BMJ Open

Social disparities in the first wave of COVID-19 incidence rates in Germany: a county-scale explainable machine learning approach

Gabriele Doblhammer ,[1,2] Constantin Reinke,[1] Daniel Kreft[1,2]

To cite: Doblhammer G, Reinke C, Kreft D. Social disparities in the first wave of COVID-19 incidence rates in Germany: a county-scale explainable machine learning approach. *BMJ Open* 2022;**12**:e049852. doi:10.1136/bmjopen-2021-049852

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-049852).

Received 08 February 2021
Accepted 24 January 2022

[1]Institute for Sociology and Demography, University of Rostock, Rostock, Germany
[2]Demographic Studies, German Center for Neurodegenerative Diseases, Bonn, Germany

Correspondence to
Dr Gabriele Doblhammer; gabriele.doblhammer@uni-rostock.de

ABSTRACT

Objectives Knowledge about the socioeconomic spread of the first wave of COVID-19 infections in Germany is scattered across different studies. We explored whether COVID-19 incidence rates differed between counties according to their socioeconomic characteristics using a wide range of indicators.

Data and method We used data from the Robert Koch-Institute (RKI) on 204 217 COVID-19 diagnoses in the total German population of 83.1 million, distinguishing five distinct periods between 1 January and 23 July 2020. For each period, we calculated age-standardised incidence rates of COVID-19 diagnoses on the county level and characterised the counties by 166 macro variables. We trained gradient boosting models to predict the age-standardised incidence rates with the macrostructures of the counties and used SHapley Additive exPlanations (SHAP) values to characterise the 20 most prominent features in terms of negative/positive correlations with the outcome variable.

Results The first COVID-19 wave started as a disease in wealthy rural counties in southern Germany and ventured into poorer urban and agricultural counties during the course of the first wave. High age-standardised incidence in low socioeconomic status (SES) counties became more pronounced from the second lockdown period onwards, when wealthy counties appeared to be better protected. Features related to economic and educational characteristics of the young population in a county played an important role at the beginning of the pandemic up to the second lockdown phase, as did features related to the population living in nursing homes; those related to international migration and a large proportion of foreigners living in a county became important in the postlockdown period.

Conclusion High mobility of high SES groups may drive the pandemic at the beginning of waves, while mitigation measures and beliefs about the seriousness of the pandemic as well as the compliance with mitigation measures may put lower SES groups at higher risks later on.

Strengths and limitations of this study

- We combine scattered information on the driving factors of the first wave of COVID-19 infections in Germany in an overall approach.
- We investigate the association between age-standardised COVID-19 incidence rates and a variety of county-specific indicators using a data-driven approach based on machine learning methods.
- Examination of macro factors associated with county-specific COVID-19 incidence rates does not allow conclusions to be drawn at the individual level.
- COVID-19 infection data may reflect diagnostic patterns rather than infection patterns.
- The social characteristics of patients with diagnosed COVID-19 infections may differ from those without a diagnosis.

INTRODUCTION

Germany had comparatively low COVID-19 incidence rates in the first wave.[1] There was a distinct south–north gradient with higher incidence rates in the south than the north,[2] and a number of different factors have been identified with this geographic spread.[2–5] Cross-country studies showed that age structure[6–8] had been shaping COVID-19 risk and in particular death from COVID-19,[9] together with coresidence patterns.[6] Early studies from the start of the pandemic in China indicate that occupational risk factors do not follow am obvious social hierarchy,[10] while for UK, the risk of COVID-19 infections varied by ethnicity and socioeconomic status (SES).[11]For an international review on SES and COVID-19 infections/deaths, see Wachtler *et al*.[12] Turning to Germany, a study using selected regional indicators related to economic, demographic, health and spatial characteristics of regions did not find a relationship with income or unemployment rate but did find a correlation with the number of employees in nursing professions.[13] Other macro-level studies using a limited set of indicators found a change in the relationship between SES and infections/deaths from high rates among high-SES groups to high

rates among low-SES groups over the course of the first pandemic.[12 14] Hospitalisation data point towards a higher risk for the unemployed.[15] No correlation was found with the density of built environment beyond the number of churches in a county[2]; however, labour market participation of the young appeared to be positively correlated with higher incidence rates.[2]

All these studies used partly different indicators and shed light on different aspects of the influence of SES on COVID-19 infections. Our aim is to consolidate these study results using an overall empirical approach. We would expect a possible SES gradient to be negative, that is, higher incidence rates in low-SES groups, because low-SES groups live in more crowded environments, putting them at higher risk for lower respiratory tract infections. They have fewer opportunities to work from home, which impairs their ability to socially distance themselves, making them less protected by lockdown measures. Poverty and associated stress may increase exposure to the virus and reduce the immune system's ability to fight it, while risk factors such as hypertension, diabetes, lung disease and heart disease are more prevalent in low-SES groups. Low-SES groups may be less able to navigate the healthcare system, and their unequal access to information, different policy preferences and attitudes toward risk may influence the processing of information and the assessment of risk.[16 17]

In Germany, the expectation of a negative SES gradient is supported by after-lockdown hotspots in abattoirs and among fruit and vegetable short-term harvest workers.[18 19] This was attributed to the low temperatures and heavy physical work in abattoirs, combined with crowded and unhygienic living conditions. Being able to work from home office is socially stratified,[4] and risk factors of poor health were present in the majority of severe COVID-19 infections.[20] However, there are also reasons for a positive social gradient, that is, higher incidence rates in high SES groups, at least at the beginning of the pandemic, as a study on social distancing responses in the US points out.[17] While the entry of SARS-CoV-2 into Germany was not exclusive from one location, the Alpine ski resort of Ischgl in Austria near the southern border of Germany was identified as a hotspot,[4] and groups with high SES were more likely having spent time there.

Given the lack of individual-level socioeconomic information on COVID-19 infections in Germany, we resorted to a macro-level study design, exploring regional correlates of COVID-19 diagnoses. Macro-level study designs are usually hampered by the myriad of possible regional indicators, which often are highly correlated, and by the limited knowledge about the possible influence factors in relation to the time course of the pandemic. We overcame this limitation by using a data-driven approach that allowed us to identify the most important indicators of a region in predicting COVID-19 incidence rates. We applied methods of explainable machine learning to five distinct periods starting with the first COVID-19 case on 23 January 2020 in Bavaria through 23 July 2020. We used 166 different regional indicators on a county level and explored: (1) to what extent the epidemiological information provided by the RKI, which is summarised further, is reflected in the regional indicators identified by the machine learning algorithms; and (2) whether there are indications of social gradients in the regional distribution of COVID-19 incidence rates.

