Infection Rates from Covid-19 in Great Britain by Geographical Units: A Model-based Estimation from Mortality Data

Hill Kulu and Peter Dorey

ESRC Centre for Population Change and
School of Geography and Sustainable Development
University of St Andrews

Abstract

This study estimates cumulative infection rates from Covid-19 in Great Britain by geographical units and investigates spatial patterns in infection rates. We propose a model-based approach to calculate cumulative infection rates from data on observed and expected deaths from Covid-19. Our analysis of mortality data shows that between 5 and 6% of people in Great Britain were infected by Covid-19 by the last third of April 2020. It is unlikely that the infection rate was lower than 3% or higher than 12%. Secondly, England had higher infection rates than Scotland and especially Wales, although the differences between countries were not large. Thirdly, we observed a substantial variation in virus infection rates in Great Britain by geographical units. Estimated infection rates were highest in the capital city of London where more than 10% of the population might have been infected and also in other major urban regions, while the lowest were in small towns and rural areas. Finally, spatial regression analysis showed that the virus infection rates increased with the increasing population density of the area and the level of deprivation. The results suggest that people from lower socioeconomic groups in urban areas (including those with minority backgrounds) were most affected by the spread of coronavirus in March and April.

Keywords: Covid-19, infectious diseases, infection rates, mortality, statistical modelling, spatial analysis

Address
ESRC Centre for Population Change, School of Geography and Sustainable Development, University of St Andrews, Irvine Building, North Street, St Andrews, KY16 9AL, United Kingdom
E-mail: Hill.Kulu@st-andrews.ac.uk
E-mail: psd3@st-andrews.ac.uk

Background

The Covid-19 pandemic has become a major public health threat in many countries. Observed infections are well documented - they vary across countries and across regions within countries (1). However, observed cases significantly underestimate the actual number of infected individuals and they cannot be easily compared across countries as they depend on the scale of testing, which varies substantially across countries. Little is still known about the actual number of infected people in Europe and other industrialised regions and their proportion of the population. This has led to speculation as to how widely coronavirus is spread and has caused debate in the media on whether the worst is over or is yet to come in the form of a second (and subsequent) wave. For example, Lourenco et al. argued that the majority of the population in the UK might have already been infected by mid-March (2). Others have shown that the virus is not widely spread, although the infected numbers are much higher than reported cases. In a US study Benavid et al. estimated some 54 thousand infected individuals in Santa Clara County (California), which was much higher than the reported cases in the country by early April (approximately one thousand). However, the estimated infection rate was still only 2.8% of the population (3). A study on Gangelt, a small German community, by Streeck et al. reported the proportion of infected individuals to be 15.5%, which was 5 times higher than reported cases (4). However, the area is one of the worst-affected areas in Germany, where the virus spread rapidly and widely during the carnival season. Most estimates of the virus prevalence from other locations lie somewhere between these limits.

National Statistical Offices increasingly provide information on individuals who have died from Covid-19. Normally the data include deaths from Covid-19 by age, sex and geographical region (5–7). An increasing number of studies from various countries also provide information on the infection fatality rates. Although the estimated infection fatality rates vary, most studies report estimates of between 0.4% and 1.3% (4,8–11). It is widely known that data on the likelihood of dying from Covid-19 exhibit a clear age pattern with the infection fatality rates low among young and middle-aged populations. They increase by age and are at their highest among those in their eighties and nineties. Mortality data by age and sex support these patterns - for example, in England and Wales 83% of people who had died by 8th May were aged 70 and older (this includes registered deaths by 8/5/20) (5). Therefore, by bringing together information on the infection fatality rates and data on the number of deaths from Covid-19 the virus infection rate can be estimated with a high degree of accuracy for countries, for various regions within countries and, increasingly, for population subgroups.

The aims of this study are threefold. Firstly, to estimate cumulative infection rates from Covid-19 in Great Britain by geographical units. Secondly, to investigate spatial patterns in infection rates and thirdly, to examine determinants of geographical variation in infection rates. We propose a model-based approach to calculate cumulative infection rates from mortality data. To the best of our knowledge this is the first study to investigate spatial variation in infection rates within a country. Previous studies have either estimated infection rates in one region or town or at the country level (4,12). To date no study has estimated cumulative infection rates using a statistical model. Research has shown that the spread of infectious diseases follows spatial patterns - they normally spread from a few places (often big cities) to other settlements and areas (13–15). Therefore, determining spatial patterns in infection rates and detecting affected areas is important in order to gain a better insight into how widely and where coronavirus has spread. In Great Britain recent mortality data published by the Office for National Statistics (for England and Wales) and the National Records of Scotland provide indirect evidence of significant variation in the virus prevalence rate across the regions (16).

Methods

Estimation of infection rates

Kulu and Dorey calculated expected deaths from Covid-19 using the following formula (17):

$$D_i = c \times \sum_g \sum_x P_{x,g,i} \times F_{x,g} \qquad (1)$$

where D_i is the number of expected deaths in geographical unit *i*, $P_{x,g,i}$ is the number of individuals aged *x* in sex *g* in a geographical unit, $F_{x,g}$ is the infection fatality rate (IFR) in age *x* by sex, which is the same for all spatial units, and *c* is the infection rate, which they assumed to be a constant (0.2) or the same across spatial units. They calculated expected deaths for each geographical unit if everyone becomes infected over time (or the virus spreads widely). Given that we have now data available on observed deaths from Covid-19 we can re-arrange the formula to find the true infection rate for each geographical unit:

$$IR_i = \frac{D_i}{\sum_g \sum_x P_{x,g,i} \times F_{x,g}} \qquad (2)$$

where IR_i is the estimated infection rate for geographical unit *i*. The reader may already have noticed that, essentially, this is a conventional formula for the Standardised Mortality Ratio (SMR) where we calculate the ratio of observed and expected deaths, which we find by applying external (i.e. 'standard') mortality rates by age and sex to our study population. Here the expected deaths are the deaths from Covid-19 assuming that the infection rate is 1 (or everyone is infected). If we can estimate how many people would die if everyone was infected by the virus and we know the observed number of deaths at time moment *t*, we can interpret the ratio of the observed and expected deaths as the Covid-19 infection rate at time *t*, minus two to three weeks, which is normally the time from infection to death.

