Geo-social gradients in predicted COVID-19 prevalence and severity in Great Britain: results from 2,266,235 users of the COVID-19 *Symptoms Tracker app*

Ruth C E Bowyer (0000-0002-6941-8160)[1*], Thomas Varsavsky(0000-0002-8624-8116)[2*], Carole H. Sudre, (0000-0001-5753-428X)[2], Benjamin A K Murray(0000-0002-2413-923X)[2], Maxim B Freidin (0000-0002-1439-6259)[1], Darioush Yarand (0000-0003-3570-3921), Sajaysurya Ganesh (0000-0002-3720-4176), Joan Capdevila (0000-0003-1658-1076)[3], Ellen J Thompson[1] (0000-0003-2118-821X), Elco Bakker (0000-0002-0863-7140)[3], M. Jorge Cardoso (0000-0003-1284-2558)[2], Richard Davies (0000-0003-2050-3994)[3], Jonathan Wolf (0000-0002-0530-2257)[3], Tim D Spector (0000-0002-9795-0365)[1] Sebastien Ourselin, (0000-0002-5694-5340)[2], Claire J Steves (0000-0002-4910-0489)[1**], Cristina Menni (0000-0001-9790-0571)[1**]

Affiliations:

[1] Department of Twin Research and Genetic Epidemiology, King's College London, Westminster Bridge Road, SE17EH London, UK

[2] School of Biomedical Engineering & Imaging Sciences, King's College London, SE1 7EU London UK

[3] Zoe Global Limited, 164 Westminster Bridge Road, London SE1 7RW, UK

* equal contribution

**equal contribution

Corresponding authors:

cristina.menni@kcl.ac.uk; claire.j.steves@kcl.ac.uk

ABSTRACT

Understanding the geographical distribution of COVID-19 through the general population is key to the provision of adequate healthcare services. Using self-reported data from 2,266,235 unique GB users of the *COVID Symptom Tracker app,* we find that COVID-19 prevalence and severity became rapidly distributed across the UK within a month of the WHO declaration of the pandemic, with significant evidence of "urban hot-spots". We found a geo-social gradient associated with disease severity and prevalence suggesting resources should focus on urban areas and areas of higher deprivation. Our results demonstrate use of self-reported data to inform public health policy and resource allocation.

The COVID-19 epidemic has led to large-scale closures and lockdown measures across the world with the British government sanctioning lockdown from March 23rd 2020.

Early in the pandemic, case distribution did not appear to be evenly spread across countries, with dense urban centres being the most affected[1]. In tandem, individuals in deprived areas have lower life expectancy[2], are more likely to have multiple underlying comorbidities, have a higher level of influenza-associated hospitalisation[3], and therefore could be more susceptible to COVID-19[2].

Based on the high COVID-19 prevalence in urban areas, and the known socioeconomic health gradient, we hypothesised that (i) individuals in deprived areas are at greater risk of both being exposed to and (ii) of experiencing adverse outcomes following contraction of COVID-19.

Understanding the geographical distribution of the virus in a socioeconomic context is key to assist adequate healthcare resourcing, particularly intensive care beds[4].

Here we investigated the GB geographical distribution of COVID-19 using self-reported data from over 2 million users of the *COVID-19 Symptom Tracker*[5] app. We then identify specific geo-social, demographic and clinical factors associated with predicted COVID-19 prevalence and severity.

We studied 2,266,235 unique GB app users reporting daily on COVID-19 symptoms, hospitalisation, RT-PCR test outcomes, demographic information and pre-existing medical conditions over 24 days immediately after major social distancing measures were introduced in the GB ("lockdown"). We computed two proxies of contracting COVID-19: a predicted prevalence score[6], (Positive Predicted Value (PPV)= 0.69[0.66; 0.71] and a predicted severity score based on symptoms associated with hospitalisation PPV=0.77(0.64;0.79).

Following aggregation of variables to local authority district level (LAD/geographic unit derived by the Office of National Statistics), we tested the geographical distribution of symptom severity and predicted prevalence using Global and Local Moran's I tests which assess for non-random spatial distribution and clustering of a feature[7].

We further employed linear regression adjusting for age, gender, obesity, co-morbidities and spatial autocorrelations (SAC)[8] to assess the association between predicted COVID-19 prevalence and severity and geo-social-health factors at eight time points across 24 days using a seven-day window. Comorbidity and demographic data were included as percentage of respondents by middle super output area (MSOA, a more granular geographic unit).