We hypothesised that the social gradient in incidence rates changed over the course of the pandemic, which started with well-off (skiing) tourists returning from winter holidays in Austria and Italy, was further spread by carnival events in South Germany but later affected workers in abattoirs and agriculture. However, it is unclear whether there was a general social gradient in terms of regions and, if so, if this gradient was positive or negative and when it occurred.

Given the reports about the large number of deaths in nursing homes in the most affected countries such as Spain, Italy and the UK, we expect to find higher incidence rates in regions with a large proportion of elderly who reside in nursing homes or who are dependent on care.

It is not clear if mobility between regions, in addition to the initial start of the disease, is of importance. The decline of mobility measured in terms of distance started on the weekend 14–15 March, and by the end of March, all federal states had agreed on common guidelines and regulations. The increase in mobility from mid-April onwards, however, did not result in increasing incidence rates, but in further decrease.[21] However, mobility had been the decisive factor at the start of the pandemic, and it might still play an important role in spreading the disease out of hotspots. In New York City, the subway system was critical for the spread of the disease from one district to another[22]; mobile phone geolocation was used to show how population outflows from Wuhan to other prefectures were related to the spread.[23]

To explore these questions, we differentiated between five time periods. The first, the initial phase, covered the time span up through 15 March and was characterised by exponentially increasing infection diagnoses from the end of February onwards, with a reproduction value (R) well above 3. The second, covered the period from 16 March to 31 March and is referred to as the first lockdown period. First lockdown measures were introduced from 12 March onwards, with full lockdown starting 16 March. This lowered R to below 1.5. The third period, called the second lockdown period, extended from 1 April to 15 April, during which R fell below 1 and reached a minimum of 0.5 around 15 April. Full lockdown was in place until around 19 April, when smaller shops (<800 m^2) and zoos/parks started to reopen. The fourth period, referred to as the easing period, extends from 16 April to 30 April, with a gradual easing of lockdown measures in all counties. Finally, the fifth period covers 1 May–23 July, a period in which R increased from roughly 0.3 up to levels fluctuating around 1, surging up in specific confined hotspots. Schools and shops started to reopen; masks became

mandatory in public places such as shops, public transport, etc. This is termed the postlockdown period.

DATA AND METHODS

Data

We used data from the RKI, which provides information on COVID-19 diagnoses ($diag_i$) in age group i (i=0–4, 5–14, 15–34, 35–59, 60–79, 80+) and county (NUTS3 region). These were downloaded on 23 July 2020 through the publicly accessible NPGEO-DE platform.[24] Patients were not involved in this study.

Population size on county level was derived from the regional database of the Statistical Offices of the Federation and the Länder at the end of the year 2018.[25] We calculated age-standardised incidence of COVID-19 diagnoses (Inc_{std}) on the county level, using the German age distribution from the year 2018 as the standard population: $Inc_{std} = \sum_{i=0-4}^{80+} \frac{N_i}{\sum_{i=0-4}^{80+} N_i} \cdot Inc_i$ where N_i is the number of persons in age group i in the selected standard population, and $Inc_i=(diag_i/N_i)*100\,000$ is the estimated incidence rate per 100 000 persons in age group i. We used age-standardised incidence rates because counties differ largely in their age distribution, and age has been identified as one of the most important risk factors for severe COVID-19 infections. Since we are not interested in the age effect, we control for it by age standardisation. For the sake of brevity, we will use the term incidence rate when referring to the age-standardised rates.

Macro variables characterise counties in nine domains: 'Demography', 'Employment', 'Politics, religion, and education', 'Income', 'Settlement structure and environment', 'Health care', '(structural) Poverty', 'Interrelationship with other regions' and 'Geography'. The data stem from the "Indikatoren und Karten zur Raum- und Stadtentwicklung"(INKAR) database (2020) of the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR),[26] latitude and longitude were defined in terms of the centres of the county capitals. Air distance of the county centres to Ischgl was calculated by applying the equation: distance in km=sqrt(dx * dx +dy * dy) with dx=111.3 * cos((lat1 +lat2) / 2 * 0.01745) * (lon1 – lon2) and dy=111.3 * (lat1 – lat2), where lat1 and lon1 were the latitude and longitude of county 1 and lat2 and lon2 were the latitude and longitude of county 2. A dichotomous variable indicating more than 100 outbound commuters from the selected early hotspots Heinsberg, Tirschenreuth, Hohenlohekreit, Olpe, Aachen, Greiz, Saarbrücken, Potsdam, Coesfeld, Rosenheim and Göttingen to the respective county stemmed from publicly available commuter flows from the Institute for Employment Research (IAB)[27] for the year 2019, the proportion of Roman-Catholics in a county from the 2011 Census (DESTATIS),[28] the emission of particulate matter with a diameter of 10 micrometres (µm) or less (PM10) from the German Environment Agency Database (UBA)[29] and the number of people in need of care from the Statutory Long-Term Care Census (SLTC) 2015/2017.[30] See data availability statement below for access to the data and the supplement (online supplemental table 1) for the list of all variables. All variables are numeric or dummies taking the values zero or one.

Analysis strategy

Using machine learning approaches we trained random forests and gradient boosting models to predict the age-standardised incidence rates with the 163 macro structures of the counties, which are termed features (figure 1). We also included the age-standardised incidence rates of the previous period (with the exception of the first period) to account for the presence of infections. For each time period, a k-fold random subsampling[31] was performed with 40 folds. The data were randomly split to fit the model to each training set (80%) and predict to the corresponding test set (20%). On the basis of each model, we calculated SHapley Additive exPlanations (SHAP) values that give the contribution of a feature value to the prediction of each individual county in every possible combination with all other features. The higher the contribution, the more important the feature. We used the SHAP procedure in Python.[32] We calculated the average R^2 over all 40 folds to evaluate how well the models fit the data. To evaluate the out-of-sample model performance, the fitted models were used to make predictions on the 40 test sets (20%) and to calculate their average Root Mean Square Error (RMSE). In addition, linear regression models were applied to explain the predictions by the actual response values from the test sets. The average of the R^2 from the linear regression models indicates how much variance from the actual response values could be explained by the predictions.

We used the random forest regressor from the Scikit-learn module in Python[33] with 5000 trees. We kept all other hyperparameters at their default values. Gradient boosting models where trained using the CatBoostRegressor from the CatBoost algorithm.[34] To identify the most important features, we selected the 10 most frequent features from each top 10 ranking of SHAP values over all subsamples. Because the county-specific COVID-19 incidence rates reflect the infection pathways in the entire German population, we fitted a final model on the entire data set based on all 401 counties using these 10 most important features. We displayed their SHAP values as means over all regional SHAP values of the specific feature indicating whether a high/low value of the predicted outcome variable is correlated with a high/low value of the feature.