Clearly, the approach raises a number of questions about its underlying assumptions. Firstly, what infection fatality rates should be used? Currently, the most reliable estimates come from a study by Verity et al., which are based on the analysis of Covid-19 mortality in China (11). Ferguson et al. have adjusted these estimates to the UK's context - they received an overall infection fatality rate of 0.9% (10). We have used the age-specific estimates provided by Ferguson et al. in our baseline model. However, we also examined how much the results would change for Great Britain with higher or lower infection fatality rates using the estimates provided by Verity et al (11). Secondly, can we assume the same IFRs across geographical units? Studies show a significant variation in health and mortality in the UK across regions (18,19). Hence, a 75-year old individual living in a region with relatively poor health is more likely to have an underlying health condition and so to die from Covid-19 rather than a 75-year living in a region with high life expectancy (as this individual is more likely to be healthy). There are several ways of adjusting IFRs to regional differences in mortality and health. We can use estimated life expectancy at age 50 or 65 by region if the data are available and the spatial units are not too small (to avoid a bias because of a small number of deaths); use age-adjusted information on self-reported health by region; or estimate an adjustment factor using a statistical model on deaths and some explanatory factors (e.g. deprivation) on lower level units if data are available. We calculated the age-standardised illness rate for individuals aged 60 and over for each geographical unit and used this as a multiplicative factor for infection fatality rates. We thus slightly modified formula 2 to adjust it to regionally varying mortality and health:

$$IR_i = \frac{D_i}{\sum_g \sum_x P_{x,g,i} \times F_{x,g} \times h_i} \qquad (3)$$

where h_i is an age-standardised coefficient to adjust infection fatality rates for geographical unit *i*. We used the 2011 census data on self-reported limiting long-term illness for the population aged 60 and over (20).

Finally, can we assume that all deaths from Covid-19 are recorded? Although this will not influence our estimates on regional differences in the Covid-19 infection rate (assuming that the same death recording practice is followed across Great Britain), it has potentially an effect on the estimated infection rate at the country level. Clearly, the Great Britain's official statistics have reported an excess of deaths from causes other than Covid-19 in the last month or so (5). Indeed, increased mortality from other causes explains most of this excess (the so-called indirect effect of the pandemic), but some increase may be directly related to Covid-19 mortality (e.g. multiple causes of deaths etc). We thus also estimated the Covid-19 infection rate in Great Britain assuming that some excess mortality from other causes is directly linked to deaths from Covid-19.

We can use formula 3 to estimate the Covid-19 infection rate by geographical units and also calculate other relevant measures, e.g. standard errors and confidence intervals for the estimates. However, we propose to estimate infection rate using the modern regression approach. As we used deaths from Covid-19 in our estimation we can apply a Poisson regression model, which is an appropriate method for count data. The general form of the Poisson regression model without any covariates is as follows:

$$\log(\lambda) = \beta_0 \qquad (4)$$

where λ is infection rate. Since $\log \lambda = \log (D) - \log (E)$ (see formula 3), then

$$\log(D) \ = \log (E) + \beta_0 \qquad (5)$$

where *D* is the observed number and *E* is the expected number of deaths (or an offset). In order to estimate the Covid-19 infection rates by geographical units we stratified the analysis by spatial units to obtain strata-specific estimates for λ_i and their standard errors. There are several advantages in using a regression framework to estimate the Covid-19 infection rates. Firstly, the model will provide an estimate for infection rate and its standard errors and confidence intervals can be easily calculated. Secondly, infection rates can be estimated for different strata, e.g. for geographical units. Further stratification is straightforward (e.g. by education, occupation or ethnic origin) provided that the data are available. Thirdly, the variation in infection rates by strata (e.g. geographical units) can be modelled including explanatory factors (e.g. population density). Finally, the model can be extended to also account for spatial autocorrelation, which is an ingredient of modelling any geographical data.

Spatial patterns in infection rates

We used Moran's *I* statistics to describe the spatial clustering of infection rates. Moran's *I* is calculated using the following formula (21,22):

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (\lambda_i - \lambda)(\lambda_j - \lambda)}{(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}) \sum_{i=1}^{n} (\lambda_i - \lambda)^2} \qquad (6)$$

where *n* is the number of spatial units, λ_i and λ_j are log infection rates for geographical units *i* and *j*, λ is the country's log infection rate and w_{ij} is a measure of the spatial proximity between spatial units *i* and *j*. We used a binary connectivity definition where w_{ij} = 1 if spatial units *i* and *j* share a common boundary, and w_{ij} = 0 otherwise. The interpretation of Moran's *I* is straightforward - the value 1 shows

the perfect spatial clustering of similar values, whereas the value 0 shows no spatial autocorrelation in the variable of interest.

Modelling spatial variation in infection rates

A Poisson regression model (5) can be extended to also include explanatory variables to investigate why infection rates vary across geographical units. For example, infection rates are likely to depend on the population density or the level of deprivation present in the area. However, conventional regression models, when applied to spatial data, violate the assumption of independence of observations. The residuals of such models are spatially autocorrelated. In order to control for spatial autocorrelation we need to apply a spatial regression model. The simplest way of extending an aspatial Poisson regression model to a spatial one would be to include the spatial lag term in the model (23). However, the auto-Poisson model suffers from severe shortcomings and its application should be avoided. A number of suggestions have been made on how to overcome these shortcomings (24), but none of the suggested improvements or models has become widely accepted by the wider research community. For the sake of simplicity we applied a spatial linear model. This approach has several advantages - firstly, it is easy to understand as it is an extension of a conventional OLS regression model. Secondly, sophisticated models have been developed in this research area in the past decade, which allow the measurement not only of direct, but also indirect effects of explanatory variables (25,26). We applied a spatial lag model, which is as follows:

$$\lambda_i = \rho \sum_{j=1}^{n} w_{ij}\lambda_j + \beta_0 + \beta X_i + \varepsilon_i \qquad (7)$$

where w_{ij} is a spatial weight (see formula 6) and ρ is a spatial autocorrelation parameter to measure the effect of the dependent variable λ of neighbouring regions; X is an explanatory variable (e.g. population density or deprivation level). The spatial effects are thus included in the model as a weighted sum of the values of neighbouring regions. For the sake of simplicity and interpretation we used rate as an outcome variable.

Data

Information on deaths from Covid-19 come from the weekly mortality statistics provided by the Office for National Statistics (England and Wales) and the National Records of Scotland. We used mortality data from weeks 10 until week 19 (8th May). The number of registered deaths from Covid-19 were 35,454 in England, 1,775 in Wales and 3,213 in Scotland (6,7). Data on mid-year population by age and sex (one-year intervals) in England, Wales and Scotland by local authority or council areas come from 2018. This is the latest year in which we have detailed information on population age-sex structure by geographical units (27,28). We applied the age-specific fatality rates provided by Ferguson et al. to calculate the number of expected deaths by geographical units (10). We multiplied these estimates by 1.2 and 0.8 to obtain age-specific fatality rates for males and females, accordingly[1]. In order to adjust our estimated number of deaths to regionally varying health conditions we used the 2011 census individual-level data on (self-reported) limiting long-term illness for the population aged 60 and over (20). Although the UK's population health has improved over the last ten years, we used rich individual-level data from 2011 to calculate an adjustment factor assuming that regional differences in health have persisted. We also used deaths from Covid-19 by middle super output areas (MSOAs) in England and Wales (16), population by age and sex (one-year intervals) by MSOA in mid-2018 and

[1] The UK experienced 41,020 registered deaths from Covid-19 by 8/5/20. The deaths of males formed 56% of the total number of deaths and those of females 44%. Multiplying the age-specific fatality rates provided by Ferguson et al. by 1.2 and 0.8, respectively, and applying the obtained rates to the UK's population will lead to a 56/44 split.

information on MSOA-level deprivation to calculate an alternative adjustment factor for regionally varying health conditions (see Table A1 in Appendix). All analyses were performed using R (29–34).

