App-based COVID-19 syndromic surveillance and prediction of hospital admissions:

The COVID Symptom Study Sweden

Beatrice Kennedy[1], Hugo Fitipaldi[2], Ulf Hammar[1], Marlena Maziarz[3], Neli Tsereteli[2], Nikolay Oskolkov[4], Georgios Varotsis[1], Camilla A Franks[3], Diem Nguyen[1], Lampros Spiliopoulos[3,5], Hans-Olov Adami[6,7,8], Jonas Björk[9,10], Stefan Engblom[11], Katja Fall[12,13], Anna Grimby-Ekman[14], Jan-Eric Litton[7], Mats Martinell[15,16], Anna Oudin[9,17], Torbjörn Sjöström[18], Toomas Timpka[19], Carole H Sudre[20,21,22], Mark S Graham[22], Julien Lavigne du Cadet[23], Andrew T. Chan[24], Richard Davies[23], Sajaysurya Ganesh[23], Anna May[23], Sébastien Ourselin[22], Joan Capdevila Pujol[23], Somesh Selvachandran[23], Jonathan Wolf[23], Tim D Spector[25], Claire J Steves[25], Maria F Gomez[3†], Paul W Franks[2†], Tove Fall[1†*]

[†]Joint senior authorship

Affiliations

1 Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Sweden

2 Genetic and Molecular Epidemiology Unit, Department of Clinical Sciences, Lund University Diabetes Centre, Lund University, Sweden

3 Diabetic Complications Unit, Department of Clinical Sciences in Malmö, Lund University Diabetes Centre, Sweden

4 Department of Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Lund University, Sweden

5 Skåne University Hospital, Malmö, Sweden

6 Clinical Effectiveness Group, Institute of Health and Society, University of Oslo, Oslo, Norway

7 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

8 Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

9 Division of Occupational and Environmental Medicine, Lund University, Sweden

10 Clinical Studies Sweden, Forum South, Skåne University Hospital, Lund, Sweden

11 Division of Scientific Computing, Department of Information Technology, Uppsala University, Sweden

12 Clinical Epidemiology and Biostatistics, School of Medical Sciences, Örebro University, Örebro, Sweden

13 Integrative Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

14 Biostatistics, School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

15 Department of Public Health and Caring Sciences, Uppsala University, Sweden

16 Primary Care and Health, Region Uppsala, Sweden

17 Department of Public Health and Clinical Medicine, Section of Sustainable health, Umeå University, Sweden

18 Novus International Group AB, Sweden

19 Department of Medical and Health Sciences, Linköping University, Sweden

20 MRC Unit for Lifelong Health and Ageing at UCL, University College London, London, UK

21 Centre for Medical Image Computing, University College London, London, UK

22 School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK

23 ZOE Limited

24 Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

25 Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

Corresponding author: Professor Tove Fall, EpiHubben, Uppsala University, 751 85 Uppsala, Sweden, tove.fall@medsci.uu.se

ABSTRACT

The app-based COVID Symptom Study was launched in Sweden in April 2020 to contribute to real-time COVID-19 surveillance. We enrolled 143,531 study participants (≥18 years) 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 model to estimate the individual probability of symptomatic COVID-19, with an AUC of 0.78 (95% CI 0.74–0.83) in an external dataset. These individual probabilities were used to estimate daily regional COVID-19 prevalence, which were in turn used together with current hospital data to predict next week COVID-19 hospital admissions. We found that this hospital prediction model demonstrated a lower median absolute percentage error (MdAPE: 25.9%) across the five most populated regions in Sweden during the first pandemic wave than a model based on case notifications (MdAPE: 30.3%). During the second wave, the error rates were similar. When applying the same model to an English dataset, not including local COVID-19 test data, we observed MdAPEs of 22.3% and 19.0%, respectively, highlighting the transferability of the prediction model.

INTRODUCTION

Real-time and accurate COVID-19 disease surveillance data is critical for adequate public health decision making and evaluation, as well as for healthcare system preparedness. The WHO guidelines for COVID-19 surveillance highlight the importance of combining data from multiple surveillance systems, and how participatory syndromic surveillance, where participants self-report symptoms of possible infection, may constitute a useful tool in early detection of disease outbreaks[1]. The European Centre for Disease Prevention and Control further notes that the utility of COVID-19 participatory syndromic surveillance may be enhanced if symptom data can be combined with information on testing[2]. Expanding knowledge on the feasibility of large-scale syndromic surveillance may thus enable tailored population-based participatory surveillance initiatives in future pandemics and epidemics.

