App-based COVID-19 surveillance and prediction:

The COVID Symptom Study Sweden

Beatrice Kennedy, Hugo Fitipaldi, Ulf Hammar, Marlena Maziarz, Neli Tsereteli, Nikolay Oskolkov, Georgios Varotsis, Camilla A Franks, Lampros Spiliopoulos, Hans-Olov Adami, Jonas Björk, Stefan Engblom, Katja Fall, Anna Grimby-Ekman, Jan-Eric Litton, Mats Martinell, Anna Oudin, Torbjörn Sjöström, Toomas Timpka, Carole H Sudre, Mark S Graham, Julien Lavigne du Cadet, Andrew T. Chan, Richard Davies, Sajaysurya Ganesh, Anna May, Sébastien Ourselin, Joan Capdevila Pujol, Somesh Selvachandran, Jonathan Wolf, Tim D Spector, Claire J Steves, Maria F Gomez[ǂ], Paul W Franks[ǂ], Tove Fall[ǂ]

[ǂ] Shared senior authorship

Corresponding authors: Paul Franks, paul.franks@med.lu.se, Maria Gomez, maria.gomez@med.lu.se, Tove Fall, tove.fall@medsci.uu.se

ABSTRACT

Background

The response of the Swedish authorities to the COVID-19 pandemic was less restrictive than in most countries during the first year, with infection and death rates substantially higher than in neighbouring Nordic countries. Because access to PCR testing was limited during the first wave (February to June 2020) and regional data were reported with delay, adequate monitoring of community disease spread was hampered. The app-based COVID Symptom Study was launched in Sweden to disseminate real-time estimates of disease spread and to collect prospective data for research. The aim of this study was to describe the research project, develop models for estimation of COVID-19 prevalence and to evaluate it for prediction of hospital admissions for COVID-19.

Methods

We enrolled 143 531 study participants (≥18 years) throughout Sweden, who contributed 10·6 million daily symptom reports between April 29, 2020 and February 10, 2021. Data from 19 161 self-reported PCR tests were used to create a symptom-based algorithm to estimate daily prevalence of symptomatic COVID-19. The prediction model was validated using external datasets. We further utilized the model estimates to forecast subsequent new hospital admissions.

Findings

A prediction model for symptomatic COVID-19 based on 17 symptoms, age, and sex yielded an area under the ROC curve of 0·78 (95% CI 0·74-0·83) in an external validation dataset of 943 PCR-tested symptomatic individuals. App-based surveillance proved particularly useful for predicting hospital trends in times of insufficient testing capacity and registration delays. During the first wave, our prediction model estimates demonstrated a lower mean error (0·38 average new daily hospitalizations per 100 000 inhabitants per week (95% CI 0·32, 0·45)) for subsequent hospitalizations in the ten most populated counties, than a model based on confirmed case data (0·72 (0·64, 0·81)). The model further correctly identified on average three out of five counties (95% CI 2·3, 3·7) with the highest rates of hospitalizations the following week during the first wave and four out of five (3·0, 4·6) during the second wave.

Interpretation

The experience of the COVID Symptom Study highlights the important role citizens can play in real-time monitoring of infectious diseases, and how app-based data collection may be used for data-driven rapid responses to public health challenges.

Funding

Swedish Heart-Lung Foundation (20190470, 20140776), Swedish Research Council (EXODIAB, 2009-1039; 2014-03529), European Commission (ERC-2015-CoG - 681742 NASCENT), and Swedish Foundation for Strategic Research (LUDC-IRC, 15-0067) to MG or PF. European Research Council Starting Grant (801965) to TF. NO was financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. ATC was supported in this work through the Massachusetts Consortium on Pathogen Readiness (MassCPR). ZOE Limited provided in-kind support for all aspects of building, running and supporting the app and service to all users worldwide. Support for this study for KCL researchers was provided by the National Institute for Health Research (NIHR)-funded Biomedical Research Centre based at Guy's and St Thomas' (GSTT) NHS Foundation Trust. This work was supported by the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value-Based Healthcare. Investigators also received support from the Wellcome Trust, Medical Research Council (MRC), British Heart Foundation (BHF), Alzheimer's Society, European Union, NIHR, COVID-19 Driver Relief Fund (CDRF) and the NIHR-funded BioResource, Clinical Research Facility and Biomedical Research Centre (BRC) based at GSTT NHS Foundation Trust in partnership with KCL. ZOE Limited developed the app for data collection as a not-for-profit endeavour. None of the funding entities had any role in study design, data analysis, data interpretation, or the writing of this manuscript.

INTRODUCTION

In the midst of the COVID-19 pandemic, this global health crisis has stimulated abundant research that might help improve our response to future public health challenges. Amongst the most promising is the use of modern app-based technologies allowing real-time monitoring and prediction of the dynamics of the pandemic. Such an approach was indeed launched almost simultaneously in the US, UK, and Sweden with the COVID Symptom Study app [1,2]. The Swedish experience of the pandemic, which has received global attention, might be particularly informative because the public health response to the pandemic was much less restrictive than in most other countries [3], and also entailed a 4·5 to 10-fold higher death rate up to December 2020 than in neighbouring Nordic countries [4].

Community transmission of SARS-CoV-2 was confirmed in Sweden in early March 2020 [5]. By late March, visits to care homes as well as public gatherings of more than 50 people had been banned, universities and upper secondary schools had initiated distance learning, and adults were encouraged to work from home if possible. Other measures recommended by the European Centre for Disease Prevention and Control (ECDC), such as household quarantine when one individual in the household had confirmed COVID-19, were not implemented until the second wave in the autumn 2020. In addition, during the first pandemic wave, PCR testing was only available for hospital patients and healthcare staff [3], and assessments of national and regional prevalence were based on a few smaller PCR surveys performed by the Public Health Agency (Folkhälsomyndigheten, FoHM). Nationwide PCR testing for symptomatic adults was only introduced in June 2020 [3] and has since suffered delays throughout Sweden during periods of high demand.

The aim of the study was to collect symptom data across Sweden, develop models for estimation of symptomatic COVID-19 prevalence and to evaluate it for prediction of hospital admissions for COVID-19.

METHODS

COVID Symptom Study Sweden

The COVID Symptom Study was launched in Sweden on April 29, 2020, to provide COVID-19 infection surveillance data and to build a large-scale repeated measures database for COVID-19 research (Figure 1). More than 166 000 participants (2.4% of the adult population) joined CSSS in the first five weeks after launch. The non-commercial mobile application used in the study was initially developed by health data science company ZOE Limited in partnership with King's College London and Massachusetts General Hospital [1,2], and adapted for use in Sweden by ZOE Limited in collaboration with Lund University and Uppsala University. The app has been used to study the contemporary burden and predict future consequences of COVID-19 [6-8].

All individuals ≥18 years of age living in Sweden with access to a smart device have been eligible to participate in the CSSS by downloading the app and providing informed consent. Participants are asked to report year of birth, sex, height, weight, postal code, if they work in the healthcare sector, and to complete a health survey including pre-existing health conditions. Subsequently, participants are asked daily (with voluntary response frequency) if they feel "healthy as normal" or not, and to report the date and result of any COVID-19 PCR or serology test. If they do not feel healthy, they are asked about an array of symptoms potentially associated with COVID-19. Questions on COVID-19 vaccinations, added on March 27, 2021, are not included in the current analysis. All questionnaire variables are presented in Supplementary Table 1, with additional information available in the Supplementary Material.

In this analysis, we included data from April 29, 2020 to February 10, 2021. Individuals were excluded if they: 1) never submitted a daily report (n=5 931), 2) had missing age or reported an age <18 or >99 (n=801), or 3) stated their sex as other/intersex (n=236) as this sample size was insufficient for a separate analysis (Supplementary Table 2). Individuals whose last report was within seven days of joining the study (n=45 483) or did not provide a valid postal code (n=7 310) were excluded from the prevalence estimation, but included in model training if they had reported a PCR test and had submitted at least one symptomatic daily report within seven days preceding or on the test date (n=967). The final study population consisted of 143 531 individuals (Figure 2 and Supplementary Figure 1).