Results

Country level

We have estimated the cumulative Covid-19 infection rates in Britain based on deaths that occurred between 2nd March and 8th May (weeks 10 to 19). With our baseline infection fatality rate about 3.4 million people had been infected by coronavirus in Britain by the second half of April (assuming that the time between infection and death is 2 to 3 weeks). This is 5.2% of Britain's population [95% confidence interval: 5.1-5.2] (Table 1). This estimated number depends on the assumptions we make on the actual infection fatality rate; if we used a higher infection fatality rate the number of infected people would be 2.0 million, which is 3.1% of the population [95% CI: 3.1-3.2]. The number of infected cases is smaller because with higher death rates from Covid-19 fewer people are needed to observe the same death counts. With a lower infection fatality rate the number of infected individuals would be 7.2 million or 11.1% of Britain's population [95% CI: 11.0-11.3]. Which ever estimate we take these figures are many times higher than reported cumulative Covid-19 cases in Britain by the last third of April (e.g. 134,638 in 22/4/20 (1)).

Table 1 shows the Covid-19 infection rate by the Great Britain's constituent countries with the IFR of 1.2%, which is our baseline scenario[2]. We see that England has an infection rate of 5.3% [95% CI: 5.3-5.4], which is the highest among the three countries. The infection levels in Scotland and Wales are lower, 4.8% [95% CI: 4.6-4.9] and 4.0% [95% CI: 3.8-4.2], correspondingly. However, the differences between countries are not substantial, suggesting that the virus has spread to all of Great Britain's constituent countries. The (small) variation between countries persists if we use a lower or higher infection fatality rate to calculate infection rates.

Table 1. Estimated infection rates (IR) from Covid-19 in Great Britain.

	Baseline scenario			Low scenario			High scenario		
Country	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**
England	5.3	5.3	5.4	3.2	3.2	3.3	11.4	11.3	11.5
Scotland	4.8	4.6	4.9	2.9	2.8	3.0	10.2	9.9	10.6
Wales	4.0	3.8	4.2	2.4	2.3	2.5	8.6	8.2	9.0
Great Britain	5.2	5.1	5.2	3.1	3.1	3.2	11.1	11.0	11.3

Notes: Baseline scenario: Ferguson et al., Table 1 (10), Low scenario: Verity et al., Table 1 (11), Upper CrI; High scenario: Verity et al. (11), Lower CrI; 95% Confidence Intervals.

Local authority level

We have estimated the cumulative Covid-19 infection rates in Great Britain by local authorities. We observe a clear spatial pattern in the spread of the virus (Figure 1). The estimated infection fatality rates are highest in the capital city of London, where 10% of the population had likely been infected by the last part of April. Infection rates are also higher than average in other major British cities and their surrounding areas, i.e. Birmingham, Manchester, Liverpool, Cardiff, Newcastle and Glasgow. The infection rates in other major urban areas varied between 7 and 10% by the last third of April. Unsurprisingly, the virus is relatively little spread outside the main urban areas, i.e. in small towns and rural areas. These are large areas of South-West England (Cornwall, Devon, Somerset and Dorset); coastal areas of South-East and Eastern England; Northern England and Central and North Wales;

[2] An overall IFR for our baseline scenario is 1.2%, which is higher than the IFR reported by Ferguson et al, which is 0.9%. We received this estimate by applying ASFRs provided by Ferguson et al. to GB's population by age and sex from 2018. We adjusted these age-specific fatality rates for males and females (see footnote 1).

Southern Scotland (Scottish Borders, South Ayrshire, Dumfries and Galloway) and the North West of Scotland (Highlands and Islands). The estimated infection rate in most of these areas is below the average for Great Britain. However, there are also a few clusters outside the main cities. For example, the virus has also spread in the Lake District where the infection rate is higher than the average for Great Britain.

Next, we calculated the values of Moran's *I*, a global measure of spatial autocorrelation using the local authority level data for England and Wales and council areas for Scotland. The value of the index is 0.49 (with $p<0.01$), which indicates a substantial spatial clustering of infection rates in Britain, which is not surprising. Figure 2 shows the estimated infection rates for selected Great Britain regions to illustrate variation within regions and the local clusters of high infection rates. The estimated infection rates in most areas of the capital city of London are above the national average (Figure 2a). The highest infection rates are observed in the Northwestern part of the city including Brent and surrounding areas. Another cluster is in the Central-Eastern part of the city extending from the North to the South with the highest estimated infection rates in Newham. By contrast, lower infection levels are observed in several boroughs on the edge of the city region including Bexley and Bromley in the Southeast and Kingston, Richmond upon Thames and Sutton in the Southwest. Interestingly, infection rates are also relatively low in some boroughs in the city centre.

In the Midlands the highest infection rates are observed in Birmingham and its surrounding areas (Sandwell and Walsall), but also in East Staffordshire and Derby (Figure 2b). Infection rates are low in Lincolnshire, Rutland and Herefordshire in the Eastern and Western corners of the Midlands accordingly. In South West England the highest levels are observed in Gloucestershire; infection rates are slightly higher in Bristol, but perhaps not as high as one would expect for a large city (Figure 2c). By contrast, infection rates are low in large areas of Cornwall and Devon, and also Dorset and Somerset. In Wales the highest levels are observed in Cardiff and Newport, the largest and third largest city of the country. They are located in the proximity of the English border and thus also helps to explain the high infection levels observed in Gloucestershire (Figure 2d). Estimated infection rates are still low in large areas of Southwestern Wales especially in Ceredigion and also in the Isle of Anglesey in North Wales.

In North West England the areas mainly affected by coronavirus are Merseyside and Greater Manchester and, as expected, the infection rates are highest in Liverpool and Manchester (Figure 2e). Interestingly, there is also a region with higher-than-average infection rates in Cumbria including the districts of Barrow-in-Furness and South Lakeland. This indicates that the virus had also spread to parts of the Lake District before the lockdown was introduced in late March. Significant spatial clustering and regional variation are also observed in Scotland. The highest infection rates are found in the Glasgow region, with the highest levels in Inverclyde (Figure 2f). The infection rates are also above the national average in other main cities including Edinburgh and Dundee. By contrast, the virus has not spread much to the Southern part of Scotland or the North West of Scotland (Highlands and Islands). Interestingly, the infection levels are also low in Aberdeen.

