

1 **Supplementary Methods**

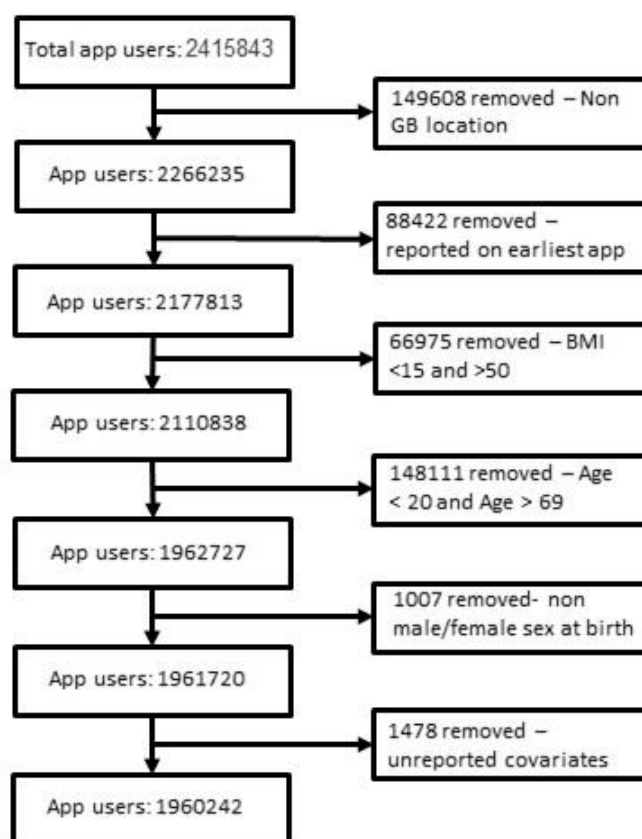
2 **Study setting and participants**

3 The COVID Symptom Study app developed by Zoe with scientific input from researchers and
4 clinicians at King's College London and Massachusetts General Hospital, was launched in GB on
5 Tuesday the 24th March 2020 (<https://covid.joinzoe.com/>) and in the 23 days (March 29th–April
6 19th) immediately after the UK lockdown ([https://www.gov.uk/government/speeches/pm-
7 statement-on-coronavirus-22-march-2020](https://www.gov.uk/government/speeches/pm-statement-on-coronavirus-22-march-2020)) was introduced, it reached 2,266,235 unique GB users,
8 making 9,108,769 assessments (e.g. an average user is included in 4 out of 8 timepoints).
9 Referrals/word of mouth, press and eventually partnerships with charities and the Welsh and
10 Scottish governments drove usage.

11 The app enables capture of self-reported information related to COVID-19 infections. On first use,
12 the app records self-reported location, age, and core health risk factors. With continued use,
13 participants provide daily updates on symptoms, health care visits, COVID-19 testing results, and if
14 they are self-quarantining or seeking health care, including the level of intervention and related
15 outcomes. Individuals without apparent symptoms are also encouraged to use the app. Through
16 direct updates, the research team can add or modify questions in real-time to capture new data to
17 test emerging hypotheses about COVID-19 symptoms and treatments. Importantly, participants
18 enrolled in ongoing epidemiologic studies, clinical cohorts, or clinical trials, can provide informed
19 consent to link data collected through the app in a Health Insurance Portability and Accountability
20 Act (HIPAA) and General Data Protection Regulation (GDPR)-compliant manner with extant study
21 data they have previously provided or may provide in the future.

22 In this study, we included 1,960,242 unique users as outlined in the flow diagram below (**Figure A**).
23 Briefly, out of 2,415,843 unique app uses who reported on the COVID-19 symptom Study App
24 between 29th March 2020 and 19th April 2020, we excluded (i) 149,608 non GB users; (ii) 88,422
25 users who only reported on the earliest app-version that did not include loss of smell and taste (the
26 strongest single predictor of COVID-19^{1,2}); (iii) 66,975 reporting BMI outside the biological range; (iv)
27 148,111 users younger than 20 or older than 69; (v) 1007 with missing biological sex at birth or who
28 were not assigned male or female as their biological sex at birth; (v) 1478 users who did not report
29 on pre-existing medical conditions (**Figure A**).

30

31 **Figure A. Flow diagram representing the study subjects' inclusion criteria.**

32

33 **Geographic clustering of COVID-19 prevalence**

34 Because we were primarily interested in understanding the geography of COVID-19 distribution, and
 35 how aspects of an area, in particular area-level deprivation, associated with COVID-19 prevalence we
 36 aggregated user data at different GB geographic areas. This was particularly of use as the geosocial
 37 variables considered (please see below) are also defined geographically and are time invariant (as
 38 they are not defined by the app users themselves but by GB geographic area).