We categorised the associations into 12 categories depicting the correlation between the feature and the outcome: 1=positive SES gradient (SES high): higher incidence rates in high SES groups; 2=negative SES gradient (SES low): higher incidence in low SES groups; 3=urban/high density gradient (urban): higher incidence in urban/high density regions; 4=rural/low density gradient

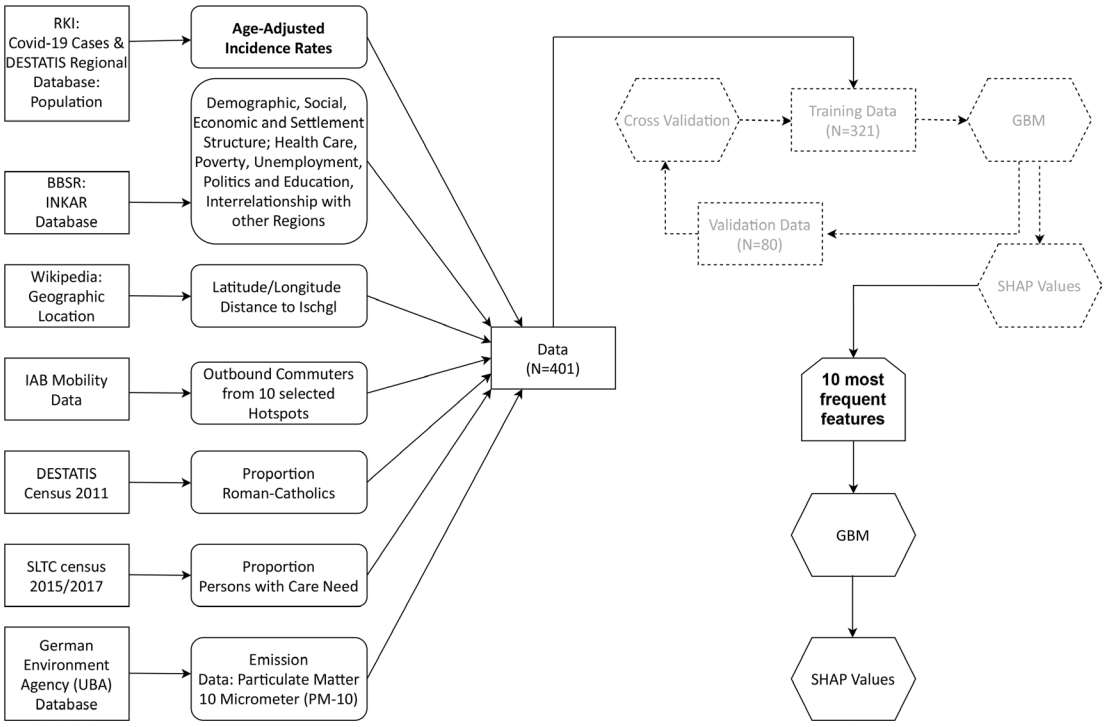


Figure 1 Analysis flow. SHAP, SHapley Additive exPlanations;GBM, gradient boosting model; RKI, Robert Koch-Institute; DESTATIS,Statistics Wiesbaden; BBSR, Federal Institute for Research on Building, Urban Affairs and Spatial Development; INKAR, "Indikatoren und Karten zur Raum- und Stadtentwicklung" ; IAB, Institute for Employment Research; SLTC, Statutory Long-Term Care; UBA, German Environment Agency Database.

(rural): higher incidence in rural/low density regions; 5=poor health gradient (poor health): higher incidence associated with poor health; 6=good health gradient (good health): higher incidence associated with good health; 7=community's connectedness low (connect low): higher incidence associated with low connectedness; 8=community's connectedness high (connect high): higher incidence associated with high connectedness; 9=international migration high (migration high): higher incidence associated with high international migration; 10=geography; 11=population characteristics; and 12=other.

In a sensitivity analysis, we identified all features with pairwise correlations smaller/larger than −0.8/+0.8 and excluded the one of the features that was more strongly correlated with the others. In an additional sensitivity analysis, we randomly selected the five periods to examine whether randomly subdividing the time would distort the interpretable ranking of the features. All analyses were performed using Stata V.16 and Python V.3.8.3.

RESULTS

Age-standardised COVID-19 incidence rates in the five periods

COVID-19 incidence rates revealed distinct geographic patterns that changed over time, as displayed in table 1 and online supplemental figure 1. In the initial period, only a few counties had high incidence rates, while 90% of all counties had rates lower than 16.73 cases per 100 000 person-years. The highest rates were registered

Table 1 Distribution of age-standardised COVID-19 incidence rates per 100 000 person-years by periods (n=401 counties)

	Mean	SD	Min	10%	25%	50%	75%	90%	Max	IQR
Periods	**Per 100 000 person-years**									
Initial period	7.71	14.4	0.0	0.9	2.5	5.0	9.2	16.7	260.0	6.7
First lockdown period	79.46	64.1	2.7	23.9	40.3	63.8	99.6	146.7	671.1	59.3
Second lockdown period	79.01	72.6	3.6	17.8	34.0	56.9	107.5	158.0	721.4	73.5
Easing period	35.32	34.2	0.0	6.5	11.9	25.5	47.3	76.4	223.5	35.4
Postlockdown period	43.94	46.5	0.9	8.6	17.5	30.9	56.7	84.9	549.7	39.2

Initial period: 1 January–15 March 2020; First lockdown period: 16 March–31 March 2020.
Second lockdown period: 1 April–15 April; Easing period: 16 April–30 April 2020.
Postlockdown period: 1 May–23 July 2020.
IQR, Inter Quartile Range; SD, Standard Deviation.