Symptom prevalences were investigated in participants with a positive (n=5 178) or negative (n=32 089) PCR test during 15 days preceding or following the test date. If a participant reported multiple PCR tests during the study period, only one randomly selected test was included.

Comparison with other data sources

We used external data sources to evaluate the representativeness of the CSSS cohort as a sample of the Swedish general population. Aggregate population demographic and socioeconomic information was obtained for all postal code areas from Statistics Sweden [9]. We calculated a neighbourhood deprivation index (NDI, lowest = most disadvantaged) for each postal code based on the proportion of adult inhabitants employed or studying, the proportion with a university education, and the median yearly net income [10] (for details, see Supplementary Material). We also obtained data from the Living Condition Surveys in Sweden from 2018-2019 [11] on current smoking status and BMI in the general population.

We compared symptom reporting from CSSS with data collected by NOVUS, a private company which conducts opinion polls and other surveys using panels recruited by random sampling of the Swedish population [12]. Since March 2020, NOVUS has carried out repeated surveys on COVID-19-related symptoms and healthcare contacts, not including PCR test results, with a response rate of approximately 70% [13]. While in the CSSS participants report symptoms on the same day they experience them, NOVUS participants report any symptoms experienced over the past 14 days even if these reflect their baseline health status (Supplementary Material).

Prediction model training

We developed a prediction model to estimate the probability of a positive PCR test in participants with symptoms. The model was based on participants who reported at least one PCR test between April 29 and December 31, 2020 and at least one reported candidate symptom within seven days before or on the test date (n=19 161, of whom 2 586 tested positive). Reports submitted during the first seven days after joining the study are excluded to reduce participation bias from increased motivation among symptomatic individuals. For participants who had not submitted daily reports, we assumed the last observation to be current for no more than seven subsequent days. If a participant submitted more than one report on a given day, all reports were combined into a single daily report; a symptom was treated as reported if it was mentioned in at least one of these reports.

We used an L1-penalized logistic regression model (LASSO) to select variables predicting symptomatic COVID-19. The shrinkage parameter was determined using ten-fold cross-validation and set to be the shrinkage that minimized the cross-validation error most. The starting set of predictors included all symptoms introduced through May 7, 2020 (excluding hay fever and chills or shivers), as well as their interaction with loss of smell and/or taste, as the latter constituted the strongest predictor of COVID-19 [6]. Predictors in the final model were: fever, persistent cough, diarrhoea, delirium, skipped meals, abdominal pain, chest pain, hoarse voice, loss of

smell and/or taste, headache, eye soreness, nausea, dizzy or lightheaded, red welts on face or lips, blisters on feet, sore throat, unusual muscle pains, fatigue (mild or severe), and shortness of breath (significant or severe), interaction terms between 14 of those and loss of smell and taste, as well as age and sex (see Supplementary Table 3 for model coefficients).

To better assess the time course of symptomatic COVID-19 retrospectively, we constructed a second time-dependent model, based on the first model with the addition of restricted cubic splines for calendar time with six knots placed according to Harrell's recommendations [14] (coefficients in Supplementary Table 4).

Prediction model validation

The models were internally evaluated in terms of discrimination and calibration using the ROC area under the curve (AUC) estimated by tenfold cross-validation within the dataset from April 29 to December 31, 2020, as well as based on CSSS data from a time period not used for model training (i.e., January 1 to February 10, 2021). Model calibration was assessed by plots with expected probabilities divided into deciles. For external validation, we used data from the CRUSH Covid study, which invites all individuals (≥18 years) to complete a symptom survey if they have conducted a COVID-19 PCR-test in the Uppsala healthcare county. Using data from October 18, 2020 to February 10, 2021, the classification ability was assessed using ROC analysis among individuals who had completed the survey on the day of the test, reported at least one symptom, and had a conclusive test result (n=943; see Supplementary Material and Supplementary Table 5).

Evaluation of prevalence estimates

The daily regional prevalence of symptomatic COVID-19 infection was estimated in real-time using a weighted mean of individual predicted probabilities for a given county weighted by age (<50 and ≥50 years) and sex. Participants not reporting any of the symptoms included in the prediction model were assigned a probability of zero. Participants with long-lasting COVID-19 symptoms were excluded after their 30th day of reporting loss of smell and/or taste to ensure that the estimates were not inflated due to COVID-19 sequelae. We then averaged the predicted probabilities and the number of participants reported for each subgroup, day, and county. Prevalence estimates were subsequently reweighed using direct standardization, yielding daily prevalence estimates for each county.

The odds ratios for all variables in the prediction model were assumed to be generalizable to the background population. Because the model was trained in a dataset with higher prevalence of COVID-19 compared to the general population, the intercept was inflated. We therefore recalibrated the model intercept until the nationwide

app-based predictions for 27 May, 2020 matched the estimated nationwide prevalence of 0·3% (95% CI 0·1-0·5%) between May 25-28, 2020 [15]. In that survey, performed by FoHM, self-sampling nasal and throat swabs with saliva samples were delivered to a random sample of 2 957 individuals (details provided in the Supplementary Material). We assumed both the sensitivity and the proportion of symptomatic COVID-19 in the FoHM survey to be 70% [16,17].

95% confidence intervals (95% CI) for predictions were generated using the function *ageadjust.direct* from the epitools package in R [18], using the method of Fay and Feuer [19]. This function accommodates the sum of the model-generated probabilities, number of participants for each of the four strata on a given day, and the total population of Sweden. The output is a weighted probability with a 95% CI. The method assumes that the sum of the model-generated probabilities is Poisson-distributed using an approximation based on the gamma distribution. Although this method may be regarded as conservative, we assume the FoHM point estimate for May 25-28, 2020 the PCR sensitivity and the proportion of asymptomatic individuals to be known quantities.

Prediction of cases and hospital visits

To compare CSSS prevalence estimates with reported confirmed cases by FoHM, we extracted a linked dataset of all COVID-19 cases from SmiNet, an electronic notification system of communicable diseases maintained by FoHM, where all confirmed cases are registered by law. To evaluate the ability to predict in- and out-patient hospital visits on a regional level, we acquired data from the National Patient Register from January 1, 2020 to January 4, 2021. Because the time lag for registrations of COVID-19 hospital visits was upto one month, we utilized data from the register until December 4, 2020 for these analyses.

We evaluated the agreement of CSSS-estimated prevalence with case notification rates and hospital trends by inspecting trend plots. We further applied a median regression model with either CSSS-estimated prevalence or case notification, modelled as linear exposures with new hospital admissions as the dependent variable, assessing the mean absolute prediction error (lower indicating higher accuracy) for both models during the first and second wave, respectively, using leave-one-out cross-validation (excluding one week from the model building for each iteration). We also calculated the Spearman correlation of CSSS prevalence and case notification to hospitalization the following week, during the first (up to July 6, 2020) and the second (from October 19, 2020) waves. We defined the end of the first wave based on the date on which there were fewer than three new ICU COVID-19 admissions per day and the start of the second wave when there were again three or more new ICU

admissions per day [20]. We also compared the average agreement between the five top-ranked counties in CSSS with the five counties with the most new cases and highest hospital notification rates the following week.

Ethical approvals

The Swedish Ethical Review Authority has approved CSSS (DNR 2020-01803 with addendums 2020-04006, 2020-04145, 2020-04451, and 2020-07080) and CRUSH Covid (DNR 2020-07080, and DNR 2020-04210 with addendum 2020-06315).

