Hunting the tiger - what can be learned from public surveillance data on the population risk caused by SARS-COV-2?

Claus Heinrich[1]

[1]Affiliation not available

September 09, 2020

Abstract

There is a large reservoir of publicly available data that could be used to better understand the population risk caused by the novel corona virus SARS-COV 2. However, important questions subject to public debate have not been sufficiently addressed with empirical data so far.

Based on data published by official sources covering the period from March 02 to May 31, the impact of general testing activity in Germany on the number of confirmed cases was investigated with a linear regression model. The model yielded an adjusted R square of .07, which was statistically significant but numerically too small to explain a substantial part of the variation observed.

For the same period, the relationship between changes in public mobility and the number of confirmed cases was analyzed. A strong correlation (-.51) was found for mobility and confirmed cases on the same day, which decreased with an increasing time lag. The correlation was stronger (-.68) when the date of reporting was used as a basis for confirmed cases rather than the date of first symptoms. These findings suggest that public mobility decreased in response to infection numbers reported rather than mobility restrictions having an impact on case numbers.

Two important sources of bias are discussed that should be considered for disease modelling based on public surveillance data. The strong initial increase of case numbers observed in some countries might be an artifact of the national testing policy. Furthermore, the numbers are subject to a strong negative selection bias which does not allow for valid conclusions on the population.

There is a continued and growing demand for representative data to arrive at a more realistic picture of the true population-based hazard potential of this novel virus

Introduction

"Dance with the tiger" - that's how leading health politicians, epidemiologists and virologists described the situation when the first, cautious steps were taken on April 20, 2020 to revive social and commercial life after the lockdown introduced in Germany one month before. Meanwhile more than nine months have passed after the first COVID 19 cases were detected in China and many countries have at least started to alleviate their restrictions. There can be no reasonable doubt about the novelty of SARS-COV-2 and its' potential to cause severe viral pneumonia with unusual features, a high fatality rate and a substantial risk for severe residues. However, this is not so much different from other respiratory viruses and classical risk analysis is based on severity AND likelihood of a potential hazard. For the latter, several simulation studies have been published , trying to justify the measures taken by comparing actual numbers against hypothetical scenarios and coming up with impressive effect estimates. These models share two important disadvantages: they are no valid replacement for empirical data, and they depend on presumptions, estimates and model parameters, that are mostly derived from the early phase of monitoring infection rates. As far as I know, those have never been re-assessed or re-validated in the light of increasing knowledge on aspects like unreported cases, infection pathways, and prevalence estimates based on antibody testing. It is my firm believe that it is time to leave those models behind and return to empiric, evidence-based science. This is my personal attempt to make a start by looking at the publicly available data from daily infection monitoring by the German

institute for disease control, the Robert Koch Institut (RKI), from a slightly different angle. My best hope is to inspire some broader discussion and research on these topics, which I think is urgently needed to make the right decisions on future strategy.

Objectives

Using public data sources, the following questions will be addressed for the daily reported numbers of confirmed COVID 19 cases in Germany:

1. To what extent are observed variations in the number of newly confirmed cases per day determined by the overall number of tests conducted?
2. Can changes in public mobility patterns - an indicator for compliance with "stay at home rules" - explain the observed decrease in daily number of confirmed cases after March 12, 2020?
3. Does the empirical curve of daily confirmed COVID 19 cases show any other indications for effects of political milestone decisions on public health measures?
4. Can the number of confirmed cases per day serve as a reliable basis for estimates of basic reproduction rate and subsequent scenario modelling? Which sources of bias remain, that cannot be statistically eliminated, and how severe would they impact the current understanding of major disease characteristics?