Table 2 R^2 and RMSE scores of gradient boosting models for all periods

	Initial period	First lockdown period	Second lockdown period	Easing period	Postlockdown period
Mean R^2 on training data	0.9996	0.9997	0.9996	0.9996	0.9994
Mean RMSE (out of sample)	13.7705	46.3976	47.2047	23.9655	43.4385
Mean R^2 (out of sample)	0.1446	0.4743	0.6365	0.4823	0.1812
R^2 final model	0.9925	0.9910	0.9924	0.9901	0.9812
RMSE final model	1.2417	6.0469	6.2973	3.4007	6.3612

Initial period: 1 January–15 March 2020; First lockdown period: 16 March–31 March 2020.
Second lockdown period: 1 April–15 April; Easing period: 16 April–30 April 2020.
Postlockdown period: 1 May–23 July 2020.

in counties in South, Southwest and West Germany. The incidence rates steeply increased during the first lockdown period, which was marked by profound clusters of high-incidence counties in South and North Bavaria, central Baden-Wurttemberg and counties in North Rhine Westphalia. These clusters remained stable in the second lockdown period, but the maximum and the between-county range of the incidence rates increased further. The easing period showed the consequences of the lockdown period. In this period, the mean, median and maximum rate and the between-county range declined. More than half of the counties had low and very low incidence rates (below 25.8 cases). Counties with the highest rates were still in Bavaria, Baden-Wurttemberg and North Rhine Westphalia. These patterns remained stable in the post-lockdown period with a slight increase in the cross-county mean, median and the range of the incidence rates but a steep increase in the maximum.

Model fitting and diagnostics

We decided to use the gradient boosting models, as displayed in table 2, because in each period they outperformed the random forests in terms of accuracy (not shown). The out-of-sample performance varied over the periods. Especially the initial phase (period 1) as well as the postlockdown period (period 5) showed a poor out-of-sample performance. For each period, the descriptive statistics of the outcome variable and the 10 most prominent features are presented in the online supplemental tables 2-6).

Model results

The change in the incidence rates over time is also reflected in the changing importance of features as indicated by the number of top 10 features for the five periods (online supplemental table 7).

Period 1: initial phase

In the initial phase, the most important feature was longitude (figure 2 and online supplemental table 8), with high incidence rates especially in hotspot regions in southwestern Germany. The second highest feature revealed a positive social gradient with higher incidence in counties with a higher 'Percentage of employed persons with academic degree in all dependently employed persons'; the third was related to regional population characteristics in terms of the 'Percentage of Roman-Catholics' with higher incidence rates. Among the first 10 features, there were three (3/10), which indicated a positive social gradient with higher incidence in wealthy counties (SES positive), and two (2/10) with a negative SES gradient (SES negative). Furthermore, there were two features with a positive gradient (2/10) with good health (good health) (figure 3).

In summary, in this period geographic location (west vs east) and a large population with Roman- Catholic denomination were the decisive factors. As expected, the latter was positively correlated with the outcome, displaying effects of the superspreading events associated with carnival. We found higher incidence rates both in wealthy counties characterised by high SES and good health, as well as in poorer counties.

Period 2: first lockdown period

Infections from the first period, the percentage of Roman-Catholics and the distance to Ischgl were among the top features with the highest importance, with declining incidence rates for increasing distance to Ischgl (figure 2 and online supplemental table 9). Longitude and latitude now indicated higher incidence rates in the east and the south. The proportion of Roman-Catholics in a county still ranked second. Less connected areas appeared to be associated with higher incidence rates. Wealthy counties were more affected with (2/10) features displaying a positive gradient with SES (figure 3) and zero a negative gradient. In summary, the geographical spread became more distinct with a focus in less connected areas. New infections were heavily influenced by the infections of the previous period in addition to the superspreading events related to carnival, as well as to Ischgl.

Period 3: second lockdown period

The most important features of the previous period are still present: previous incidence, distance to Ischgl, longitude and percentage of Roman-Catholics. Low connectedness, rurality and low population density of a county were still correlated with high incidence rates

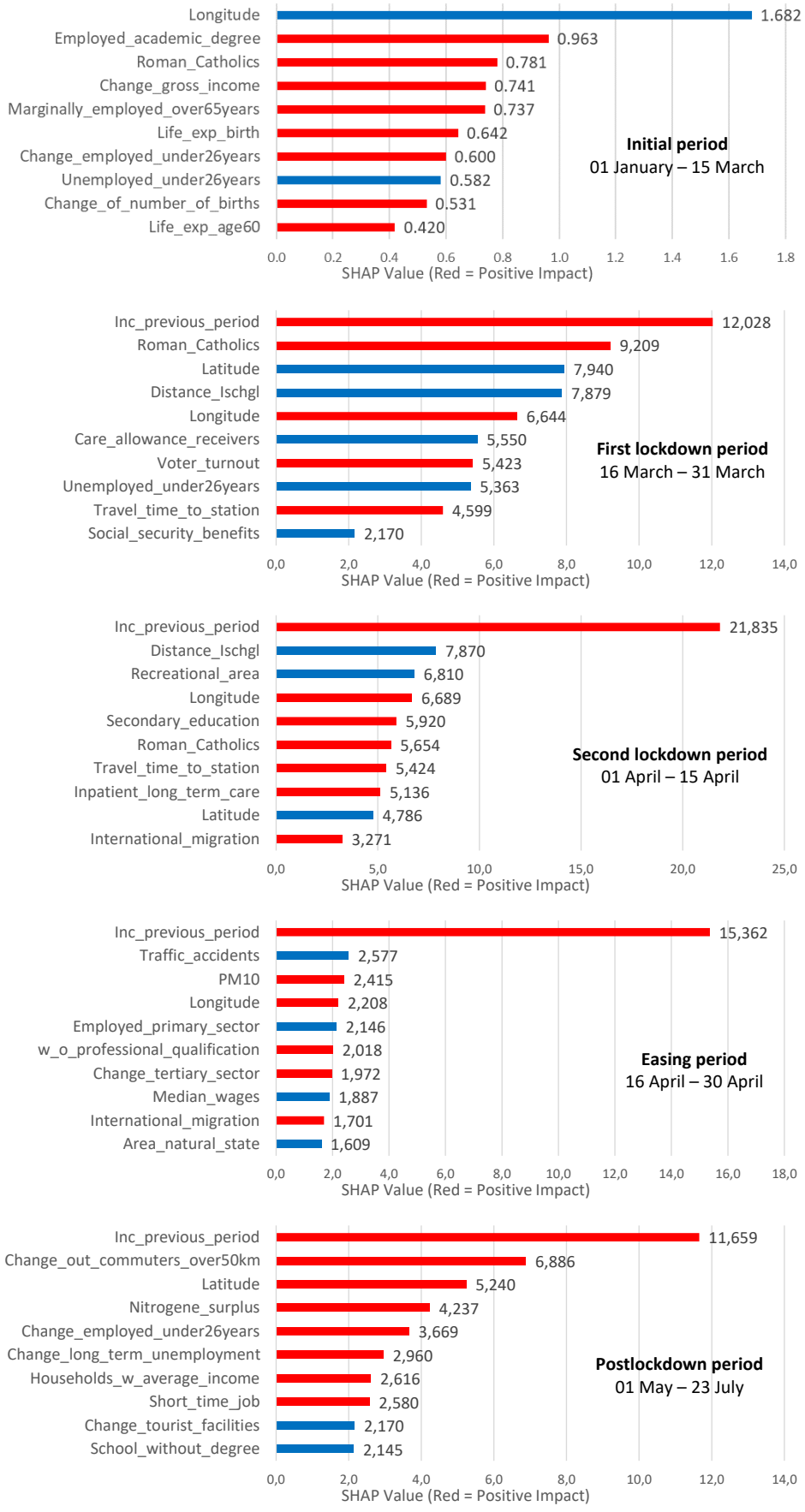
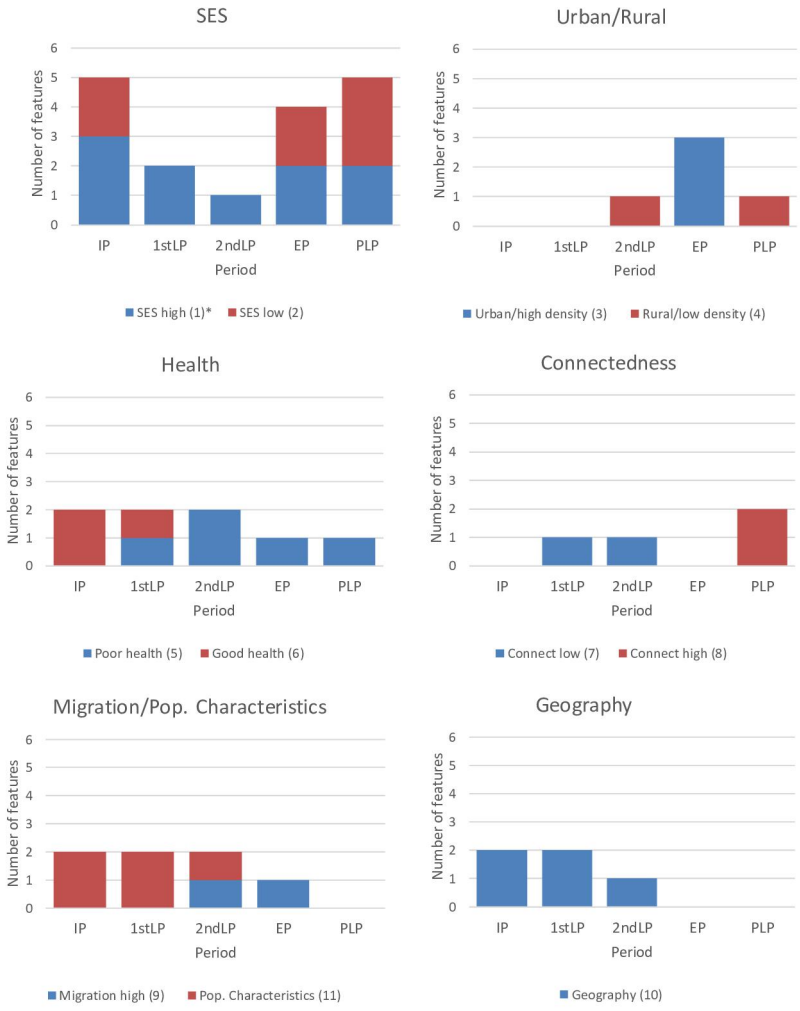