39 The maps (**Figure 1, S2**) were created using a shapefile of Local Authority Districts (LADs) from the
 40 Office for National Statistics (ONS) using the geopandas package in Python. Overlaid on the map are
 41 statistically significant 'hot-spots' and 'cold-spots' at LAD level. To assess the significance of these
 42 regions, we used Local Moran's I test, as introduced below. In order to do this, spatial weights were

43 calculated to create a spatially lagged COVID-19 prevalence variable for each LAD. Because our
44 geographical units share borders we assume a queen criterion, which assumes equal weights of
45 neighbouring areas, which is appropriate for defining these. Islands were considered to have zero
46 neighbours. We adjusted for multiple testing using the Benjamini & Hochberg method ('p.adjust')
47 and used the 'spdep' package in R for the Local Moran's I and calculation of the spatial lag. This
48 approach of calculating the spatial lag was repeated at the middle super output area level (MSOA)
49 level (below).

50 **Hotspot and Coldspot definition**

51 Predicted prevalence hotspots at LAD levels were defined using Local Moran's I. The Moran's I
52 statistic gives a value indicating the spatial clustering of a variable relative to its neighbours. Where
53 there are significant (false discovery rate (FDR)adjusted $p < 0.05$) high positive local Moran's I in high
54 value neighbourhood (i.e. where the significant area also had a predicted prevalence greater than
55 the mean predicted prevalence and greater than the mean of the lagged variable, which effectively
56 represents how similar COVID-19 prevalence is to the areas that surround it) this implies the area
57 can be considered a 'hotspot'³. This method ensures we do not consider areas as hotspots where
58 they may have higher predicted prevalence to the surrounding areas but are lower than average for
59 the UK, although it might miss areas that are surrounded on all borders by other areas which would
60 be considered hotspots. A coldspot is assessed similarly using Local Moran's I, but where the area is
61 less than the mean and mean of the lagged variable.

62 **Sources of geographic data**

63 ***Index of Multiple Deprivation (IMD)***

64 The IMD was downloaded from the relevant government websites as below, and the most recent
65 IMD available at time of analysis was used:

- 66 • English (2019): [https://www.gov.uk/government/statistics/english-indices-of-deprivation-
67 2019](https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019)
- 68 • Scottish (2016): <https://www2.gov.scot/Topics/Statistics/SIMD>
- 69 • Welsh (2019): [https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-
70 Inclusion/Welsh-Index-of-Multiple-Deprivation/WIMD-2019](https://statswales.gov.wales/Catalogue/Community-Safety-and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation/WIMD-2019)

71 Because the IMD is calculated in each devolved administration using slightly different methodology,
72 and because of the different number of areas in each country, ranks are not directly comparable.

73 Therefore, we used within-country defined deciles. As the IMD is calculated for smaller area
74 geographies than MSOA, we calculated the average IMD per MSOA. This was then categorised into
75 quintiles where 1 is the least deprived and 5 is the most deprived.

76 ***Rural-urban gradient (RUC)***

77 The RUC was downloaded from the relevant government websites as below:

- 78 • England and Wales RUC (2011): <https://data.gov.uk/dataset/9c0e093d-d267-4eb8-90d8-54475ab4d1ff/rural-urban-classification-2011-of-middle-layer-super-output-areas-in-england-and-wales>
- 80 <https://data.gov.uk/dataset/9c0e093d-d267-4eb8-90d8-54475ab4d1ff/rural-urban-classification-2011-of-middle-layer-super-output-areas-in-england-and-wales>
- 81 • Scotland RUC (8 fold classification):
- 82 <https://www2.gov.scot/Topics/Statistics/About/Methodology/UrbanRuralClassification>

83 The resulting scale runs from 1 – 8, where 1 is the most urban and 8 is the least.

84 ***Nitrogen Oxide (NOx) data***

85 We used NOx pollution data from the Department of Environment, Food and Rural Affairs
86 (<https://uk-air.defra.gov.uk/data/>) for England, Scotland and Wales from 2018. Data is provided with
87 Ordinance Survey 1km² grid resolution which was used to calculate per MSOA air pollution by taking
88 the area-weighted average of the readings.

89 ***General Practitioners (GPs)/MSOA***

90 GPs addresses were used to derive the number of GPs from each MSOA, from the following data
91 sources:

- 92 • England & Wales: <https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data>
- 93 <https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data>
- 94 • Scotland: <https://www.opendata.nhs.scot/ne/dataset/general-practitioner-contact-details/resource/b092b69f-0838-408e-bb89-082562f0e1cd>
- 95 <https://www.opendata.nhs.scot/ne/dataset/general-practitioner-contact-details/resource/b092b69f-0838-408e-bb89-082562f0e1cd>

96 ***Average household number***

97 This figure was derived from data by dividing the number of houses with at least one usual occupant
98 with the total population for the same area.

