Original research

# Diet quality and risk and severity of COVID-19: a prospective cohort study

Jordi Merino ,[1,2,3] Amit D Joshi ,[4,5] Long H Nguyen ,[4,5,6] Emily R Leeming ,[7] Mohsen Mazidi,[7] David A Drew ,[4,5] Rachel Gibson,[8] Mark S Graham,[9] Chun-Han Lo ,[4,5] Joan Capdevila,[10] Benjamin Murray,[9] Christina Hu,[10] Somesh Selvachandran,[10] Alexander Hammers,[9,11] Shilpa N Bhupathiraju,[3,12] Shreela V Sharma,[13] Carole Sudre,[9] Christina M Astley,[2,14] Jorge E Chavarro,[12,15,16] Sohee Kwon,[4,5] Wenjie Ma,[4,5] Cristina Menni ,[7] Walter C Willett,[12,15,16] Sebastien Ourselin,[9] Claire J Steves,[7] Jonathan Wolf,[10] Paul W Franks,[12,17] Timothy D Spector,[8] Sarah Berry,[8] Andrew T Chan[4,5,18]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2021-325353).

For numbered affiliations see end of article.

**Correspondence to**
Dr Andrew T Chan, Harvard Medical School, Boston, USA; achan@mgh.harvard.edu

JM, ADJ and LHN are joint first authors.
TDS, SB and ATC are joint senior authors.

Received 7 June 2021
Accepted 19 August 2021
Published Online First 6 September 2021



**To cite:** Merino J, Joshi AD, Nguyen LH, *et al*. *Gut* 2021;**70**:2096–2104.

## ABSTRACT

**Objective** Poor metabolic health and unhealthy lifestyle factors have been associated with risk and severity of COVID-19, but data for diet are lacking. We aimed to investigate the association of diet quality with risk and severity of COVID-19 and its interaction with socioeconomic deprivation.

**Design** We used data from 592 571 participants of the smartphone-based COVID-19 Symptom Study. Diet information was collected for the prepandemic period using a short food frequency questionnaire, and diet quality was assessed using a healthful Plant-Based Diet Score, which emphasises healthy plant foods such as fruits or vegetables. Multivariable Cox models were fitted to calculate HRs and 95% CIs for COVID-19 risk and severity defined using a validated symptom-based algorithm or hospitalisation with oxygen support, respectively.

**Results** Over 3 886 274 person-months of follow-up, 31 815 COVID-19 cases were documented. Compared with individuals in the lowest quartile of the diet score, high diet quality was associated with lower risk of COVID-19 (HR 0.91; 95% CI 0.88 to 0.94) and severe COVID-19 (HR 0.59; 95% CI 0.47 to 0.74). The joint association of low diet quality and increased deprivation on COVID-19 risk was higher than the sum of the risk associated with each factor alone ($P_{interaction}$=0.005). The corresponding absolute excess rate per 10 000 person/months for lowest vs highest quartile of diet score was 22.5 (95% CI 18.8 to 26.3) among persons living in areas with low deprivation and 40.8 (95% CI 31.7 to 49.8) among persons living in areas with high deprivation.

**Conclusions** A diet characterised by healthy plant-based foods was associated with lower risk and severity of COVID-19. This association may be particularly evident among individuals living in areas with higher socioeconomic deprivation.

### SIGNIFICANCE OF THIS STUDY

#### WHAT IS ALREADY KNOWN ON THIS SUBJECT?

⇒ Poor metabolic health and unhealthy lifestyle behaviours have been associated with higher risk and severity of COVID-19.
⇒ Improved nutrition, especially in the context of socioeconomic deprivation, has been shown to reduce the burden of certain infectious diseases in the past. Evidence on the association of diet quality with susceptibility and progression of COVID-19 is lacking.

#### WHAT ARE THE NEW FINDINGS?

⇒ A dietary pattern characterised by healthy plant-based foods was associated with lower risk and severity of COVID-19.
⇒ We found evidence of a synergistic association of poor diet and increased socioeconomic deprivation with COVID-19 risk that was higher than the sum of the risk associated with each factor alone.
⇒ The beneficial association of diet with COVID-19 risk seems particularly relevant among individuals living in areas of higher socioeconomic deprivation.

#### HOW MIGHT IT IMPACT ON CLINICAL PRACTICE IN THE FORESEEABLE FUTURE?

⇒ Our study suggests that efforts to address disparities in COVID-19 risk and severity should consider specific attention to improve nutrition as a social determinants of health.

## INTRODUCTION

Poor metabolic health linked to conditions such as obesity, type 2 diabetes or hypertension[1 2] has been associated with increased risk and severity of COVID-19, and excess adiposity or preexisting liver disease might be causally associated with increased risk of death from COVID-19.[3 4] Underlying these conditions is the contribution of a diet, which may be independently associated with COVID-19 risk and severity.

On the basis of prior scientific evidence, Diet Quality Scores (DQSs) have been developed to evaluate the healthfulness of dietary patterns.[5–7] Dietary patterns capture the complexity of food intakes better than any one individual food item and offer the advantage of describing usual consumption of foods in typical diets.[8] One such diet score is the healthful Plant-Based Diet Index (hPDI), which emphasises intake of healthy plant foods, such as fruits, vegetables and whole grains, and has been associated with lower risk of fatty liver, type 2 diabetes, and coronary artery disease.[5 9 10]

Adherence to healthful dietary patterns may also be a proximal manifestation of distal social determinants of health.[11–13] Addressing adverse social determinants of health, such as poor nutrition, has been shown to reduce the burden of certain infectious diseases in the past,[14] supporting calls for prioritising social determinants of health in the public health response to COVID-19. A previous study including ~3000 healthcare workers from 6 countries showed that plant-based or pescatarian diets were associated with lower odds of moderate-to-severe COVID-19.[15] However, evidence on the association between diet quality and the risk and severity of COVID-19 in a general population is lacking, especially in the context of upstream social determinants of health. To address this evidence gap, we analysed data for 592 571 UK and USA participants from the smartphone-based COVID-19 Symptom Study,[16] to prospectively investigate the association of diet quality with risk and severity of COVID-19 and its intersection with socioeconomic deprivation.

## MATERIALS AND METHODS

### Study design and participants

The COVID-19 Symptom Study is a smartphone-based study conducted in the UK and USA. Study design and sampling procedures have been published elsewhere.[16] In brief, members of the general public were recruited through general media, social media outreach and direct invitations from investigators from the COronavirus Pandemic Epidemiology Consortium,[16] a multinational collaboration including several large clinical and epidemiological cohort studies. This analysis included participants recruited from 24 March 2020 and followed until 2 December 2020. Participants who reported any symptoms related to COVID-19 prior to start of follow-up, or reported symptoms that classified them as having predicted COVID-19 within 24 hours of first entry, or who tested positive for COVID-19 at any time prior to start of follow-up or 24 hours after first entry were excluded. We also excluded participants younger than 18 years old, pregnant, and participants who logged only one daily assessment during follow-up. At enrolment, we obtained informed consent to the use of volunteered information for research purposes and shared relevant privacy policies and terms of use agreements.

### Data collection procedures

Information on demographic factors was collected through standardised questionnaires at baseline,[16] including age, sex, race, zip code or postcode, healthcare worker status, personal medical history including lung disease, diabetes, cardiovascular disease, cancer, kidney disease, and use of medications, and self-reported of a COVID-19 positive test or any COVID-19-related symptoms. During a follow-up, daily prompts queried for updates on interim symptoms, healthcare visits and COVID-19 testing results. Through software updates, a survey to examine usual diet and lifestyle habits during a prepandemic time frame (reference time February 2020) and during the pandemic (reference time July/August 2020) was launched between August and September 2020. Details about the diet and lifestyle survey are available in online supplemental file and published elsewhere.[17] For this study, we used participant's recall of their diet during February 2020, reflecting the period before the pandemic. A graph illustrating how and when diet and symptoms information was collected is provided in online supplemental figure 1.

### Assessment of diet quality

Diet quality was assessed using information obtained from an amended version of the Leeds Short Form Food Frequency Questionnaire[18] that included 27 food items (online supplemental methods). The rationale to use this short-form food frequency questionnaire and not a full food frequency questionnaire was to limit participant burden by reducing time for completion. The accuracy and reliability of the short food frequency questionnaire has been assessed against a 217-item food frequency questionnaire and suggest that the short-form food frequency questionnaire is a reliable method of assessing diet quality.[18]

Participants were asked how often on average they had consumed one portion of each item in a typical week. The responses had eight frequency categories ranging from 'rarely or never' to 'five or more times per day'. Diet quality was quantified using the validated hPDI score.[5] To compute the hPDI, the 27 food items were combined into 14 food groups (online supplemental table 1). The original hPDI score included 18 food groups but nuts, vegetable oils, tea or coffee and animal fat were not specifically queried. Food groups were ranked into quintiles and given positive (healthy plant food groups) or reverse scores (less healthy plant and animal food groups). With positive scores, participants within the highest quintile of a food group received a score of 5, following on through to participants within the lowest quintile who received a score of 1. With reverse scores, this pattern of scoring was inverted. All component scores were summed to obtain a total score ranging from 14 (lowest diet quality) to 70 (highest) points. Criteria for generation of the hPDI are provided in online supplemental table 2.

As an additional method to quantify diet quality based on available diet information, we used the DQS.[18] The DQS is a score for adherence to UK dietary guidelines and was computed from five broad categories including fruits, vegetables, total fat, oily fish and non-milk extrinsic sugars. Each component was scored from 1 (unhealthiest) to 3 (healthiest) points, with intermediate values scored proportionally (online supplemental table 3). All component scores were summed to obtain a total score ranging from 5 (lowest diet quality) to 15 (highest) points (online supplemental table 4).

### Assessment of COVID-19 risk and severity

The primary outcome of this analysis was COVID-19 risk defined using a validated symptom-based algorithm,[19] which provides similar estimates of COVID-19 prevalence and incidence as those reported from the Office for National Statistics Community Infection Survey.[20] Details on the symptoms included in the predictive algorithm and corresponding weights are provided in online supplemental methods. In brief, the symptom-based approach uses an algorithm to predict whether a participant has been infected with SARS-CoV-2 on the basis of their reported symptoms, age and sex. To validate our case ascertainment, a subset of individuals who had reported symptoms in the COVID-19 Symptom Study application were invited to provide a copy of the test results. Among 235 participants, we found that self-reported COVID-19 testing yielded a positive predictive value of 88% and a negative predictive value of

94% for confirmed medical record results. The rationale for symptom-based classifier as a primary outcome was due to widespread difficulties obtaining testing during the early stages of the pandemic.[21] Secondary outcomes were confirmed COVID-19 based on a self-report of a reverse transcription PCR positive test and COVID-19 severity. COVID-19 severity was ascertained based on a report of the need for a hospital visit which required (1) non-invasive breathing support, (2) invasive breathing support and (3) administration of antibiotics combined with oxygen support (online supplemental methods).

### Statistical analysis

We elaborated a prespecified protocol including quality control procedures, definitions of exposures, outcomes and covariates, and statistical analysis plan prior to data analysis (online supplemental file). We summarised continuous measurements by using medians and percentiles 25th and 75th, and present categorical observations as frequency and percentages. The methods for classifying a priori selected covariates are provided in online supplemental file. Based on zip code (USA) or post code (UK) of residence, participants were assigned to country-specific community-level socioeconomic measures including socioeconomic deprivation and population density. For UK participants, we retrieved the education and income measures for the indices of multiple deprivation calculated by the Office of National Statistics from 2019 data aggregated to the 2011 Lower Super Output Areas.[22] For participants in the USA, socioeconomic measures were generated using aggregated census data for up to 25 characteristics that have been used consistently to approximate neighborhood-level environments.[23] Principal component analysis was used for census data reduction and seven variables were retained for the generation of the index. Loadings were obtained from the principal component analysis. The deprivation index was then standardised to have a mean of 0 and an SD of 1. Multiple imputations by chained equations with five imputations were used to impute missing values. Details about missingness prior to imputation are provided in online supplemental figure 2. All covariates in the primary analysis were included in the multiple imputation procedure, and estimates generated from each imputed dataset were combined using Rubin's rules.[24]

Because diet information was collected for the prepandemic period, we conducted prospective analyses using Cox regression, in which follow-up time for each participant started 24 hours after first log-in to the time of predicted COVID-19 (or to time of secondary outcomes) or date of last entry prior to 2 December 2020, whichever occurred first. We modelled the DQS as a continuous variable and generated categories of the score based on quartiles of the distribution (quartile 1, low diet quality; quartiles 2–3, intermediate diet quality; quartile 4, high diet quality). Cox regression models stratified by calendar date at study entry, country of origin and 10-year age group were used to calculate HR and 95% CI for COVID-19 risk and severity (age-adjusted model 1). Model 2 was further adjusted for sex, race/ethnicity, index of multiple deprivation, population density and healthcare worker status. Model 3 was further adjusted for presence of comorbidities (diabetes, cardiovascular disease, lung disease, cancer, kidney disease), body mass index (BMI), smoking status and physical activity. A directed acyclic graph depicting a possible scenario that could explain the association between diet quality and COVID-19 risk and severity is provided in online supplemental figure 3. We verified the proportional hazards assumption of the Cox model by using the Schoenfeld residuals technique.[25] Absolute risk was calculated as the percentage of COVID-19 cases occurring per 10 000 person-months in a given group. We used restricted cubic splines with four knots (at the 2.5th, 25th, 75th and 97.5th percentiles) to assess for non-linear associations between diet quality and COVID-19 risk.

