Mortality attributed to Covid19 and excess deaths – a statistical analysis

Tomas Hrycej
University of St. Gallen
Institute of Computer Science I Data Science

1 Goal of investigation and available data

A systematic monitoring of data concerning the extent of Covid19 infection in the population is a crucial part of the strategic treatment of this pandemic. A widespread source are the data collected by the Johns Hopkins University (Johns Hopkins University, Coronavirus Resource Center, 2020), with a download possibility at (The Humanitarian Data Exchange, 2020). The largest attention is paid to the counts of daily confirmed or registered infections. This is justified by the fact that it is the earliest indicator. However, these counts are massively influenced by testing volume and strategy as well as the type of the test used. All these factors are varying in time and space. This is why these data cannot be compared between different countries and time periods and can thus be misleading if used to justify countermeasures.

Another important data type are the deaths attributed to Covid19. Their definitions are also varying in the way the death cause is assigned to Covid19 or not. But they are generally expected to provide a more objective picture of the instantaneous seriousness of the pandemic. Their drawback is a delay of approximately two weeks against the confirmed cases.

One of the most objective criterions to assess the instantaneous impact of Covid19, unbiased by varying definitions, is a comparison of two sources: the reported Covid19 mortality on one hand and overall mortality, in particular with the excess deaths on the other hand. Excess deaths are defined as instantaneous death surplus over the long term average of the given time period (typically a calendar week). Weekly morality data are collected and provided for download by (Eurostat Database, 2020). Regular summaries are presented by (EuroMOMO, European mortality monitoring activity, 2020).

The comparison can be done for European countries included in the Eurostat database. Unfortunately, only a part of these countries provides a data extent sufficient for this goal. Some countries such as Germany, France, Italy and United Kingdom did not supply a sufficiently long history (at least since 2011) to assess the average weekly mortality. Others such as Czech Republic and Slovakia have a large delay in supplying new data so that at this moment, they do not cover the second Covid19 wave beginning in September 2020 sufficiently. Czech Republic has been included in this investigation because of the large extent of its second wave, in spite of this delay. So the countries subject to the following investigations are Czech Republic, Poland, Hungary, Austria, Switzerland, Spain, Sweden, the Netherlands and Belgium, with week scopes as listed in Table 1.1.

```
Last week of mortality data:
     Czechia || 39
      Poland || 40
     Hungary || 41
     Austria || 41
 Switzerland || 43
       Spain || 43
      Sweden || 44
 Netherlands || 42
     Belgium || 43
```

Table 1.1. Weeks until which mortality data are available.

The initial point of this investigation and the diversity encountered can be illustrated with help of three European countries: Hungary, Switzerland and the Netherlands (Figure 1.1). The numbers are normed to a million of country population.