The descriptive characteristics of the study population across the 8 time points are presented in **Table 1**. The number of predicted COVID-19 positive individuals ranged between 79,378

and 15,991 , while the average severity (on a case 1-100) ranged from 4.16 and 1.47. Using local Moran's I, we found that predicted COVID-19 prevalence and severity significantly clusters in urban areas across GB when considered as a proportion of the population per LAD (**Figure 1** and **Figure 1S**) adjusting for multiple testing. We also found that predicted prevalence and severity decreased over time, likely as a result of "lockdown" (**Figure 1** and **Figure 2 a** and **b**) (Pairwise Wilcoxon rank sum tests, Prevalence: all time points accept T2:T3 and T1:T4 , $P<0.001$, Severity all time points $P < 0.001$). However, some hot-spots remained, suggesting in some areas social distancing measures/compliance may not have been equally effective.

In the more granular analysis, we find that urbanicity, area-level deprivation and average household size were positively associated to higher predicted COVID-19 prevalence and severity ($P<0.01$) across all time points; with stronger associations for severity than for prevalence (**Figure 2 c** and **d**). This suggests that people in deprived and/or urban areas remained at higher risk of more severe symptoms. Moreover, we see a positive trend between NOx pollution and COVID-19 prevalence and to a lesser extent severity.

Finally, as expected, we find the association of obesity, smoking and lung disease to be stronger with predicted COVID-19 severity than prevalence.

Here, we observe that predicted COVID-19 prevalence and severity is significantly higher in urban areas compared to rural, and in more deprived areas compared to less deprived. This could reflect the likelihood of individuals in more deprived areas to work in vocations, or live with people who work in vocations, where they are unable to work from home and are thus more likely to be exposed to circulating COVID-19. Accumulation of socio-environmental exposures across the life course are known to contribute to a greater health deficit and disease burden[2]; our results suggest that COVID-19 is no exception.

Moreover, our study illustrates how app data could be used to successfully monitor COVID-19 over time and identify hotspots as the viral pandemic progresses and social distancing measures are implemented or eased. Using this method, we detected a geo-social gradient associated with disease severity and prevalence in the context of COVID-19 suggesting resources should focus on urban areas, areas with highest deprivation, higher average household number and higher air pollution.

Our study has some limitations and assumptions. First, we used self-reported data. Second, volunteers using the app are a self-selected group that may not be fully representative of the

general population. Third, our data on COVID-19 incidence is not from confirmed tests via RT-PCR, but rather from two variables predicted based on user responses. Additionally, we assume that people who suffer from COVID-19 are equally likely to use the app as those who do not. We also assume that people will report the symptoms in the same way. Finally, we aggregated data at MSOA level as we did not have enough respondents for more granular geography.

Future work could seek to integrate this data with data on area-level morbidity, extended pollution data and ethnicity. Indeed higher mortality has been observed amongst minority ethnic groups[9] and disentangling the environmental and biological factors contributing to greater disease burden in both deprived areas and among ethnic minorities is an essential focus of future work to ensure resources and intervention are better assigned.

Conceived and designed the experiments: CJS, TDS, SO, CM; **Analysed the data:** RCEB, TV.

Contributed reagents/materials/analysis tools: MF, CHS, BM, MBF, DY, SG, JC, EJT,EB, MJC, RD, JW

Wrote the manuscript: RCEB, TV, CM **Revised the manuscript**: all

Competing interests: TDS is a consultant to Zoe Global Ltd ("Zoe"). SG, JC, EB, RD, JW are or have been employees of Zoe Global Limited. Other authors have no conflict of interest to declare.

Acknowledgements:

This work was supported by Zoe Global. The Department of Twin Research is funded by the Wellcome Trust, Medical Research Council, European Union, Chronic Disease Research Foundation (CDRF), Zoe Global Ltd and the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. CM is funded by the Chronic Disease Research Foundation and by the MRC Aim-Hy project grant. CHS is an Alzheimer's Society Junior Fellowship AS-JF-17-011; SO and MJC are funded by the Wellcome/EPSRC Centre for Medical Engineering (WT203148/Z/16/Z), Wellcome Flagship Programme (WT213038/Z/18/Z).