In secondary analyses, we used a self-report of a positive test to define COVID-19 risk. For these analyses, we used inverse probability-weighted Cox models to account for predictors of obtaining country-specific testing. Inverse probability-weighted analyses included presence of COVID-19-related symptoms, interaction with a person with COVID-19, occupation as a healthcare worker, age group and race. Inverse probability-weighted Cox models were stratified by 10-year age group and date with additional adjustment for the covariates used in previous models. For severe COVID-19 analyses, we adjusted for the same covariates used in previous models. As an additional method to quantify diet quality we used the DQS and tested for associations between diet quality and COVID-19 risk and severity. In addition, we censored our analyses to cases that occurred after completing the diet survey to investigate potential bias due to time-varying confounding.

In subgroup analyses, we assessed the association between diet quality and COVID-19 risk according to comorbidities, demographic and lifestyle characteristics. We also classified participants according to categories of the DQS and socioeconomic deprivation (nine categories based on thirds of DQS and deprivation index) and conducted joint analyses for COVID-19 risk. The joint analysis was conducted to quantitatively estimate the combined association of diet and deprivation simultaneously with risk of COVID-19. We tested for additive interactions by assessing the relative excess risk due to interaction (RERI), and further examined the COVID-19 risk proportions attributable to diet, deprivation and to their interaction (online supplemental methods).[26]

We conducted sensitivity analyses to account for regional differences in the effective reproductive number ($R_t$) or other risk mitigating behaviours such as mask wearing. The $R_t$ parameter denotes the number of additional persons infected, on average, by a single case and has been used as a measure of how quickly the virus is spreading and serves as a virulence proxy. For $R_t$ analyses, we extracted US state-level information on $R_t$ from the COVID-19 Tracking Project (https://covidtracking.com), for the period between March 2020 and January 2021. For the UK, we calculated $R_t$ time-series for Scotland, Wales, and each of the National Health Service (NHS) regions in England, using a previously published methodology from our group.[20] For these analyses, we defined community peak and nadir $R_t$ time-windows as the period between 1 week before and 2 weeks after $R_t$ was all-time high or low. Using censored time-windows, we tested the association between diet quality and COVID-19 risk after adjusting for the same confounders as included in model 3. For mask wearing analyses, we used survey data launched between June 2020 and until September 2020 on whether participants had worn a face mask when outside the house in the last week. Responses were categorised into two categories: participants who wore masks 'none of the time or sometimes', and those who reported wearing masks 'most of time/always' at least once. For mask wearing analyses, we included the same covariates as included in model 3.

Further, structural equation models were implemented to conduct a mediation analysis of BMI. For this analysis, diet quality and BMI were used as continuous variables. We estimated the relative contribution of BMI to the association between diet quality and COVID-19 risk and computed the proportion of total effect that was explained by indirect effects of BMI. Indirect

effects were estimated by taking the product of the effect of the exposure on the mediator and the effect of the mediator on the outcome. To calculate the proportion of the mediated effect, we divided the indirect effect by the total effect. The direct effect, which is defined as the association of diet quality on COVID-19 risk through mechanisms independent of mediation, was estimated from regressing COVID-19 on diet quality.

Two-sided values of $p<0.05$ were considered statistically significant for main analyses. All statistical analyses were performed using R software, V.4.0.3 (R Foundation).

### Patient and public involvement

No patients were directly involved in designing the research question or in conducting the research. No patients were asked for advice on interpretation or writing up the results. Results of the study will be shared with the public and patients directly via the ZOE symptom app, blog posts hosted on the https://covid.joinzoe.com/ website and webinars.

## RESULTS

Self-reported diet quality was evaluated in 647 137 survey responders, of which 54 566 were excluded due to prevalent COVID-19 (n=1555), presence of any symptoms at baseline (n=47 594), logged only once (n=1201), pregnancy (n=1129) or age under 18 years (n=3087; online supplemental figure 4). Baseline characteristics of the 592 571 participants included in this study according to categories of the hPDI score are shown in table 1. Participants in the highest quartile of the diet score (reflecting a healthier diet) were more likely than participants in the lowest quartile to be older, female, healthcare workers, of lower BMI, engage in physical activities ≥5 days/week and less likely to reside in areas with higher socioeconomic deprivation. Characteristics of participants from the COVID-19 Symptom Study according to diet survey participation are presented in online supplemental table 5. The hPDI score was normally distributed (online supplemental figure 5).

Over 3 886 274 person-months of follow-up, 31 815 COVID-19 cases were documented. Crude COVID-19 rates per 10 000 person-months were 72.0 (95% CI 70.4 to 73.7) for participants in the highest quartile of the diet score and 104.1 (95% CI 101.9 to 106.2) for those in the lowest quartile. The corresponding age-adjusted HR for COVID-19 risk was 0.80 (95% CI 0.78 to 0.83, table 2). Differences in the risk of COVID-19 persisted after adjustment for potential confounders. In fully adjusted models, the multivariable-adjusted HR for COVID-19 risk was 0.91 (95% CI 0.88 to 0.94) when we compared participants with high diet quality to those with low diet quality. We observed non-linear decreasing trends in the risk of COVID-19 with higher diet quality ($p<0.001$ for non-linearity), in which COVID-19 risk plateau among individuals with a DQS >50 (online supplemental figure 6). The association between diet quality and COVID-19 risk was consistent but attenuated in secondary analyses using the DQS score (HR 0.92; 95% CI 0.89 to 0.95; online supplemental table 6), and became non-significant in fully adjusted models (HR 1.00; 95% CI 0.97 to 1.03). We also investigated whether our primary findings were consistent in an analysis censored to cases that occurred after the completion of the diet survey. These analyses showed that high diet quality, compared with low diet quality, was associated with lower COVID-19 risk (multivariable-adjusted HR 0.88; 95% CI 0.83 to 0.93; online supplemental table 7).

In secondary analyses for COVID-19 risk based on a positive test, we showed that crude COVID-19 incidence rates per 10 000 person-months were 12.9 (95% CI 12.2 to 13.6) for individuals with high diet quality and 16.4 (95% CI 15.5 to 17.2) for individuals with low diet quality. The corresponding multivariable-adjusted HR for risk of COVID-19 was 0.82 (95% CI 0.78 to 0.86; table 2). For risk of severe COVID-19, crude incidence rates were lower for individuals reporting high diet quality compared with those with low diet quality (1.6 (95% CI 1.3 to 1.8) vs 2.1 (95% CI 1.9 to 2.5; per 10 000 person-months) table 2). In the fully adjusted model, high diet quality, as compared with low diet quality, was associated lower risk of severe COVID-19 with an an HR of 0.59 (95% CI 0.47 to 0.74; table 2).

In stratified analyses, the inverse association between diet quality and COVID-19 risk was more evident in participants living in areas of high socioeconomic deprivation and those reporting low physical activity levels ($p<0.05$; table 3). We found no significant effect modification for other characteristics such age, BMI, race/ethnicity or population density. When diet quality and socioeconomic deprivation were combined, there was a risk gradient with low diet quality and high socioeconomic deprivation. Compared with individuals living in areas with low socioeconomic deprivation and high diet quality, the multivariable-adjusted HR for risk of COVID-19 for low diet quality was 1.08 (95% CI 1.03 to 1.14) among those living in areas with low socioeconomic deprivation, 1.23 (95% CI 1.17 to 1.29) for those living in areas with intermediate socioeconomic deprivation, and 1.47 (95% CI 1.38 to 1.52) for those living in areas with high socioeconomic deprivation (figure 1). The joint association of diet quality and socioeconomic deprivation was higher than the sum of the risk associated with each factor alone (RERI=0.05 (95% CI 0.02 to 0.08); $P_{interaction}=0.005$; online supplemental table 8). The proportion of contribution to excess COVID-19 risk was estimated to be 31.9% (95% CI 18.2% to 45.6%) to diet quality, 38.4% (95% CI 26.5% to 50.3%) to socioeconomic deprivation, and 29.7% (95% CI 2.1% to 57.3%) to their interaction. The absolute excess rate of COVID-19 per 10 000 person-months for lowest versus highest quartile of the diet score was 22.5 (95% CI 18.8 to 26.3) among persons living in areas with low socioeconomic deprivation and 40.8 (95% CI 31.7 to 49.8) among individuals living in areas with high deprivation (online supplemental figure 7).

We conducted a series of sensitivity analyses to further account for variation in $R_t$, mask wearing. For peak $R_t$ censored analyses, crude COVID-19 rates per 10 000 person-months were 148.1 (95% CI 139.9 to 156.8) among participants with low diet quality and 92.9 (95% CI 86.6 to 99.5) for participants with high diet quality. The corresponding multivariable-adjusted HR was 0.84 (95% CI 0.76 to 0.92, figure 2). The same trend was observed for nadir $R_t$ censored analyses, in which crude COVID-19 rates per 10 000 person-months were 67.1 (95% CI 61.7 to 73.0) among participants with low diet quality and 45.8 (95% CI 41.3 to 50.5) for participants with high diet quality (multivariable-adjusted HR 0.89; 95% CI 0.80 to 1.00, figure 2). We further adjusted our models for mask wearing. This analysis showed that high diet quality, as compared with low diet quality, was associated with lower risk of COVID-19 with an adjusted HR of 0.88 (95% CI 0.83 to 0.94; online supplemental table 9).

In a mediation analysis of BMI, we observed that BMI mediated 37% (95% CI 30% to 44%; $p<0.001$) of the effect of diet quality on COVID-19 risk, and that there was also evidence of a direct effect of diet and COVID-19 risk (HR 0.98: 95% CI 0.97 to 0.98; per 1 SD increase in hPDI; online supplemental table 10).

**Table 1** Baseline characteristics of study participants according to categories of the diet quality score

| | All participants (n=592 571) | Low hPDI (Q1; n=148 143) | Intermediate hPDI (Q2-Q3; n=296 286) | High hPDI (Q4; n=148 142) |
|---|---|---|---|---|
| hPDI score, median ($P_{25}$–$P_{75}$) | 50 (47–54) | 45 (43–47) | 51 (49–52) | 56 (55–58) |
| Demographic characteristics | | | | |
| Age, years | 56 (44–65) | 52 (41–62) | 57 (45–66) | 57 (45–65) |
| ≥18–24 | 14 397 (2.4) | 5146 (3.4) | 5846 (2.0) | 3405 (2.3) |
| 25–34 | 52 922 (8.9) | 16 535 (11.2) | 23 150 (7.8) | 13 237 (8.9) |
| 35–44 | 86 251 (14.6) | 26 907 (18.2) | 40 145 (13.5) | 19 199 (13.0) |
| 45–54 | 125 802 (21.2) | 34 890 (23.6) | 62 491 (21.1) | 28 421 (19.2) |
| 55–64 | 158 637 (26.8) | 34 279 (23.1) | 81 837 (27.6) | 42 521 (28.7) |
| ≥65 | 153 810 (26.0) | 30 215 (20.4) | 82 413 (27.8) | 41 182 (27.8) |
| Missing | 752 (0.1) | 171 (0.1) | 404 (0.1) | 177 (0.1) |
| Sex, no (%) | | | | |
| Male | 187 450 (31.6) | 58 199 (39.3) | 93 162 (31.4) | 36 089 (24.4) |
| Female | 404 126 (68.2) | 89 706 (60.5) | 202 605 (68.4) | 111 815 (75.5) |
| Prefer not to say | 995 (0.2) | 238 (0.2) | 519 (0.2) | 238 (0.2) |
| Race*, no (%) | | | | |
| White | 568 770 (96.0) | 141 365 (95.4) | 284 804 (96.1) | 142 601 (96.3) |
| Black | 4328 (0.7) | 1466 (1.0) | 2053 (0.7) | 809 (0.5) |
| Asian | 10 435 (1.8) | 2954 (1.9) | 5043 (1.7) | 2438 (1.6) |
| Other | 7228 (1.2) | 1925 (1.3) | 3463 (1.2) | 1840 (1.2) |
| Missing | 1810 (0.3) | 433 (0.3) | 923 (0.3) | 454 (0.3) |
| Country, no (%) | | | | |
| UK | 543 984 (91.8) | 135 360 (91.4) | 272 494 (92.0) | 136 130 (91.9) |
| USA | 48 587 (8.2) | 12 783 (8.6) | 23 792 (8.0) | 12 012 (8.1) |
| Index of deprivation, no (%)† | | | | |
| Most deprived, decile 1 | 13 416 (2.3) | 4696 (3.1) | 6163 (2.1) | 2557 (1.7) |
| Least deprived, decile 10 | 103 608 (17.5) | 23 122 (15.6) | 53 652 (18.1) | 26 834 (18.1) |
| Missing | 40 759 (6.9) | 10 489 (7.1) | 20 249 (6.8) | 10 021 (6.8) |
| Population density, $km^2$, no (%)† | | | | |
| <500 | 119 782 (20.2) | 28 139 (19.0) | 61 230 (20.7) | 30 413 (20.5) |
| 500–1999 | 90 541 (15.3) | 23 631 (16.0) | 45 902 (15.5) | 21 008 (14.2) |
| 2000–4999 | 94 345 (15.9) | 24 813 (16.7) | 47 233 (15.9) | 22 299 (15.1) |
| ≥5000 | 244 295 (41.2) | 60 156 (40.6) | 120 319 (40.6) | 63 820 (43.1) |
| Missing | 43 608 (7.4) | 11 404 (7.7) | 21 602 (7.3) | 10 602 (7.2) |
| Healthcare worker, yes, no (%) | 41 141 (6.9) | 10 633 (2.3) | 20 183 (6.8) | 10 325 (7.0) |
| Lifestyle characteristics | | | | |
| Smoking status, no (%) | | | | |
| Never | 475 347 (81.9) | 113 165 (79.0) | 238 192 (81.9) | 123 990 (84.6) |
| Former | 87 901 (15.1) | 22 683 (15.8) | 44 771 (15.4) | 20 447 (13.9) |
| Current | 17 401 (3.0) | 7402 (5.2) | 7837 (2.6) | 2162 (1.5) |
| Physical activity | | | | |
| <1 day/week | 106 294 (17.9) | 37 258 (25.2) | 50 713 (17.1) | 18 323 (12.4) |
| 1–2 days/week | 224 606 (37.9) | 59 325 (40.0) | 113 749 (38.4) | 51 532 (34.8) |
| 3–4 days/week | 143 548 (24.2) | 30 009 (20.3) | 73 601 (24.8) | 39 938 (27.0) |
| ≥5 days/week | 117 007 (19.7) | 21 164 (14.3) | 57 701 (19.5) | 38 142 (25.7) |
| Missing | 1116 (0.2) | 387 (0.3) | 522 (0.2) | 207 (0.1) |
| Body mass index, $kg/m^2$ | 25.1 (22.6–28.7) | 26.6 (23.6–30.7) | 25.2 (22.7–28.5) | 24.0 (21.8–26.9) |
| <18.5 | 12 004 (2.0) | 2680 (1.8) | 5540 (1.9) | 3784 (2.6) |
| 18.5–24.9 | 277 536 (46.8) | 52 109 (35.2) | 138 503 (46.7) | 86 924 (58.7) |
| 25–29.9 | 189 197 (31.9) | 51 517 (34.8) | 97 919 (33.0) | 39 761 (26.8) |
| ≥30 | 113 056 (19.1) | 41 655 (28.1) | 53 909 (18.2) | 17 492 (11.8) |
| Missing | 778 (0.1) | 182 (0.1) | 415 (0.1) | 181 (0.1) |
| Mask wearing, no (%)‡ | | | | |
| Most of the time/always | 437 782 (73.9) | 113 202 (76.4) | 218 402 (73.7) | 106 178 (71.6) |
| Never/sometimes | 152 551 (25.7) | 34 240 (23.1) | 76 809 (25.9) | 41 502 (28.0) |
| Missing | 2238 (0.4) | 701 (0.5) | 1075 (0.4) | 462 (0.3) |
| Clinical history, yes, no (%) | | | | |