RESULTS

Descriptive characteristics

Table 1 shows characteristics of the 143 531 study participants. Compared with the general population, the study cohort included a larger proportion of women and fewer smokers and people aged ≥65; the cohort had a similar prevalence of obesity to the national average. The median duration of study participation was 151 days (IQR 52-252) with a median of 43 days with submitted reports (IQR 13-119). Thirty percent of participants reported at least one COVID-19 PCR test during the study period and 20% at least one serology test (Table 1). Six % of women and 4% of men reported a positive PCR test during the study period. CSSS participants resided in postal code areas with a higher median NDI, a similar proportion of inhabitants with foreign background, and a higher population density than the general population. The number of study participants per capita is depicted by county in Supplementary Figure 2.

The majority of CSSS participants with confirmed COVID-19 experienced loss of smell and/or taste, with headache, fever, and sore throat constituting other common symptoms (Figure 3a). Among participants who tested negative, the most common symptoms were headache and sore throat, whereas loss of smell and/or taste was rarely reported (Figure 3b). The non-adjusted prevalence of different symptoms was considerably higher in NOVUS than in CSSS, with the exception of loss of smell and/or taste, but temporal trends were similar (Supplementary Figure 3).

Training and validation of the prediction model

The final model selected by LASSO included 17 symptoms and sex, as well as 2-way interactions between loss of smell and/or taste and 14 symptoms and a 2-way interaction between loss of smell and/or taste and sex. The AUC for the main model was 0·76 (95% CI 0·75-0·78) during the training period and 0·72 (95% CI 0·69-0·75) during the evaluation period from January 1 to February 10, 2021. The AUC for the time-dependent model was 0·84 (95% CI 0·83-0·85) and 0·72 (95% CI 0·69-0·75) for the two time periods, respectively.

Out of the 2 116 participants in the CRUSH Covid study, 943 completed the survey and the COVID-19 test on the same day, reported at least one of the symptoms included in the CSSS model training, and had a conclusive test result (144 positive (15.3%)). The AUC for the main model was 0·78 (95% CI 0·74-0·83) and the AUC for the time-dependent model was 0·75 (95% CI 0·70-0·79). All calibration graphs are available in Supplementary Figure 4.

Prevalence estimates and prediction of cases and hospitalizations

The prevalence estimates of symptomatic COVID-19 based on data from the CSSS depicted in real-time the first and second waves of COVID-19 (Figure 4a). In contrast, the SmiNet data on laboratory-confirmed cases of COVID-19 did not detect the first wave (Figure 5a). Retrospective comparisons with national and regional register data on COVID-19 hospitalizations showed trends similar to CSSS estimates (Figure 5b, Supplementary Figures 5a and 5b), with a higher agreement on a national level observed for the retrospective time-dependent model (Figure 4b).

Overall, the average daily hospitalization rate per week ranged from 0-5 new patients per 100 000 inhabitants (≥18 years) in the five largest counties in Sweden during the study period (Supplementary figure 5b). During the first wave, our prediction model estimates demonstrated a lower mean error (0·38 average new daily hospitalizations per 100 000 inhabitants per week (95% CI 0·32, 0·45)) for subsequent hospitalizations in the ten most populated counties, than a model based on case notifications from SmiNet (0·72 (0·64, 0·81)). During the autumn, mean errors were similar (Table 2, Supplementary Table 6). The rank-based correlation of CSSS main model prevalence with next week hospital admission rate was 0·43 (0·24, 0·62) during the spring and 0·70 (0·49, 0·90) during the autumn of 2020 (Table 2). The main model further successfully identified three (95% CI 2.3, 3.7) out of five counties with the highest rates of hospitalizations the following week during the spring and four out of five (3.0, 4.6) during the autumn.

We observed a higher estimated prevalence of symptomatic COVID-19 in women than in men across the entire study period, which was most apparent in those aged ≤64 years (Figure 6a). Post-hoc analyses revealed that this difference was mainly driven by participants who were healthcare professionals, where women were over-represented (Figure 6b).

DISCUSSION

In this study, our main findings were two-fold. Firstly, app-based prediction estimates allowed for monitoring of COVID-19 prevalence on county level in Sweden before the general PCR testing programme was initiated and during gaps in reporting, and secondly, the prediction model could be used to forecast new COVID-19 hospitalizations.

Syndromic surveillance of a novel virus also enables study participants to report an array of symptoms which allows detection of new disease-specific symptoms [6] and symptom clusters associated with disease severity [21] or duration [7]. In the CSSS data, we observed that the most prevalent symptom in participants with PCR-confirmed COVID-19 was loss of smell and/or taste, which was rare in participants who tested negative, thus expanding and confirming other reports on COVID-19 symptomatology. In the event that novel SARS-CoV-2 variants are associated with other symptoms or symptom clustering, this change will also be captured in CSSS app data.

Covid Symptom Study data from the UK successfully identified hotspots in England, with findings validated in UK Government test data [8]. The CSSS identified a large proportion of the Swedish counties with highest hospital admissions the following week and predicted new hospital admissions with moderate accuracy. Similar to the UK, we found higher agreement to external hospital data sources during autumn with regards to ranking of regions compared with spring 2020. However, only the last part of the spring wave 2020 was captured because the study was initiated in late April.

The CSSS data collection method allowed rapid data analysis and dissemination of results. National and regional CSSS prediction estimates from interim models were shared daily with the public via the CSSS dashboard [22]. In addition, weekly summaries have been sent to health care leaders across Sweden since May 2020. In contrast, we estimate that the time interval between the first presentation of symptoms, confirmation via PCR testing, and the reporting of COVID-19 test data on the county and municipality level from the FoHM has taken at least 10 days throughout the study period, with larger delays during weekends, holidays and problems with the national SmiNet register during suspected data breaches. Furthermore, CSSS data have proven valuable when testing capacity is suddenly compromised, as happened in the fourth most populated county in Sweden during November 2020 [23]. Prevalence estimates derived from the CSSS data were valid during that period, as confirmed by the concurrent pattern of COVID-19 hospitalizations in the county. This highlights the need of multiple data sources for optimal surveillance.

More than 166 000 participants (2.4% of the adult population) joined CSSS in the first five weeks after launch, supporting the feasibility of large-scale app-based surveillance, which can be rapidly scaled-up without needing additional staff or costly resources. The highest rates of participation were in areas where the study's two founding universities are located (Supplementary Figure 5). Run as an academic initiative, the CSSS received some media attention in national and local press, but it was only explicitly endorsed by a handful of public health leaders. In a survey, only 18% of adults tested for COVID-19 in one of Sweden's larger cities, Uppsala, had heard about the CSSS app. However, among those aware of the CSSS, about half said they were already participants.

CSSS sought to fulfil the core principles of citizen science [24] by allowing participation in multiple stages of the scientific process, including defining research questions and study design, gathering and analysing data, and communicating results. Participant feedback through the CSSS Facebook page, email, and natural language processing of free-text symptom reporting was used to improve the content and design of the app and to expand the scope of the research questions. The CSSS dashboard received >8 000 visits per month, and our CRAN R package *covidsymptom* [25] for downloading aggregate data was downloaded >900 times (as of April 2021) and has been used by public health decision makers.

Although the use of a smart device app is intended to minimize barriers to enrolment [26], a lesser proportion of CSSS participants were male, aged ≥65 years or smokers, and had lower prevalence of comorbidities as compared with the general population, indicating overrepresentation of healthy individuals. Even though the rates of daily reporting in male and female CSSS participants were comparable, women were also more likely to report a PCR SARS-CoV-2 test, which aligns with other Swedish testing patterns (NOVUS [27] and CRUSH), as well as with international testing data from the UK and Canada [28,29]. Furthermore, when we applied the prediction model to the general population, we observed a higher estimated prevalence of symptomatic COVID-19 in women than in men, which was partially explained by higher COVID-19 risk among healthcare professionals that are more often female.