Several novel eHealth solutions aimed at real-time monitoring and prediction of the dynamics of COVID-19 transmission were introduced early in the pandemic[3, 4, 5, 6], with app-based technologies quickly recognized as a potentially powerful surveillance tool. One of these technologies was the ZOE COVID Symptom Study app, designed to collect baseline health data as well as daily reports on symptoms and test results from study participants. The app was launched in the United Kingdom and in the United States in late March 2020[7, 8, 9].

Community transmission of SARS-CoV-2 was confirmed in Sweden in early March 2020. However, during the first pandemic wave in the spring of 2020, PCR testing was only available for hospital inpatients and healthcare workers[10] in Sweden and assessments of national COVID-19 prevalence were based on two PCR surveys performed by the Public Health Agency of Sweden in April (n=2,571) and May (n=2,957)[11]. Nationwide PCR testing for symptomatic adults was later introduced in June 2020[10], but suffered from various issues such as long analysis times during periods of high demand[12]. In response to the limited surveillance during the first pandemic wave, the COVID Symptom Study was launched in Sweden in April 2020.

The aims of this study were to develop and evaluate a syndromic surveillance-based framework to estimate the regional prevalence of COVID-19 and to evaluate if these could be used to accurately predict subsequent trends in COVID-19 hospital admissions. We showed that a model trained on symptoms and test data could provide informative prevalence estimates, and contribute to predictions of hospital burden the following week. Without using any additional test data, the hospital prediction model further performed well outside Sweden in a second country, England.

RESULTS

Descriptive characteristics

In this nationwide study on COVID-19 during the first year of the pandemic, we included data from 143,531 COVID Symptom Study Sweden (CSSS) participants ≥18 years from April 29, 2020 to February 10, 2021, who had contributed with at least one daily report in the app (Table 1). The median duration of study participation was 151 days (inter-quartile range [IQR] 52–252), with a median of 43 days with submitted reports (IQR 13–119). Of all participants, 30% reported at least one COVID-19 PCR test, and 6% of women and 4% of men reported at least one positive test result. The cohort included a larger proportion of women, fewer people ≥65 years, and fewer smokers than the general population, while the prevalence of obesity was similar. Participants further resided in postal code areas with less deprivation, similar proportions of inhabitants with foreign background, and higher population densities, as compared to the general population. The frequency of participants across regions is depicted in Supplementary Figure 1. CSSS was led by researchers at Lund University and Uppsala University, and the highest CSSS participation rates were observed in the regions of Skåne and Uppsala where these universities are located, as well in the most populated region in Sweden, Stockholm (Supplementary Table 1).

The most common symptoms reported in participants with PCR-confirmed COVID-19 were loss of smell and/or taste, headache, fever, and sore throat (Figure 1a). The prevalence of loss of smell and/or taste peaked at five days after the test date. Among participants who tested negative, headache and sore throat were most common, whereas loss of smell and/or taste was rarely reported (Figure 1b). The non-adjusted prevalence of different symptoms was considerably higher in the symptom data collected by the Swedish company NOVUS than in CSSS, with the exception of loss of smell and/or taste, but temporal trends were similar (Supplementary Figure 2). This discrepancy was likely due to the different approach in CSSS, where participants were asked questions about symptoms only if they responded negative to the first gate-keeper question, namely whether they felt healthy as normal.

Step 1. Training of model for estimation of individual probability of COVID-19

Our analysis strategy consisted of five steps, as illustrated in Figure 2. In the first step, we developed a model to estimate individual probability of symptomatic COVID-19, utilizing information from 19,161 CSSS participants who reported at least one PCR test (of whom 2,586 were COVID-19 infection positive) between April 29 and December 31, 2020; these individuals also reported at least one candidate symptom within seven days before or on the test date. 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 this main model was 0.76 (95% CI 0.75–0.78) during the training period April 29 to December 31, 2020; n=19,161) and 0.72 (95% CI 0.69–0.75) during the evaluation period (January 1 to February 10, 2021; n=1,753) and 0.78 (95% CI 0.74–0.83) in the external dataset of 943 symptomatic individuals from the CRUSH Covid survey (144 positive; test positivity 15.3%; October 18, 2020 to February 10, 2021). Calibration graphs are available in Supplementary Figure 3.