Continued

**Table 1** Continued

| | All participants (n=592 571) | Low hPDI (Q1; n=148 143) | Intermediate hPDI (Q2-Q3; n=296 286) | High hPDI (Q4; n=148 142) |
|---|---|---|---|---|
| Diabetes | 20 058 (3.4) | 6079 (4.1) | 10 158 (3.4) | 3821 (2.6) |
| Heart disease | 20 376 (3.4) | 5200 (3.5) | 10 660 (3.6) | 4516 (3.0) |
| Cancer | 6559 (1.9) | 1643 (1.8) | 3348 (1.9) | 1568 (1.8) |
| Lung disease | 62 999 (10.6) | 17 534 (11.8) | 31 227 (10.5) | 14 238 (9.6) |
| Kidney disease | 5134 (0.9) | 1492 (1.0) | 2594 (0.9) | 1048 (0.7) |

Values are median ($P_{25}$–$P_{75}$) for continuous variables; numbers and (percentages) for categorical variables.
hPDI ranges from 0 to 70, with higher scores indicating higher adherence to a healthy plant-based diet.
*Race was self-reported by the participants.
†Index of deprivation and population density were generated using zipcode or postcode information linked with census track data. Country-specific deprivation indices were generated (online supplemental file).
‡Mask wearing information was collapsed into two categories: participants who wore masks 'none of the time or sometimes', and those who reported wearing masks 'most of time/always'.
hPDI, healthful Plant-Based Diet Index.

## DISCUSSION

In this large survey among UK and US participants prospectively assessing risk and severity of COVID-19 infection, we found that a dietary patterns characterised by healthy plant foods was associated with lower risk and severity of COVID-19. We observed a risk gradient of poor diet quality and increased socioeconomic deprivation that departed from the additivity of the risks attributable to each factor separately, suggesting that the beneficial association of diet with COVID-19 may be particularly evident among individuals with higher socioeconomic deprivation.

Our findings are aligned with preliminary evidence showing that improving nutrition could help reduce the burden of infectious diseases.[12 14 27] Previous studies have shown that the administration of arachidonic or linoleic acid partially suppresses SARS-CoV-1 and coronavirus 229E viral replication,[28] and that specific nutrients or dietary supplements associate with modest

**Table 2** Adjusted HRs of COVID-19 risk and severity according to hPDI scores

| | Low hPDI (Q1; n=148 143) | Intermediate hPDI (Q2–Q3; n=296 286) | High hPDI (Q4; n=148 142) | P for trend |
|---|---|---|---|---|
| hPDI score, median ($P_{25}$–$P_{75}$) | 45 (43–47) | 51 (49–52) | 56 (55–58) | |
| COVID-19 risk | | | | |
| No of events/person-months | 8739/839 747 | 15 733/2 026 824 | 7359/1 022 078 | — |
| Incidence rate (10 000 person-months; 95% CI) | 104.1 (101.9 to 106.2) | 77.6 (76.4 to 78.8) | 72.0 (70.4 to 73.7) | — |
| Age-adjusted model | 1.00 (Ref) | 0.85 (0.82–0.87) | 0.80 (0.78–0.83) | <0.001 |
| Multivariable model 2 | 1.00 (Ref) | 0.85 (0.83–0.87) | 0.81 (0.78–0.83) | <0.001 |
| Multivariable model 3 | 1.00 (Ref) | 0.91 (0.89–0.93) | 0.91 (0.88–0.94) | <0.001 |
| COVID-19 risk (positive test) | | | | |
| No of events/person-months | 1423/869 664 | 2829/2 081 970 | 1350/1 046 887 | — |
| Incidence rate (10 000 person-months; 95% CI) | 16.4 (15.5 to 17.2) | 13.6 (13.1 to 14.1) | 12.9 (12.2 to 13.6) | — |
| Age-adjusted model* | 1.00 (Ref) | 0.86 (0.83–0.90) | 0.79 (0.75–0.83) | <0.001 |
| Multivariable model 2* | 1.00 (Ref) | 0.87 (0.84–0.91) | 0.80 (0.76–0.84) | <0.001 |
| Multivariable model 3* | 1.00 (Ref) | 0.88 (0.85–0.92) | 0.82 (0.78–0.86) | <0.001 |
| Severe COVID-19 | | | | |
| No of events/person-months | 187/871 995 | 390/2 086 790 | 163/1 049 476 | — |
| Incidence rate (10 000 person-months; 95% CI) | 2.1 (1.9 to 2.5) | 1.9 (1.7 to 2.1) | 1.6 (1.3 to 1.8) | — |
| Age-adjusted model | 1.00 (Ref) | 0.66 (0.56–0.77) | 0.45 (0.36–0.57) | <0.001 |
| Multivariable model 2 | 1.00 (Ref) | 0.66 (0.57–0.78) | 0.45 (0.36–0.57) | <0.001 |
| Multivariable model 3 | 1.00 (Ref) | 0.77 (0.66–0.91) | 0.59 (0.47–0.74) | <0.001 |

HRs and 95% CI for COVID-19 risk and severity. COVID-19 risk defined using a validated symptom-based model. COVID-19 or an RT-PCR positive test report. COVID-19 severity was defined based on hospitalisation with requirement of oxygen support (methods, online supplemental file).
Cox proportional hazards models were stratified by calendar date at study entry, country of origin and 10-year age group (age-adjusted model).
Multivariable model 2 was further adjusted for sex (male, female), race/ethnicity (white, black, Asian, other), index of multiple deprivation (most deprived <3, intermediate deprived 3–7, less deprived >7), population density (<500 individuals/$km^2$, 500–1999 individuals/$km^2$, 2000–4999 individuals/$km^2$ and ≥5000 individuals/$km^2$) and healthcare worker status (yes with interaction with patients with COVID-19, yes without interaction with patients with COVID-19, no).
Model 3 was further adjusted for presence of comorbidities (diabetes (yes, no), cardiovascular disease (yes, no), lung disease (yes, no), cancer (yes, no), kidney disease (yes, no)), body mass index (<18.5 kg/$m^2$, 18.5–24.9 kg/m2, 25.0–29.9 kg/$m^2$ and ≥30 kg/$m^2$), smoking status (yes, no) and physical activity (<1 day/week, 1–2 days/week, 3–4 days/week, ≥5 days/week).
*Inverse probability-weighted analyses were conducted to account for predictors of obtaining RT-PCR testing (presence of COVID-19-related symptoms, interaction with a COVID-19 case, healthcare worker, age group and race). Inverse probability-weighted Cox proportional hazards models were stratified by 10-year age group and date with additional adjustment for the covariates used in previous models.
hPDI, healthful Plant-Based Diet Index; RTPCR, reverse transcription PCR.

**Table 3** Adjusted HRs of COVID-19 risk according to healthful Plant-Based Dietary Index scores stratified by sociodemographic and clinical characteristics

| Factor | No of events/person-months* | HR per 1 SD increase in diet quality score | P value |
|---|---|---|---|
| Age | | | |
| <60 | 25 329/2 285 329 | 0.94 (0.93–0.95) | |
| ≥60 | 6486/1 600 945 | 0.94 (0.92–0.97) | 0.63 |
| Sex | | | |
| Male | 9338/1 232 656 | 0.95 (0.93–0.97) | |
| Female | 22 428/2 647 254 | 0.96 (0.95–0.98) | 0.21 |
| Race | | | |
| White | 30 335/3 736 972 | 0.96 (0.95–0.97) | |
| Non-white | 1480/149 303 | 0.96 (0.91–1.01) | 0.95 |
| Socioeconomic deprivation† | | | |
| High | 5244/456 271 | 0.94 (0.91–0.96) | |
| Intermediate | 13 172/1 567 516 | 0.96 (0.94–0.98) | |
| Low | 13 399/1 862 489 | 0.97 (0.95–0.99) | 0.04 |
| Population density, km$^2$ no | | | |
| <2000 | 10 581/1 490 084 | 0.96 (0.94–0.98) | |
| ≥2000 | 21 234/2 396 190 | 0.96 (0.94–0.97) | 0.74 |
| Healthcare worker | | | |
| Yes | 2908/140 087 | 0.95 (0.92–0.99) | |
| No | 28 907/3 638 588 | 0.96 (0.95–0.97) | 0.78 |
| Body mass index, kg/m$^2$ | | | |
| <25 | 13 989/1 905 517 | 0.96 (0.94–0.97) | |
| 25–30 | 9854/1 252 222 | 0.96 (0.94–0.98) | |
| ≥30 | 7972/728 536 | 0.96 (0.94–0.98) | 0.73 |
| Physical activity | | | |
| <1 day/week | 6751/683 562 | 0.94 (0.91–0.96) | |
| 1–4 days/week | 19 476/2 425 198 | 0.96 (0.94–0.97) | |
| ≥5 dasy/week | 5588/777 515 | 0.99 (0.96–1.01) | 0.01 |

Association between predicted COVID-19 and diet quality according to sociodemographic and clinical characteristics.
*Number of observations varies among imputations.
†Socioeconomic deprivation categories were based on deciles of the deprivation index (methods). Cox models were adjusted for the same covariates as previous model 3. P values obtained using the Q test for heterogeneity.

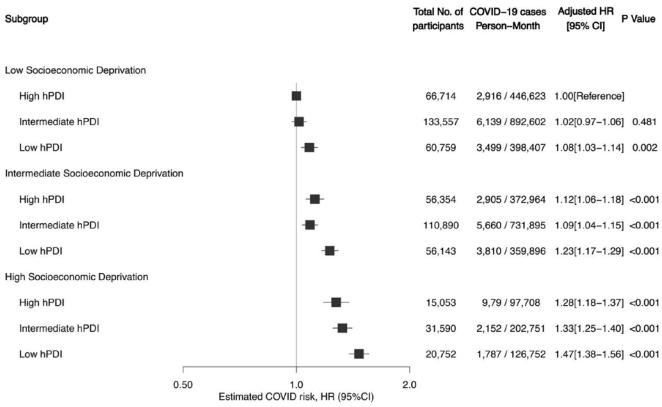


**Figure 1** Risk of COVID-19 according to diet quality and socioeconomic deprivation. Shown are adjusted HRs and 95% CI of the estimate for predicted COVID-19 according to categories of diet quality and socioeconomic deprivation. Cox model stratified by calendar date at study entry, country of origin and 10-year age group, and adjusted for sex, race/ethnicity, index of multiple deprivation, population density, presence of diabetes, cardiovascular disease, lung disease, cancer, kidney disease, healthcare worker status, body mass index, smoking status and physical activity. In these comparisons, participants with high-quality diet and low socioeconomic deprivation served as the reference group. hPDI, healthful Plant-Based Diet Index.

reductions in COVID-19 risk.[29] Trace elements, vitamins (A, $B_6$, $B_{12}$, C, D, and E, and folate), amino acids, long-chain omega-3 fatty acids (docosahexaenoic and eicosapentaenoic) and non-nutrient-bioactive such as polyphenols have key roles in immune system function and cytokine release,[30] and might partially explain some of the observed associations. Results from this observational study could expand previous single nutrient observations and highlight the beneficial association of healthy dietary patterns, which was most pronounced for risk of severe COVID-19. Our findings also concur with a comparative risk assessment study suggesting that a 10% reduction in the prevalence of diet-related conditions such as obesity and type 2 diabetes would have prevented ~11% of the COVID-19 hospitalisations that have occurred among US adults since November 2020.[31]

The association of healthy diet with lower COVID-19 risk appears particularly evident among individuals living in areas of higher socioeconomic deprivation. Our models estimate that nearly a third of COVID-19 cases would have been prevented if one of two exposures (diet and deprivation) were not present. While the population attributable risk is a population-specific calculation that is dependent on the prevalence of the exposure and its association with disease risk with an assumed causal effect, we acknowledge that our estimation of population attributable risk has several limitations. First, as with all the observational research, our estimates did not necessarily indicate causal effects. Second, estimated attributable risks are likely to change over time with the prevailing SARS-CoV-2 infection rate. However, our observations are consistent with data from ecological studies showing that people living in regions with greater social inequalities are likely to have