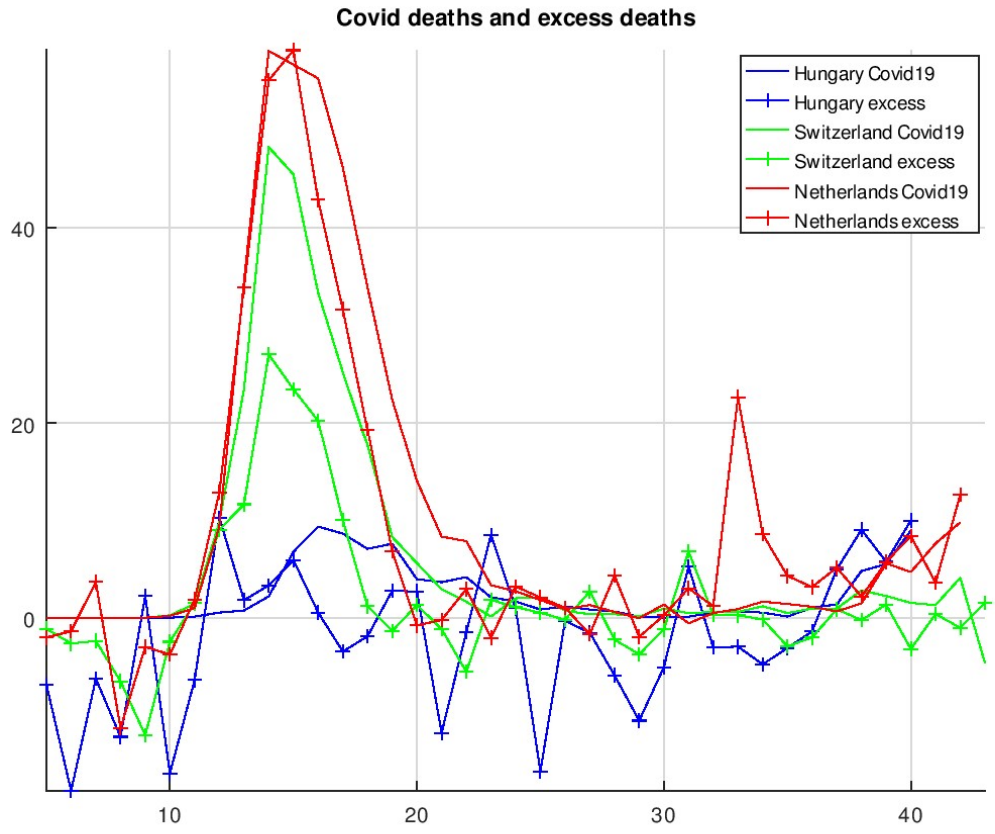

Figure 1.1. Covid19 deaths and excess deaths per million inhabitants of some European countries.

Some differences are obvious. The Netherlands and Switzerland show similarly large peaks of Covid19 deaths in the first wave in the late spring 2020. However, the Swiss excess mortality is substantially lower. There is no significant mortality in Switzerland at the considered part of the second wave in the early autumn 2020, neither in Covid19 deaths nor in excess mortality. There is a slight increase in the Netherlands which is preceded by a larger excess death peak without Covid19 causality. A small increase of both statistics can be observed in Hungary, while the first wave is not reflected in excess deaths of this country.

Viewing the plots like this may suggest some correct (and some wrong) conclusions concerning the relationship of Covid19 deaths to overall excess mortality. To put them on a firm foundation, they will be analyzed with the means of standard regression analysis.

2 Remarks to statistical methods

Widespread standard statistical methods to assess the relationship between observed variables are correlation and regression analysis. The latter group is appropriate if one variable is assumed to (at least partially) explain the movement of the other variable. In our case, the two variables are

- Deaths attributed to Covid19 and
- Total excess deaths, independent of the cause

The mortality development throughout a year in the years before 2020 without Covid19 can be viewed, for the goals of this investigation, as "normal". In every calendar week, they exhibit a mean level as well as fluctuations around this level. The size of these fluctuations is described by standard deviation observed in this week. In 2020, Covid19 deaths occur additionally to the normal level.

To investigate the influence of Covid19 deaths to the total excess deaths, a simple regression model can be formulated. For the i-th week of 2020, a functional equation of the following form is assumed:

$$t_i = a_i + b_i c_i + e_i \qquad (2\text{-}1)$$

With

- t_i being the total excess deaths in the *i*-th week of 2020,
- c_i the reported Covid19 deaths and
- e_i the excess deaths with no relationship to the Covid19 deaths.

The constant b_i is the regression coefficient expressing how many of total excess deaths can be explained by the number of deaths attributed to Covid19 in this week. The constant a_i corresponds to the basic level of excess deaths in a given period.

Excess deaths e_i not explainable by Covid19 deaths correspond to a normal excess death fluctuation around the mean mortality level of the week. So they are, in fact, a random variable not explained by the model. In years before 2020, this random variable are the "normal" excess deaths.

To summarize, total excess deaths t_i of the *i*-th week are decomposed to the part $b_i c_i$ that can be explained by some proportional effect of reported Covid19 deaths and to the part a_i+e_i that corresponds to the normal random fluctuations of mortality.

The regression model (2-1) is frequently formulated under the assumptions of *homoscedasticity* that the random components e_i

1. have the same variance (i.e., "size of fluctuation") and
2. are independent of each other (i.e., do not correlate with each other)

These assumptions are clearly not satisfied in our case. The variability of excess deaths is particularly large in winter season where infection waves typically occur. An example is given in Figure 2.1., depicting weekly standard deviations as a measure of variability, over the week range covered by 2020 data. Raw figures (blue curve) are obviously too variable themselves, due to their computation from only nine years (2011-2019) provided in (Eurostat Database, 2020). This is why they have been smoothed by a five-week gliding averaging.