Data

The statistical part of this manuscript is based on daily estimates of confirmed COVID 19 cases in Germany using the dataset available on https://npgeo-corona-npgeo-de.hub.arcgis.com. The data analyzed cover a timeframe from 02. March to 31 May 2020. Due to the daily updates from local health authorities the numbers are sometimes adjusted retrospectively. I have made my last update of the dataset on July 27 and the numbers for the period mentioned are stable in the meantime. For most calculations, the "disease onset date" variable from that dataset was used, which is thought to be less affected by administrative reporting delays than the "report date". Only where no clinical information on date of symptom onset is available, the RKI substitutes by the date of report. Data on daily mobility tracking based on mobile phone data were manually transcribed from the "Mobility Monitor", published by the working group on their homepage "http//covid-19-mobility.org". The time series presented with the mobility monitor starts March 12, the period before that date was set to "100 percent" mobility. This should be acceptable, since the data are presented as percentage deviation from baseline and the official reports published by the working group show that mobility was at previous year levels before that date. Information on PCR test activities per calendar week was derived from the regular publications issued by the RKI on that topic within the institutes regular journal "Epidemiological Bulletin" . Every Wednesday cumulative data on test activities are presented within that bulletin, with a break-out by calendar week starting from week 11 and summarizing all reported tests up to week 10. The numbers used were first published with edition 23 on 04 June 2020. Data for CW 10 were included with half of the test recorded up to that week in absence of a more precise approximation.

Methods

Linear regression models were calculated to assess and quantify the effects of general testing activity and public mobility on the number of confirmed cases per day. Pearson correlation coefficients were calculated to further explore the type of correlation between mobility and number of confirmed cases.

To visualize a potential impact of major political decisions on the trend in newly confirmed cases, residuals were obtained from the linear regression model including background test activity. The residuals were submitted to time series decomposition to examine seasonal variation per calendar week and the remaining error variation. The trend curve, defined as moving average over seven days, was used to identify trend changes in plausible temporal relationship to tightening or loosening of governmental restrictions.

Results

Can the increase in confirmed cases be explained by the number of tests conducted?

The argument, that the increase in registered SARS-COV-2 infections would mainly be driven by a mere increase in the number of test performed has been repeatedly used by against political countermeasures . There is a continuous debate about correct interpretation of rising and falling case numbers in view of varying test activity, yet without any substantial data provided on that aspect so far.

Unfortunately, neither the RKI nor any other institution in Germany are in the position to report on the exact number of tests conducted behind the daily number of laboratory-confirmed COVID 19 cases. The RKI publishes the number of tests and the rate of positive results reported by laboratories participating in their surveillance system on a weekly basis. However, the number of reporting labs was not constant over time, since obligatory reporting was introduced with some delay. In addition, not every test corresponds to a new patient tested, since multiple testing of e.g. persons under quarantine are included. Still, the data available can be used to estimate the overall testing activity in Germany, assuming that the surveillance labs have been selected by the RKI for representativity. The weekly test data were adjusted for the number of labs reporting and divided by the number of days to obtain test rates on a per-day-basis.

A linear regression model was built with test activity as explaining and number of confirmed cases as dependent variable. The model yielded an adjusted R square of .07 ($F = 7.70$ at 89 DF, $p = .007$), meaning a numerically weak, but still statistically significant effect.

Can changes in the number of confirmed cases be explained by changes in public mobility?

It has been observed by various parties that observed infection rates in Germany already started to decline before implementation of lockdown measures. This is typically explained with the assumption that many people were concerned and voluntarily followed the public appeals to avoid unnecessary movements in public long before the official lockdown was announced. In this context, the "COVID 19 mobility project" was initiated to measure public mobility based on mobile phone data, which should serve as an indicator for compliance with the social distancing policy. The project had met with some initial resistance and concerns regarding data protection, but it was considered as important by the RKI.

Data on public mobility as transcribed from the website were included in the linear model described above as a second predictor for number of confirmed cases. The effect was impressive, the combined model yielded an adjusted R squared of .25 ($F = 15.04$ at 88 DF, $p = .000002$), with a highly significant effect for mobility, while the "number of test" effect was no longer visible. However, what seems to prove that mobility reduction prevented infections at first sight is quite surprising in fact. The outcome implies a strong correlation of the number of confirmed cases, aggregated by date of first symptoms, with mobility on the same(!) day. This is not to be expected given an incubation period of about 5 to 6 days . To further investigate if there is also a predictive value for the number of confirmed cases at a later point in time, correlation coefficients were calculated with a time lag increasing from 0 to 8 days. As it can be seen in Figure 1, the correlation coefficients show a linear decrease with increasing time lag.