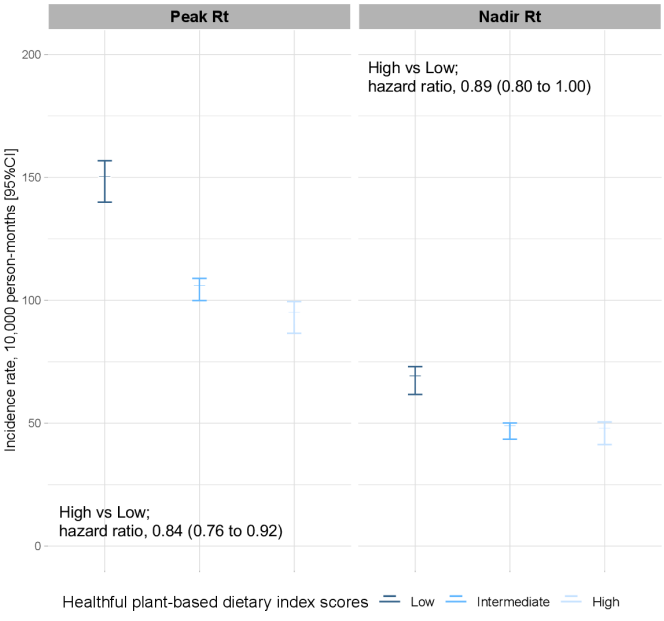


**Figure 2** Risk of COVID-19 according to community transmission rate and diet quality. COVID-19 incidence rate per 10 000 person-month and 95% CI of the estimate based on different community transmission rate and diet quality categories. Peak $R_t$ and nadir $R_t$ were defined using (methods). Adjusted HRs and 95% CI of the estimate for risk of COVID-19 were obtained from fully adjusted Cox models.

higher rates of COVID-19 incidence and deaths.[32] By generating a granular deprivation index based on zip code information our study adds to previous country-level ecological studies. In addition, recent studies on the impact of socioeconomic status on COVID-19 have shown that community-level deprivation indices are strongly associated with COVID-19 risk and mortality.[33 34] However, it is still possible that differences in deprivation exists within communities. Further studies including information about household characteristics, built environment, or access to healthy foods are needed to expand these initial associations.

Our study adds to knowledge by investigating the association between diet quality and risk and severity of COVID-19 in a general population and in the context of social determinants of health. While our study supports the beneficial association of diet quality with COVID-19 risk and severity, particularly among individuals with higher deprivation, we cannot completely rule out the potential for residual confounding. Individuals who eat healthier diets are likely to share other features that might be associated with lower risk of infection such as the adoption of other risk mitigation behaviours, better household conditions and hygiene, or access to care. However, it is reassuring that our findings were consistent despite controlling for additional surrogate markers of SARS-CoV-2 infection such as mask wearing or community transmission rate, two of the most relevant factors associated with virus transmission and COVID-19 risk.[35] These findings suggest that efforts to address disparities in COVID-19 risk and severity should consider specific attention to access to healthy foods as a social determinants of health.

We acknowledge several limitations. First, as an observational study, we are unable to confirm a direct causal association between diet and COVID-risk or infer specific mechanisms. Second, our study population was not a random sampling of the population. Although we tried to minimise potential selection bias, we recognise our participants are mainly white and less likely to live in deprived areas than the general population. Thus, the generalisability of our findings need to be confirmed in additional studies. Third, our results could be biased due to the time lapse between the dietary recalls, administered a few months after the relevant period of exposure (prepandemic). However, our sensitivity analyses in which we censored cases that had occurred before the administration of the diet survey showed consistent results. Fourth, the self-reported nature of the diet questionnaire is prone to measurement error and bias, and the use of a short food frequency survey could have further reduced the resolution of dietary data collected. More accurate dietary intake assessment methods such as the use of dietary intake biomarkers would be valuable in future studies,[36] but also difficult to implement in large-scale and time-sensitive investigations. Fifth, outcomes rely on self-reported data. While it is possible that misclassification exists, there is an excellent agreement between self-report and test reports with 88% sensitivity and 94% specificity. In addition, the symptom-based algorithm provides similar estimates of COVID-19 prevalence and incidence as those reported from the Office for National Statistics Community Infection Survey. Sixth, we defined risk of severe COVID-19 according to reports of hospitalisation with oxygen support, which may not have captured more severe or fatal cases. We acknowledge that we were unable to include participants who may have died of COVID-19 before the administration of the diet questionnaire.

## CONCLUSIONS

In conclusion, our data provide evidence that a healthy diet was associated with lower risk of COVID-19 and severe COVID-19 even after accounting for other healthy behaviours, social determinants of health and virus transmission measures. The joint association of diet quality with socioeconomic deprivation was greater than the addition of the risks associated with each individual factor, suggesting that diet quality may play a direct influence in COVID-19 susceptibility and progression. Our findings suggest that public health interventions to improve nutrition and poor metabolic health and address social determinants of health may be important for reducing the burden of the pandemic.

**Author affiliations**
[1]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[2]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA
[3]Department of Medicine, Harvard Medical School, Boston, MA, USA
[4]Clinical and Translational Epidemiological Unit, Massachusetts General Hospital, Boston, MA, USA
[5]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[6]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[7]Department of Twin Research, King's College London, London, UK
[8]Department of Nutritional Sciences, King's College London, London, UK
[9]School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK
[10]Zoe Limited, London, UK
[11]King's College London & Guy's and St Thomas' PET Centre, King's College London, London, UK
[12]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[13]Department of Epidemiology, Human Genetics, and Environmental Sciences, UT Health School of Public Health, Houston, Texas, USA
[14]Division of Endocrinology & Computational Epidemiology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
[15]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[16]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[17]Department of Clinical Sciences, Genetic and Molecular Epidemiology Unit, Lund University, Lund, Sweden
[18]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, USA

**Twitter** Jordi Merino @Riudecanyenc, Emily R Leeming @EmilyLeemingRD and David A Drew @DADrewPhD

**Acknowledgements** We express our sincere thanks to all of the participants who entered data into the app, including study volunteers enrolled in cohorts within the Coronavirus Pandemic Epidemiology (COPE) consortium. We thank the staff of Zoe Global, the Department of Twin Research at King's College London and the Clinical and Translational Epidemiology Unit at Massachusetts General Hospital for tireless work in contributing to the running of the study and data collection. This work was conducted using the Short Form FFQ tool developed by Cleghorn as reported in https://doi.org/10.1017/S1368980016001099 and listed in the Nutritools (www.nutritools.org) library.

**Contributors** JM, ADJ, LHN, ERL, TDS, SB and AC conceived the study design. JM, ADJ, ERL, MSG, JC, BM and SS contributed to the statistical analysis. All authors were involved in acquisition, analysis or interpretation of data. JM, LHN and DAD wrote the first draft of the manuscript. DAD, WW, SO, CJS, JW, PWF, TDS, SB and AC obtained funding. JM, ADJ and LHN provided administrative, technical or material support. TDS, SB and AC jointly supervised this work. All authors contributed to the critical revision of the manuscript for important intellectual content and approved the final version of the manuscript. The corresponding authors attest that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. TDS, SB and AC are joint last authors.

**Funding** National Institutes of Health (K01DK110267, K01DK120742, K23DK120899, K23DK125838, P30DK046200, P30DK40561, U01HL145386, R24ES028521), National Institute for Health Research (MR/M016560/1), UK Medical Research Council/Engineering and Physical Sciences Research Council (T213038/Z/18/Z), Wellcome Trust (WT212904/Z/18/Z, WT203148/Z/16/Z, T213038/Z/18/Z), Massachusetts Consortium on Pathogen Readiness (MassCPR-003), American

Gastroenterological Association (AGA2021-5102), American Diabetes Association (7-21-JDFM-005) and Alzheimer's Society (AS-JF-17-011).

**Disclaimer** The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**Competing interests** JW, CH, SS and JC are employees of Zoe Ltd. TDS, ERL and SB, area consultant to Zoe Ltd. DAD, JM and AC previously served as investigators on a clinical trial of diet and lifestyle using a separate mobile application that was supported by Zoe Ltd.

**Patient consent for publication** Not required.

**Ethics approval** The study protocol was approved by the Mass General Brigham Human Research Committee (protocol 2020P000909) and King's College London Ethics Committee (REMAS ID 18210, LRS-19/20–18210).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** The diet quality data used for this study are held by the department of Twin Research at Kings' College London. The data can be released to bona fide researchers using our normal procedures overseen by the Wellcome Trust and its guidelines as part of our core funding (https://web.www.healthdatagateway.org/dataset/fddcb382-3051-4394-8436-b92295f14259). Zoe Platform data used in this study is available to researchers through UK Health Data Research using the following link:https://web.www.healthdatagateway.org/dataset/fddcb382-3051-4394-8436-b92295f14259. The diet quality data used for this study are held by the department of Twin Research at Kings' College London. The data can be released to bona fide researchers using our normal procedures overseen by the Wellcome Trust and its guidelines as part of our core funding. We receive around 100 requests per year for our datasets and have a meeting three times a month with independent members to assess proposals. Application is via https://twinsuk.ac.uk/resources-for-researchers/access-our-data/. This means that the data needs to be anonymized and conform to GDPR standards.



**ORCID iDs**
Jordi Merino http://orcid.org/0000-0001-8312-1438
Amit D Joshi http://orcid.org/0000-0001-7581-6934
Long H Nguyen http://orcid.org/0000-0002-5436-4219
Emily R Leeming http://orcid.org/0000-0002-0531-4901
David A Drew http://orcid.org/0000-0002-8813-0816
Chun-Han Lo http://orcid.org/0000-0001-8202-4513
Cristina Menni http://orcid.org/0000-0001-9790-0571

## REFERENCES

1 Cummings MJ, Baldwin MR, Abrams D, *et al*. Epidemiology, clinical course, and outcomes of critically ill adults with COVID-19 in New York City: a prospective cohort study. *Lancet* 2020;395:1763–70.
2 The Lancet Diabetes Endocrinology. Metabolic health: a priority for the post-pandemic era. *Lancet Diabetes Endocrinol* 2021;9:189.
3 Singh S, Khan A. Clinical characteristics and outcomes of coronavirus disease 2019 among patients with preexisting liver disease in the United States: a multicenter research network study. *Gastroenterology* 2020;159:768–71.
4 Leong A, Cole JB, Brenner LN, *et al*. Cardiometabolic risk factors for COVID-19 susceptibility and severity: a Mendelian randomization analysis. *PLoS Med* 2021;18:e1003553.
5 Satija A, Bhupathiraju SN, Spiegelman D, *et al*. Healthful and Unhealthful Plant-Based Diets and the Risk of Coronary Heart Disease in U.S. Adults. *J Am Coll Cardiol* 2017;70:411–22.
6 Chiuve SE, Fung TT, Rimm EB, *et al*. Alternative dietary indices both strongly predict risk of chronic disease. *J Nutr* 2012;142:1009–18.
7 Ley SH, Hamdy O, Mohan V, *et al*. Prevention and management of type 2 diabetes: dietary components and nutritional strategies. *Lancet* 2014;383:1999–2007.
8 Willett WC, Stampfer MJ. Current evidence on healthy eating. *Annu Rev Public Health* 2013;34:77–95.
9 Satija A, Bhupathiraju SN, Rimm EB, *et al*. Plant-Based dietary patterns and incidence of type 2 diabetes in US men and women: results from three prospective cohort studies. *PLoS Med* 2016;13:e1002039.
10 Mazidi M, Kengne AP. Higher adherence to plant-based diets are associated with lower likelihood of fatty liver. *Clin Nutr* 2019;38:1672–7.
11 Marmot MG, Smith GD, Stansfeld S, *et al*. Health inequalities among British civil servants: the Whitehall II study. *Lancet* 1991;337:1387–93.
12 Belanger MJ, Hill MA, Angelidi AM, *et al*. Covid-19 and disparities in nutrition and obesity. *N Engl J Med* 2020;383:e69.
13 Rehm CD, Peñalvo JL, Afshin A, *et al*. Dietary intake among US adults, 1999-2012. *JAMA* 2016;315:2542.
14 Storm I, den Hertog F, van Oers H, *et al*. How to improve collaboration between the public health sector and other policy sectors to reduce health inequalities? - A study in sixteen municipalities in the Netherlands. *Int J Equity Health* 2016;15:97.
15 Kim H, Rebholz CM, Hegde S, *et al*. Plant-Based diets, pescatarian diets and COVID-19 severity: a population-based case-control study in six countries. *BMJ Nutr Prev Health* 2021;4:257–66.
16 Drew DA, Nguyen LH, Steves CJ, *et al*. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* 2020;368:1362–7.
17 Mazidii M, Leming E, Merino J. Impact of COVID-19 on health behaviours and body weight: a prospective observational study in a cohort of 1.1 million UK and US individuals. *Research Square [Preprint]* 2021.
18 Cleghorn CL, Harrison RA, Ransley JK, *et al*. Can a dietary quality score derived from a short-form FFQ assess dietary quality in UK adult population surveys? *Public Health Nutr* 2016;19:2915–23.
19 Menni C, Valdes AM, Freidin MB, *et al*. Real-Time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med* 2020;26:1037–40.
20 Varsavsky T, Graham MS, Canas LS, *et al*. Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study. *Lancet Public Health* 2021;6:e21–9.
21 Botti-Lodovico Y, Rosenberg E, Sabeti PC. Testing in a Pandemic - Improving Access, Coordination, and Prioritization. *N Engl J Med* 2021;384:197–9.
22 Indices of deprivation. Available: https://www.gov.uk/government/publications/english-indices-of-deprivation-2019-technical-report [Accessed 18 Jan 2021].
23 Messer LC, Laraia BA, Kaufman JS, *et al*. The development of a standardized neighborhood deprivation index. *J Urban Health* 2006;83:1041–62.
24 Toutenburg H. Rubin, D.B.: multiple imputation for nonresponse in surveys. *Statistical Papers* 1990;31:180.
25 Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982;69:239–41.
26 VanderWeele TJ, Tchetgen Tchetgen EJ. Attributing effects to interactions. *Epidemiology* 2014;25:711–22.
27 Calder PC. Nutrition, immunity and COVID-19. *BMJ Nutr Prev Health* 2020;3:74–92.
28 Yan B, Chu H, Yang D, *et al*. Characterization of the lipidomic profile of human coronavirus-infected cells: implications for lipid metabolism remodeling upon coronavirus replication. *Viruses* 2019;11:73.
29 Louca P, Murray B, Klaser K, *et al*. Modest effects of dietary supplements during the COVID-19 pandemic: insights from 445 850 users of the COVID-19 symptom study APP. *BMJ Nutr Prev Health* 2021;4:149–57.
30 Eiser AR. Could dietary factors reduce COVID-19 mortality rates? Moderating the inflammatory state. *J Altern Complement Med* 2021;27:176–8.
31 O'Hearn M, Liu J, Cudhea F, *et al*. Coronavirus disease 2019 hospitalizations attributable to cardiometabolic conditions in the United States: a comparative risk assessment analysis. *J Am Heart Assoc* 2021;10:e019259.
32 Wise J. Covid-19: highest death rates seen in countries with most overweight populations. *BMJ* 2021;372:n623.
33 Mena GE, Martinez PP, Mahmud AS, *et al*. Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *Science* 2021;372:eabg5298.
34 Feldman JM, Bassett MT. The relationship between neighborhood poverty and COVID-19 mortality within racial/ethnic groups (Cook County, Illinois). *medRxiv*2020:2020.10.04.20206318.
35 Brooks JT, Butler JC, Redfield RR. Universal masking to prevent SARS-CoV-2 Transmission-The time is now. *JAMA* 2020;324:635–7.
36 Savolainen O, Lind MV, Bergström G, *et al*. Biomarkers of food intake and nutrient status are associated with glucose tolerance status and development of type 2 diabetes in older Swedish women. *Am J Clin Nutr* 2017;106:ajcn152850–10.

