

---

# The Severity of the COVID-19 Pandemic in Terms of Excess Mortality and the Influence of Vaccinations

---

Joschka Strüber

Matrikelnummer 5980381

[joschka.strueber@student.uni-tuebingen.de](mailto:joschka.strueber@student.uni-tuebingen.de)

[https://github.com/the-klingspor/data\\_literacy\\_project](https://github.com/the-klingspor/data_literacy_project)

## Abstract

The COVID-19 pandemic is the defining event of our decade. To estimate its impact, we analyze global excess mortality data. Linear Regression is used to draw a connection between demographic features and excess mortality. The alleviating effect of the vaccination campaign is shown using Kernel Density Estimation.

## 1 Introduction

In December 2019 a novel coronavirus emerged in China (Zhou et al., 2020). Due to its genetic relationship to the Severe Acute Respiratory Syndrome (SARS), which caused a smaller pandemic at the beginning of the millennium, it became known as SARS-CoV-2. Without natural immunity in the population, this virus outbreak quickly emerged into the global COVID-19 pandemic and destroyed the lives and livelihoods of millions of people worldwide.

The goal of this paper is to evaluate the severity of the pandemic based on excess mortality and analyze the influence of countries' demographics and the vaccination campaign. Excess mortality is defined as the occurrence of more fatalities during a period of time than the previously empirically predicted amount. All data analyzes in this report are based on the COVID-19 data sets from Our World in Data (Appel et al., 2020). The death statistics mostly come from the World Mortality Dataset, while the vaccination rates and demographics were collected from various official sources.

Using excess mortality to measure the pandemic severity has several advantages. First, it does not depend on a country's test strategy. This increases comparability and is especially important for countries where tests are rare. Further, excess mortality also really measures severity. While large amounts of mild cases can be a burden on general practitioners, they do not have an effect on society as lasting and severe as casualties. However, this statistic also has a few downsides: Excess mortality is time-delayed or still often undercounted. The used projected deaths are estimates based on statistics from previous years and trends. It also does not consider other influences on mortality, such as non-pharmaceutical interventions reducing the spread of other communicable diseases, famine, or war. All of these factors have to be kept in mind when interpreting excess mortality data.

## 2 Using P-scores to measure excess mortality

Excess mortality is defined as the difference between reported and expected deaths. While it gives an intuition for the situation, it does not take the country's population size into account. This makes comparisons between different countries difficult. For this reason, Appel et al. (2020) introduced the P-score, which normalizes the excess mortality with the amount of projected deaths:

$$\text{P-score} = \frac{\text{ReportedDeaths} - \text{ProjectedDeaths}}{\text{ProjectedDeaths}} \cdot 100 \quad (1)$$

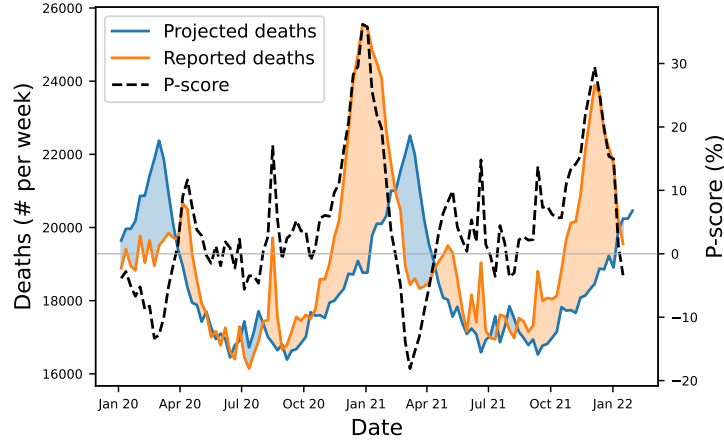


Figure 1: Visualization of the weekly reported and projected deaths in Germany as well as the resulting P-scores over time. The color of the shaded region indicates the P-score’s sign.

As an example, we consider the P-scores of Germany during the course of the pandemic as shown in Fig. 1. The figure depicts the weekly projected and reported deaths. The difference is shaded in blue, if there were more projected than reported deaths, and orange, if there was excess mortality. Normalizing and rescaling it yields the P-score. The large excess mortality during the two winter waves is clearly visible. One can also see how the reported deaths were below projections during both springs. The main reason was likely the effect of non-pharmaceutical interventions on the flu waves (Olsen et al., 2020). In total, there is a moderate excess mortality in Germany of 4.4% since the start of the pandemic. This does not compare to many other countries, e.g. Peru with up to 258% per week and twice as many casualties as expected since the start of the pandemic (Appel et al., 2020).