With that, the data do not support the idea, that changes in public mobility impact the number of confirmed COVID 19 cases. The correlations rather suggest the opposite, it might be the daily communication of rising or falling infection rates, that had an impact on peoples' compliance with the "stay at home" requests. If that would be the case, it might be expected that the mobility index correlates stronger with the number of cases based on date of report than with the number based on date of first symptoms. And indeed, the correlation was found to be -.51 for the former and -.68 for the latter. The correlation between mobility and the number of confirmed case five days later, the estimated mean incubation time, was just -.26.

Do changes in the number of newly confirmed COVID 19 cases per day occur in a plausible relationship to public health measures taken by the German government?

Residuals derived from the linear regression model calculated in section 2 were used for further analyses, thus eliminating the influence of varying test activity. The residuals were submitted to a timeline decomposition on calendar week basis, using a standard procedure of the analysis software package. Figure 2 shows the results for trend (i.e. the 7 days moving average), seasonal trend (i.e. the observed regular fluctuations

per weekday) and the remaining error or random component. The random component indicates some more dynamic developments between calendar week 12 and 16 and, at the lower end, around calendar week 21/22. The seasonal curve shows a pronounced periodicity, which is probably owing to less testing activity on weekends and delayed reporting, even though the effect is reduced by using the date of disease onset instead of date of reporting. The moving average over 7 days clearly accounts for most of the variation.

For a more detailed inspection the trend curve was isolated in Figure 3 with line markers added for landmark decisions taken by the German Government. The timepoints for major interventions in March ("a" to "c") were chosen in alignment with the modeling studies by the MPI mentioned earlier, timepoints "d" (first opening steps on April 20) and "e" (law on extended testing for asymptomatic patients became effective, May 22) were added on top. The timepoints "a" to "d" were marked with a 5 days lag from actual date of intervention to account for the estimated incubation time, as these were assumed to have a direct impact on infections. Timepoint "e" was marked with the date, the law took effect, since testing by itself should not influence infection risk but might impact the chance of detecting cases.

The curve shows a steep and remarkably smooth increase which starts to bend slightly before line mark "a", the day when the first measures – cancelling of large public events on March 09 - would be supposed to take effect. After a peak at the beginning of calendar week 12 it starts to decline, without any visible effect of line marker "b", the closure of schools and nurseries (March 16). Between calendar week 14 and 16, i.e. for approximately two weeks after line mark "c", the full implementation of lock down measures in Germany on March 23, the curve shows a slightly accelerated decline and returns to the previous trend thereafter. There is no discernible change at line marker "d", the day when Germany started to alleviate some of the lockdown measures, while the curve shows a small peak that corresponds to the date, when the updated testing policy took effect, allowing for more tests of asymptomatic patients under certain conditions.

The only political measure implemented in a plausible temporal relationship to a turning point of the curve was the decision to cancel large public events on March 09. None of the following decisions on strengthening or loosening restrictions show a significant impact, even though there are two weeks when the numbers went down a bit faster. This probably reflects the true effect of the shutdown measures, which would not be sustainable and not in a reasonable relationship to their dramatic economic and social impact. Interestingly, what seems to have caused an immediate reaction was a slight modification of the national COVID 19 test strategy put into effect on May 22. The potential bias brought in by the test strategy will be topic of the next section.

What further sources of bias impede valid conclusions from the number of confirmed cases per day?

Figure 3 still shows an exponential increase in confirmed cases during the first weeks, even after the effect of varying test activity has been eliminated. However, it remains questionable if this observed increase truly reflects the spread of infection in the German population. Typical model simulation studies that were trying to estimate the effect of our measures on number of infected and number of deaths have two implicit assumptions:

"patient zero" was correctly identified, i.e. it is known when the virus hit the country.

the number of confirmed cases is always linearly related to the total number of infected patients.

Both assumptions may need to be challenged in the light of today's knowledge. While the first cases in Europe have been officially confirmed in January this year, genetic analyses, sewage water assays and retrospective analyses of frozen blood samples provide more and more evidence that the virus might have circulated at least in France and Italy much earlier, at least mid-December, maybe November, or still earlier . The authors of the respective papers already indicated that this could imply some bias in the assumptions regarding disease progression, but they did not expand on that and for what I know nobody has picked the topic up.