Figure 2 Mean SHAP values of the first 10 features identified by the gradient boosting models by period. SHAP, SHapley Additive exPlanations.



Initial period (IP): -15 March 2020; First lockdown period (1stLP): 16 March -31 March 2020;
Second lockdown period (2ndLP): 1 April - 15 April; Easing period (EP): 16 April - 30 April 2020;
Postlockdown period (PLP): 1 May - 23 July 2020

Figure 3 Number of top 10 features according to their type of correlation with COVID-19 incidence by period. SES, socioeconomic status; Pop., population.

(figure 2 and online supplemental table 10). A total of 2/10 features pointed towards higher rates in counties with poor health, most notably towards counties with a large 'Proportion of persons in inpatient long-term care among all persons in long-term care' (figure 3). With 1/10 features, we continue to observe a positive gradient with SES; at the same time, incidence is elevated in counties with a high 'International net-migration'. In summary, in addition to a persistent positive SES gradient, there is first evidence of vulnerability in counties with a high proportion of nursing home residents among those in need of help and with a high international net- migration.

Period 4: easing period

Incidence rates of the previous period still ranked first, while the next two highest ranking features indicated an urban/high density gradient (figure 2and online supplemental table 11). Poorer counties were affected with 2/10 features indicating a negative social gradient, but also 2/10a positive social gradient (figure 3). In summary, the relationship between SES and the urban/rural association with incidence rates continued to change during the easing period: low SES counties were increasingly less protected and rural/low density counties were better protected than urban/high density counties.

Period 5: postlockdown period

The trends of the easing period were re-enforced: poorer counties showed higher incidence rates (3/10), which is also true for rural/less dense (1/10) and in particular agricultural areas as indicated by the positive correlation with the feature 'Nitrogen surplus' (figure 2 and online supplemental table 12). A county's connectedness in terms of '% outbound commuters/change in outbound commuters' becomes an important feature ranking second, and overall (2/10) features related to connectedness show a positive correlation with incidence rates. In summary, while the negative SES gradient persisted (figure 3), the infections moved back to rural/low density and agricultural areas, and transmission indicated by high

connectedness between countries became an important pathway of spreading the disease.

The sensitivity analysis excluding one feature of highly correlated pairs did come to similar results (online supplemental table 13).

The sensitivity analysis with random assignment of time splits on 27 February, 12 April, 14 June and 1 July revealed an attenuated relationship with the incidence rate of the previous period and the absence of important features such as the distance to Ischgl or the proportion of the population that was Roman-Catholic (online supplemental figure 2).

DISCUSSION

Due to the lack of socioeconomic information of COVID-19 infections in Germany, we resorted to a cross-sectional macro-level study design with regional variables on county-level possibly showing associations with infections. By using machine learning techniques, we neither imposed our expectations on the analysis model, nor did we preselect possible characteristics of the counties. We explored: (1) whether the results reflected our knowledge about the epidemiological situation in the first wave of the pandemic as published in summary bulletins by the RKI and the literature cited above; (2) whether indicators of SES can be identified; and (3) whether these changed over time. Our study shows that in the absence of individual-level data, explainable machine learning methods based on regional data can help shed more light on COVID-19 infection pathways in Germany and better understand the changing nature of the drivers of the pandemic. Explainable machine learning are able to corroborate findings that are already known, but scattered in individual studies, by bringing them together in an empirical data-driven approach.