A limitation of the CSSS app is that, owing to limited resources, it is only available in Swedish, which precludes inclusion of non-Swedish speakers, who may be at high risk of COVID-19 infection [30]. It is also possible that participants were more likely to join the study and report daily if they experienced symptoms associated with COVID-19 than if they were healthy, potentially inflating COVID-19 prevalence estimates. We sought to reduce

this bias by excluding the first seven days of data collected for each participant. However, the regional ranking of hospital admissions was comparable or slightly less strongly correlated to those based on testing.

We observed a peak in app-based COVID-19 prevalence estimates in mid-September 2020 with no corresponding peak in any disease-specific COVID-19 national register data. The prevalence of loss of smell and/or taste, sore throat, and headache were similarly elevated in NOVUS. The FoHM also noted acute respiratory infections symptom reporting at this time [31]. Weekly laboratory analyses of respiratory viruses later indicated a high incidence of common colds caused by rhinoviruses in September 2020 [32]. Hence, the specificity of the CSSS data was compromised when prevalence of other pathogens with similar symptomatology to COVID-19 was elevated. We therefore developed a prediction model, which permitted time-varying coefficients conditional on the PCR test results during a given period. This model yielded results more consistent with the national COVID-19 incidence data. Because of the delay inherent in this type of analysis, the time-dependent model is not suitable for real-time COVID-19 surveillance; it is also ineffective when test positivity varies greatly across counties. A possible extension of this model would be to use seasonality of symptoms from other causes to improve the model. Exceptionally few cases of seasonal influenza were confirmed in Sweden in the winter season of 2020/2021 compared with previous years [31,32], which rendered the lower specificity during this period less problematic.

Conclusion

Citizen science represents a powerful and rapid asset when combatting public health emergencies. Our experience with CSSS suggests that app-based technologies should be incorporated into national research and public health efforts to understand and predict the impact of disease.

ACKNOWLEDGEMENTS

We thank all COVID Symptom Study participants whose participation, engagement, and feedback were essential to the study. We also thank ZOE Limited for excellent collaboration and development and maintenance of the app. We thank NOVUS for generously sharing their data with us.

We thank the CSSS team for their dedication and engagement during the set up and running of the study. In particular, we thank Anna-Maria Dutius Andersson and Mattias Borell for administrative and technical support; Ulrika Blom-Nilsson, Pernilla Siming, Diem Nguyen, and Riia Sustarsic for project management assistance; Sara Liedholm, Johanna Sandahl, Lars Uhlin, Caroline Runéus, and Katrin Ståhl, along with LU and UU communication teams, for dissemination and outreach; Jacqueline Postma and LU and UU legal teams provided valuable advice regarding the legal aspects of this project; Erik Renström and Stacey Ristinmaa Sörensen also provided valuable advice during the planning phase of the study. Koen Dekkers is thanked for valuable support with computational programming. The computations were enabled by resources in project sens2020559 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

This work was funded by Swedish Heart-Lung Foundation (20190470, 20140776), Swedish Research Council (EXODIAB, 2009-1039; 2014-03529), European Commission (ERC-2015-CoG - 681742 NASCENT), and Swedish Foundation for Strategic Research (LUDC-IRC, 15-0067) to MG or PF. European Research Council Starting Grant (801965) to TF. NO was financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. ATC was supported in this work through the Massachusetts Consortium on Pathogen Readiness (MassCPR). ZOE Limited provided in-kind support for all aspects of building, running and supporting the app and service to all users worldwide. Support for this study for KCL researchers was provided by the National Institute for Health Research (NIHR)-funded Biomedical Research Centre based at Guy's and St Thomas' (GSTT) NHS Foundation Trust. This work was supported by the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value-Based Healthcare. Investigators also received support from the Wellcome Trust, Medical Research Council (MRC), British Heart Foundation (BHF), Alzheimer's Society, European Union, NIHR, COVID-19 Driver Relief Fund (CDRF) and the NIHR-funded BioResource, Clinical Research Facility and Biomedical Research Centre (BRC) based at GSTT NHS Foundation Trust in partnership with KCL. ZOE Limited

developed the app for data collection as a not-for-profit endeavour. None of the funding entities had any role in study design, data analysis, data interpretation, or the writing of this manuscript.

DATA SHARING

Primary data in this study were collected by ZOE Limited and provided to CSSS under a data-sharing agreement. Additional data originated from the National Board of Health and Welfare, FoHM, Statistics Sweden, and NOVUS. Restrictions apply to the availability of data, which were used under license and ethical approval and are not publicly available. Data are, however, available from the authors upon reasonable request and with written permission from the Swedish Ethical Review Authority. Aggregate data is available for download at https://github.com/csss-resultat/covidsymptom [25].

CONTRIBUTORS

MG, PF, TF, BK, and UH conceived and designed the analysis. UH created and trained the prediction model, and together with HF, MM, NT, and NO also performed the data analyses. BK wrote the first draft of the manuscript. All authors together interpreted the findings, and reviewed, edited, and approved the final article. MG, PF, TF are the principal investigators, have had full access to all the data in the study, and accept final responsibility for the decision to submit for publication.

DECLARATION OF INTERESTS

CSSS is a strictly non-commercial research project. PF consults for and has stock options in ZOE Limited relating to the PREDICT nutrition studies, which are entirely separate from the COVID Symptom Study app development and COVID-19 research. ATC previously served as an investigator on the PREDICT nutrition studies. TDS is a consultant to ZOE Limited. TS is the current CEO and a shareholder at Novus International Group AB, Sweden. RD, JW, JCP, SG, AM, SS and JLC work for ZOE Limited. All other authors declare that they have no competing interests.

REFERENCES

1. Chan AT, Drew DA, Nguyen LH, et al. The COronavirus Pandemic Epidemiology (COPE) Consortium: A Call to Action. *Cancer Epidemiol Biomarkers Prev* 2020; **29**(7): 1283-9.
2. Drew DA, Nguyen LH, Steves CJ, et al. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* 2020; **368**(6497): 1362-7.
3. Ludvigsson JF. The first eight months of Sweden's COVID-19 strategy and the key actions and actors that were involved. *Acta Paediatr* 2020; **109**(12): 2459-71.
4. Claeson M, Hanson S. COVID-19 and the Swedish enigma. *Lancet* 2021; **397**(10271): 259-61.
5. Folkhalsomyndigheten. Flera tecken på samhällsspridning av covid-19 i Sverige. https://www.folkhalsomyndigheten.se/nyheter-och-press/nyhetsarkiv/2020/mars/flera-tecken-pa-samhallsspridning-av-COVID-19-i-sverige/ (accessed April 29, 2020).
6. Menni C, Valdes AM, Freidin MB, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med* 2020; **26**(7): 1037-40.
7. Sudre CH, Murray B, Varsavsky T, et al. Attributes and predictors of long COVID. *Nat Med* 2021; **27**(4): 626-31.
8. 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**(1): e21-e9.
9. Statistics Sweden. Statistics Sweden. https://www.scb.se/en/ (accessed April 29, 2021).
10. Messer LC, Laraia BA, Kaufman JS, et al. The development of a standardized neighborhood deprivation index. *J Urban Health* 2006; **83**(6): 1041-62.
11. Statistics Sweden. Undersökningarna av levnadsförhållanden (ULF/SILC). https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/undersokningarna-av-levnadsforhallanden-ulf-silc/ (accessed April 29, 2021).
12. NOVUS. Novus Sverigepanel. https://novus.se/metoder/sverigepanel/ (accessed April 29, 2021).
13. NOVUS. Novus Coronastatus 210226. https://novus.se/coronastatus-210226/ (accessed April 29, 2021).
14. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis: Springer; 2015.
15. Folkhalsomyndigheten. Förekomsten av covid-19 i Sverige 21–24 april och 25–28 maj 2020. https://www.folkhalsomyndigheten.se/contentassets/fb47e03453554372ba75ca3d3a6ba1e7/forekomstren-covid-19-sverige-21-24-april-25-28-maj-2020_2.pdf (accessed April 29, 2021).
16. Centers for Disease Control and Prevention. COVID-19 Pandemic Planning Scenarios. https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html (accessed March 26, 2021).
17. Woloshin S, Patel N, Kesselheim AS. False Negative Tests for SARS-CoV-2 Infection - Challenges and Implications. *N Engl J Med* 2020; **383**(6): e38.
18. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.
19. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. *Stat Med* 1997; **16**(7): 791-801.
20. Svenska Intensivvårdsregistret. Data & resultat. https://www.icuregswe.org/data--resultat/ (accessed April 29, 2021).
21. Sudre CH, Lee KA, Lochlainn MN, et al. Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app. *Sci Adv* 2021; **7**(12).
22. COVID Symptom Study Sverige. COVID Symptom Study Sverige - Dashboard. https://csss-resultat.shinyapps.io/csss_dashboard/ (accessed April 29, 2021).
23. Region Östergötland. Region Östergötland utökar provtagningen – personer över 70 först ut. https://www.regionostergotland.se/Om-regionen/Pressrum/Nyheter/region-ostergotland-utokar-provtagningen--personer-over-70-forst-ut/ (accessed May 5, 2021).
24. Robinson LD, Cawthray JL, West SE, Bonn A, Ansine J. Ten principles of citizen science. In: Hecker S, Haklay M, Bowser A, Makuch Z, Vogel J, Bonn A, eds. Citizen science: Innovation in open science, society and policy: UCL Press; 2018: 1-23.
25. Fitipaldi H. covidsymptom. 2021.
26. Pilemalm S, Timpka T. Third generation participatory design in health informatics--making user participation applicable to large-scale information system projects. *J Biomed Inform* 2008; **41**(2): 327-39.
27. NOVUS. Novus Coronastatus. https://novus.se/novus-coronastatus/ (accessed March 16, 2021).
28. Department of Health and Social Care. Demographic data for coronavirus testing (England): 28 May to 26 August https://www.gov.uk/government/publications/demographic-data-for-coronavirus-testing-england-28-may-to-26-august (accessed March 16, 2021).