We express our sincere thanks to all the participants of the COVID Symptom Tracker app. We thank the staff of Zoe Global Limited, the Department of Twin Research for their tireless work in contributing to the running of the study and data collection.

Data Sharing: Anonymised research data will be shared with third parties via the centre for Health Data Research UK (HDRUK.ac.uk). Data updates can be found on https://covid.joinzoe.com

Ethics: The Ethics for the app has been approved by KCL ethics Committee and all users provided consent for non-commercial use. An informal consultation with TwinsUK members over email and social media prior to the app having been launched found that they were overwhelmingly supportive of the project.

References

1. Stier A, Berman M, Bettencourt L. COVID-19 Attack Rate Increases with City Size. *Preprint at https://papersssrncom/sol3/paperscfm?abstract_id=3564464* 2020
2. Marmot M. Health equity in England: the Marmot review 10 years on. *BMJ* 2020;368:m693. doi: 10.1136/bmj.m693 [published Online First: 2020/02/26]
3. Hungerford D, Ibarz-Pavon A, Cleary P, et al. Influenza-associated hospitalisation, vaccine uptake and socioeconomic deprivation in an English city region: an ecological study. *BMJ Open* 2018;8(12):e023275. doi: 10.1136/bmjopen-2018-023275 [published Online First: 2018/12/24]
4. Blumenshine P, Reingold A, Egerter S, et al. Pandemic influenza planning in the United States from a health disparities perspective. *Emerg Infect Dis* 2008;14(5):709-15. doi: 10.3201/eid1405.071301 [published Online First: 2008/04/29]
5. Drew DA, Nguyen L, Steves CJ, et al. Rapid implementation of mobile technology for real-time epidemiology of COVID-19, 2020.
6. Menni C, Valdes AM, Freidin MB, et al. Loss of smell and taste in combination with other symptoms is a strong predictor of COVID-19 infection. *Preprint at https://wwwmedrxivorg/content/101101/2020040520048421v1* 2020
7. Rezaeian M, Dunn G, St Leger S, et al. Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. *J Epidemiol Community Health* 2007;61(2):98-102. doi: 10.1136/jech.2005.043117 [published Online First: 2007/01/20]
8. Anselin L, Griffith DA. Do spatial effects really matter in regression analysis? *Papers - Regional Science Association* 1988
9. Khunti K, Singh AK, Pareek M, et al. Is ethnicity linked to incidence or outcomes of covid-19? *BMJ* 2020;369:m1548. doi: 10.1136/bmj.m1548 [published Online First: 2020/04/22]
10. Bivand RS, Wong DWS. Comparing implementations of global and local indicators of spatial association. *Test-Spain* 2018;27(3):716-48. doi: 10.1007/s11749-018-0599-x

Figure Legends

Figure 1. Geographical Distribution of predicted COVID-19 prevalence across eight time points

The map was created using a shapefile of Local Authority District level data from the ONS using the geopandas package and Python. Overlaid over the map are statistically significant 'hot-spots' and 'cold-spots' at LAD level. To create these regions, a 'neighbours list was calculated from a shapefile of the LADs using a queen contiguity condition. Spatial weights were calculated assuming equal weight of neighbouring areas, and spatially lagged values which represent the average neighbouring COVID-19 prevalence/severity for each LAD. We adjusted for multiple testing using the Benjamini & Hochberg method and used the 'spdep' package for the spatial components of our analysis[10]. The severity score was calculated using a weighted sum of symptoms a patient had, normalised between 0 and 1. The weights were found by using the frequency of symptoms amongst a training set of 993 respondents who reported visiting hospital for a COVID-19 related issue and having a positive RT-PCR test. Predicted severity= chest pain x 0.115 + severe or significant persistent cough x 0.104 + hoarse voice x 0.097 + skipped meals x 0.096 + loss of smell x 0.092 + severe fatigue x 0.089 + shortness of breath x 0.0855 + delirium x 0.083 + fever x 0.0814 + diarrhoea x 0.081 + abdominal pain x 0.076

Predicted Covid-19 +ve cases in GB with highlighted spatially significant hotspots