Restricting our analysis to the first 10 risk factors identified by the variable importance, we conclude that both social gradients, positive and negative, were present in COVID-19 infections right from the beginning; however, they changed over time. Distinguishing five time periods between February and mid-July 2020, we show that the first COVID-19 wave started as a disease in wealthy rural counties in southern Germany and ventured into poorer urban and agricultural counties during the course of the first wave. The negative social gradient became more pronounced from the second lockdown period onwards, when wealthy counties appeared to be better protected than counties with a large proportion of people living in nursing homes or with high net-migration. However, both negative and positive SES gradients were present over the full period. This course of the pandemic is consistent with findings from the USA, where wealthier areas had higher mobility before the pandemic.[17] In Germany, this is reflected in the high feature importance of the distance to Ischgl, an international skiing resort in the Alps, which was one of the hotspots of infections at the beginning of the pandemic. Return mobility from the skiing resort may have contributed to thousands of COVID-19 infections all over Europe,[4] with high SES groups being more likely having spent time there. The positive SES gradient remained strong until the first lockdown period, while from the second lockdown period, a negative gradient began to appear. Again, this is consistent with findings from USA, where the wealthier areas decreased mobility significantly more than poorer areas.[17] Features related to international migration started to play an important role, again an indication of a negative social gradient with migrants being highly represented in occupations with system relevance and thus a higher potential exposition to the virus, such as cleaning workers, workers in food production or nursing of elderly.[35]

Superspreading events have been identified as an important driver of the pandemic, among them the carnival festivals in southern Germany,[36] which most probably are reflected in the feature '%Roman-Catholics' in a county and which is among the most important features until the second lockdown period. They contributed to the positive SES gradients because counties in southern Germany have higher SES and better health profiles. However, superspreading events were also related to the emergence of the negative SES gradient[37] in the easing period, due to poor and little protected working and housing conditions in abattoirs and among agricultural workers. These outbreaks have been attributed to the predominance of migrant workers in these occupations, who often lack social security and easy access to healthcare and may therefore be less likely to report illness or self-isolate.[38]

The spread of the disease in nursing homes during the first lockdown period was often concentrated in a few small facilities, with nursing home staff also at increased risk of infection, which was about six times higher in residential care facilities and twice as high in ambulatory care services than in the general population.[39] While these infections accounted for 60% of all COVID-19 related deaths in Germany, they were responsible for only 8.5% of all registered COVID-19 infections.[39]

Population density per se does not appear to be a risk factor, which is supported by a regional analysis of COVID-19 prevalence in the USA,[40] as well as by Scarpone for Germany.[2] It may be explained by the fact that cities have both the most healthiest population group, whose members benefit from better infrastructure and better access to healthcare, but also the least healthy groups, who have a higher burden of disease and lower life expectancy due to behavioural risk factors and exposure to environmental risk factors.[41] Only in the postlockdown period did connectedness become an important regional characteristic correlated with higher infections, which may reflect the increase in mobility after the lockdown.[21]

High PM10 emission did only play an important role in our study in the easing period, which may be explained by the coarse nature of the county-level data. While one review highlights the possible role of particulate matter (PM) in the spread of COVID-19 in Italian cities,[42] the role of PM in the transmission of SARS-CoV-2 remains unclear.[43] Upregulation of ACE2 receptor by PM is a possible mechanism that is frequently discussed.[42 43]

Features related to economic and educational characteristics of the young population in a county played an important

role at the beginning of the pandemic up to the first lockdown phase. Thus, our results suggest that as early as the first wave, the young population may have considerably contributed to the spread of the virus. Again this is supported by Paul *et al*,[40] who concluded that the infections spread more easily among the elderly in regions where the population is younger. It is also supported by Del Fava *et al*,[44] who showed that social contacts decreased more rapidly among the older than the younger population.

We divided our periods into four 2-week timeslots, which mainly reflect lockdown and easing measures, followed by a longer fifth period over more than 1.5 months, when infection rates were low. Our choice of period duration is supported by Dehning *et al*[45] in their change point analysis of the spread of COVID-19 in Germany, in which they found that change points in the spreading rate affected the confirmed case numbers with a delay of about 2 weeks. They observed three change points, which are: (1) the cancellation of large events with >1000 participants (around 9 March 2020), (2) the closing of schools, childcare centres and most stores (in effect 16 March 2020) and (3) the contact ban and closing of all non-essential stores (in effect 23 March 2020). These three change points fall into the first two time periods of our study, where we observed a positive social gradient and a positive gradient with good health. From our third period onwards, 2 weeks after the contact ban and the closing of non-essential stores, a strong negative social gradient emerged in our analysis, hence suggesting that these restrictions were more likely to protect high SES counties than low ones. This is consistent with a study of the work-from-home capacity in Germany before the pandemic,[46] which was lower among low-skilled and low-wage earners. Our selection of periods is further supported by a sensitivity analysis in which the division of periods was random, leading to inconsistent associations with the incidence of the previous period, as well as an absence of features such as the distance to Ischgl and the proportion of the Roman Catholic population, which have already been confirmed in previous studies.

STUDY LIMITATIONS

Our study is hampered by a series of limitations. First, resorting to county-level data does only permit to interpret results on an aggregate level; any interpretation on the individual level would be misleading. Second, county-level data might be either too course or too finely graded to detect important features driving the pandemic, a problem generally referred to as modifiable areal unit.[47] Third, the data are limited to Germany and do not reflect if or how infections are acquired locally or internationally, with the exception of the variable 'Distance to Ischgl'. Fourth, true infection rates are unknown for COVID-19 because of asymptomatic individuals, regional eligibility criteria for testing leading to different testing rates, as well as differences in reporting of the local 'Gesundheitsämter' to the RKI. To further complicate analyses, data from the RKI do not report the time of infection but rather of diagnosis, and by mid-April, the date of the start of the illness was only known for 62% of the cases.[48] Of these 50% were reported to the RKI within 7 days, on 21 March it took 6.6 days, on 31 March it was 9.9 and in April it took 7.6 days. However, it has been shown that infected individuals are most contagious 2–3 days before symptoms start. In addition, there was a strong weekday effect with lower numbers reported on weekends. Our 14-day time period averages over these various delays, yielding an average picture of infections in the time period. In addition, we included information on infections in the previous period. Fifth, we did not include information on regional health profiles reflecting well-known comorbidities of severe COVID-19 cases such as hypertension, diabetes, cardiac arrhythmia, renal failure, heart failure and chronic pulmonary disease.[20] These comorbidities are more common among persons with low SES and may be one pathway responsible for the negative social gradient observed in this study. However, we included general health measures such as (remaining) life expectancy and premature mortality, both of which are closely related to the chronic diseases mentioned above. Furthermore, we found positive gradients with both good and poor health measures as well as positive and negative SES gradients. This suggests that the relationship between chronic disease and (severe) COVID-19 infections is non-linear and that mitigation measures play an important role. Sixth, we did not use mobile phone data to explore whether changes in mobility account for changes in incidence rates. Seventh, results from the use of machine learning algorithms to identify features and their importance depend on several factors, among them on the procedures implemented, and this may produce spurious splits. We used both random forests (results available on request) and gradient boosting algorithms, which led to similar conclusions. We relied on the latter because of better fit to the data in terms of R^2 and RMSE. Nevertheless, one has to keep in mind that the SHAP values interpreted explain the model rather than the data. Our out-of-sample model fit was poor for both the initial and the postlockdown periods, which reflects the low number of incidence and the huge regional heterogeneity in infections at that time. It was high for periods with high incidence rates in a large number of counties.

