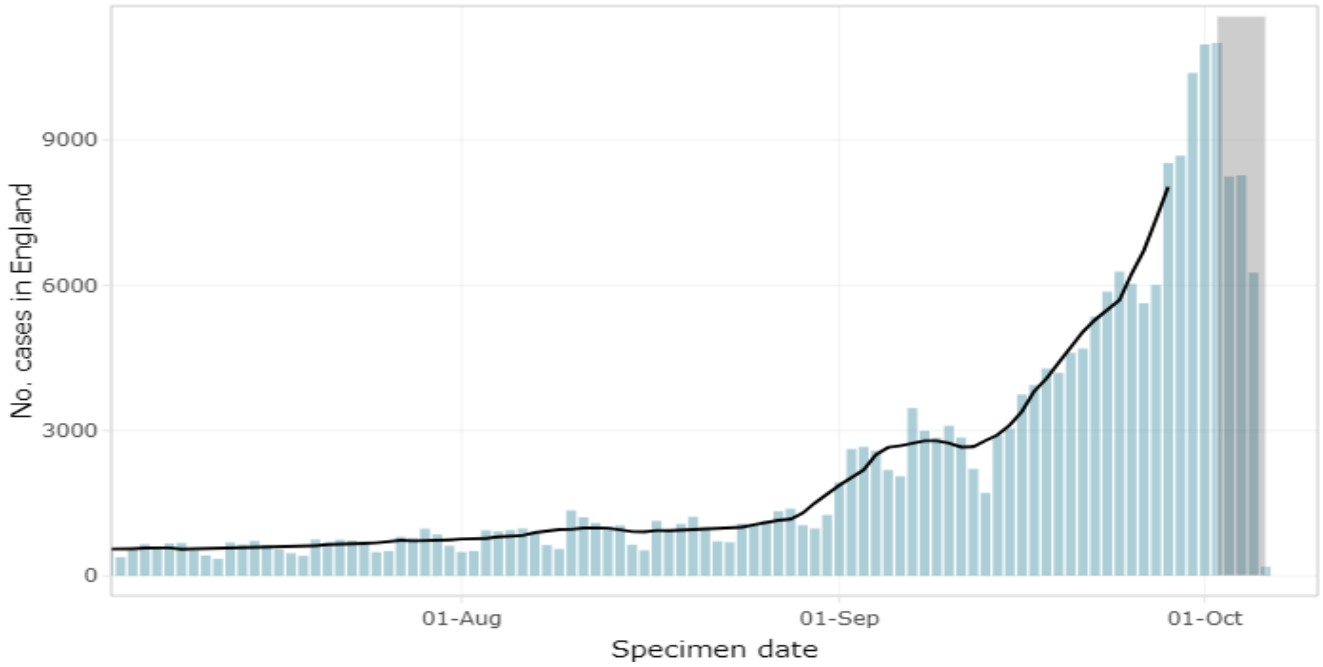
[1] https://coronavirus.data.gov.uk/cases

Figure 1 - England COVID-19 Case numbers by Reporting Date



The 7-day moving average, represented by the black line, is distorted by the "catching-up" seen on the 3rd Oct. Moreover, it is difficult to identify whether the high volumes since are a further catch-up, or representative of the underlying position. The corresponding graph by Specimen Date (up to and including data reported on 7th Oct – so that the most recent Specimen Date is the previous day, 6th Oct), is as follows:

Figure 2 - England COVID-19 Case numbers by Specimen Date



We can immediately see that the latest few days (shaded) dip down, not due to the recent data issue, however, simply to the inherent delay in all cases being reported for any Specimen Date. For this reason, representations of this graph, including that on the government website, usually ignore the last few days in calculating a moving average, to avoid a false impression being given of a declining trend. (The explanation of this on the government site is included in the accompanying notes, although not readily apparent.)

The delays in reporting cases would suggest that the data for approximately the last 5 Specimen Dates needs to be disregarded. This means that potentially important data is ignored for several days, with a 5 day lag before the trend in case numbers by Specimen Date can be properly examined and interpreted. This could lead to a delay in making important policy decisions in response to these trends.

Noting these caveats, the Specimen Date graph shows

- A recent sharp increase in case numbers up to 11,006 on Oct 2nd
- A decline beyond that date, with for example 8,243 on the 3rd, which we might expect to increase beyond due to the natural delays described earlier, but we cannot be sure
- Only 189 on the most recent day (6th), so this is telling us nothing yet about the recent trend in case numbers.

By collating the Specimen Date data into a "triangle" format it is possible to better understand the trend by Specimen Date. In particular, the use of well-established techniques used in insurance can be applied to this form of data so that an estimate can be made of the ultimate number of cases for recent Specimen Dates at a much earlier stage.