99 Data sources for occupancy data were downloaded from the following sources:

- 100 • England & Wales (table PHP01 2011): <https://www.nrscotland.gov.uk/statistics-and->
101 [data/statistics/statistics-by-theme/households/household-estimates/small-area-statistics-](https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/households/household-estimates/small-area-statistics-on-households-and-dwellings)
102 [on-households-and-dwellings](https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/households/household-estimates/small-area-statistics-on-households-and-dwellings)
103 • Scotland: <https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by->
104 [theme/households/household-estimates/small-area-statistics-on-households-and-dwellings](https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/households/household-estimates/small-area-statistics-on-households-and-dwellings)

105 **MSOA-level mixed-effects models**

106 We employed multivariable mixed-effects models to understand the relationship of predicted
107 COVID-19 prevalence at MSOA level with deprivation. As a reminder, these models were ran at
108 MSOA-level rather than individual-level. This included the following variables:

109 The Index of Multiple Deprivation, our primary explanatory variable (IMD, categorised into quintiles
110 generated on the average IMD within each MSOA, where 1 is most deprived and 5 is least, and
111 considered as a continuous variable).

112 Other considered geosocial factors included a rural-urban gradient (RUC, considered as a continuous
113 variable where 1 is the most urban and 8 is the most rural), General practitioners per population in
114 MSOA (GPs/MSOA, where a higher number indicates more GPs per individual by MSOA), average
115 household number (calculated as number of inhabited dwellings/MSOA population, where a higher
116 number indicates a higher average number of individuals per household). Because it was on a very
117 different scale to the rest of the predictor variables, GPs/MSOA was scaled to have mean 0 and 1 SD
118 prior to model inclusion.

119 We additionally adjusted for the following variables derived from app response data, considered as
120 percentage of responders within the MSOA: those who reported having kidney, heart or lung
121 disease, and who are diabetic, a smoker or obese (calculated as BMI<30). We derived mean-adjusted
122 age and sex variables to partially adjust for response bias (i.e. the extent responders in an MSOA
123 represented the demographic of that MSOA). This was calculated as the difference of the expected
124 mean/ratio of age/sex in the MSOA (derived from ONS population data) and the observed
125 mean/ratio of age/sex amongst respondents.

126 We included a spatial lagged variable of the COVID-19 prevalence outcome. Inclusion of the lagged
127 variable is one method that accounts for spatial autocorrelation (SAC)⁴. It attempts to adjust for
128 spatial autocorrelation by capturing the variance explained by the influence of neighbouring regions
129 on the value of interest – in this case COVID-19 severity/prevalence. The lagged variable is calculated
130 at MSOA level by applying a spatial weights matrix (calculated in this instance under queen's

131 contiguity) to the outcome variable (in this case COVID-19 prevalence) and computing the lag using
132 the function `lag.listw` in the 'spdep' R package. This variable is then included as a covariate within
133 the model.

134 Data from eight time points were analysed , calculating the covariates (derived from app
135 responders) and spatial lag at each time point, a dummy variable adjusting for the different sample
136 times was included in the model as a random effect (allowing for a random intercept). MSOA was
137 also included to allow for a random intercept to account for the repeat observations over the eight
138 time periods, along with country as a fixed effect to account for difference in methodology in
139 creation of IMD and RUC.