### 3 Demographic data as predictor of the pandemic’s severity

Linear Regression is one of the most established models for regression (Galton, 1886). While it often lacks in precision, it has the advantage of being unusually easy to interpret, provided the features are not too highly correlated and their scale is similar. We chose the amount of people aged more than 65 years old, population density, amount of people living in extreme poverty, amount of deaths related to cardiovascular diseases, diabetes prevalence and the amount of hospital beds per 1,000 inhabitants as features. Those were selected, because they are available for fairly many countries and are not highly correlated. For comparability they are normalized to zero mean and variance one.

As target the cumulative P-scores from March 1st, 2020 to December 31st, 2020 are chosen. We chose a large cumulative time frame to reduce variations. One problem of this period is that it is not a full year long. This can be problematic, because SARS-CoV-2 has a high seasonality and we consider the full winter on the southern hemisphere, but only half of it in the north (Gavenčiak et al., 2021). This likely leads towards a bias underestimating the severity in the often highly developed countries of the northern temperate region. Nevertheless, it appears to be the best option, because many countries have not reported mortality data for 2021 yet, later data is influenced by vaccines and prior to March 2020 COVID-19 was not globally relevant. In total, the data set contains 62 countries.

The resulting coefficients as well as plots showing the relationship between each feature and the target are presented in Fig. 2. To rule out overfitting, we compared the results to Lasso and Ridge Regression using cross validation. Both models behaved similar and the coefficients did not explode. We observed a positive relationship of extreme poverty and the amount of cardiovascular deaths as well as a negative of diabetes prevalence and the amount of hospital beds. The size of the elder population and population density were no factor.

In general, the predictive quality of the model is non-optimal with an average RMSE of 16.5, which was to be expected. Taking a look at the correlation plots on the right, there seems to be no clear linear relationship between any feature and the P-Scores. However, features that correlate with high

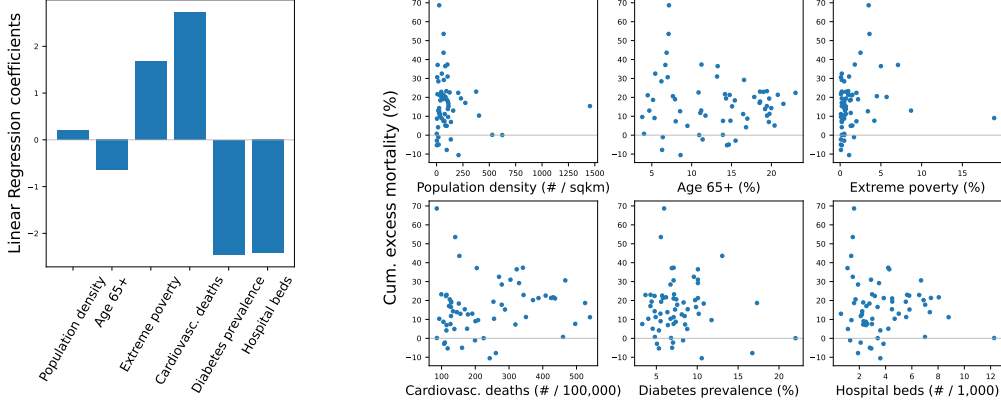


Figure 2: The left plot shows the coefficients of a Linear Regression model for each feature. On the right, the correlations between features and the cumulative excess mortality as targets are visualized.

development of a country, such as age, quality of the healthcare system and prevalence of diseases of civilization, appear to imply a much smaller variance in excess mortality.

#### 4 The effect of the global vaccination campaign

The COVID-19 vaccines introduced one year ago have proven to be the most powerful tool against the pandemic. They are estimated to have saved the lives of almost half a million people in Europe alone and probably millions worldwide (Meslé et al., 2021). To model the relationship between the rate of fully vaccinated people and excess mortality, we approximate their joint distribution for all data points since the start of the global vaccination campaign on December 15th, 2020.

We decided to approximate the distribution using Kernel Density Estimation (Parzen, 1962), because previous attempts with Bayesian models were not successful. The model assumes that the pdf of the true distribution is higher in close proximity to our observations  $\{\mathbf{x}_i\}_{1 \leq i \leq n}$ . This is realized by centering a stochastic kernel  $K$  on each of the data points. A bandwidth  $h$  determines the variance of the kernel distribution. To predict the density of a new point  $\mathbf{x}$ , the results of all kernels are averaged:

$$p_K(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K(\mathbf{x} - \mathbf{x}_i; h) \quad (2)$$

The results of Kernel Density Estimation are shown in Fig. 3 with lighter regions indicating a higher pdf and observations in white. We used a Gaussian Kernel and cross validation to optimize the bandwidth. To get the conditional distribution for a given vaccination rate we marginalized the joint distribution. This allowed us to estimate the mean and 95% confidence intervals, which are depicted in orange. The mean is very stable between five and eight percent and slightly increases with larger vaccination rates. The distribution is very asymmetrical with a much larger variance in the direction of high excess mortality. The variance clearly behaves inversely proportional to the vaccine rate.