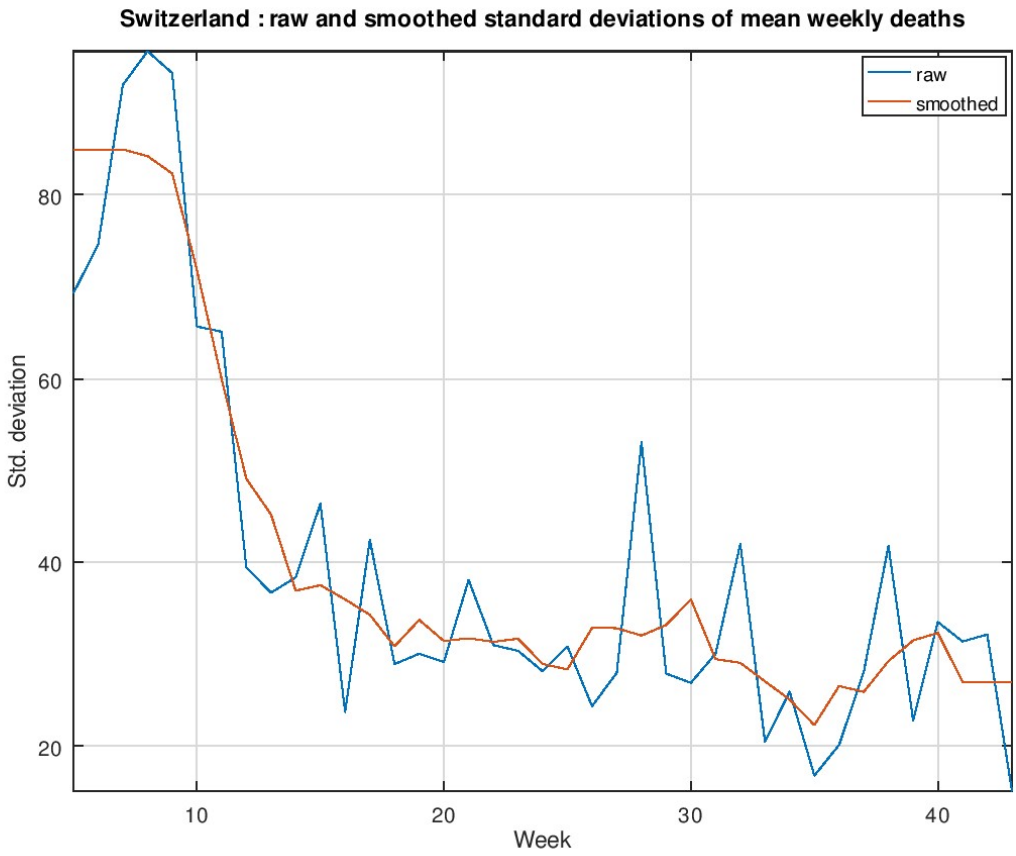

Figure 2.1. Variability of weekly excess deaths in Switzerland, raw and smoothed weekly standard deviations.

Also the second homoscedasticity assumption of independence between individual weekly figures does not apply. There is a substantial *autocorrelation* between neighbor week figures e_i and e_{i+1}. This reflects the fact that infections occur in waves of several weeks extension. Once such wave has started and leads to excess deaths e_i in the *i*-th week, a similar count of excess deaths can be expected in the *(i+1)*-st week. Autocorrelation coefficients (theoretically ranging from -1 to 1) for countries considered are given in Table 2.1. Autocorrelations between excess deaths of neigboring weeks in considered countries.

```
Autocorrelation:
     Czechia ||  0.65
      Poland ||  0.80
     Hungary ||  0.61
     Austria ||  0.37
 Switzerland ||  0.62
       Spain ||  0.86
      Sweden ||  0.52
 Netherlands ||  0.80
     Belgium ||  0.70
```

Table 2.1. Autocorrelations between excess deaths of neigboring weeks in considered countries.

Neglecting autocorrelations may lead to wrongly attributing some effect to the explaining variable (here: Covid19 deaths) while they are influenced by neighbor values of the explained, dependent variable.