Finally, we have estimated the Covid-19 infection rates by area-type for England and Wales using the ONS urban-rural classification of local authority districts. We have modified the ONS classification by also distinguishing Inner and Outer London. Although the classification is based on local authority districts rather than lower (i.e. LSOA or MSOA) level area classification, and it does not capture all regional variation observed in the UK, it does provide a good summary of the spread of coronavirus in the country. The analysis reveals a clear urban-rural gradient in the spread of Covid-19. The highest levels are observed in Inner and Outer London where the infection rate is 9.7% [95% CIs: 9.3-10.1] and 10.2% [95% CIs: 9.9-10.5], accordingly, followed by other major cities with an infection

rate of 6.8% [95% CIs: 6.7-7.0]. The lowest levels are observed in the areas classified as largely and mainly rural, 3.6% [95% CIs: 3.5-3.7] and 3.0% [95% CIs: 2.8-3.1], respectively.

Table 2. Estimated infection rates from Covid-19 in England and Wales by area type.

Urban-Rural Classification	Population (%)	IR	Lower CI	Upper CI
Inner London	6	9.7	9.3	10.1
Outer London	10	10.2	9.9	10.5
Urban with Major Conurbation	20	6.8	6.7	7.0
Urban with Minor Conurbation	4	5.1	4.8	5.3
Urban with City and Town	26	4.8	4.7	4.9
Urban with Significant Rural	13	4.5	4.4	4.6
Largely Rural	12	3.6	3.5	3.7
Mainly Rural	9	3.0	2.8	3.1

Notes: Baseline scenario.

Regression analysis

Finally, we have fitted a regression model to explain spatial variation in infection rates across local authority districts and council areas. We included in analysis the following explanatory variables: population density (persons per square km) and the level of deprivation measured by the Index of Multiple Deprivation (between 0 and 100). The rationale for choosing these variables is as follows. Infection rates vary significantly between urban and rural areas, and population density is a good measure of the level of urbanicity. Further, more densely populated areas are more likely to bring together different people and thus promote the spread of infectious diseases. Deprivation is believed to be associated with increased infection rates; this may be related to poorer housing conditions (e.g. living in flats) and overcrowding. Furthermore, it is also an indicator of social class and occupation. People from lower socioeconomic groups are more likely to work in occupations exposed to infections (e.g. bus drivers, shop assistants) and are also less able to protect themselves than those from higher socioeconomic groups who can often work from home. In preliminary analysis we also examined the percentage of ethnic minorities (or non-white population). However, the variable has a strong correlation with the level or deprivation and population density so we decided to exclude it because of the issue of multicollinearity.

In a first model we included in analysis population density and the level of deprivation separately (not shown). Both variables showed a significant relationship with the Covid-19 infection rates. In a second model we included both variables simultaneously (Table 3). The coefficient changed only slightly for population density but reduced for the level of deprivation indicating that part of the deprivation effect is explained by population density. Nevertheless, both variables display a significant effect on the virus infection rate. Clearly, infection rates increase with increasing levels of population density and deprivation. (Table 3 reports the coefficients of a spatial lag model. The direct, total and indirect effects, which are required to calculate the exact effect of explanatory variables are displayed in Table A2 in Appendix.) We also observed a significant impact of the spatially lagged dependent variable. The estimate for rho is 0.69 (i.e. it is significantly different from zero) suggesting that infection rates of neighbouring areas are closely related. In substantive terms, the results are largely consistent with the idea of the spread of a virus as a spatial process where spatial proximity and spillover effects play an important role.

Table 3. Results of a spatial lag model on the Covid-19 infection rate.

Variable		**Coefficient**	**Std. Error**	**p-value**
Intercept		-0.02165	0.00351	<0.01
Log Population Density		0.00529	0.00088	<0.01
Index of Multiple Deprivation		0.00022	0.00011	0.036
Rho		0.69476	0.08875	<0.01
N=365		**Residuals**		
Min	*Q1*	*Median*	*Q3*	*Max*
-0.06041	-0.00920	-0.00190	0.00698	0.05755

Conclusions

The aim of this study was to estimate cumulative infection rates from Covid-19 in Great Britain by geographical units. To the best of our knowledge this is the first study to investigate spatial variation in Covid-19 infection rates. We proposed an integrated model-based approach to estimate cumulative infection rates by geographical units and to study determinants of spatial variation in the spread of the virus. Statistical agencies increasingly provide data on the number of deaths from Covid-19 in countries by geographical units and the research community has provided reliable information on infection fatality rates by age, which may vary across geographical units. Our study demonstrates how the cumulative virus infection rates can be estimated with a high degree of accuracy by applying a statistical model to existing mortality data.

Our analysis showed the following. Firstly, based on mortality data up to 8th May we estimated that about 3.4 million people might have been infected by Covid-19 in Britain by the last part of April, which is 5.2% of the population [95% CIs: 5.1-5.2]. Secondly, England exhibited a higher infection rate than Scotland or especially Wales, although the variation between the countries was relatively small. Thirdly, we observed a significant variation in the virus infection rates by geographical units, especially by the level of urbanicity. Estimated infection rates were highest in London and surrounding areas, followed by other major British cities. By contrast, virus infection rates were below average for Great Britain in small towns and rural areas, which included large areas of South-West England, coastal areas of the South-East and East of England, Northern England, Central and North Wales, Southern Scotland and the North West of Scotland. Finally, regression analysis showed a significant effect of population density and levels of deprivation on Covid-19 infection rates. The virus infection rates were higher in areas with higher population densities and deprivation levels.

How much uncertainty is there in our estimates? We quantified uncertainty by using different scenarios and confidence intervals around the estimates for each scenario. In our baseline scenario we used the infection fatality rate of 1.2%, which is an adjusted estimate for Great Britain based on the analysis of Chinese data (10,11). We also used lower and higher infection fatality rates based on uncertainty in the Great Britain estimate and on the recent studies from Germany, Italy and France (4,8,9). Our baseline scenario provided the cumulative infection rate of 5.2%; our low and high scenario gave estimates of 3.1% (95 CIs: 3.1-3.2] and 11.1% [95% CIs: 11.0-11.3] (see also Table A3 in Appendix). How likely are the lower or higher estimates? Lower or higher estimates for infection rates in Great Britain are possible only if our current knowledge of infection fatality rates by age is seriously biased. We also assumed that some excess deaths from causes other than coronavirus in Great Britain in recent months are actually deaths from Covid-19 (e.g. the actual number of deaths from Covid-19 was 10% higher than the reported number); however, the infection levels increased only by a percent point. But equally it is possible that some deaths from causes other than coronavirus have been

recorded as Covid-19 deaths. Recent report shows that the number of deaths from seasonal flu and pneumonia are below the five-year average. From a geographical point of view, regional differences persist whatever infection rate we use.