29. Stall NM, Wu W, Lapointe-Shaw L, Fisman DN, Hillmer MP, Rochon PA. Sex-specific differences in COVID-19 testing, cases and outcomes: a population-wide study in Ontario, Canada. *medRxiv* 2020: 2020.04.30.20086975.
30. Drefahl S, Wallace M, Mussino E, et al. A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden. *Nature Communications* 2020; **11**(1): 5097.
31. Folkhalsomyndigheten. Veckorapport om covid-19, vecka 39. https://www.folkhalsomyndigheten.se/globalassets/statistik-uppfoljning/smittsamma-sjukdomar/veckorapporter-covid-19/2020/covid-19-veckorapport-vecka-39-final.pdf (accessed March 1, 2021).
32. Karolinska Universitetslaboratoriet. Luftvägspatogener. https://www.karolinska.se/globalassets/global/2-funktioner/funktion-kul/klinisk-mikrobiologi/epidemiologi/rapport-influensa--och-rs-virus-och-andra-luftvagspatogener.pdf (accessed March 1, 2021).

Table 1. Study population characteristics in COVID Symptom Study Sweden.

		All	Women	Men
N (%)[1]		143 531 (100)	89 545 (62·4)	53 986 (37·6)
Age, years*		48 (37, 59)	47 (35, 57)	50 (38, 61)
	Aged ≥65 (%)	22 272 (15·5)	11 641 (13·0)	10 631 (19·7)
Pregnant (%)		..	1251 (1·4)	..
BMI, kg/m^2*		25 (23, 28)	25 (22, 28)	26 (24, 28)
	Obese, BMI ≥30 kg/m^2 (%)	25 131 (17·6)	16 355 (18·4)	8 776 (16·3)
Current smoker (%)		9 291 (6·5)	6 650 (7·4)	2 641 (4·9)
Cardiovascular disease (%)		6 950 (4·8)	3 100 (3·5)	3 850 (7·1)
Antihypertensive medication (%)		23 526 (16·4)	12 168 (13·6)	11 358 (21·0)
Kidney disease (%)		1107 (0·8)	594 (0·7)	513 (1·0)
Diabetes mellitus (%)				
	Yes, type 1	941 (0·7)	515 (0·6)	426 (0·8)
	Yes, type 2	3 432 (2·4)	1 423 (1·6)	2 009 (3·7)
	Yes, gestational	9 (<1)	9 (<1)	0 (0·0)
	Yes, other	107 (0·1)	60 (0·1)	47 (0·1)
	Yes, type not specified	838 (0·6)	341 (0·4)	497 (0·9)
Lung disease (%)				
	Yes, asthma only	13 787 (9·6)	10 022 (11·2)	3 765 (7·0)
	Yes, both asthma and lung disease	913 (0·6)	664 (0·7)	249 (0·5)
	Yes, lung disease only	1 389 (1·0)	828 (0·9)	561 (1·0)
	Yes, type not specified	2 056 (1·4)	1444 (1·6)	612 (1·1)
Current cancer (%)		13 787 (9·6)	10 022 (11·2)	3 765 (7·0)
Immunosuppressive medication[2] (%)		5 817 (4·1)	3 926 (4·4)	1 891 (3·5)
Health care professional (%)				
	Interacts with patients	15 120 (10·5)	12 816 (14·3)	2 304 (4·3)
	Does not interact with patients	6 742 (4·7)	5 539 (6·2)	1 203 (2·2)
Months entering the study (%)				
	April-May 2020	122 765 (85·5)	76 039 (84·9)	46 726 (86·6)
	June-July 2020	11 016 (7·7)	7 307 (8·2)	3 709 (6·9)
	August-September 2020	22 29 (1·6)	1 455 (1·6)	774 (1·4)
	October-November 2020	5 761 (4·0)	3 638 (4·1)	2 123 (3·9)
	December 2020-January 2021	1 726 (1·2)	1 089 (1·2)	637 (1·2)
	February 2021	28 (<1)	14 (<1)	14 (<1)
Number of daily reports*		43 (13, 119)	43 (14, 116)	43 (13, 124)

Duration of study participation, days[3]*	151 (52, 252)	154 (53, 253)	147 (50, 252)
PCR test[4] (%)	43 501 (30·3)	30 702 (34·3)	12 799 (23·7)
Antibody test[4] (%)	29 208 (20·3)	19 216 (21·5)	9 992 (18·5)
NDI*	0·36 (-0·25, 1·02)	0·34 (-0·27, 1·00)	0·39 (-0·22, 1·07)
Foreign background, %*	19 (13, 27)	19 (12, 27)	19 (13, 27)
Population density, inhabitants/km^2*	1706 (357·5244)	1686 (334, 5190)	1729 (389, 5340·)

[1]Row percentage, [2]Corticosteroids, methotrexate and/or biological agents (treatment of cancer and/or rheumatic disease), [3]From first to last daily report, [4]At any time during follow-up, *Median (first and third quartile); BMI: Body Mass Index;

Table 2. Predictive capacity of weekly CSSS and case notification data for the following week hospitalizations in ten most populated counties* in Sweden during first and second wave, respectively.

Model		**First wave**	**Second wave**
Mean absolute deviation (lower better) from a median regression model (daily new hospital admissions per 100 000 inhabitants (≥18 years))			
	CSSS main model	0·38 (0·32, 0·45)	0·65 (0·52, 0·79)
	CSSS time-dependent model	0·38 (0·31, 0·45)	0·70 (0·56, 0·85)
	Official case notification	0·72 (0·64, 0·81)	0·55 (0·43, 0·66)
Spearman correlation of ranking of counties with following week hospitalization ranking			
	CSSS main model	0·43 (0·24, 0·62)	0·70 (0·49, 0·90)
	CSSS time-dependent model	0·44 (0·25, 0·62)	0·66 (0·44, 0·88)
	Official case notification	0·59 (0·42, 0·76)	0·73 (0·53, 0·92)
Ability to predict top-5 counties for next week hospitalization (number of correct, possible range 0-5)			
	CSSS main model	3 (2.3, 3.7)	4 (3.0, 4.6)
	CSSS time-dependent model	3.1 (2.4, 3.8)	4 (3.0, 4.6)
	Official case notification	4 (3.3, 4.5)	3 (2.1, 3.8)

*Stockholm, Västra Götaland, Skåne, Östergötland, Uppsala, Gävleborg, Jönköping, Västmanland, Värmland, Halland

Figure 1. COVID Symptom Study Sweden data flow.