Step 2. Estimation of daily individual probability of COVID-19 in CSSS

We applied the model from Step 1 to estimate the daily individual probability of symptomatic COVID-19 in all CSSS study participants, including non-tested individuals, across the entire study period from May 10, 2020, through February 10, 2021. Non-symptomatic individuals were assigned a probability of 0 for that day.

Step 3. Daily regional COVID-19 prevalence estimates in the general population

The individual probabilities from Step 2 were then used to estimate the daily regional COVID-19 prevalence in the general population in Sweden, accounting for differences in age and sex distributions of the participants as compared to the general population in each region. We calibrated the intercept of the model generated in step 1 so that the estimated prevalence of May 27, 2020 matched the estimated prevalence from a national COVID-19 prevalence survey (n=2,957). The resulting CSSS prevalence estimates of symptomatic COVID-19 showed similar waves as the first and

second waves of COVID-19 hospitalization (Figure 3a, Supplementary Figure 4a). In contrast, data from SmiNet, the national Swedish register on laboratory-confirmed cases of COVID-19, did not capture the first wave (Figure 3b, Supplementary Figure 4b).

During the autumn of 2020, we observed a peak in CSSS-based COVID-19 prevalence estimates in September 2020 with no corresponding peak in other COVID-19 national case notification rates or hospital admission data. We therefore constructed a retrospective time-dependent model for individual probability of symptomatic COVID-19, based on the main model in Step 1 and validated correspondingly (additional information and calibration graphs available in the Supplementary Material). Retrospective national COVID-19 prevalence estimates based on the time-dependent model showed higher concordance with national COVID-19 case notification and hospital admission trends than the main model (Figures 3c and 3d).

We further 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 ≤64 years (Supplementary Figure 5a). Post-hoc analyses revealed that this difference was mainly generated by participants who were healthcare professionals, where women were over-represented (Supplementary Figure 5b).

Step 4. Prediction of regional COVID-19 hospital admissions the following week

We developed an iterative time-updated prediction model to assess if the regional prevalence estimates from Step 3 could be used together with current hospital data for prediction of regional COVID-19 hospital admissions seven days ahead. The parameters were estimated on available data through June 1, 2020 with larger weights applied to more recent observations, to predict admissions on June 8, 2020. This procedure was repeated to calibrate the coefficients throughout the study period. Weights and model specifications were based on exploratory analyses using data from May 11 to November 29, 2020. Overall, 16,752 individuals (≥18 years) were admitted to hospital with a diagnosis of COVID-19 from May 11 through November 29, 2020, and the number of daily new COVID-19 hospital admissions ranged from 0–104 across the 21 Swedish healthcare regions.

Across the five most populated regions in Sweden, the CSSS hospital prediction model demonstrated a median absolute percentage error (MdAPE) of 25.9% between daily predicted and observed number of hospital admissions for the first pandemic wave (May 11 to July 3, 2020; Figure 4 and Table 2), while the MdAPE for the second wave (October 19 to November 29, 2020) was 26.8%, which was lower than, or similar to, the predictions from a similar prediction model combining daily case notifications from SmiNet with hospital admissions yielded MdAPEs of 30.3% and 25.9% for the five most populated regions (first and second wave, respectively; Supplementary Table 2).

The MdAPEs were lowest in the most populated region in Sweden (Stockholm, population ≥18 years n=1.85 million) with 12.2% and 16.6% (SmiNet-based MdAPEs 13.5% and 24.5%) for the first and second waves, respectively. When we pooled data from all 21 Swedish regions, MdAPEs for the first and second waves were higher for both the CSSS hospital prediction model (37.0% and 42.4%, respectively; Supplementary Figure 6) and the SmiNet-based model (38.7% and 38.5%, respectively). Overall, we noted that the accuracy of the hospital prediction model as measured on the relative scale was lower when regional daily number of hospital admissions was low (Supplementary Figure 7a).