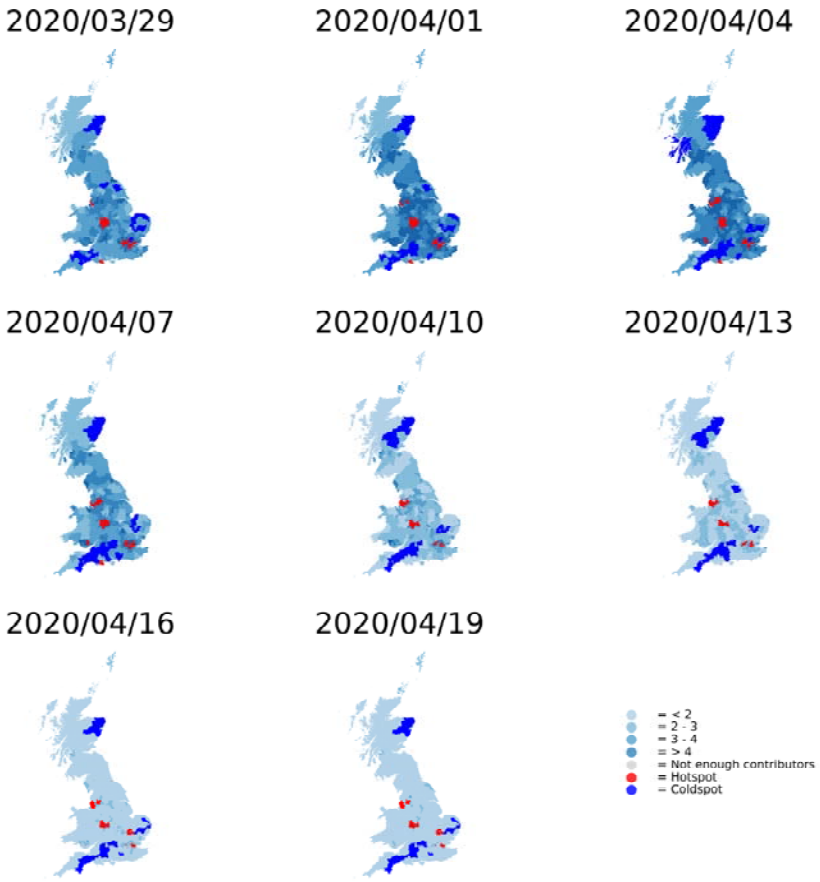

Figure 2. Predicted COVID- 19 A) prevalence and B) severity in the UK across eight time points.

Results of multivariate linear regression models for each of the time points, where the two COVID measures C) prevalence and D) severity at MSOA level were regressed against geosocial factors: the Index of Multiple Deprivation (IMD, binned into quintiles generated on the average IMD within each MSOA, where 1 is most deprived and 5 is least), a rural-urban gradient (RUC, considered as a continuous variable where 1 is the most urban and 8 is the most rural), General practitioners per population in MSOA (GPs/MSOA, where a higher number indicates more GPs per individual by MSOA), average household number (calculated as number of inhabited dwellings/MSOA population, where a higher number indicates a higher average number of individuals per household) and comorbidity and demographic data derived from app responses representing percentage of respondents by MSOA who reported having kidney, heart or lung disease, and who are diabetic, a smoker or obese (calculated as BMI<30). Finally, we derived age and sex variables to adjust for response bias and considered as demographic factors: these were calculated as the difference of the expected mean/ratio of age/sex in the MSOA (derived from ONS population data) and the observed mean/ratio of age/sex amongst respondents. Therefore, a positive number suggests we had *less* males/*younger* respondents than would be expected were our population to be the same as the average population by MSOA, and thus a positive association with COVID prevalence/severity suggests greater association with this reporting bias. A lagged variable was included to account for spatial autocorrelation. Only MSOAs where at least 20 individuals were considered (n = 8097). Standardised coefficients are reported. Analysis conducted in RStudio v1.1.423 and R v3.6.3.