Our study shows that coronavirus is still not as widely spread in Great Britain as some believe, supporting similar research elsewhere in Europe. On the one hand, this may not be such good news to those who hope that 'herd immunity' will be achieved rapidly (without a heavy death toll). On the other hand, if coronavirus is not widely spread then its suppression and control is still possible with various public health measures before a cure and vaccine become available. Our analysis showed that coronavirus is mostly spread in big cities with a younger-than-average population. This may have reduced the number of deaths in comparison with what would have happened if the virus had spread widely in areas with an older population, although, as we know, Covid-19 has hit hard some pockets of elderly population in cities (e.g. care homes). Our analysis also showed that the virus has hit harder in areas of higher deprivation in cities, exacerbating existing social and spatial inequalities in Great Britain. Many of these areas have an above-average share of ethnic minorities. Although various factors related to living conditions may explain higher infection rates (poor housing conditions, overcrowding, etc.) we believe that the main reason is occupational structure. Many people from lower socioeconomic groups and minority backgrounds work in occupations directly exposed to infections (e.g. bus drivers, shop assistants). These are less able to protect themselves than those from higher socioeconomic groups who can often work from home. A recent analysis by the ONS of deaths from Covid-19 by occupation and ethnicity seems to provide indirect support for this argument (35,36).

It is needless to emphasise that policy-makers should learn from these findings. Firstly, to mitigate the effects that Covid-19 has already had among people in the cities from lower socioeconomic and ethnic minority backgrounds. And secondly, to ensure that people who are exposed to virus infections due to their employment are properly protected, including those in occupations outside the National Health Services. Looking ahead it is also important to ensure that after easing the lockdown the virus should not spread rapidly from the cities to rural areas and small towns with older populations. If the virus spreads rapidly and widely in Great Britain (e.g. during a possible second wave) the effects could be devastating to remote rural communities with an elderly population. Some of these areas in England, Wales and Scotland are strongholds of minority languages and cultures.

References

1. John - Hopkins Coronavirus Resource Center [Internet]. Available from: https://coronavirus.jhu.edu/

2. Lourenco J, Paton R, Ghafari M, Kraemer M, Thompson C, Simmonds P, et al. Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the SARS-CoV-2 epidemic. medRxiv [Internet]. 2020;2020.03.24.20042291. Available from: https://www.medrxiv.org/content/10.1101/2020.03.24.20042291v1

3. Bendavid E, Mulaney B, Sood N, Shah S, Ling E, Bromley-Dulfano R, et al. COVID-19 Antibody Seroprevalence in Santa Clara County, California. medRxiv [Internet]. 2020;2020.04.14.20062463. Available from: https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v1?fbclid=IwAR3NrK8oRFFOVulmb1_4hMJpOUvKUgC6MuS7vi7jPvNyy2xeTPlZoaYmlxA

4. Streeck H, Schulte B, Kümmerer BM, Richter E, Höller T, Fuhrmann C, et al. Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event.

5. ONS. Deaths registered weekly in England and Wales, provisional [Internet]. 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales

6. ONS. Death registrations and occurrences by local authority and health board [Internet]. 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/datasets/deathregistrationsandoccurrencesbylocalauthorityandhealthboard

7. NRS. Deaths involving coronavirus (COVID-19) in Scotland [Internet]. 2020. Available from: https://www.nrscotland.gov.uk/covid19stats

8. Rinaldi G, Paradisi M. An empirical estimate of the infection fatality rate of COVID-19 from the first Italian outbreak. medRxiv. 2020;2020.04.18.20070912.

9. Roques L, Klein E, Papaix J, Sar A, Soubeyrand S. Using early data to estimate the actual infection fatality ratio from COVID-19 in France. medRxiv [Internet]. 2020 Apr 7 [cited 2020 Apr 16];2020.03.22.20040915. Available from: http://medrxiv.org/content/early/2020/04/07/2020.03.22.20040915.abstract

10. Neil M Ferguson et al. COVID-19 reports | Faculty of Medicine | Imperial College London. 2020;(March). Available from: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/news--wuhan-coronavirus/

11. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019 : a model-based analysis. Lancet Infect Dis. 2020;3099(20):1–9.

12. Bohk-Ewald, C., Dudel, C., Myrskyla, M. A demographic scaling model for estimating the total number of COVID-19 infections. medRxiv [Internet]. 2020; Available from: https://www.medrxiv.org/content/10.1101/2020.04.23.20077719v2

13. He D, Dushoff J, Day T, Ma J, Earn DJD. Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales. Proc R Soc B Biol Sci. 2013;280(1766).

14. Langford C. The age pattern of mortality in the 1918-19 influenza pandemic: an attempted explanation based on data for England and Wales. Med Hist. 2002;46(1):1–20.

15. Trilla A, Trilla G, Daer C. The 1918 “Spanish Flu” in Spain. Clin Infect Dis. 2008;47(5):668–73.

16. ONS. Deaths involving COVID-19 by local area and deprivation [Internet]. 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/deathsinvolvingcovid19bylocalareasanddeprivation

17. Kulu H, Dorey P. The Contribution of Age Structure to the Number of Deaths from Covid-19 in the UK by Geographical Units. medRxiv [Internet]. 2020;2020.04.16.20067991. Available from: http://medrxiv.org/content/early/2020/04/22/2020.04.16.20067991.abstract

18. Popham F. Is there a “Scottish effect” for self reports of health? Individual level analysis of the 2001 UK census. BMC Public Health [Internet]. 2006 Dec 21 [cited 2018 Sep 14];6(1):191. Available from: http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-6-191

19. Allan R, Williamson P, Kulu H. Gendered mortality differentials over the rural-urban continuum: The analysis of census linked longitudinal data from England and Wales. Soc Sci Med [Internet]. 2019;221(September 2018):68–78. Available from: https://doi.org/10.1016/j.socscimed.2018.10.005

20. UK Data Service. Census microdata [Internet]. 2020. Available from: https://census.ukdataservice.ac.uk/get-data/microdata.aspx

21. Fotheringham, A Stewart, Chris Brunsdon MC. Quantitative geography: perspectives on spatial data analysis [Internet]. 2010 [cited 2020 Apr 13]. Available from: https://uk.sagepub.com/en-gb/eur/quantitative-geography/book207480

22. Rogerson PA. Statistical Methods for Geography | SAGE Publications Ltd [Internet]. 2014 [cited 2020 Apr 13]. Available from: https://uk.sagepub.com/en-gb/eur/statistical-methods-for-geography/book243249

23. Besag J. Spatial Interaction and the Statistical Analysis of Lattice Systems. J R Stat Soc Ser B. 1974;36(2):192–225.