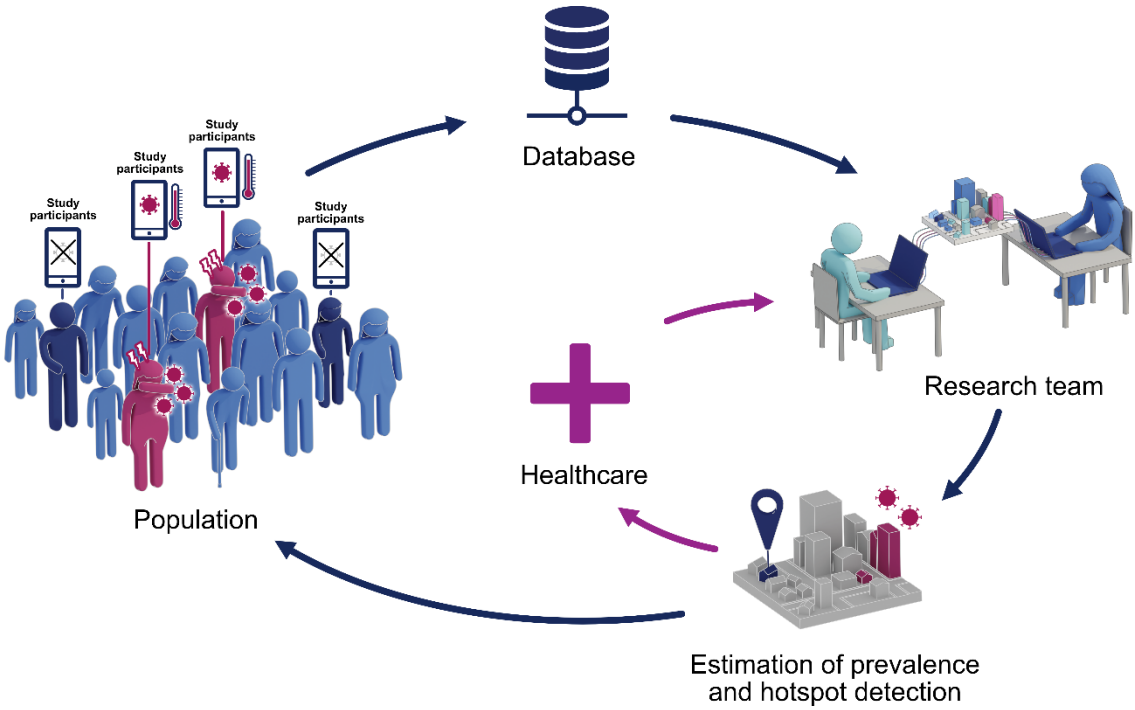
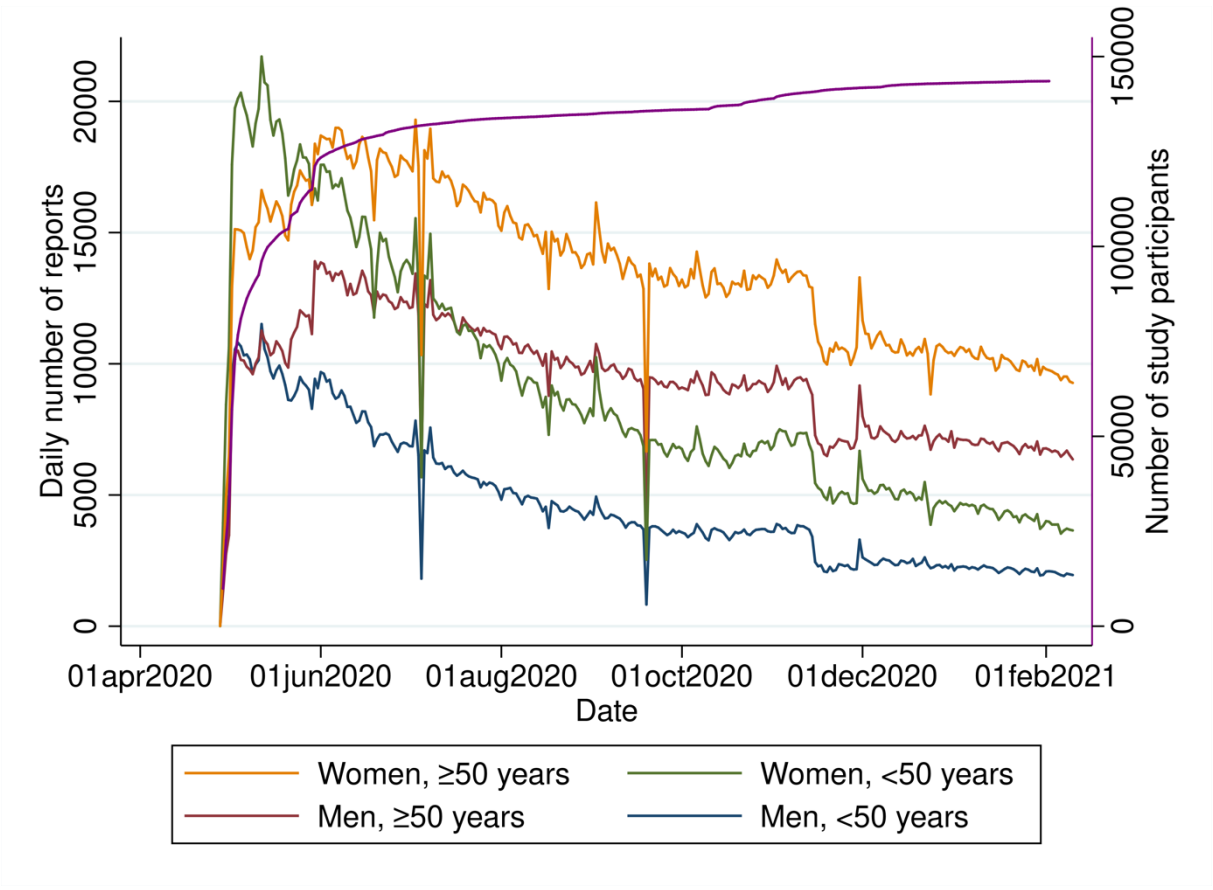


Figure 2. Cumulative number of study participants (n=143 531) after initial exclusions[ǂ], and number of daily reports from these participants stratified by sex and age (<50 and ≥50 years), in COVID Symptom Study Sweden during the study period April 29, 2020 to February 10, 2021.



ǂ Excluding individuals who: 1) did not submit any daily reports even at day of registration (n=5 931), 2) did not state age or self-reported an age at start of participation outside the range of 18-99 years (n=801), 3) stated their sex as other (n=200) or intersex (n=36).

* Temporary halt in data collection due to a technical issue in the COVID Symptom Study app

Figure 3. The prevalence of symptoms reported by participants in COVID Symptom Study Sweden with a) a negative PCR test for COVID-19 (n=32 089) and b) a positive PCR test for COVID-19 (n=5 178).