CONCLUSION

Lessons for future waves are that there appear to be no unique SES drivers of the pandemic, and dependent on the phase of the pandemic, different social groups are more or less affected. High mobility of high SES groups may drive the spread of the pandemic at the beginning of waves, while mitigation measures and beliefs about the seriousness of the pandemic as well as the compliance with mitigation measures[49] may put lower SES groups at higher risks later on. To further substantiate this finding, we urgently need individual-level data on the socioeconomic background of patients with COVID-1[50] in Germany as well as internationally.

Acknowledgements We would like to thank Stefan Simm and the five anonymous reviewers for their insightful comments, Anna Victoria-Holtz for her assistance with formatting the manuscript and Renee Lüskow-Flibotte for English proofreading.

Contributors GD: substantial contribution to conception and design; interpretation of the data; drafting and revising the article; responsible for the overall content as the guarantor; and final approval of the version to be published. CR: substantial contribution to conception and design; analysis and interpretation of data; revising the article; and final approval of the version to be published. DK: acquisition of data; interpretation of the data; revising the article; and final approval of the version to be published.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data may be obtained from a third party and are not publicly available. The following datasets were derived from sources in the public domain:Robert Koch Institute, ESRI. RKI Corona Landkreise. https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets.Statistische Ämter des Bundes und der Länder. Bevölkerung nach Geschlecht. https://www.regionalstatistik.de/genesisDESTATIS Census 2011: https://ergebnisse.zensus2011.dehttps://ergebnisse.zensus2011.deINKAR Database: Federal Institute for Research on Building, Urban Affairs, and Spatial Development. INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung 2020. https://www.inkar.de/ Institut für Arbeitsmarkt und Berufsforschung (IAB): https://statistik.arbeitsagentur.de/Navigation/Statistik/Statistik-nach-Themen/Beschaeftigung/Beschaeftigte/Beschaeftigte-Nav.html The following datasets are available on request from the data holder:Statutory long-term care census 2015/2017: http://www.forschungsdatenzentrum.de/de/gesundheit/pflege Emission data: German Environment Agency Database (UAB): https://www.umweltbundesamt.de/en

ORCID iD

Gabriele Doblhammer http://orcid.org/0000-0001-7746-0652

REFERENCES

1 Vestergaard LS, Nielsen J, Richter L, *et al*. Excess all-cause mortality during the COVID-19 pandemic in Europe – preliminary pooled estimates from the EuroMOMO network, March to April 2020. *Eurosurveillance* 2020;25.
2 Scarpone C, Brinkmann ST, Große T, *et al*. A multimethod approach for county-scale geospatial analysis of emerging infectious diseases: a cross-sectional case study of COVID-19 incidence in Germany. *Int J Health Geogr* 2020;19:32.
3 Steiger E, Mussgnug T, Kroll LE. Causal analysis of COVID-19 observational data in German districts reveals effects of mobility, awareness, and temperature. *medRxiv* 2020.
4 Felbermayr G, Hinz J, Chowdhry S. Apres-ski: the spread of coronavirus from ischgl through Germany. *Covid Economics: Vetted and Real-Time Papers* 2020;22:177–204 https://www.ifw-kiel.de/fileadmin/Dateiverwaltung/IfW-Publications/Gabriel_Felbermayr/Apres-ski__The_Spread_of_Coronavirus_from_Ischgl_through_Germany/coronavirus_from_ischgl.pdf
5 Frank C, Lewandowsky M, Saad N. Der erste Monat MIT COVID-19-Fällen Im Landkreis Wittenberg. *Sachsen-Anhalt* 2020 https://edoc.rki.de/bitstream/handle/176904/6731/20_2020_korr_DOI_Cluster%20Jessen_15%2005%202020.pdf?sequence=1&isAllowed=y
6 Esteve A, Permanyer I, Boertien D. National age and co-residence patterns shape covid-19 vulnerability. *medRxiv* 2020.
7 Dowd JB, Andriano L, Brazel DM, *et al*. Demographic science AIDS in understanding the spread and fatality rates of COVID-19. *Proc Natl Acad Sci U S A* 2020;117:9696–8.
8 Dudel C, Riffe T, Acosta E, *et al*. Monitoring trends and differences in COVID-19 case-fatality rates using decomposition methods: contributions of age structure and age-specific fatality. *PLoS One* 2020;15:e0238904.
9 Kulu H, Dorey P. The contribution of age structure to the number of deaths from Covid-19 in the UK by geographical units. *medRxiv* 2020.
10 Koh D. Occupational risks for COVID-19 infection. *Occup Med* 2020;70:3–5.
11 Prats-Uribe A, Paredes R, Prieto-Alhambra D. Ethnicity, comorbidity, socioeconomic status, and their associations with COVID-19 infection in England: a cohort analysis of UK Biobank data. *medRxiv* 2020.
12 Wachtler B, Michalski N, Nowossadeck E. Socioeconomic inequalities and COVID-19–A review of the current international literature. *Journal of Health Monitoring*2020.
13 Ehlert A. The socio-economic determinants of COVID-19: a spatial analysis of German County level data. *Socioecon Plann Sci* 2021;78:101083.
14 Plümper T, Neumayer E. The pandemic predominantly hits poor neighbourhoods? SARS-CoV-2 infections and COVID-19 fatalities in German districts. *Eur J Public Health* 2020;30:1176–80.
15 Wahrendorf M, Rupprecht CJ, Dortmann O, *et al*. Erhöhtes Risiko eines COVID-19-bedingten Krankenhausaufenthaltes für Arbeitslose: Eine Analyse von Krankenkassendaten von 1,28 Mio. Versicherten in Deutschland. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* 2021;64:314–21.
16 Patel JA, Nielsen FBH, Badiani AA, *et al*. Poverty, inequality and COVID-19: the forgotten vulnerable. *Public Health* 2020;183:110–1.
17 Weill JA, Stigler M, Deschenes O, *et al*. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc Natl Acad Sci U S A* 2020;117:19658–60.
18 Neef A. Legal and social protection for migrant farm workers: lessons from COVID-19. *Agric Human Values* 2020;1:641–2.
19 Yapici S. Labor and the love of Asparagus: a German panic. *Gastronomica* 2020;20:97.
20 Karagiannidis C, Mostert C, Hentschker C, *et al*. Case characteristics, resource use, and outcomes of 10 021 patients with COVID-19 admitted to 920 German hospitals: an observational study. *Lancet Respir Med* 2020;8:853–62.
21 Bönisch S, Wegscheider K, Krause L, *et al*. Effects of coronavirus disease (COVID-19) related contact restrictions in Germany, March to may 2020, on the mobility and relation to infection patterns. *Front Public Health* 2020;8:619.
22 Harris JE. The subways seeded the massive coronavirus epidemic in New York City. *NBER Working Paper* 2020:w27021 https://www.nber.org/papers/w27021
23 Jia JS, Lu X, Yuan Y, *et al*. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* 2020;582:389–94.
24 Robert Koch Institute, ESRI. RKI corona Landkreise, November 18, 2020. Available: https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/917fc37a709542548cc3be077a786c17_0?selectedAttribute=cases_per_population
25 Statistische Ämter des Bundes und der Länder. Bevölkerung nACh Geschlecht – Stichtag 31.12. – regionale Tiefe: Kreise und krfr. Städte, 2018. Available: https://www.regionalstatistik.de/genesis/online?operation=statistic&levelindex=0&levelid=1605698370989&code=12411&option=table&info=off#abreadcrumb [Accessed 20 Nov 2020].
26 Federal Institute for Research on Building, Urban Affairs, Spatial Development (BBSR). INKAR - Indikatoren und Karten zur Raum- und Stadtentwicklung, 2020. Available: https://www.inkar.de/ [Accessed 25 Nov 2020].
27 Institut für Arbeitsmarkt und Berufsforschung (iab), 2020. Available: https://statistik.arbeitsagentur.de/Navigation/Statistik/Statistik-nach-Themen/Beschaeftigung/Beschaeftigte/Beschaeftigte-Nav.html