In addition, there is increasing and consistent evidence from Germany and all over the world, that the number of unreported cases is at least tenfold higher compared to the laboratory confirmed cases. We have learned that a much larger proportion of cases than expected – estimates vary between 40% and 85% remain

completely asymptomatic or at best with signs of a common cold, while the RKI, at least in earlier days, had expected almost every infected person to become symptomatic, sooner or later.

For most of the time covered in this paper the official recommendations on PCR testing remained unchanged. Only patients with clinical symptoms like cough or fever should be tested if they had been in contact with a confirmed case before. Until they were showing symptoms, contact persons identified as suspects by the local health authorities were kept under closer surveillance.

If this test strategy is implemented at the beginning of an outbreak the number of cases identified will indeed give an idea of what happens in the population, even if some asymptomatic cases are missed out. The number of unreported cases will be in linear relationship to the number of confirmed. However, what happens if testing starts while there already is a small but substantial share of infected patients in the region where the first case is detected, e.g. some 2 to 3 percent? This is a realistic scenario, estimates from one of the first German hotspots have even been in the region of 15% .

The trigger case could be someone ending up on ICU with severe pneumonia. Following up on contacts will yield an average of 30 to 40 persons under closer surveillance . This already gives a reasonable chance to find the next infected, who just might have his slight cold and normally would not have been tested. Another 40 contacts under surveillance, another one, an ever-growing group of suspects and 2 to 3 cases per hundred detected, since this is prevalence rate assumed. This pattern of confirmed cases begetting confirmed cases constitutes a classic exponential increase, but it does not involve mutual infection. It rather represents a sort of calibration curve for the test strategy. After a very short time – some two or three weeks –the point would be reached when the detectable share of patients has been identified – those with symptoms at a certain point in time – and only from this time on the observations truly reflect the increase or decrease of infection rates in the wider population. Still this does not mean that new confirmed cases have been infected by previously confirmed cases - they all might have caught their virus from various asymptomatic spreaders who never showed up in the statistics.

This is not the only potential bias in public surveillance date, the other one is a negative sampling bias affecting the perceived risk of severe course of disease and death. Basically, there are two ways to get confirmed as a COVID 19 patient, either as a contact of a confirmed case as described above, or as a patient hospitalized for severe respiratory syndrome. With that the group of confirmed cases represents a mix of negative selection (patients with symptoms) and a highly negative selection (patients requiring hospitalization). Indeed, this can also be seen directly from the data. One of the graphs provided within the daily situation update published by the RKI shows how the age groups are represented within the number of confirmed cases compared to the age distribution in total population. What can be seen is a more or less evenly distribution between 20 and 59 years, the working population, and a tendential decrease in numbers for the Younger and Older, which fits to the results showing children to be more resilient, while the older often have reduced social contacts and might thus be less exposed to potential infections. But then patients older than 70 years are clearly over-represented compared to all other age groups. About 17% of all confirmed COVID-cases belong to this group and they account for 85% of all deaths reported. Interestingly, the percentage of hospitalized patients has likewise been in the region 17% for most of the time.

This clearly looks like a bimodal distribution and it could be interesting to analyze the way how these 17% hospitalized patients have been identified. My prediction would be that a large share of those patients was not detected by contact tracing, but they were delivered directly to hospital with severe pneumonia and then tested. The high-risk group should not just be viewed as a share of all COVID 19 patients but as a distinct sub-population with specific features – one of them obviously high age – which makes them vulnerable for a severe course of disease. Any backward conclusion from the currently registered COVID patients to the general risk in an elderly population is meaningless since it is unknown how many of them – despite advanced age – do not have any severe consequences. We observe a number x of patients who require hospitalization and a number y of patients who do not, but there does not seem to be a real connection between those groups. With the current massive expansion of test activities, we observe a continuously decreasing hospitalization and death rate, which is in line with this assumption.