To account for both unequal variability and autocorrelation, *generalized regression*, also called *heteroscedastic* model, has to be used. Generalized regression assumes the variability of e_i being described by a covariance matrix *W*. In our case, the matrix elements are

$$w_{ij} = s_i s_j d^{|i-j|}$$

(2-2)

with s_i being the standard deviation of the *i*-th week and autocorrelation coefficient *d*. The $|i\text{-}j|$-th power of the autocorrelation coefficient expresses the fact that weekly figures being |i-j| apart in time correlate with a coefficient of $d^{|i\text{-}1|}$.

3 Computing results

With help of the regression model of Section 1, the data of countries considered have been analyzed. Because of obvious differences between both Covid19 waves, the waves have been analyzed separately. The first wave data are those until week 26, that is, until end of June 2020. The second wave starts in week 31 (August 2020). The so defined time scopes include some period before and after the increased infection to contain also some Covid19 free data.

The results are summarized in Table 3.1.

```
WAVE 1
     Country     regr. heterosced.    R2 heterosced.   regr. homosced.      R2 homosced.
     Czechia ||  0.47  <-1.36, 2.30>  0.01 (0.40)  ||  0.55  <-0.26, 1.35>  0.06 (0.74)
      Poland ||  0.06  <-1.19, 1.31>  0.33 (1.00)  ||  1.33  < 0.63, 2.04>  0.71 (1.00)
     Hungary ||  0.26  <-1.19, 1.71>  0.01 (0.41)  ||  0.82  <-0.08, 1.71>  0.21 (0.97)
     Austria ||  0.64  < 0.16, 1.13>  0.33 (1.00)  ||  0.84  < 0.42, 1.25>  0.43 (1.00)
 Switzerland ||  0.56  < 0.42, 0.70>  0.71 (1.00)  ||  0.60  < 0.51, 0.69>  0.88 (1.00)
       Spain ||  0.62  < 0.41, 0.82>  0.59 (1.00)  ||  0.97  < 0.84, 1.10>  0.92 (1.00)
      Sweden ||  0.62  < 0.53, 0.70>  0.92 (1.00)  ||  0.64  < 0.60, 0.68>  0.98 (1.00)
 Netherlands ||  0.86  < 0.68, 1.05>  0.77 (1.00)  ||  0.90  < 0.79, 1.01>  0.93 (1.00)
     Belgium ||  0.42  < 0.33, 0.51>  0.78 (1.00)  ||  0.55  < 0.47, 0.63>  0.90 (1.00)
WAVE 2
     Country     regr. heterosced.    R2 heterosced.   regr. homosced.      R2 homosced.
     Czechia ||  0.56  <-0.25, 1.36>  0.57 (0.98)  ||  0.48  <-0.29, 1.24>  0.83 (1.00)
      Poland || -1.77  <-4.36, 0.83>  0.49 (0.98)  || -1.19  <-3.05, 0.68>  0.93 (1.00)
     Hungary ||  0.08  <-0.62, 0.77>  0.21 (0.85)  ||  0.66  < 0.10, 1.22>  0.43 (0.97)
     Austria || -0.60  <-1.98, 0.78>  0.53 (0.99)  || -0.19  <-1.60, 1.22>  0.61 (1.00)
 Switzerland || -0.27  <-0.84, 0.30>  0.11 (0.73)  || -0.30  <-0.94, 0.33>  0.06 (0.59)
       Spain ||  0.03  <-0.53, 0.58>  0.56 (1.00)  || -0.01  <-0.25, 0.23>  0.95 (1.00)
      Sweden || -2.02  <-5.36, 1.31>  0.30 (0.96)  || -4.05  <-7.07,-1.02>  0.60 (1.00)
 Netherlands ||  0.58  <-1.29, 2.46>  0.08 (0.62)  ||  0.33  <-0.67, 1.32>  0.59 (1.00)
     Belgium ||  0.78  <-0.04, 1.61>  0.24 (0.91)  ||  0.37  <-0.38, 1.12>  0.31 (0.95)
```

Table 3.1. Regression coefficient and coefficients of determination for heteroscedastic and homoscedastic model.

The most important figure for a country and a Covid19 wave is the regression coefficient. It expresses the fraction of Covid19 deaths that is reflected in excess deaths. For example, for a regression coefficient of 0.80, 100 deaths attributed to Covid19 correspond to 80 excess deaths of the given week. Values of regression coefficient exceeding unity can be interpreted so that there are some more deaths with Covid19 as a cause but with this cause unidentified. Values below unity may imply that a part of deaths attributed to Covid19 is due to other causes.