# Supplementary Material

**Supplement to: Diet quality and risk and severity of COVID-19: a prospective cohort study**

Jordi Merino,[1,2,3#] Amit D. Joshi,[4,5#] Long H. Nguyen,[4,5,6#] Emily R. Leeming,[7] Mohsen Mazidi,[7] David A Drew,[4,5] Rachel Gibson,[8] Mark S. Graham,[9] Chun-Han Lo,[4,5] Joan Capdevila,[10] Benjamin Murray,[9] Christina Hu,[10] Somesh Selvachandran,[10] Sohee Kwon,[4,5] Wenjie Ma,[4,5] Cristina Menni,[7] Alexander Hammers,[9,11] Shilpa N. Bhupathiraju,[3,12] Shreela V. Sharma,[14] Carole Sudre,[9] Christina M. Astley,[2,13] Walter C. Willet,[12,15,16] Jorge E. Chavarro,[12,15,16] Sebastien Ourselin,[9] Claire J. Steves,[7] Jonathan Wolf,[10] Paul W. Franks,[12,17] Tim D. Spector, MBBS,[8*] Sarah E. Berry,[8*] Andrew T. Chan,[4,5*]

**Correspondence:** Andrew T. Chan, M.D., M.P.H. Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, 100 Cambridge Street Boston, MA 02114
achan@mgh.harvard.edu

## Table of Contents:





Merino J, *et al. Gut* 2021;0:1–9. doi: 10.1136/gutjnl-2021-325353

## Supplementary Methods

### Dietary intake assessment

Habitual dietary intake information was collected through an amended version of the Leeds Short Form Food Frequency Questionnaire (LSF-FFQ).[1] In brief, the LSF-FFQ includes 20 food items with reference to fruit, vegetables, fibre-rich foods, high fat and high-sugar foods, meat, meat products and fish. Seven additional food items were added to capture broader dietary intake information including refined carbohydrates (e.g. white rice, white pasta and white bread), eggs, fast food and live probiotic or fermented foods (e.g. live yogurt, kefir and kimchi). Participants were asked how often on average they had consumed one portion of each food in a typical week. The responses had eight frequency categories ranging from "rarely or never" to "five or more times per day. Further detail on the development, dissemination and procedures of the diet and lifestyle survey to UK and US participants is described elsewhere.[2]

### Outcome ascertainment

Predicted COVID-19 definition: We used a symptom-based classifier developed by our group to predict COVID-19.[3] To build the prediction model, UK participants were randomly divided into a training set and a test set (ratio: 80:20). Based on the training set, a logistic model generated to predict symptomatic COVID-19 was: Log odds (Predicted COVID-19) = -1.32 - (0.01 x age) + (0.44 x male sex) + (1.75 x loss of smell or taste) + (0.31 x severe or significant persistent cough) + (0.49 x severe fatigue) + (0.39 x skipped meals). The prediction model achieved a sensitivity of 0.65 (95% CI 0.62-0.67) and specificity of 0.78 (95% CI 0.76-0.80) in the test set. In additional validation in the U.S. participants, the prediction model achieved a sensitivity of 0.66 (95% CI 0.62-0.69) and specificity of 0.83 (95% CI 0.82-0.85).

Severe COVID: To ascertain severe COVID-19, we used responses to the question "*What treatment did you receive while in the hospital / What treatment are you receiving right now?*" Participants had the option to respond a) None, b) Oxygen and fluids breathing support administered through an oxygen mask, no pressure applied, c) Non-invasive ventilation breathing support administered through an oxygen mask, which pushes oxygen into your lungs, d) Invasive ventilation breathing support administered through an inserted tube. People are usually asleep for this procedure, e) Other. COVID-19 severity was ascertained based on a report of the need for a hospital visit which required 1) non-invasive breathing support, 2) invasive breathing support, and 3) administration of antibiotics combined with oxygen support.

### Covariate classification

Covariates were selected *a priori* based on putative confounders and risk factors for COVID-19 and included sex (male, female), race/ethnicity (White, Black, Asian, Other), index of multiple deprivation (most deprived <3, intermediate deprived 3 to 7, less deprived >7), population density (<500 individuals/km$^2$, 500 to 1,999 individuals/km$^2$, 2,000 to 4,999 individuals/km$^2$, and ≥ 5,000 individuals/km$^2$), healthcare worker status (yes with interaction with COVID-19 patients, yes without interaction with COVID-19 patients, no), presence of comorbidities [diabetes (yes, no), cardiovascular disease (yes, no), lung disease (yes, no), cancer (yes, no), kidney disease (yes, no)], body mass index (<18.5 kg/m2, 18.5 to 24.9 kg/m2, 25.0 to 29.9 kg/m2, and ≥30 kg/m2), smoking status (yes, no), and physical activity (<1 day/week, 1 to 2 days/week, 3 to 4 days/week, ≥5 days/week).

### Interactions between diet quality and deprivation on COVID-19 risk

We tested for additive interactions by assessing the relative excess risk due to interaction, and further examined the risk proportions attributable to diet quality alone, to deprivation alone, and to their interaction. For these analyses, we considered diet quality and socioeconomic deprivation as continuous variables. We assessed the relative excess risk due to interaction as an index of additive interaction using the following formula (RERI = $RR_{11}$ - $RR_{10}$ - $RR_{01}$ + 1)[4], and further examined the decomposition of the joint effect, which is the proportion attributable to genetic risk alone, to diet quality alone, and to their interaction (i.e., AP= RERI/ $RR_{11}$).[4]

**References**

1 Cleghorn CL, Harrison RA, Ransley JK, et al. Can a dietary quality score derived from a short-form FFQ assess dietary quality in UK adult population surveys? Public Health Nutr 2016;19:2915–2923

2 Mazidi M, Leming E, Merino J, et al. Impact of COVID-19 on health behaviours and body weight: A prospective observational study in a cohort of 1.1 million UK and US individuals. Research Square. [Preprint]. Cited 2021 April 25.

3 Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med 2020;26:1037–1040.

4 VanderWeele TJ, Tchetgen Tchetgen EJ. Attributing effects to interactions. Epidemiology 2014;25:711–22.

# Pre-specified protocol

## Purpose and meaning

The main objective of the present proposal is to use self-reported individual-level data from up to 1.1 million volunteers included in the COVID Symptom Study to evaluate the associations between diet quality and COVID-19 risk and severity. In addition we will investigate its intersection with deprivation.. Findings from this study have the potential to identify susceptible individuals to increased COVID-19 risk and severity and inform public health strategies to reduce the burden of the COVID-19 pandemic.

## 1. General methodological considerations

Dataset preparation: We will collect daily app responses to generate a dataset that contains follow-up data from March 24, 2020 and followed until December 2, 2020. We will obtain information on demographic factors, self-reported COVID-19 or any COVID-19 related symptoms and personal and medical history including lung disease, diabetes, cardiovascular disease, cancer, kidney disease, and use of medications.

Quality Control: For this study we will include individuals who responded to the diet and lifestyle survey for the pre-pandemic period. We will filter out multiple records for the same participant and time point, records without an indication of whether they were recorded pre- or peri- pandemic, and records not linked to UK or US participants.

Main exclusions: Prevalent COVID-19 prior to start of follow-up. Presence of symptoms that classify them as having predicted COVID-19 within 24 hours of first entry. Participants younger than 18 years old. Pregnant women. Participants who logged only one daily assessment during follow-up.

## 1.1 Primary outcome and exposures

Outcome: The main outcome will be COVID-19 risk defined using a validated symptom based-algorithm developed by our research team.

Secondary outcomes will include:

> COVID-19 risk base on report of a positive COVID-19 test by RT-PCR.
>
> COVID-19 severity will be defined based on the risk of hospitalization and the need of oxygen requirements based on responses to the following question "*What treatment did you receive while in the hospital / What treatment are you receiving right now?"*

Exposure definitions: Diet quality will be quantified using diet quality indices. We will generate the healthful plant-based diet index (hPDI) and the Diet Quality Score (DQS). To generate these scores we will use the items and weighting criteria used in previous studies.

## 1.2 Covariates

Covariates will be selected a priori based on putative confounders and risk factors for COVID-19. Modes will be adjusted for 10-year age group, country of origin (UK, US), sex (male, female), race/ethnicity (White, Black, Asian, Other), index of multiple deprivation (most deprived <3, intermediate deprived 3 to 7, less deprived >7), population density (<500 individuals/km2, 500 to 1,999 individuals/km2, 2,000 to 4,999 individuals/km2, and ≥ 5,000 individuals/km2), healthcare worker status (yes with interaction with COVID-19 patients, yes without interaction with COVID-19 patients, no), presence of comorbidities [diabetes (yes, no), cardiovascular disease (yes, no), lung disease (yes, no), cancer (yes, no), kidney disease (yes, no)], body mass index (<18.5 kg/m2, 18.5 to 24.9 kg/m2, 25.0 to 29.9 kg/m2, and ≥30 kg/m2), smoking status (yes, no), and physical activity (<1 day/week, 1 to 2 days/week, 3 to 4 days/week, ≥5 days/week).

## 1.3 Unit of analysis



Merino J, *et al. Gut* 2021;0:1–9. doi: 10.1136/gutjnl-2021-325353

Estimated effect sizes will be reported per change in diet quality category (low diet quality would be the reference) or 1SD increase.

**1.4 Subgroup analysis**

We will assess the association between diet quality and COVID-19 risk according to comorbidities, demographic, and lifestyle characteristics. We will also classified participants according to categories of the diet quality score and socioeconomic deprivation and conducted joint analyses. We will also test for additive interactions by assessing the relative excess risk due to interaction, and further examined the COVID-19 risk proportions attributable to diet, deprivation, and to their interaction.

**2. Statistical analyses**

1. Multiple imputations by chained equations with five imputations will be used to handle missing data.

2. Follow-up time for each participant will start 24 hours after first log-in to the time of predicted COVID-19 (or to time of secondary outcomes) or date of last entry prior to December 2, 2020, whichever occurred first.

3. Cox regression models will be stratified by calendar date at study entry, country of origin, and 10-year age group (age-adjusted model 1). Model 2 will be further adjusted for sex, race/ethnicity, index of multiple deprivation, population density, and healthcare worker status. Model 3 will be further adjusted for presence of diabetes, cardiovascular disease, lung disease, cancer, kidney disease, body mass index, smoking status, and physical activity.

4. Absolute risk will be calculated as the percentage of COVID-19 cases occurring per 10,000 person-months.

5. We will use restricted cubic splines with four knots (at the 2.5th, 25th, 75th, and 97.5$^{th}$ percentiles) to assess for non-linear associations between diet quality and COVID-19 risk.

6. For COVID-19 risk defined by a positive test we will use inverse probability-weighted Cox models to account for predictors of btaining country-specific testing. Inverse probability-weighted analyses will include presence of COVID-19-related symptoms, interaction with a person with COVID-19, occupation as a healthcare worker, age group, and race. These models will be adjusted for the same confounders as before.

7. Subgroup analysis will be based according to comorbidities, demographic, and lifestyle characteristics; Age (<60, ≥60), Sex (Male, Female), Race (White, Non-white), Deprivation (Low, Intermediate, High), Population density (<2,000, ≥2,000), Healthcare worker (yes, no), BMI (<25, 25-30, ≥ 30), physical activity (<1d/wk, 1-4 d/wk, ≥5d/wk). Cox models will be adjusted for the same covariates as previous model 3. In a sensitivity analysis we will use the DQS score to investigate associations between diet quality and COVID-19 risk and severity. These models will be adjusted for the same confounders as before.