In populations with no prior immunization, the pandemic usually proceeds in waves: very steep peaks with high case and death counts are followed by longer periods of low numbers. The results are few very high P-scores and many low ones, often negative because of mortality displacement. Since the vaccine’s protection against death is outstanding, countries with high vaccination rates, as seen on the right of the figure, register both less excess mortality and mortality displacement.

However, the data has caveats that one has to consider and our simplistic one-to-one analysis leaves out many other contributing factors. A possible explanation of the upward trend of the mean is the general course of the vaccination campaign. Mortality displacement caused by the previous winter wave and the absence of the flu wave resulted in many entries of negative P-scores during March 2021 (Olsen et al., 2020). This was right before many European countries reached 5% of full vaccinations and likely biases those P-scores down. A similar effect might cause the slightly higher average excess mortality for very high vaccinations rates above 80%. These rates were only reached once infants

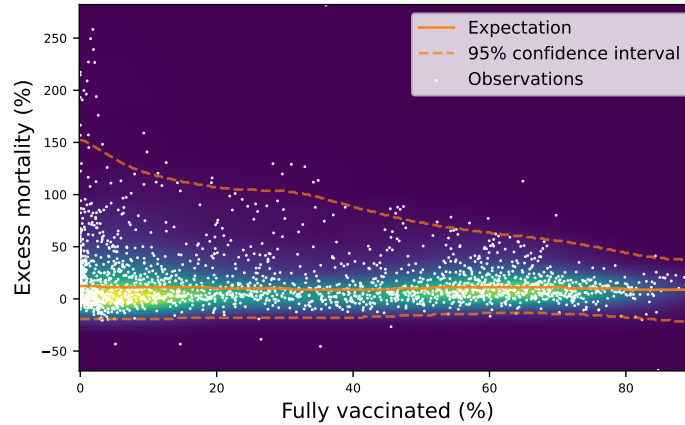


Figure 3: The figure shows the joint distribution of the rate of fully vaccinated people and P-scores. It was estimated with Kernel Density Estimation using a Gaussian Kernel. The orange lines show the expectation and 95% confidence interval of the conditional distributions given a vaccination rate.

were able to be vaccinated. This means that all entries on the very right coincide with the large Delta and Omicron waves in winter 2021/22, which caused many casualties.

Previous infections can also offer protection against severe disease processes via T cell immunity. Therefore it would make sense to consider rates of immunization instead of vaccination rates alone. Unfortunately, this data is very difficult to collect, because of reinfections, breakthrough infections and undercounted infections. Some countries have tried to estimate the immunization rate using serological studies, e.g. SIREN, but these often suffer from selection bias (Hall et al., 2021).

## 5 Conclusion

We have introduced the P-score as measure of excess mortality and analyzed the influence of various demographic features as well as vaccination rates. Predicting the severity of the pandemic has proven to be very difficult, because the situation is highly dynamic and it is impossible to take all factors into consideration. However, we have been able to show that while statistics correlating with prosperity and vaccination rates do not imply lower P-scores on average, they are very likely to be a factor in decreasing variance and acting as an upper bound on excess mortality.

## References

- Appel, C. et al. Data on COVID-19 (coronavirus) by Our World in Data. <https://github.com/owid/covid-19-data/tree/master/public/data>, 2020.
- Galton, F. Regression towards mediocrity in hereditary stature. *Anthropological Miscellanea*, 1886.
- Gavenčiak, T. et al. Seasonal variation in sars-cov-2 transmission in temperate climates. *medRxiv*, 2021.
- Hall, V.J. et al. SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (SIREN). *The Lancet*, 397(10283): 1459–1469, 2021.
- Meslé, M.M. et al. Estimated number of deaths directly averted in people 60 years and older as a result of COVID-19 vaccination in the WHO European Region, December 2020 to November 2021. *Eurosurveillance*, 26(47), 2021.
- Olsen, S.J. et al. Decreased influenza activity during the COVID-19 pandemic—United States, Australia, Chile, and South Africa, 2020. *Am. J. Transplant*, 20(12):3681–3685, 2020.
- Parzen, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33 (3):1065 – 1076, 1962.
- Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579 (7798):270–273, 2020.