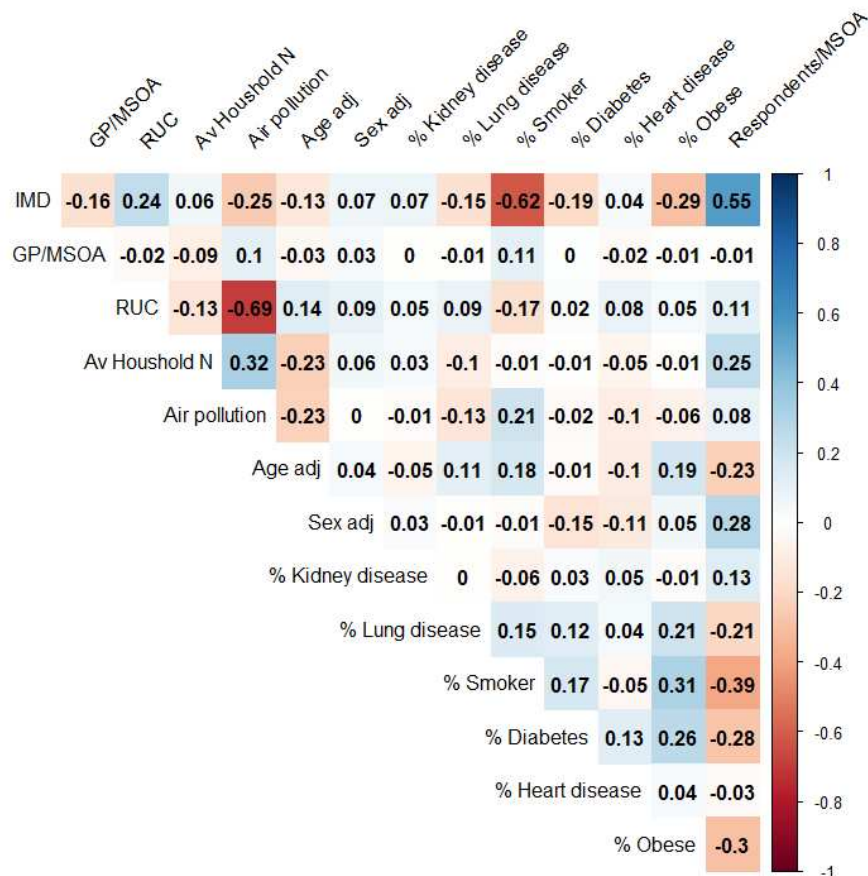
140 The users' distribution across GB is not uniform but all analyses took this into account by considering
141 only middle super output areas (MSOAs) with at least 20 individuals reporting on the app (n = 8097,
142 n removed = 387), and we included as a covariate the proportion of responders per MSOA at each
143 time point, in order to adjust for differences in responders by MSOA. Analysis was conducted in
144 RStudio v1.1.423 and R v3.6.3.

145 Variables were checked for multicollinearity before model inclusion using Spearman's correlation,
146 (see **Figure B**) with the *a priori* threshold of $> (+/-) 0.7$ indicating a variable should be removed.

147

7

148 **Figure B. Assessment of collinearity between the variables included in the MSOA-level mixed-**
 149 **effects models. Each cell of the matrix displays Spearman's correlation between two. The table is**
 150 **colour coded according to the Spearman's correlation, with blue denoting a positive correlation**
 151 **and red denoting a negative correlation. GP/MSOA= General Practitioners per middle super**
 152 **output area level; RUC= Rural-urban gradient; Av Household N= average household number.**



153

154 The model approach was therefore as follows:

- 155
- Model 1 (M1): Linear regression of the estimated COVID-19 prevalence and the IMD
 - 156 • Model 2 (M2): Linear mixed effects model (LMM) of estimated COVID-19 prevalence and the
 - 157 IMD, adjusted for country, and allowed a random effect of MSOA ID and time (assuming
 - 158 random intercept for both)
 - 159 • Model 3 (M3): Linear mixed effects model of estimated COVID-19 prevalence and the IMD,
 - 160 adjusted as above in M2, with additional adjustment for spatial autocorrelation (SAC) via
 - 161 inclusion of a spatial lag.

7

- 162 • Model 4 (M4): Linear mixed effects model as in M3, with the inclusion of geosocial
163 mediators and confounders and proportion of MSOA population who were app users.
- 164 • Model 5 (M5): Linear mixed effects model as in M4, with the inclusion of aggregated co-
165 morbidities as the % of respondents in MSOA with diabetes, kidney, lung or heart disease,
166 who are obese or are smokers.
- 167 • Model 6 (M6): Covariate + mean-adjusted LMM – Linear mixed effects model as in M6, with
168 the inclusion of mean-adjusted age and sex variables

169 **Supplementary References**

- 170 1. Menni C, Sudre CH, Steves CJ, et al. Quantifying additional COVID-19 symptoms will save lives.
171 *Lancet* 2020;395(10241):e107-e08. doi: 10.1016/s0140-6736(20)31281-2 [published Online
172 First: 2020/06/09]
- 173 2. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict
174 potential COVID-19. *Nat Med* 2020;26(7):1037-40. doi: 10.1038/s41591-020-0916-2
175 [published Online First: 2020/05/13]
- 176 3. Zhang C, Luo L, Xu W, et al. Use of local Moran's I and GIS to identify pollution hotspots of Pb in
177 urban soils of Galway, Ireland. *Sci Total Environ* 2008;398(1-3):212-21. doi:
178 10.1016/j.scitotenv.2008.03.011 [published Online First: 2008/04/29]
- 179 4. Diniz-Filho JA, Nabout JC, de Campos Telles MP, et al. A review of techniques for spatial modeling
180 in geographical, conservation and landscape genetics. *Genet Mol Biol* 2009;32(2):203-11.
181 doi: 10.1590/S1415-47572009000200001 [published Online First: 2009/04/01]