28 DESTATIS census, 2011. Available: https://ergebnisse.zensus2011.de [Accessed 25 Nov 2020].
29 German environment agency database (UAB), 2020. Available: https://www.umweltbundesamt.de/en [Accessed 25 Nov 2020].
30 Statutory long-term care census 2015/2017, 2020. Available: http://www.forschungsdatenzentrum.de/de/gesundheit/pflege
31 Ranganathan S. *Encyclopedia of bioinformatics and computational biology*. Amsterdam, Boston, Heidelberg: Elsevier, 2019: 542–5.
32 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017 https://arxiv.org/abs/1705.07874
33 Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: machine learning in python. *Journal of machine Learning research* 2011;12:2825–30 https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
34 Prokhorenkova L, Gusev G, Vorobev A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Process Syst* 2018 https://arxiv.org/abs/1706.09516
35 Statistisches Bundesamt (DESTATIS). Pressemitteilung NR. 279 vom 28, 2020. Available: https://www.destatis.de/DE/Presse/Pressemitteilungen/2020/07/PD20_279_12511.htm [Accessed 25 Nov 2020].
36 Streeck H, Schulte B, Kümmerer BM, *et al*. Infection fatality rate of SARS-CoV2 in a super-spreading event in Germany. *Nat Commun* 2020;11:1–12.
37 Althouse BM, Wenger EA, Miller JC, *et al*. Superspreading events in the transmission dynamics of SARS-CoV-2: opportunities for interventions and control. *PLoS Biol* 2020;18:e3000897.
38 International Labour Organization (ILO). ILO sectoral brief: COVID-19 and its impact on working conditions in the meat processing sector, 2021. Available: https://www.ilo.org/wcmsp5/groups/public/-ed_dialogue/-sector/documents/briefingnote/wcms_769864.pdf [Accessed 30 Jul 2021].
39 Rothgang H, Domhoff D, Friedrich A-C, *et al*. Pflege in Zeiten von corona: Zentrale Ergebnisse einer deutschlandweiten Querschnittsbefragung vollstationärer Pflegeheime. *Pflege* 2020;33:265–75.
40 Paul A, Englert P, Varga M. Socio-Economic disparities and COVID-19 in the USA. *arXiv preprint arXiv* 2009;4935:2020 https://arxiv.org/abs/2009.04935
41 Rydin Y, Bleahu A, Davies M, *et al*. Shaping cities for health: complexity and the planning of urban environments in the 21st century. *Lancet* 2012;379:2079–108.
42 Comunian S, Dongo D, Milani C, *et al*. Air Pollution and COVID-19: The Role of Particulate Matter in the Spread and Increase of COVID-19's Morbidity and Mortality. *Int J Environ Res Public Health* 2020;17:4487.
43 Tung NT, Cheng P-C, Chi K-H, *et al*. Particulate matter and SARS-CoV-2: a possible model of COVID-19 transmission. *Sci Total Environ* 2021;750:141532.
44 Del Fava E, Cimentada J, Perrotta D. The differential impact of physical distancing strategies on social contacts relevant for the spread of COVID-19. *medRxiv* 2020.
45 Dehning J, Zierenberg J, Spitzner FP, *et al*. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science* 2020;369. doi:10.1126/science.abb9789. [Epub ahead of print: 10 07 2020].
46 Alipour J-V, Falck O. IZA DP No. 13152: Germany's capacities to work from home, 2020. Available: https://www.iza.org/publications/dp/13152/germanys-capacities-to-work-from-home
47 Kirby RS, Delmelle E, Eberth JM. Advances in spatial epidemiology and geographic information systems. *Ann Epidemiol* 2017;27:1–9.
48 ander Heiden M, Hamouda O. Schätzung Der aktuellen Entwicklung Der SARS-CoV-2-Epidemie in Deutschland–Nowcasting. *Epid Bulletin* 2020;17:10–15.
49 Galasso V, Pons V, Profeta P, *et al*. Gender differences in COVID-19 attitudes and behavior: panel evidence from eight countries. *Proc Natl Acad Sci U S A* 2020;117:27285–91.
50 Khalatbari-Soltani S, Cumming RC, Delpierre C, *et al*. Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *J Epidemiol Community Health* 2020;74:jech-2020-214297.