The issues that arise with the delay in reporting of COVID-19 case numbers are directly comparable to those that arise in general insurance in relation to reporting of claims. These issues also arise with reporting of COVID-19 deaths, where the techniques described here have already been used by actuaries and others to estimate the number of such deaths by date of death.

Estimating partly developed Specimen Dates

To apply these techniques, data is needed that allows the pattern of emergence of case numbers by Specimen Date to be observed over time. To achieve this, in the absence of historical data, it has been necessary to save a daily download of the relevant data from the UK Government's Application Programming Interface (API).

To illustrate how the data can be organized to properly monitor the trends by Specimen Date, the data by Specimen Date was downloaded from the API on each day from 3 September onwards. This data can then be organized into a two-way table showing Specimen Date case numbers by delay to report (known as a "triangle" format in an insurance context). This is shown in Figure 3. This triangle has data for 3/9 to 17/9 inclusive only, so as to exclude the distorting effect of the recent data issue referred to above. It is also "old" data in the sense that we do of course know how the case numbers for these Specimen Dates has developed since 17/9, but that is intentional so as to enable us to compare the actual outturn with the results from applying the techniques.

Figure 3 - England COVID-19 Case numbers by Specimen Date - Incremental data triangle

	Days Since Specimen Date														
Specimen Date	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Thu 03/09/20	12	686	972	218	261	353	106	35	4	6	4	0	-2	5	-1
Fri 04/09/20	20	1119	776	145	74	226	118	43	20	11	26	4	0	-1	
Sat 05/09/20	22	1127	720	114	34	54	39	18	14	31	9	2	0		
Sun 06/09/20	16	617	877	399	59	56	9	3	5	2	8	0			
Mon 07/09/20	13	716	1208	897	414	126	49	8	4	12	0				
Tue 08/09/20	15	459	1355	533	310	253	40	22	6	2					
Wed 09/09/20	19	559	1186	507	365	150	49	2	4						
Thu 10/09/20	30	687	1319	461	435	123	26	13							
Fri 11/09/20	34	498	882	661	450	226	86								
Sat 12/09/20	24	186	1096	634	165	78									
Sun 13/09/20	28	170	1130	214	107										
Mon 14/09/20	29	945	1070	581											
Tue 15/09/20	28	1015	1167												
Wed 16/09/20	41	1674													
Thu 17/09/20	60														

This shows the incremental number of positive cases for each Specimen Date, as reported on successive days after the Specimen Date. So, for Monday 14 September, there were 29 cases in the data reported on Tuesday 15 September, and the next day 945 more positive cases for this date were reported. On Friday 18 September, some four days after the date the test was done, there were a further 581 reported. Organising the data in this way enables clear identification of the pattern of reporting by Specimen Date.

If we apply standard actuarial techniques to the cumulative version of data (specifically, the "Chain Ladder" method), then we can derive factors that estimate, for any day, the number of cases yet to be reported. We can then simulate what we might have estimated on 17th Sept, and compare it with the actual out-turn, as shown in Figure 4.

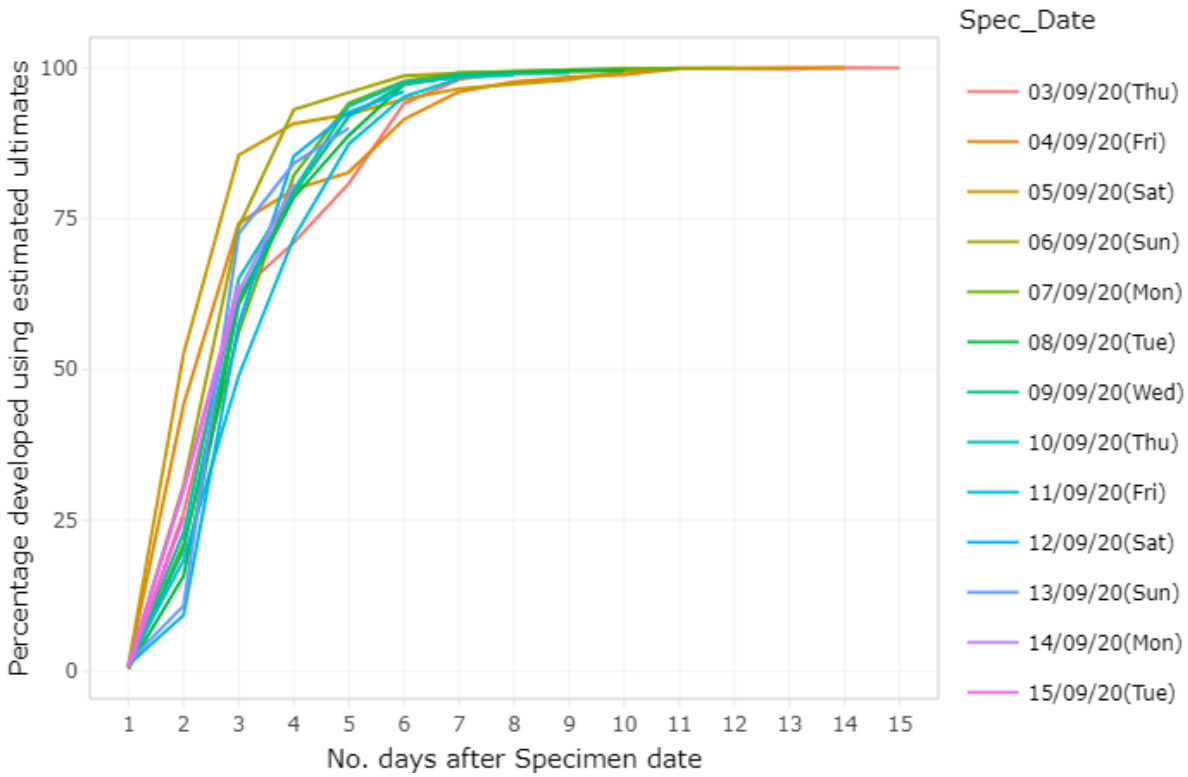
Figure 4 - England COVID-19 Case numbers by Specimen Date – Illustrative results