9. Sensitivity analysis to censor cases that occurred after completing the diet survey. These models will be adjusted for the same confounders as before.

10. Sensitivity analysis to account for regional differences in the effective reproductive number (Rt) or mask wearing. For Rt analyses, we will extract US state-level information from the COVID Tracking Project for the period between March 2020 and January 2021. For the UK we will calculate $R_t$ time-series for Scotland, Wales, and each of the NHS regions in England, using a previously published methodology from our group. For these analyses, we will define community peak and nadir $R_t$ time-windows as the period between one week before and two weeks after $R_t$ was all-time high or low. Using censored time-windows, we will test the association between diet quality and COVID-19 risk after adjusting for the same confounders as included in model 3. For mask wearing analyses, we will use survey data launched on June/September 2020 on whether participants had worn a face mask when outside the house in the last week. Responses will be categorized into never, sometimes, most of the time, or always. Mask wearing analyses will include the same covariates as included in model 3.

**Research team:** Jordi Merino, Amit D. Joshi, Long H. Nguyen, Emily Leeming, Sarah E. Berry, Andrew T. Chan. This research proposal was approved by the research team on 11/02/2020.

**Supplementary table 1: Grouping and components of the hPDI**

| hPDI component | FFQ items |
|---|---|
| Wholegrains | Fibre-rich breakfast cereal, like Weetabix, Fruit 'n Fibre, Porridge, Muesli; Wholemeal bread or chapattis |
| Fruits | Fruit (tinned / fresh) |
| Vegetables | Salad (not garnish added to sandwiches); Vegetables (tinned / frozen / fresh but not potatoes) |
| Nuts | N/A |
| Legumes | Beans or pulses like baked beans, chick peas, dahl |
| Vegetable oils | N/A |
| Tea and Coffee | N/A |
| Fruit Juice | Fruit juice (not cordial or squash) |
| Refined Grains | Crisps / savoury snacks; pasta; Refined breakfast cereals (e.g. rice krispies, cornflakes, coco pops); rice; white bread |
| Potatoes | Chips / fried potatoes |
| Sugar Sweetened Beverages | Nonalcoholic fizzy drinks/pop (not sugar free or diet) |
| Sweets and desserts | Sweet biscuits, cakes, chocolate, sweets |
| Animal fats | N/A |
| Dairy | Cheese / yoghurt; Ice cream / cream; live probiotic or fermented food products (e.g. yoghurt, kefir, kimchi) |
| Egg | Eggs - as boiled, fried, scrambled, etc |
| Fish and seafood | White fish in batter or breadcrumbs – like 'fish 'n chips'; White fish not in batter or breadcrumbs; Oily fish – like herrings, sardines, salmon, trout, mackerel, fresh tuna (not tinned tuna) |
| Meat | Beef, Lamb, Pork, Ham - steaks, roasts, joints, mince or chops; Chicken or Turkey – steaks, roasts, joints, mince or portions (not in batter or breadcrumbs); Sausages, bacon, corned beef, meat pies/pasties, burgers; Chicken/turkey nuggets/twizzlers, turkey burgers, chicken pies, or in batter or breadcrumbs |
| Miscellaneous | Fast food |

**Table legend:** FFQ items constituting the 18 food groups originally used to generate the healthy plant-based diet index in Satija et al. JACC 2017. Four out of the 18 food groups originally considered were not included for the calculation of the healthy plant-based diet index in this study as they were not available (N/A). FFQ = food frequency questionnaire; hPDI = healthful plant-based diet index

**Supplementary table 2: Criteria for scoring each component of the hPDI**

| Component | Criteria for min score of 1 | Criteria for max score of 5 |
|---|---|---|
| Whole grain | Lowest quintile of intake | Highest quintile of intake |
| Fruits | Lowest quintile of intake | Highest quintile of intake |
| Vegetables | Lowest quintile of intake | Highest quintile of intake |
| Nuts | N/A | N/A |
| Legumes | Lowest quintile of intake | |
| Vegetable oils | N/A | N/A |
| Tea and coffee | N/A | N/A |
| Fruit juices | Highest quintile of intake | Lowest quintile of intake |
| Refined grains | Highest quintile of intake | Lowest quintile of intake |
| Potatoes | Highest quintile of intake | Lowest quintile of intake |
| Sugar sweetened beverages | Highest quintile of intake | Lowest quintile of intake |
| Sweets and desserts | Highest quintile of intake | Lowest quintile of intake |
| Animal fat | N/A | N/A |
| Dairy | Highest quintile of intake | Lowest quintile of intake |
| Egg | Highest quintile of intake | Lowest quintile of intake |
| Fish or seafood | Highest quintile of intake | Lowest quintile of intake |
| Meat | Highest quintile of intake | Lowest quintile of intake |
| Miscellaneous | Highest quintile of intake | Lowest quintile of intake |

**Table legend:** Criteria for scoring the 18 food groups originally used to generate the healthy plant-based diet index in Satija et al. JACC 2017. Food groups were ranked into quintiles, and given positive (healthy plant food groups) or reverse scores ( less healthy plant food groups and animal food groups). With positive scores, participants above the highest quintile of a food group received a score of 5, following on through to participants below the lowest quintile who received a score of 1. With reverse scores, this pattern of scoring was inverted. All component scores were summed to obtain a total score ranging from 0 (lowest diet quality) to 70 (highest diet quality) points.

**Supplementary table 3: Grouping and components of the DQS**

| DQS component | FFQ items |
|---|---|
| Fruits | Fruit (tinned / fresh) |
| Vegetables | Salad (not garnish added to sandwiches)<br>Vegetables (tinned / frozen / fresh but not potatoes) |
| Oily fish | Oily fish – like herrings, sardines, salmon, trout, mackerel, fresh tuna (not tinned tuna) |
| Total fat | Fruit (tinned / fresh)<br>Fruit juice (not cordial or squash)<br>Salad (not garnish added to sandwiches)<br>Vegetables (tinned / frozen / fresh but not potatoes)<br>Chips / fried potatoes<br>Beans or pulses like baked beans, chick peas, dahl<br>Fiber-rich breakfast cereal, like Weetabix, Fruit 'n Fiber, Porridge, Muesli<br>Whole-meal bread or chapattis; Cheese / yoghurt; Crisps / savory snacks<br>Sweet biscuits, cakes, chocolate, sweets<br>Ice cream / cream<br>Nonalcoholic fizzy drinks/pop (not sugar free or diet)<br>Beef, Lamb, Pork, Ham - steaks, roasts, joints, mince or chops<br>Chicken or Turkey – steaks, roasts, joints, mince or portions (not in batter or breadcrumbs)<br>Processed meats/ meat products<br>Sausages, bacon, corned beef, meat pies/pasties, burgers<br>Chicken/turkey nuggets/twizzles, turkey burgers, chicken pies, or in batter or breadcrumbs<br>White fish in batter or breadcrumbs – like 'fish 'n chips'<br>White fish not in batter or breadcrumbs |
| Non-milk extrinsic sugars | Fruit (tinned / fresh)<br>Fruit juice (not cordial or squash)<br>Salad (not garnish added to sandwiches)<br>Vegetables (tinned / frozen / fresh but not potatoes)<br>Chips / fried potatoes<br>Beans or pulses like baked beans, chick peas, dahl<br>Fiber-rich breakfast cereal, like Weetabix, Fruit 'n Fiber, Porridge, Muesli<br>Whole-meal bread or chapattis; Cheese / yoghurt; Crisps / savory snacks<br>Sweet biscuits, cakes, chocolate, sweets<br>Ice cream / cream<br>Nonalcoholic fizzy drinks/pop (not sugar free or diet)<br>Beef, Lamb, Pork, Ham - steaks, roasts, joints, mince or chops<br>Chicken or Turkey – steaks, roasts, joints, mince or portions (not in batter or breadcrumbs)<br>Processed meats/ meat products<br>Sausages, bacon, corned beef, meat pies/pasties, burgers<br>Chicken/turkey nuggets/twizzles, turkey burgers, chicken pies, or in batter or breadcrumbs<br>White fish in batter or breadcrumbs – like 'fish 'n chips'<br>White fish not in batter or breadcrumbs |

**Table legend:** FFQ items constituting the 5 food components originally used to generate the DQS score from Cleghorn et al., listed in the Nutritools library (nutritools.org). FFQ = food frequency questionnaire; DQS = diet quality score.

**Supplementary table 4: Criteria for scoring each component of the DQS**

| DQS Component | Criteria for score of 1 | Criteria for score of 2 | Criteria for score of 3 |
|---|---|---|---|
| Fruit | ≤ 2 servings/week | >2 servings/week and <2 servings/d | ≥2 servings/d |
| Vegetables | ≤ 1 servings/d | 1-3 servings/d | ≥ 3 servings/d |
| Oily Fish | No intake | 0-200g/week | 200g/week |
| Total Fat | ≥1.5 x UK recommendations ( ≥127.5g/d) | 1-1.5 x UK recommendations | ≤ UK recommendations (≤ 85g/d) |
| Non-Milk-Extrinsic Sugars | ≥1.5 x UK recommendations ( ≥ 90g/d) | 1-1.5 x UK recommendations | ≤ UK recommendations (≤ 60g/d) |

**Table legend:** Criteria for scoring the 5 food groups originally used to generate the diet quality score from Cleghorn et al., listed in the Nutritools library (nutritools.org). Each component was scored from 1 (unhealthiest) to 3 (healthiest) points, with intermediate values scored proportionally. All component scores were summed to obtain a total score ranging from 5 (lowest diet quality) to 15 points(highest diet quality) points.

**Supplementary Table 5: Demographic, lifestyle, and clinical characteristics according to diet and lifestyle survey participation**

| | **Included participants** (n=592,571) | **Participants who did not respond to the diet survey** (n= 3,289,680) |
|---|---|---|
| Age, years | 56 (44-65) | 43 (32-56) |
| ≥18-24 | 14,397 (2.4) | 312,927 (9.5) |
| 25-34 | 52,922 (8.9) | 715,438 (21.7) |
| 35-44 | 86,251 (14.6) | 721,725 (21.9) |
| 45-54 | 125,802 (21.2) | 635,749 (19.3) |
| 55-64 | 158,637 (26.8) | 482,356 (14.7) |
| ≥65 | 153,810 (26.0) | 421,485 (12.8) |
| Sex, No. (%) | | |
| Male | 187,450 (31.6) | 1,311,439 (39.9) |
| Female | 404,126 (68.2) | 1,974,754 (60.1) |
| Race[ε], No. (%), | | |
| White | 568,770 (96.0) | 2,214,416 (67.3) |
| Black | 4,328 (0.7) | 35,932 (1.1) |
| Asian | 10,435 (1.8) | 85,605 (2.6) |
| Other/Prefer not to say | 9,038 (1.5) | 953,727 (28.9) |
| Country, No. (%) | | |
| UK | 543,984 (91.8) | 3,026,997 (92.0) |
| US | 48,587 (8.2) | 262,683 (8.0) |
| Index of deprivation, No. (%)[¶] | | |
| Most deprived, decile 1 | 1,3416 (2.3) | 128875 (3.9) |
| Least deprived, decile 10 | 103,608 (17.5) | 408310 (12.4) |
| Population density, $km^2$, No. (%)[¶] | | |
| <500 | 119,782 (20.2) | 133,740 (4.1) |
| 500-1,999 | 90,541 (15.3) | 534,421 (16.2) |
| 2,000 4,999 | 94,345 (15.9) | 874,726 (26.6) |
| ≥5,000 | 244,295 (41.2) | 1,213,590 (36.9) |
| Healthcare worker, No. (%) | | |
| Yes | 41,141 (6.9) | 274,052 (8.3) |
| Body mass index, $Kg/m^2$ | 25.1 (22.6-28.7) | 25.7 (22.8-29.6) |
| <18.5 | 12,004 (2.0) | 98,815 (3.1) |
| 18.5-24.9 | 277,536 (46.8) | 1,366,485 (41.5) |
| 25-29.9 | 189,197 (31.9) | 1,054,239 (32.0) |
| ≥30 | 113,056 (19.1) | 769,687 (23.4) |
| Diabetes | 20,058 (3.4) | 99,807 (3.0) |
| Heart disease | 20,376 (3.4) | 89,260 (2.7) |
| Cancer | 6,559 (1.9) | 28,545 (1.4) |

| | | |
|---|---|---|
| Lung disease | 62,999 (10.6) | 346,150 (10.5) |
| Kidney disease | 5,134 (0.9) | 24,251 (0.9) |

**Table Legend:** Values are median ($P_{25}$-$P_{75}$) for continuous variables; numbers and (percentages) for categorical variables.

[ε] Race was self-reported by the participants.

[¶] Index of deprivation and population density were generated using zip code or postcode information linked with census track data.