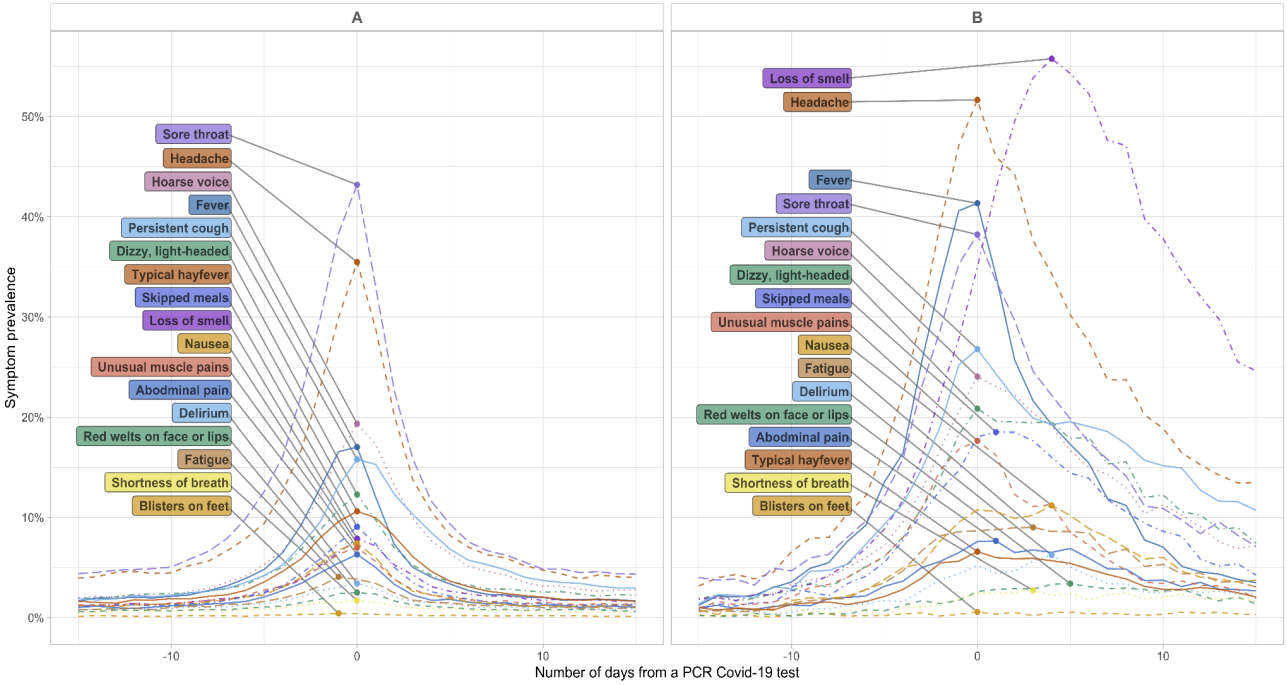
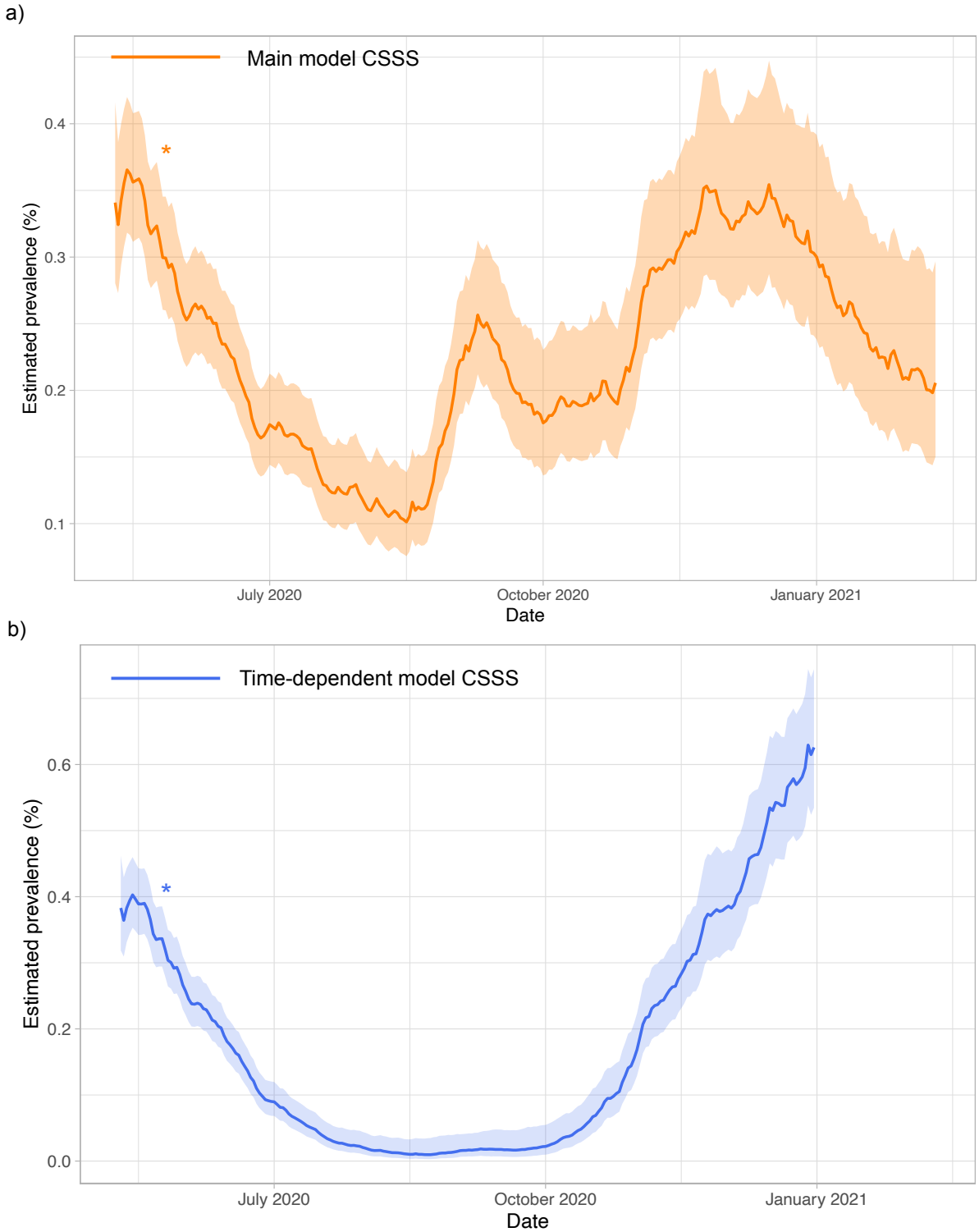


Figure 4. National prevalence estimates, with 95% confidence interval, of symptomatic COVID-19 in COVID Symptom Study Sweden, a) Main model (utilized for real-time prediction estimates), and b) Time-dependent model.



* Time-point for recalibration of CSSS estimated nationwide prevalence using national point prevalence survey findings from FoHM

Figure 5. National prevalence estimates, with 95% confidence interval, of symptomatic COVID-19 in COVID Symptom Study Sweden, combined with retrospective data on a) daily number of new COVID-19 cases registered in SmiNet, per 100 000 inhabitants, and b) daily number of new hospital admissions registered in the Patient Register, per 100 000 inhabitants.