Specimen Date	**Cases reported**	**Estimated %developed**	**Estimated Future**	**Estimated Final**	**Actual to 7/10**	**Actual / Estimated**
03/09/2020	2659	100.0%	0	2659	2664	100%
04/09/2020	2581	100.0%	0	2581	2588	100%
05/09/2020	2184	100.0%	0	2184	2188	100%
06/09/2020	2051	100.0%	0	2051	2056	100%
07/09/2020	3447	99.9%	3	3450	3465	100%
08/09/2020	2995	99.6%	12	3007	2996	100%
09/09/2020	2841	99.2%	23	2864	2852	100%
10/09/2020	3094	98.9%	34	3128	3101	99%
11/09/2020	2837	98.2%	52	2889	2856	99%
12/09/2020	2183	96.1%	89	2272	2211	97%

13/09/2020	1649	90.0%	183	1832	1711	93%
14/09/2020	2625	80.8%	624	3249	2892	89%
15/09/2020	2210	64.1%	1238	3448	3044	88%
16/09/2020	1715	25.5%	5010	6725	3745	56%
17/09/2020	60	0.8%	7440	7500	3941	53%

The percentage developed column shows the proportion of cases that are estimated to be reported at 1, 2, 3…. etc days after the Specimen Date. The pattern of development can be seen more clearly graphically, as in Figure 5, which shows the cumulative development by Specimen Date, with each day's data scaled to the corresponding Estimated Final values shown above, so that each line progresses towards 100%.

Figure 5 - England COVID-19 Case numbers by Specimen Date – Development data scaled to estimated ultimate



This suggests that the reporting of positive tests is nearly complete by around 5-7 days, but with considerable variability of development pattern between dates, particularly in the early days after the Specimen Date.

We can observe from Figure 4 that the method is reasonably accurate (albeit slightly cautious) for all but the latest two Specimen Dates. This suggests that if the method were being applied to current data, rather than just this illustrative dataset, then unless the pattern of development across Specimen Dates stabilizes, any results for the last two Specimen Dates need to be ignored or treated with caution.

If the speed of test and laboratory processing were to improve, then this should mean that the method could potentially be used for these more recent dates. The triangle format described here