The regression coefficient as estimated from some limited set of data is subject to uncertainty. This uncertainty can be quantified. The estimate of the regression coefficient given in the first column can be made more precise by its *confidence interval* (the next two columns in brackets). This interval gives the margins within which the regression coefficient is expected to be in 95 per cent of cases.

A complementary measure is the *coefficient of determination* R2 (the fourth column). It states what fraction of the variability of the dependent variable (here: total excess deaths) can be explained by the explaining variable (here: Covid19 deaths). For example, a value of 0.71 states that 71 per cent of the variability of total excess deaths can be explained by the influence of Covid19 deaths. The figure in

parentheses in the fifth column is the *significance level* of R2. Its values substantially below unity indicate that the influence of the explaining variable can be pure chance, without a deterministic substance. A usual significance threshold for such aims is 0.95.

In addition to the results of heteroscedastic model in the left five columns, analogical figures for homoscedastic model neglecting the different variabilities of individual calendar weeks and the autocorrelation between neighboring weeks. In significant cases, homoscedastic figures are similar to heteroscedastic ones, the difference being arbitrarily large in cases of low significance.

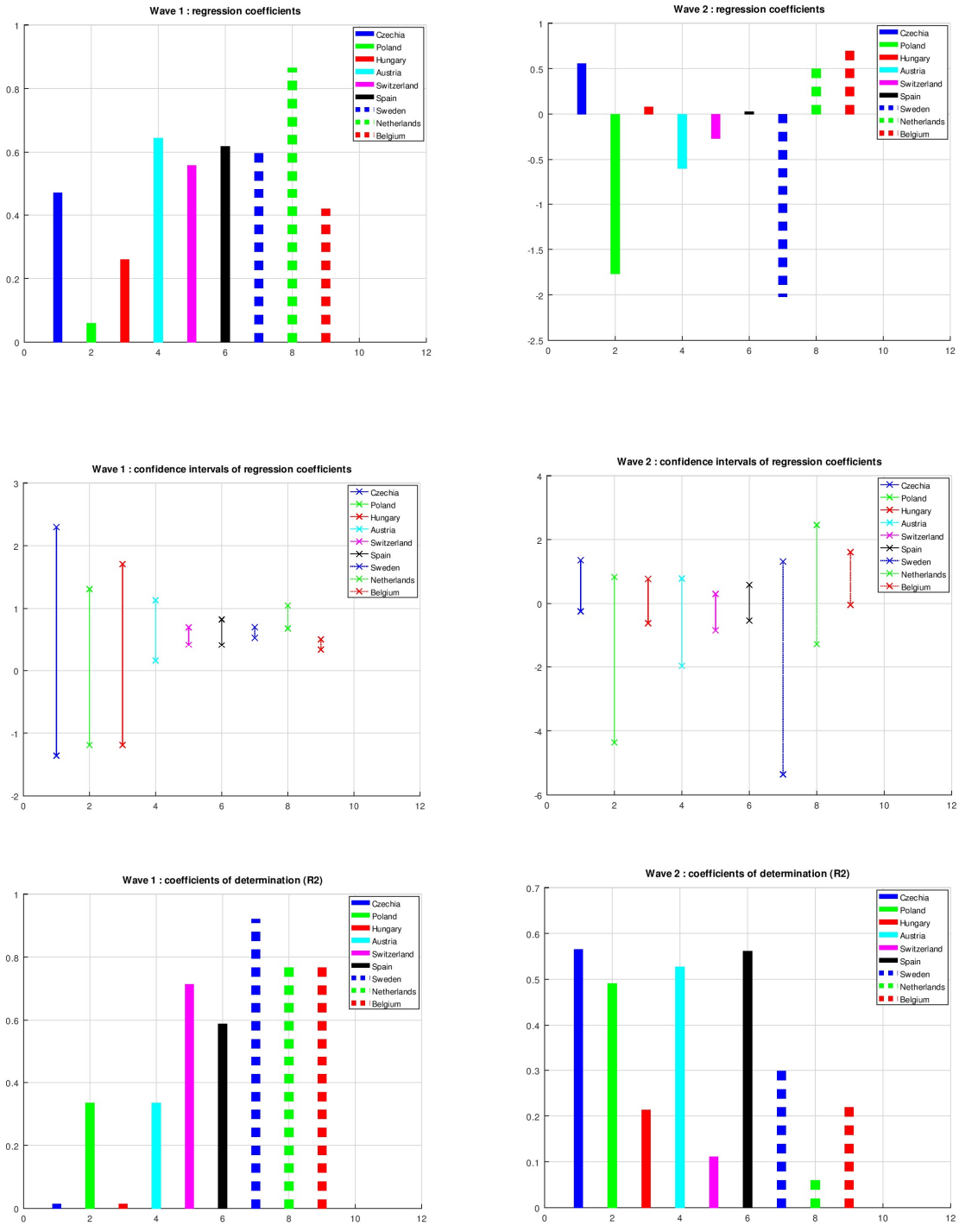

Figure 3.1. Results of regression analysis.

The results are also presented in a graphical form in Figure 3.1. The left column of panes presents wave 1, the right one wave 2. Top panes depict the regression coefficients as colored columns. In the middle row, their confidence intervals are shown by line segments. Bottom panes show the corresponding coefficients of determination as colored columns.

4 Interpretation and conclusions

The following conclusions are of statistical nature. Proposed explanations are only ideas that should be subject to epidemiological scrutiny.

Individual countries show a broad variety of characteristic values. In the first wave, Switzerland, Spain, Sweden, the Netherlands and Belgium exhibit substantial regression coefficients with narrow confidence intervals. They explain between 59 (Spain) and 92 (Sweden) per cent of the excess death variance. The regression coefficients of Switzerland, Spain and Sweden are around 0.6. It can be interpreted as 60 percent of declared Covid19 deaths being really caused by Covid19, the remaining 40 per cent being due to other causes with accidental occurrence of positive Covid19 testing. The Netherlands exhibit a high value of 0.86, testifying a credible assignment of Covid19 deaths. By contrast, the low value of 0.42 for Belgium accounts for the practice of generous assignment of deaths to the Covid19 cause.