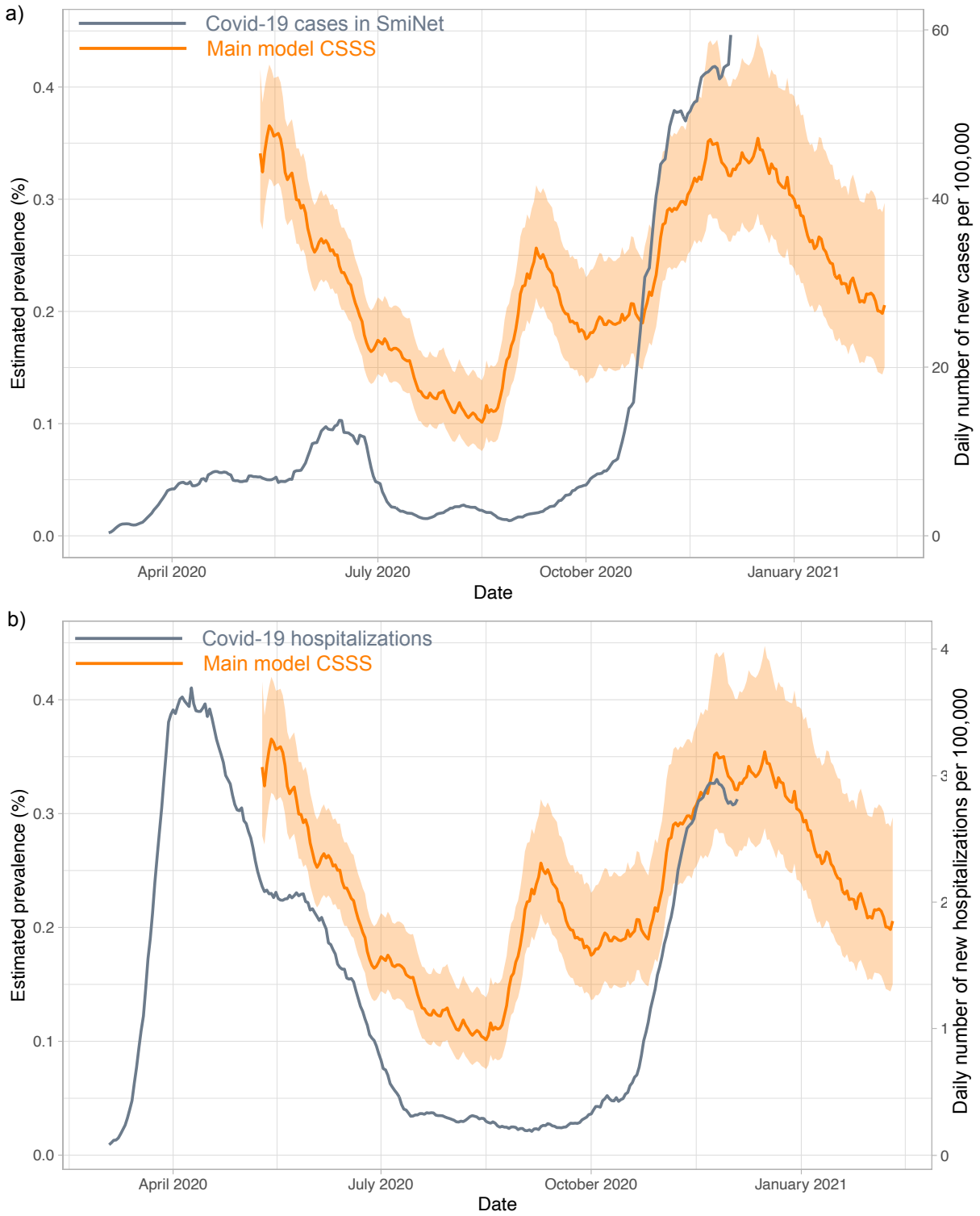
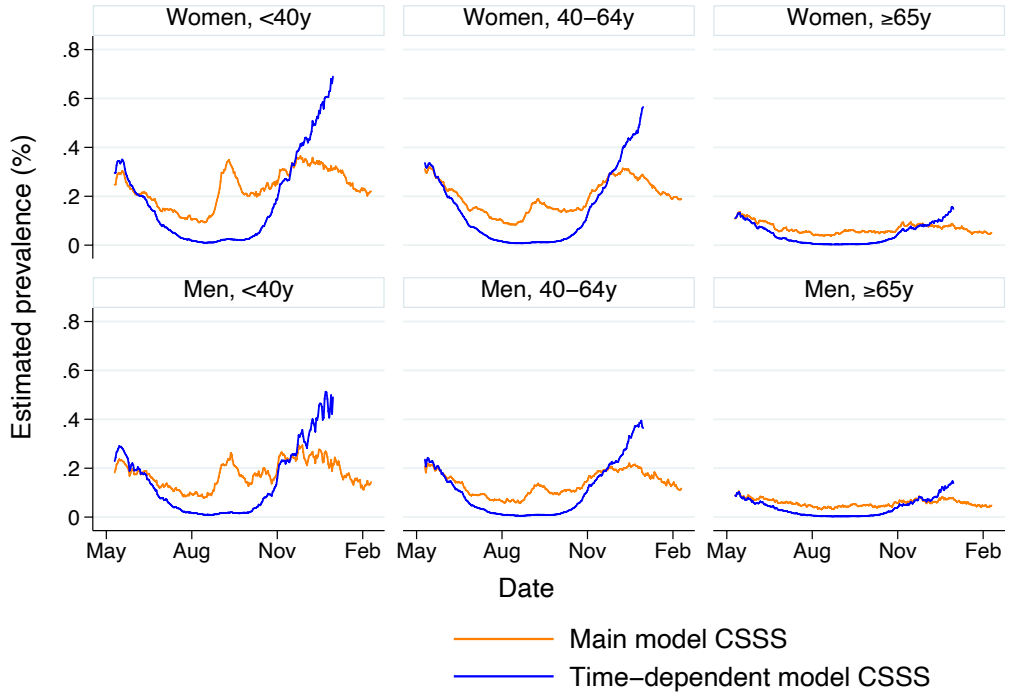
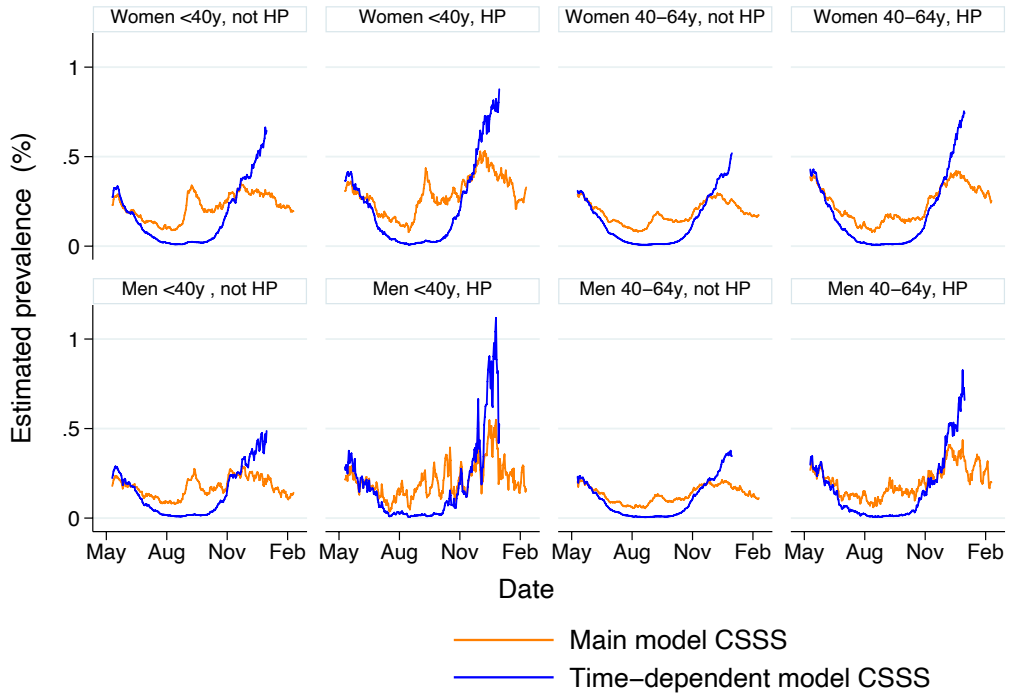


Figure 6. National prevalence estimates of symptomatic COVID-19 in COVID Symptom Study Sweden, depicting main model and time-dependent model, stratified by a) sex and age (18-39, 40-64, and ≥65 years) and b) sex and age (18-39, 40-64, and ≥65 years) and health care professional (HP).

a)



b)