Discussion

There is increasing empirical evidence that our current view on contagiousness and hazard potential of COVID 19 is heavily biased by test strategy and selective sampling. Just to name a few:

- Based on retrospective blood analyses the date of infection for patient zero in France was estimated to mid-December. With a serial interval of four days as suggested by the RKI virus profile and a conservative estimate for the basic reproduction factor of only 2, this would mean that by March 12, when the French government started to get serious on lockdown measures, the country would already have had some eight million inhabitants infected. Likewise, with the virus already present in France and Italy in November / December, it is a very likely scenario that it also had reached Germany much earlier than suspected, which has not been systematically investigated so far.
- Only some countries, like Germany or Italy, show a rapid increase of infections and only during the first two to three weeks. Countries, that are internationally criticized for their management of the pandemic like Brazil or the Sweden record high infection rates, but the increase is far from the projected exponential catastrophe assuming a basic reproduction rate of 2 to 3.
- Despite lack of resources and limited possibilities of infection control in many countries we do not see many reports of an actual crisis or collapse of national health systems so far. There have been local or regional shortages, of course. On the other hand, the region in Northern Italy as well as the city of New York, which suffered considerably from a lack of capacities, had difficult situations with respiratory diseases of other origin, which indicates that there might be some local vulnerabilities in those areas .

The "true" basic reproduction rate of SARS-COV-2 probably never was anywhere close to the dimension of 2.5 to 3, as it is still reported in official information . However, model simulation studies have been typically fitted exactly to these biased numbers. For SIR (Susceptible – Infected – Recovered) models, due to lack of representative data we do not have any valid basis for estimating the number of infected nor the number of recovered patients, nor the susceptible population. There is emerging evidence on cross-immunity with harmless Corona viruses that may already be in the region of 30 to 40 percent or even 80 percent according to some very recent results on T-cell activity . This would for example explain findings like couples sharing the same cabin on a cruise ship or living with a symptomatic COVID patient in the same household but not getting infected. It would also fit to the fact that we see the infection rate in Sweden subside, although they should not have had much more than 15 percent of their population contracting the virus so far, which would not be sufficient for herd immunity.

In summary, public surveillance data as published by the RKI do not confirm a strong impact of public health measures on SARS COV 2 spread in the German population. Specifically, public mobility as an indicator of compliance with social distancing rules seems to be influenced by published infection rates rather than being a moderating factor for infection. Data from the initial phase of the pandemic could be heavily biased and thus provide a severely distorted picture of contagiousness and risk of severe course of disease. Model simulations building on these data will likewise be biased to exaggerated outcomes. There is an ongoing need for solid and large-scale epidemiological studies that not only address currently active infections but also consider resent findings on t-cell based immunity and cross-immunity. The demand for more empirical data has been raised early e.g. by the German society for evidence-based medicine, and it is still valid. Rather than focusing on infection containment in the broad population it might be more reasonable to focus on correctly identifying and protecting the high-risk population, which is probably much smaller than currently perceived.

Figures

Hosted file

`image1.emf` available at https://authorea.com/users/718509/articles/703714-hunting-the-tiger-what-can-be-learned-from-public-surveillance-data-on-the-population-risk-caused-by-sars-cov-2

Figure 1: Pearson correlation coefficients between public mobility and number of confirmed cases; varying time lags from 0 (d)ays to 8 (d)ays

Hosted file

`image2.emf` available at https://authorea.com/users/718509/articles/703714-hunting-the-tiger-what-can-be-learned-from-public-surveillance-data-on-the-population-risk-caused-by-sars-cov-2

Figure 2: Time series decomposition of residuals derived from linear model "number of confirmed cases ~ number of tests"; x-axis labeled with calendar weeks

Hosted file

`image3.emf` available at https://authorea.com/users/718509/articles/703714-hunting-the-tiger-what-can-be-learned-from-public-surveillance-data-on-the-population-risk-caused-by-sars-cov-2

Figure 3: Trend curve (7 days moving average) for confirmed cases corrected for number of tests; x-axis labeled with calendar weeks; a = Cancelling large events; b = school closures; c = restrictions on public mobility; d = partial alleviation of restrictions; e = change of test policy