**Supplementary table 6: Adjusted hazard ratios of COVID-19 risk and severity for diet quality in the COVID Symptom Study**

| | Low DQS | Intermediate DQS | High DQS | *P* for trend |
|---|---|---|---|---|
| Diet quality score, median (IQR) | 9 (8-10) | 11 (11-11) | 13 (12 -13) | |
| **COVID-19 risk** | | | | |
| No. of events/person-months | 13,996 / 1,467,205 | 12,641 / 1,701,799 | 5,178 / 717,270 | — |
| Incidence rate (10,000 person-months; 95% CI) | 95.4 (93.8-97.0) | 74.3 (73.0-75.6) | 72.2 (70.3-74.2) | — |
| Age-adjusted model | 1.00 (Ref) | 0.90 (0.87-0.92) | 0.92 (0.89-0.95) | <0.001 |
| Multivariable model 2 | 1.00 (Ref) | 0.90 (0.88-0.92) | 0.92 (0.89-0.95) | 0.019 |
| Multivariable model 3 | 1.00 (Ref) | 0.95 (0.93-0.98) | 1.00 (0.97-1.03) | 0.216 |
| **COVID-19 risk (positive test)** | | | | |
| No. of events/person-months | 2,341 / 1,515,004 | 2,309 / 1,746,982 | 952 / 736,535 | — |
| Incidence rate (10,000 person-months; 95% CI) | 15.5 (14.8-16.1) | 13.2 (12.7-13.8) | 12.9 (12.1-13.7) | — |
| Age-adjusted model[$] | 1.00 (Ref) | 0.94 (0.91-0.98) | 0.92 (0.87-0.96) | <0.001 |
| Multivariable model 2[$] | 1.00 (Ref) | 0.95 (0.91-0.99) | 0.93 (0.89-0.98) | 0.006 |
| Multivariable model 3[$] | 1.00 (Ref) | 0.96 (0.93-1.00) | 0.95 (0.91-1.00) | 0.047 |
| **COVID-19 severity** | | | | |
| No. of events/person-months | 313 / 1,518,980 | 317 / 1,750,786 | 110 / 738,495 | — |
| Incidence rate (10,000 person-months; 95% CI) | 2.1 (1.8-2.3) | 1.8 (1.6-2.0) | 1.5 (1.2-1.8) | — |
| Age-adjusted model | 1.00 (Ref) | 0.82 (0.70-0.96) | 0.67 (0.54-0.84) | <0.001 |
| Multivariable model 2 | 1.00 (Ref) | 0.82 (0.70-0.97) | 0.68 (0.54-0.85) | <0.001 |
| Multivariable model 3 | 1.00 (Ref) | 0.93 (0.79-1.09) | 0.83 (0.66-1.04) | 0.141 |

**Table legend:** Hazards ratios and 95% CI for COVID-19 risk and severity. Sensitivity analysis using the DQS to quantify diet quality. Cox proportional hazards models were stratified by calendar date at study entry, country of origin, and 10-year age group (Age-adjusted model). Multivariable model 2 was further adjusted for sex (male, female), race/ethnicity (White, Black, Asian, Other), index of multiple deprivation (most deprived <3, intermediate deprived 3 to 7, less deprived >7), population density (<500 individuals/km$^2$, 500 to 1,999 individuals/km$^2$, 2,000 to 4,999 individuals/km$^2$, and ≥ 5,000 individuals/km$^2$), and healthcare worker status (yes with interaction with COVID-19 patients, yes without interaction with COVID-19 patients, no). Model 3 was further adjusted for presence of comorbidities [diabetes (yes, no), cardiovascular disease (yes, no), lung disease (yes, no), cancer (yes, no), kidney disease (yes, no)], body mass index (<18.5 kg/m2, 18.5 to 24.9 kg/m2, 25.0 to 29.9 kg/m2, and ≥30 kg/m2), smoking status (yes, no), and physical activity (<1 day/week, 1 to 2 days/week, 3 to 4 days/week, ≥5 days/week).

[$] Inverse probability-weighted analyses were conducted to account for predictors of obtaining RT-PCR testing (presence of COVID-19-related symptoms, interaction with a COVID-19 case, healthcare worker, age group, and race). inverse probability-weighted Cox proportional hazards models were stratified by 10-year age group and date with additional adjustment for the covariates used in previous models.

**Supplementary table 7: Association between diet quality and COVID risk - censored to cases that occurred after completing the diet questionnaires**

| | Low hPDI | Intermediate hPDI | High hPDI | *P* for trend |
|---|---|---|---|---|
| **COVID-19 risk** | | | | |
| Incidence rate (10,000 person-months; 95% CI) | 116.2 (110.9-120.3) | 84.4 (82.0-86.7) | 74.1 (71.5-78.6) | — |
| Age-adjusted model | 1.00 (Ref) | 0.82 (0.78-0.86) | 0.79 (0.75-0.83) | <0.001 |
| Multivariable model 2 | 1.00 (Ref) | 0.83 (0.79-0.87) | 0.79 (0.75-0.84) | <0.001 |
| Multivariable model 3 | 1.00 (Ref) | 0.87 (0.83-0.92) | 0.88 (0.83-0.93) | <0.001 |
| **COVID-19 risk (positive test)** | | | | |
| Incidence rate (10,000 person-months; 95% CI) | 42.1 (40.2-44.3) | 33.5 (32.1-35.0) | 29.1 (27.2-31.0) | — |
| Age-adjusted model[$] | 1.00 (Ref) | 0.84 (0.77-0.92) | 0.77 (0.70-0.86) | <0.001 |
| Multivariable model 2[$] | 1.00 (Ref) | 0.85 (0.78-0.93) | 0.79 (0.72-0.88) | <0.001 |
| Multivariable model 3[$] | 1.00 (Ref) | 0.86 (0.79-0.94) | 0.80 (0.72-0.89) | <0.001 |

**Table legend:** Hazards ratios and 95% CI for COVID-19 risk. Sensitivity analysis censored to cases that occurred after completing the diet questionnaires (September 21st, 2020). Cox proportional hazards models were stratified by calendar date at study entry, country of origin, and 10-year age group (Age-adjusted model). Multivariable model 2 was further adjusted for sex (male, female), race/ethnicity (White, Black, Asian, Other), index of multiple deprivation (most deprived <3, intermediate deprived 3 to 7, less deprived >7), population density (<500 individuals/km$^2$, 500 to 1,999 individuals/km$^2$, 2,000 to 4,999 individuals/km$^2$, and ≥ 5,000 individuals/km$^2$), and healthcare worker status (yes with interaction with COVID-19 patients, yes without interaction with COVID-19 patients, no). Model 3 was further adjusted for presence of comorbidities [diabetes (yes, no), cardiovascular disease (yes, no), lung disease (yes, no), cancer (yes, no), kidney disease (yes, no)], body mass index (<18.5 kg/m2, 18.5 to 24.9 kg/m2, 25.0 to 29.9 kg/m2, and ≥30 kg/m2), smoking status (yes, no), and physical activity (<1 day/week, 1 to 2 days/week, 3 to 4 days/week, ≥5 days/week).

[$] Inverse probability-weighted analyses were conducted to account for predictors of obtaining RT-PCR testing (presence of COVID-19-related symptoms, interaction with a COVID-19 case, healthcare worker, age group, and race). inverse probability-weighted Cox proportional hazards models were stratified by 10-year age group and date with additional adjustment for the covariates used in previous models.



Merino J, *et al. Gut* 2021;0:1–9. doi: 10.1136/gutjnl-2021-325353

**Supplementary table 8: Attributing associations to additive interaction between diet quality and socioeconomic deprivation on risk of COVID-19 infection**

| | **Predicted COVID-19 infection** |
|---|---|
| **Main effects** | |
| Diet quality, per 10 units decrease | 1.05 (1.01-1.09) |
| Deprivation index, per category decrease | 1.06 (1.01-1.12) |
| Joint effect | 1.15 (1.09-1.21) |
| **Relative excess risk due to interaction** | |
| Relative excess risk due to interaction | 0.05 (0.02-0.08) |
| *P* | 0.005 |
| **Attributable proportion, %** | |
| Diet quality | 31.9 (18.2-45.6) |
| Deprivation index | 38.4 (26.5-50.3) |
| Additive interaction | 29.7 (2.1-57.3) |

**Table Legend:** Multivariable-adjusted risk of predicted COVID-19 infection estimated from fully adjusted Cox models. The relative excess risk due to interaction was calculated using the following formula ($RERI_{RR} = RR_{11} - RR_{10} - RR_{01} + 1$). The decomposition of the joint effect, which is the proportions attributable to diet quality alone, to deprivation index alone, and to their interaction, was calculated using the following formula (i.e., $AP = RERI / RR_{11}$).

**Supplementary table 9: Association between diet quality and risk of COVID-19 infection after accounting for mask wearing**

| | Low hPDI | Intermediate hPDI | High hPDI | *P* for trend |
|---|---|---|---|---|
| **COVID-19 risk** | | | | |
| No. of events/person-months | 2,574 / 222,426 | 4,669 / 555,918 | 2,092 / 283,975 | — |
| Incidence rate (10,000 person-months; 95% CI) | 114.6 (110.2-119.0) | 84.0 (81.6-86.4) | 73.7 (70.6-76.9) | — |
| Multivariable Model 3 + Mask wearing | 1.00 (Ref) | 0.88 (0.83 to 0.92) | 0.88 (0.83-0.94) | <0.001 |
| **COVID-19 risk (positive test)** | | | | |
| No. of events/person-months | 989 / 233,564 | 1,907 / 576,267 | 874 / 293,760 | |
| Incidence rate (10,000 person-months; 95% CI) | 42.3 (39.8-45.1) | 33.1 (31.6-34.6) | 29.8 (27.8-31.8) | |
| Multivariable Model 3 + Mask wearing[$] | 1.00 (Ref) | 0.86 (0.79-0.94) | 0.80 (0.72-0.89) | <0.001 |

**Table legend:** Hazards ratios and 95% CI for COVID-19 risk after accounting for mask wearing. These analyses were left censored to September 21st 2020. Mask wearing analyses included 524,825 participants. For confirmed COVID-19 analyses, inverse probability-weighted analyses were conducted to account for predictors of obtaining RT-PCR testing[$].

**Supplementary table 10: Total, direct, and indirect effects of diet quality on COVID-19 risk**

| | COVID-19 risk | |
|---|---|---|
| | HR (95% CI) | **P value** |
| Total effect, 1SD increase in hPDI | 0.96 (0.96-0.97) | <0.001 |
| Direct effect | 0.98 (0.97-0.98) | <0.001 |
| Indirect effect | 0.99 (0.98-0.99) | <0.001 |
| Proportion mediated | 37% (30-44) | <0.001 |

**Table legend:** Structural equation models were implemented to conduct a mediation analysis of BMI using the "lavaan" package in R. For this analysis, diet quality and BMI were used as continuous variables. We estimated the relative contribution of BMI to the association between diet quality and COVID-19 risk and computed the proportion of total effect that was explained by indirect effects of BMI. Indirect effects were estimated by taking the product of the effect of the exposure (diet quality) on the mediator (BMI) and the effect of the mediator (BMI) on the outcome (COVID-19 risk). The direct effect is defined as the association of diet quality on COVID-19 risk through mechanisms independent of mediation and was estimated from regressing COVID-19 on diet quality. To calculate the proportion of the mediated effect we divided the indirect effect by the total effect.

**Supplementary figure 1: Diet and symptom data collection among participants of the COVID Symptom Study**

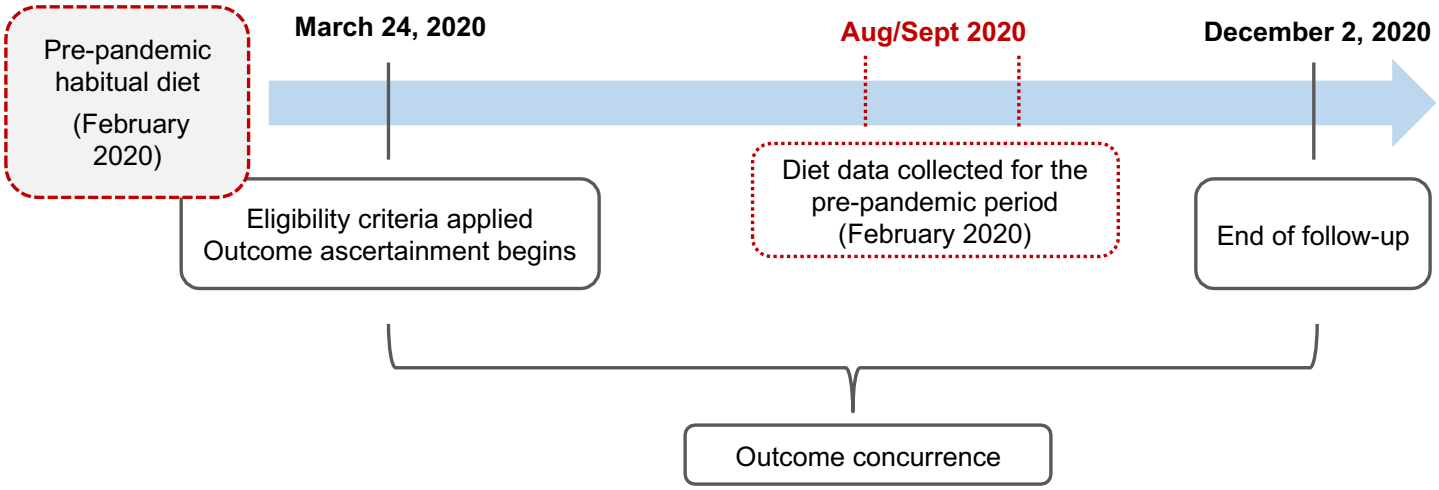


**Figure legend:** Schematic representation showing how and when diet and symptom data was collected. The diet survey was launched in August / September 2020 and queried about participant's habitual diet before (based on a time frame of February 2020) and during the pandemic (based on a time frame of July / August 2020). For the primary analysis, we used diet data deemed pre-pandemic. During follow-up, daily prompts queried for updates on interim symptoms, health care visits, and COVID-19 testing results.