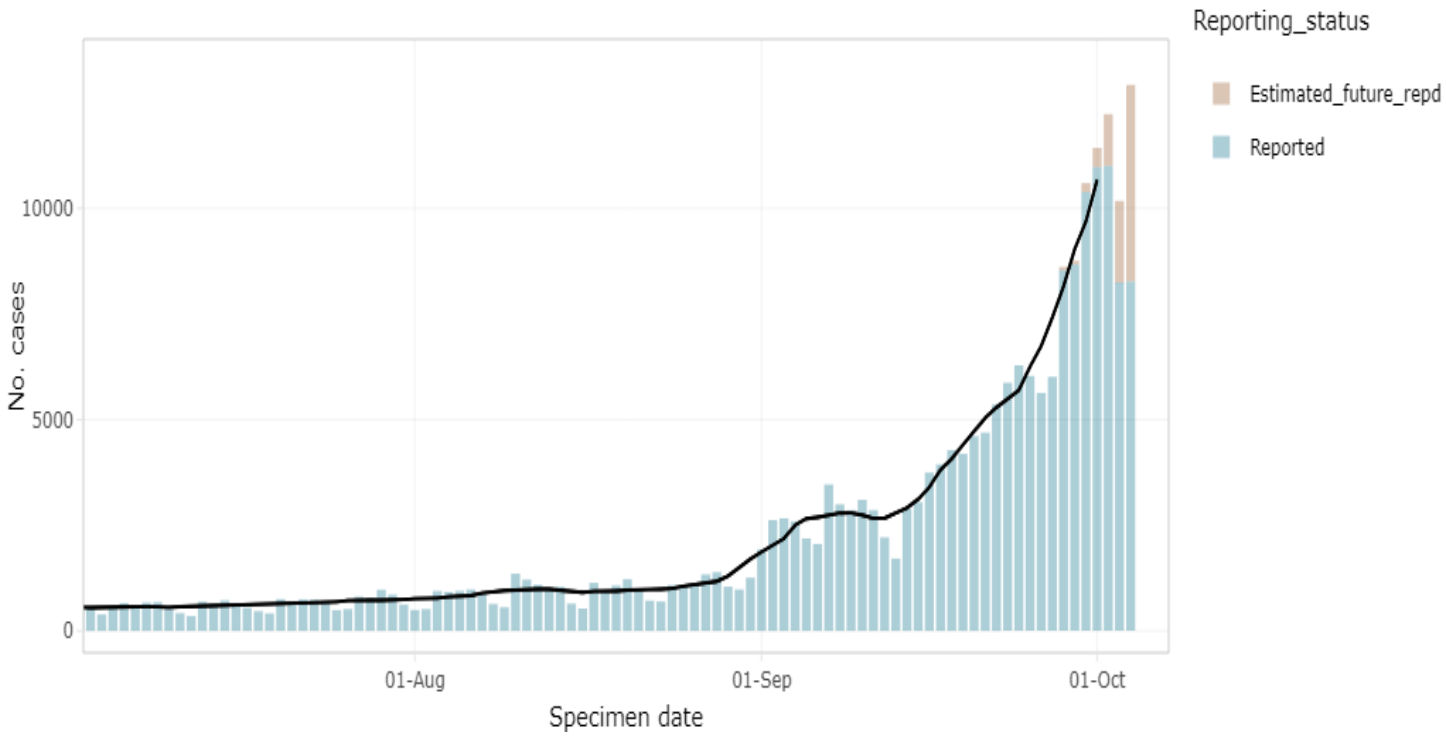
will allow any changes in the speed of processing to be monitored over time, as the factors are reviewed as part of the process.

Using the method for current data

If we want to use the method to estimate case numbers for more recent partly developed Specimen Dates, we will ideally need to wait until the data issue referred to earlier has fully worked its way through the data, as recent days will not show a typical trend in emergence of cases. In the interim, we could use a fixed set of factors from immediately prior to the distortion, until monitoring of the data in the triangle format shows that it is appropriate to revert to the standard methodology.

We can therefore apply the illustrative development pattern derived in the previous section to a data triangle using more recent data. This is merely to show an example of how the graph by Specimen Date presented at the start of this report can be enhanced to show the more recent trend. If we do this for data reported to 7/10 inclusive (but excluding the latest two Specimen Dates for the reasons outlined above), the results are given in Figure 6, which is in the same form as Figure 2, except that the estimated future reported has now been added to the recent Specimen Dates, and so the unrepresentative "dip down" is no longer evident for these dates This graph does not represent our estimate of case numbers for the recent Specimen Dates, not least because it is based on the method applied to "old" data, and because there is likely to still be some distorting effects for the days where the estimation is most material.

Figure 6 - England COVID-19 Case numbers by Specimen Date – Example graph with estimates added



Conclusion

We have demonstrated above that a standard actuarial technique can be used to provide faster insight into the trend of new cases being reported. Although in describing the method the paper has, of necessity, mentioned the data error issue which was discovered recently, the technique is not designed to compensate for such errors. However, it may have led to an earlier indication that something was awry, through unusual results emerging inconsistent with expectations.

Over the coming days we will continue to monitor the emergence of test results, with a view to fine tuning the analysis, and making the estimates publicly available, once we are satisfied as to their robustness. We will also investigate other underlying features of the data, such as any changes in the overall pattern of reporting by Specimen Date and whether there is a possible "weekend" effect.

Accessibility to pandemic related data through API is very welcome, as it enables many interested parties (journalists, scientists, etc) to analyse trends, adding to the wealth of knowledge and encouraging wider debate as to effective ways to manage the current situation. The lack of historical data to enable this analysis to be performed is an omission that should be easy to address, and result in additional insight for policy-makers and others. We therefore encourage government to facilitate its availability at the earliest opportunity.

When the techniques are applied in an insurance context, the accuracy of the results is usually improved if they are used by practitioners who understand the data and the underlying business issues that impact on the results. The same is likely to apply if these techniques are used in a COVID-19 context, so that relevant health and COVID testing specialists should preferably be involved in the use of the techniques, where possible.