24. Lambert DM, Brown JP, Florax RJGM. A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application. Reg Sci Urban Econ [Internet]. 2010;40(4):241–52. Available from: http://dx.doi.org/10.1016/j.regsciurbeco.2010.04.001

25. Golgher AB, Voss PR. How to Interpret the Coefficients of Spatial Models: Spillovers, Direct and Indirect Effects. Vol. 4, Spatial Demography. 2016. 175–205 p.

26. Elhorst JP. Relever le niveau de l'économetrie spatial appliquée. Spat Econ Anal. 2010;5(1):9–28.

27. Population estimates - Office for National Statistics [Internet]. [cited 2020 Apr 16]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates

28. Population Estimates | National Records of Scotland [Internet]. [cited 2020 Apr 16]. Available from: https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates

29. R: The R Project for Statistical Computing [Internet]. [cited 2020 Apr 16]. Available from: https://www.r-project.org/

30. Kassambara A. "ggplot2" Based Publication Ready Plots [R package ggpubr version 0.2.5].

31. Arnold JB. Extra Themes, Scales and Geoms for "ggplot2" [R package ggthemes version 4.2.0].

32. Tiedemann F. Visualizing Social Science Data with "ggplot2" [R package ggpol version 0.0.6].

33. Pebesma E. Simple features for R: Standardized support for spatial vector data. R J. 2018 Jul 1;10(1):439–46.

34. Wickham H, Averick M, Bryan J, Chang W, D' L, Mcgowan A, et al. RStudio 2 cynkra 3 Redbubble 4 Erasmus University Rotterdam 5 Flatiron Health 6 Department of Integrative Biology. J Open Source Softw. 2019;4(43):1686.

35. ONS. Coronavirus (COVID-19) related deaths by occupation, England and Wales [Internet]. 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/causesofdeath/datasets/coronaviruscovid19relateddeathsbyoccupationenglandandwales

36. ONS. Odds ratios for risk of coronavirus-related deaths by ethnic group, England and Wales [Internet]. 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/oddsratiosforriskofcoronavirusrelateddeathsbyethnicgroupenglandandwales

37. Idler EL, Benyamini Y. Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies. J Health Soc Behav [Internet]. 1997 Mar [cited 2019 Oct 28];38(1):21. Available from: http://www.jstor.org/stable/2955359?origin=crossref

38. Young H, Grundy E, O'Reilly D, Boyle P. Self-rated health and mortality in the UK: results from the first comparative analysis of the England and Wales, Scotland, and Northern Ireland Longitudinal Studies. Popul Trends. 2010;(139):11–36.

39. Abel GA, Barclay ME, Payne RA. Adjusted indices of multiple deprivation to enable comparisons within and between constituent countries of the UK including an illustration using mortality rates. BMJ Open. 2016;6(11).

Funding

This research was supported by Economic and Social Research Council grant ES/K007394/1 and carried out in the ESRC Centre for Population Change (CPC).

Contributions

HK conceptualised the study. HK and PD both designed it. PD conducted data analysis and HK prepared a manuscript, which PD revised.

Competing interest

HK and PD have nothing to declare.

Figure 1. Estimated infection rates from Covid-19 in Great Britain by local authority districts.

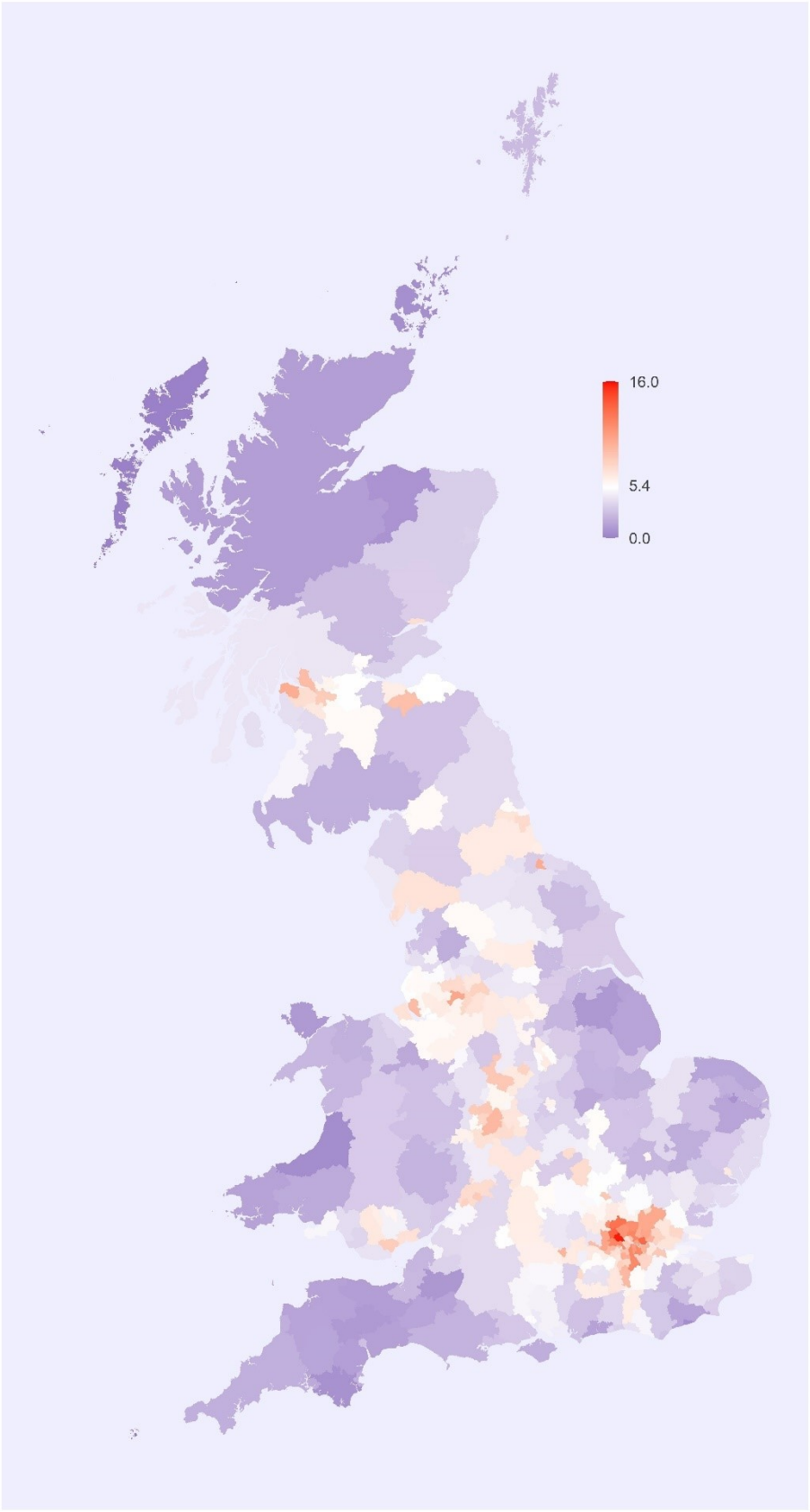
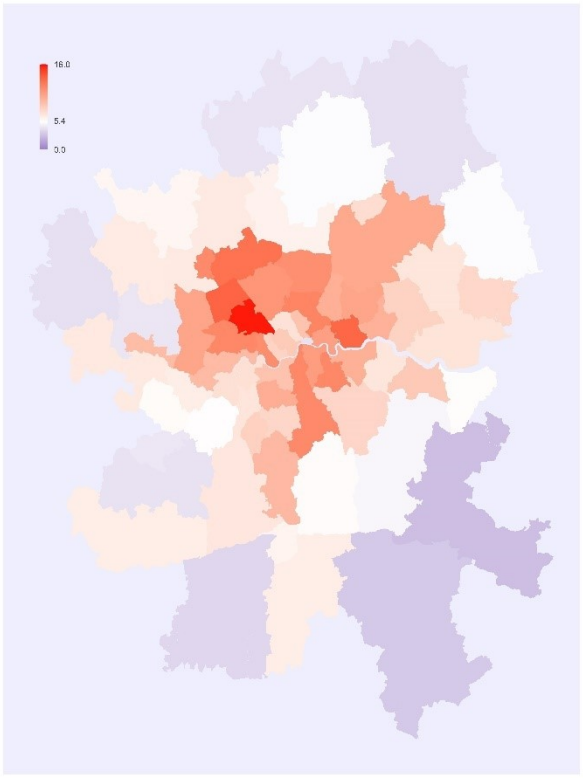