Merino J, *et al. Gut* 2021;0:1–9. doi: 10.1136/gutjnl-2021-325353

**Supplementary figure 2: Pattern of missing data before multiple imputation**

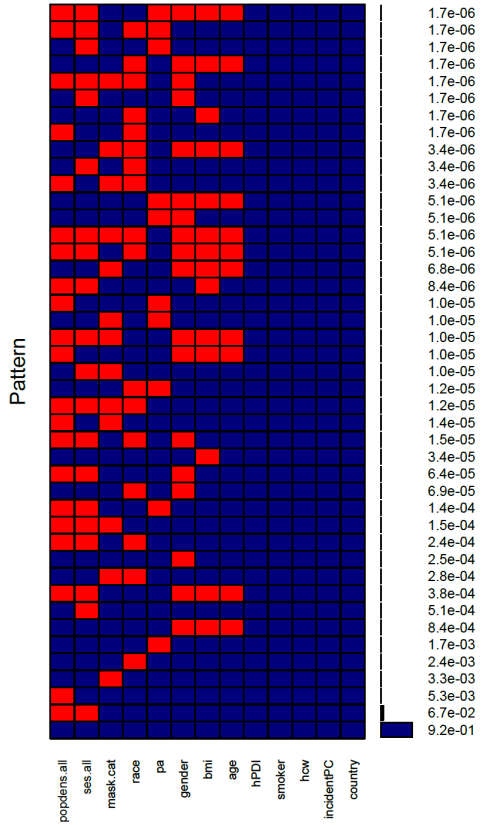


***Figure legend:*** *The plot shows the pattern of missing data across all variables and individuals included in this study. The values on the left side indicate the % of participants with missing data for combinations of variables. About 92% of included participants had complete information.*



Merino J, *et al. Gut* 2021;0:1–9. doi: 10.1136/gutjnl-2021-325353

**Supplementary figure 3: Directed acyclic graph depicting a possible scenario that could explain the association between diet quality and COVID-19 risk and severity.**

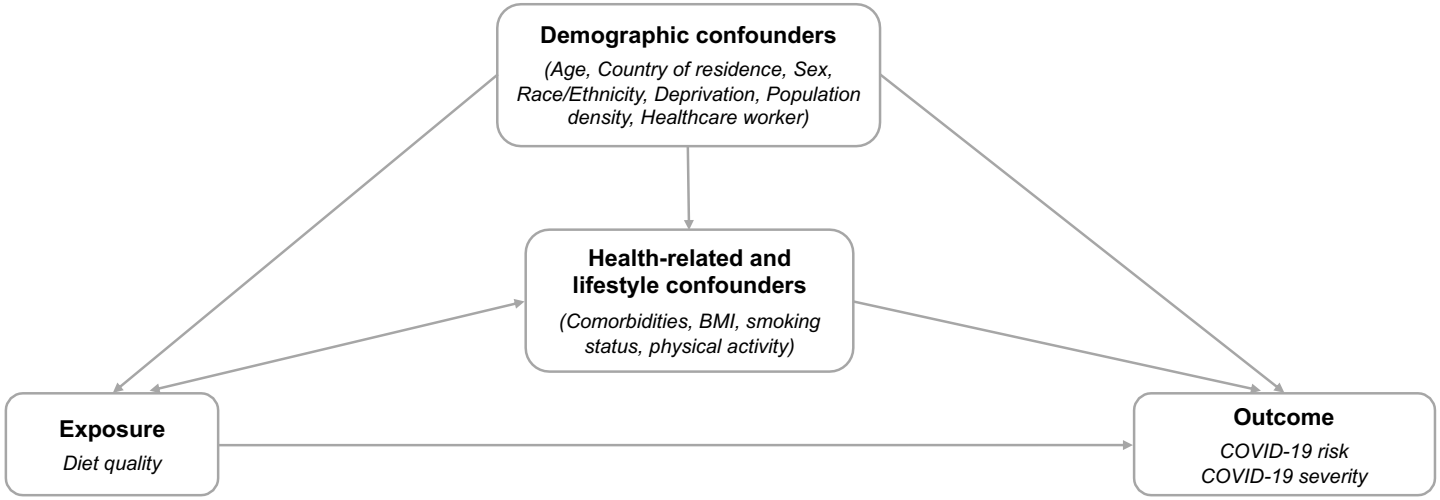


**Figure legend:** Directed acyclic graph showing the potential association between diet quality and COVID-19 risk and severity. Demographic confounders were included in the age-adjusted model and model 2. Model 3 was further adjusted for health-related and lifestyle confounders. In subgroup and sensitivity analyses, we investigated whether diet quality interacts with deprivation, and the extent to which BMI mediated the association between diet quality and COVID-19 risk.

**Supplementary figure 4: Flow diagram**

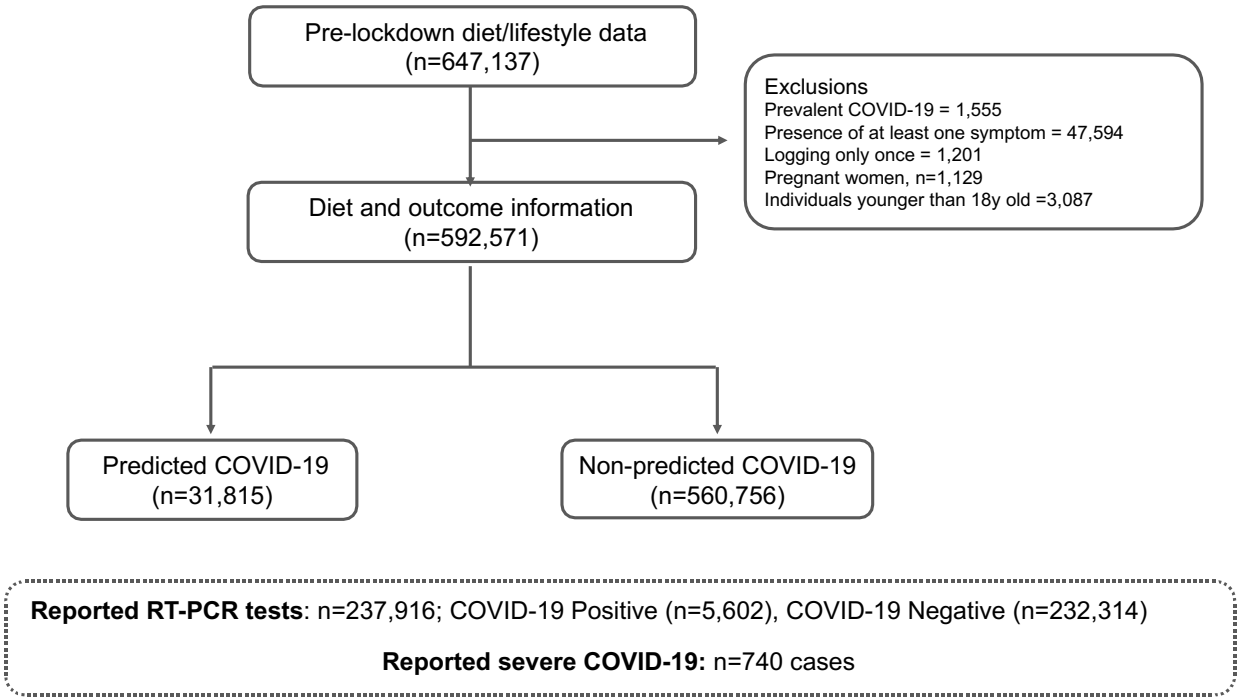


**Figure legend:** Identification of participants with diet and lifestyle data at baseline who met the eligibility criteria for this study. Number of cases and controls identified until the end of follow-up (December 2nd, 2020)

**Supplementary figure 5: Distribution of the hPDI score**

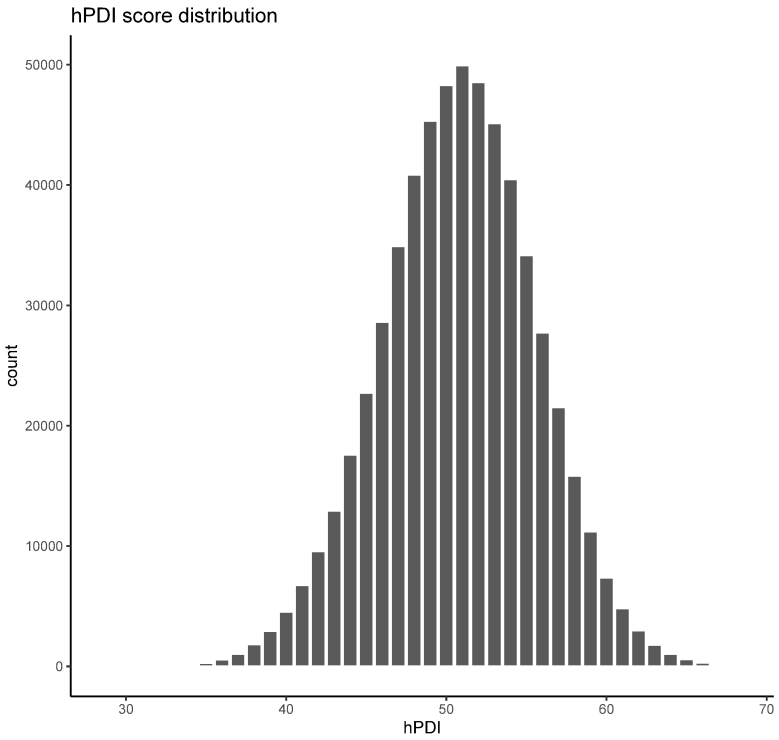


**Figure legend:** Distribution of the healthy plant-based diet index.

**Supplementary figure 6: Dose-response associations between diet quality and risk of COVID-19 infection.**

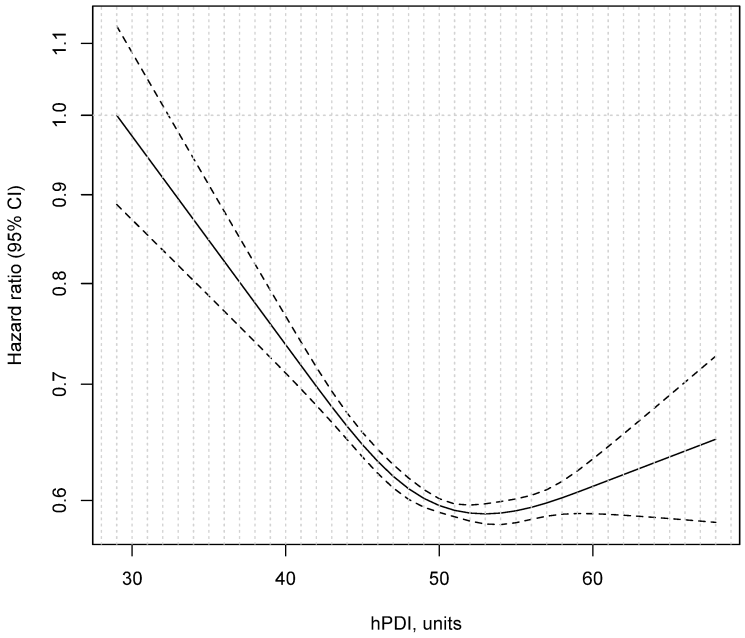


**Figure legend:** Dose-response associations between diet quality and risk of COVID-19 infection were calculated using restricted cubic splines with four knots (at the 2.5th, 25th, 75th, and 97.5th percentiles; methods). Cox models were adjusted for all confounders previously included in model 3. P for non-linearity <0.001.

**Supplementary figure 7: Absolute excess rate of COVID-19 per 10,000 person-months according to socioeconomic deprivation and diet quality**

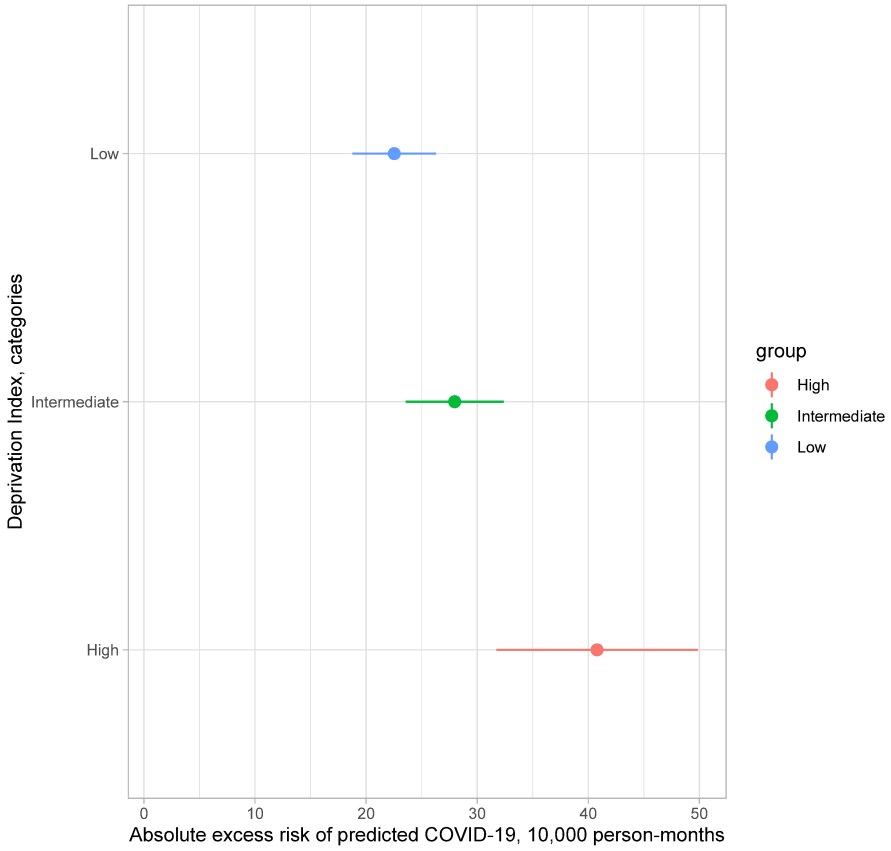


**Figure legend:** Absolute excess risk of COVID-19 per 10,000 person-months for lowest vs highest quartile of the diet score according to socioeconomic deprivation. Absolute excess risk was calculated based on the incidence rate per 1,000 person-months in each diet quality score and socioeconomic deprivation category using the "epiR" package in R.