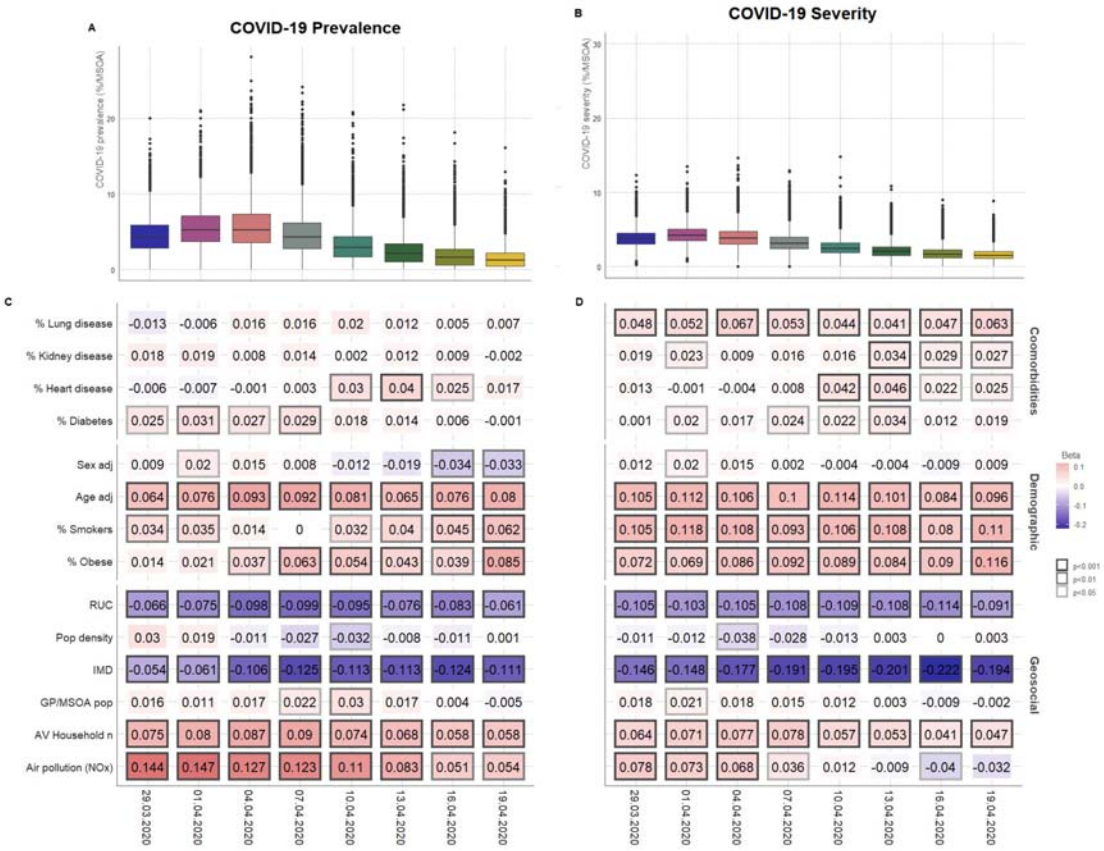

Table 1. Demographic characteristics of the study population at 8 time points. At each time point, we only consider users who have made an assessment in the last 7 days.

	29/03/2020	01/04/2020	04/04/2020	07/04/2020	10/04/2020	13/04/2020	16/04/2020	19/04/2020
N	1,324,843	1,431,515	1,142,923	1083601	995157	985860	980608	1164262
Predicted COVID-19 (n/%)	60827 (4.59%)	79378 (5.55%)	62508 (5.47%)	48418 (4.47%)	30132 (3.03%)	22352 (2.27%)	16586 (1.69%)	15991 (1.37%)
Average predicted COVID-19 severity score (max 100)	3.72	4.16	3.77	3.08	2.38	2	1.65	1.47
Age, yrs	41.85(12.88)	42.09(12.94)	43.56(13.03)	44.15(13.01)	44.75(12.98)	45.13(12.94)	45.41(12.91)	44.89(12.97)
Male: Female	426,923: 897,920	459,620: 971,895	365,078: 777,845	353,233: 730,368	327,608: 667,549	327,620: 658,240	327,114: 653,494	388,378: 775,884
Kidney Disease	0.60%	0.60%	0.60%	0.70%	0.70%	0.70%	0.70%	0.70%
Lung Disease	16.20%	16.60%	15.90%	15.60%	15.30%	15.30%	15.30%	15.80%
Diabetes	2.90%	3.10%	3.20%	3.20%	3.30%	3.40%	3.40%	3.40%
% smokers	13.80%	14.00%	12.10%	11.50%	10.80%	10.60%	10.50%	11.10%
Heart Disease	1.70%	1.70%	1.90%	1.90%	2.00%	2.00%	2.10%	2.10%

Figure S1. Geographical Distribution of predicted COVID-19 severity across eight time points