Figure 2. Estimated infection rates from Covid-19 in the UK for selected regions.

a) London



b) Midlands

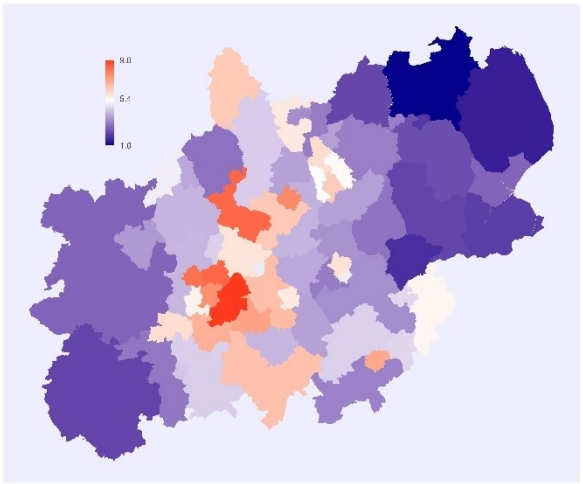
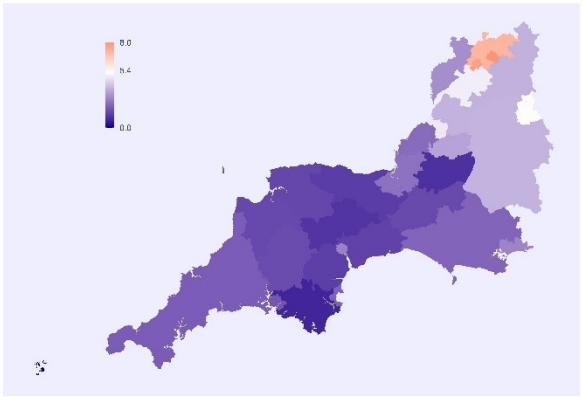


Figure 2. Continued.

c) South West



d) Wales

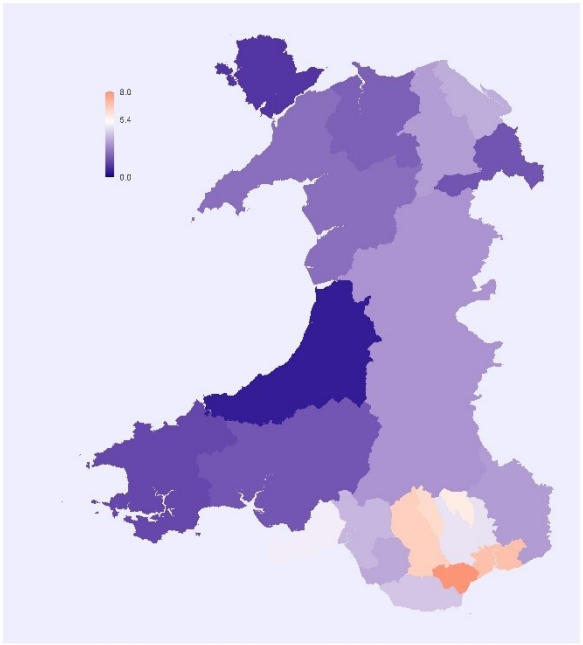
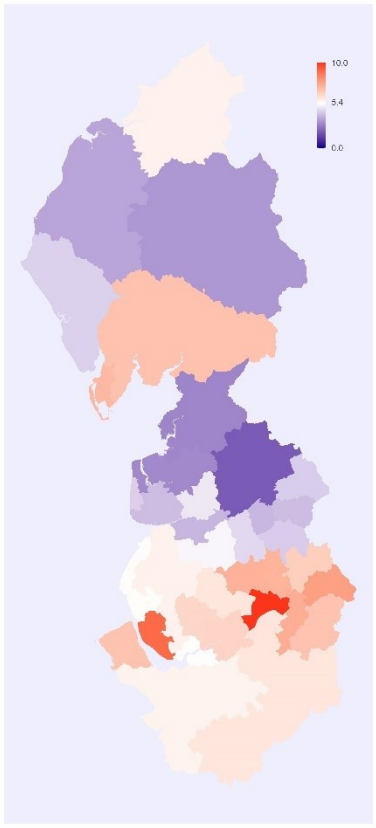
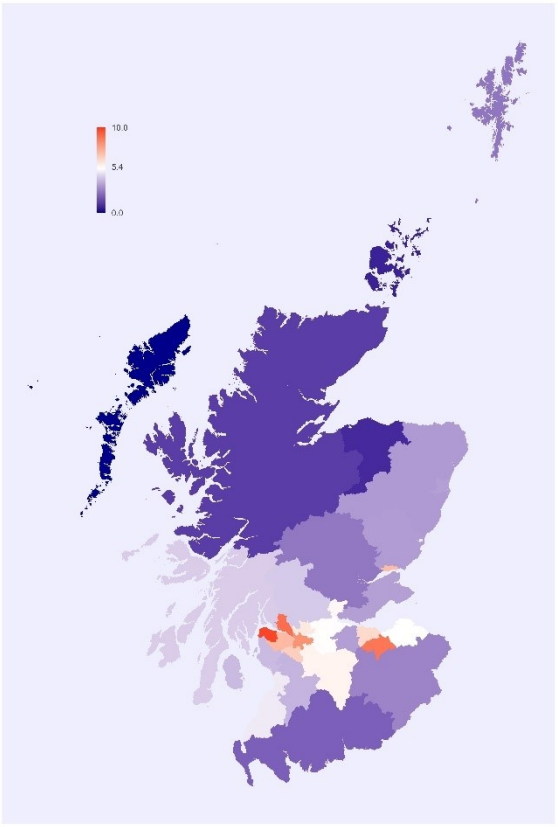


Figure 2. Continued.

e) North West



f) Scotland



Appendix.