Czech Republic, Poland, Hungary have values of low significance. Regression coefficients are not only small but have also very broad confidence intervals, including a sign change (negative low margins). This is equivalent to a very low significance which is also reflected in small values of R2. This is caused by a low intensity of the first Covid19 wave.

Austria is placed between these two groups.

The results of the second wave are much less significant. Czech Republic has one of the narrower (although still broad if compared with the first wave in Western Europe) confidence intervals of regression coefficients. Its mean of 0.56 is close to typical values of the first wave in Western Europe. Also the coefficient of determination of 57 per cent variability share suggest a serious extent of Covid19 relatively to the excess death variance. However, this variance is lower in early autumn than in the spring so that it is a large share of a small number.

West European countries show values of low significance. For Switzerland, Spain, Sweden and the Netherlands, the values of regression coefficients are either small or have a very broad confidence interval. The regression coefficient confidence interval of Belgium is somewhat narrower and its mean is 0.78 (larger than in wave 1) but Covid19 explains only 24 per cent of excess death variance – there other, more important seasonal mortality causes.

To summarize, there is little statistical evidence for the hypothesis that the second wave is comparably serious as the first one. In this relation, Czech Republic is the most pronounced from the countries considered. A possible explanation for this is the relatively negligible first wave in Central Eastern Europe, including Czech Republic. Consequences of this may be a lower immunization or absence of preceding deaths of risk groups. However, the results for the second wave are only preliminary since the wave is still in progress. With incoming data, the results may be recomputed. Expectations for significant results would then be enhanced.

These conclusions is what the data show. Definite causality is nothing that can be proven by statistical analyses. Their ultimate interpretation is to be delivered by epidemiologists. For building causal hypotheses, statistical support and validity check may be a fruitful source.

References

EuroMOMO, European mortality monitoring activity. (2020). *https://www.euromomo.eu/.*

Eurostat Database. (2020). *https://ec.europa.eu/eurostat/en/web/population-demography-migration-projections/data/database.*

Johns Hopkins University, Coronavirus Resource Center. (2020). *https://coronavirus.jhu.edu/.*

The Humanitarian Data Exchange. (2020). *https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases.*