Table A1. Estimated infection rates (IR) from Covid-19 in Great Britain.

	Unadjusted			Health-adjusted			Deprivation-adjusted		
Country	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**
England	5.3	5.2	5.3	5.3	5.3	5.4	5.3	5.3	5.4
Scotland	4.8	4.6	5.0	4.8	4.6	4.9	4.9	4.7	5.0
Wales	4.2	4.0	4.4	4.0	3.8	4.2	4.0	3.8	4.2
Great Britain	5.2	5.1	5.2	5.2	5.1	5.2	5.2	5.2	5.3

Notes: Baseline scenario; 95% Confidence Intervals.

We calculated infection rates (IR) by LADs adjusted to regionally varying health and deprivation:

$$IR_i = \frac{D_i}{\sum_g \sum_x P_{x,g,i} \times F_{x,g} \times h_i}$$

where h_i is a coefficient to adjust infection fatality rates for geographical unit *i*. We explored three options:
A.
Remove h_i from the formula and calculate unadjusted IRs (the first three columns in Table A1).

B.
Calculate health-adjusted IRs (the next three columns in Table A1). This was conducted in two steps. Firstly, using the 2011 census data we calculated age-standardised (or weighted) proportion of people with a limiting long-term illness (LLTI) for people aged 60 and older (ASIP) for each LAD (20):

$$ASIP_i = \sum_{60}^{90+} \frac{LLTI_{x,i}}{P_{x,i}}$$

We then calculated a ratio between expected deaths for LAD *i* and that of Great Britain given that individuals with a LLTI have 2.5 times higher mortality rate than individuals without a LLTI (37,38).Therefore, the share of individuals with a LLTI in a LAD determines whether mortality there is higher or lower than average for GB:

$$h_i = \frac{(1-ASIP_i)+2.5*ASIP_i}{(1-ASIP)+2.5*ASIP}$$

where ASIP is the age-standardised (or weighted) proportion of people with a LLTI for Great Britain.

C.
Calculate IMD-adjusted IRs (the last three columns in Table A1). IMD takes into account multiple factors related to mortality including health, income, overcrowding, access to healthcare and other services. We used population estimates at LSOA level for England and Wales and at Datazone level for Scotland. We also used information on the number of deaths from Covid-19 at MSOA level for England and Wales. We calculated IMDs for Scotland and for England and Wales using the methodology developed by Abel et al (39).

We follow the methodology of Kulu and Dorey to estimate the expected number of deaths for each MSOA given 100% infection (17). We use Poisson regression to model deaths in an MSOA *j* within a local authority i:

$$\log(D_{ij}) = \log(Di\sum j * \frac{E_{ij}}{E_{i\sum j}}) + \beta_1(X_{ij} - X_i) + \varepsilon_{ij}$$

where *D* is observed deaths, *E* is the expected number of deaths and X_{ij} is the population weighted mean of GB comparable IMD for all LSOA *k* in MSOA *j* and X_i is the population weighted mean of IMD for all MSOA *j*. We get a point estimate for β_1 with CIs.

Rearranging the above equation and letting now *j* be local authorities within Great Britain i we get the following:

$$\log(D_{ij}) = \log\left(\frac{D_{i\sum j}}{E_{i\sum j}}\right) + \log\left(E_{ij} * \exp\left(\beta_1(X_{ij} - X_i\right)\right) + \varepsilon_{ij}$$

where $\log\left(\frac{D_{i\sum j}}{E_{i\sum j}}\right)$ is simply the intercept of any model to be estimated thus giving us a deprivation adjustment factor of $\exp\left(\beta_1(X_{ij} - X_i\right)$ where X_{ij} is the IMD for local authority j and X_i is the population weighted IMD for Great Britain.

Overall, our analysis shows that differences in the estimated infection rates across countries are not large (Table A1). However, a closer look at the results shows that the unadjusted IR overestimates infection rates in Wales; the health- and IMD-adjusted infection rates are smaller. The reason for that is that there are more people aged 60 and older with an LLTI in Wales than in England and Scotland and more people living in deprived areas. We need to consider this when calculating expected deaths in formula for IR. (Scotland's health record in those ages and IMD are not that different from those of England.)

Table A2. Full results of a spatial lag model on the Covid-19 infection rate.

Variable	Direct Effect	Simulated Std. Error	Simulated p-value	Indirect Effect	Simulated Std. Error	Simulated p-value	Total Effect	Simulated Std. Error	Simulated p-value
Log Population Density	0.00615	0.00079	<0.01	0.01116	0.00937	0.190	0.01732	0.00940	0.050
Index of Multiple Deprivation	0.00026	0.00013	0.042	0.00047	0.00071	0.400	0.00073	0.00079	0.280

Iterations=5000

Table A3. Estimated infection rates (IR) from Covid-19 in Great Britain.

`	Verity et al: medium			Verity et al: low			Verity et al: high		
Country	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**
England	**6.2**	6.1	6.2	**3.2**	3.2	3.3	**11.4**	11.3	11.5
Scotland	**5.5**	5.3	5.7	**2.9**	2.8	3.0	**10.2**	9.9	10.6
Wales	**4.6**	4.4	4.8	**2.4**	2.3	2.5	**8.6**	8.2	9.0
Great Britain	**6.0**	6.0	6.1	**3.1**	3.1	3.2	**11.2**	11.0	11.3

	Rinaldi et al: medium			Rinaldi et al: low			Rinaldi et al: high		
Country	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**	**IR**	**Lower CI**	**Upper CI**
England	**7.3**	7.2	7.4	**4.5**	4.5	4.6	**10.8**	10.6	10.9
Scotland	**6.7**	6.4	6.9	**4.2**	4.0	4.3	**9.9**	9.5	10.2
Wales	**5.4**	5.2	5.7	**3.4**	3.3	3.6	**8.0**	7.6	8.3
Great Britain	**7.1**	7.1	7.2	**4.4**	4.4	4.5	**10.5**	10.4	10.6

Notes: Verity et al., Table 1 (11); Rinaldi and Paradisi, Table 2 (8); 95% Confidence Intervals.