Step 5. Validation of the hospitalization prediction model in England

We sought to validate the CSSS-based hospitalization prediction model in England by repeating Steps 2 and 3 and parts of Step 4. The English dataset encompassed daily reports from 2,638,536 ZOE COVID Study (CSS UK) English study participants from March 30, 2020 to January 31, 2021 (study population characteristics and regional participation rates are available in Supplementary Tables 3 and 4). We extracted information on all COVID-19 hospital admissions (n=318,232) in individuals ≥18 years across the seven English healthcare regions from April 6, 2020 through February 7, 2021 from National Health Service England data. The number of new daily COVID-19 hospital admissions ranged from 0–958 across the English regions during this period.

We applied the exact same model that was developed in Step 1 in CSSS to estimate daily individual probability of COVID-19 in the English dataset, and then estimated the daily age- and sex-weighted COVID-19 prevalence across the English regions. We further applied the same iterative time-updated prediction model as in the Swedish dataset to predict hospital admissions the following week in the seven English regions. We used available outcome data up to April 27, 2020 to tune the parameters and to predict admissions a week later on May 4, 2020 and then repeated this daily throughout the study period (until February 7, 2021).

Across the seven English regions, we observed an MdAPE of 22.3% for the part of the first English pandemic wave captured in the data (April 6 to June 19, 2020) and an MdAPE of 19.0% for the second English wave (September 20, 2020 to February 7, 2021; Figure 5, Supplementary Table 5). As in Sweden, we observe lower error in the most populated English healthcare region (West and East Midlands; population ≥18 years n=10.8 million) with MdAPE of 16.1% and 14.0%, respectively. Overall, the predicted number of hospital admissions were overestimated when daily regional hospital admissions were low (Supplementary Figure 7b).

DISCUSSION

Adequate and continuous regional COVID-19 surveillance is challenging and requires multiple sources of data. Here, we developed an app-based framework that allowed for syndromic surveillance of COVID-19 at the national and regional level in Sweden across the first two pandemic waves. We found that CSSS prevalence estimates could be used to monitor COVID-19 disease trends, and that they were particularly informative in times of limited PCR testing capacity. The accuracy of the prevalence estimates was, however, lower in September 2020 when other respiratory infections peaked. We further showed that combining app-based regional prevalence estimates with previously recorded hospital data could, with moderate accuracy, predict regional levels of COVID-19 hospital admissions the following week both in Sweden and in England.

A previous study from ZOE COVID Study demonstrated how app data from the first pandemic wave from March through September 2020 could be utilized to successfully identify emerging COVID-19 hotpots in England, with findings validated in UK Government test data[13]. The present study confirmed the utility of app-based COVID-19 syndromic surveillance, encompassing the full second pandemic wave in the Swedish population and expanding the scope to include predictions of subsequent hospital admissions. The validation of the CSSS-based hospital prediction model in English data highlights the potential transferability of our approach, without using any PCR test data in the English data. Syndromic surveillance of COVID-19 may thus provide early warnings of surges in hospital admissions, thereby helping guide the allocation of precious healthcare resources in times of crisis.

A strength of the CSSS was that the syndromic surveillance using daily reports from study participants, which enables rapid data analysis and quick dissemination of results. The prevalence estimates were continuously disseminated to the study participants and the general public via the CSSS dashboard[14]. In contrast, the official COVID-19 disease surveillance in Sweden has suffered delays even after PCR testing was made available to the general public in June 2020. The time interval from booking a PCR test to confirmation of test result exceeded six days across several

regions during our study period[12], with additional time delays in the reporting of COVID-19 notification rates on regional and municipality levels by the Public Health Agency of Sweden[15].

The accuracy of the CSSS-based hospital prediction model during the first and second wave was higher for more populated regions in Sweden. Moreover, when the Swedish model was applied in England across healthcare regions, MdAPEs were lower than those derived in the Swedish setting. Although we cannot discern the separate influences of larger total population size, higher absolute number of study participants, and higher study participation rates, it is likely that all these factors enhance the accuracy of the hospital prediction model.

When we compared the CSSS-based hospital prediction model with the SmiNet-based (PCR test-based) hospital prediction model, we observed that the accuracy of the CSSS-based model was higher during the first wave, while the SmiNet-based model was similar in the second wave. The higher accuracy of CSSS during the first pandemic wave is likely due to the limited availability of PCR testing in Sweden at that time, when tests where only available to hospital inpatients and healthcare workers.[10] We conclude that in addition to the expansion of the national PCR testing programme introduced in June 2020 and subsequent delays in PCR testing, local factors may also influence how well the CSSS app and PCR testing efforts reflect regional and/or local outbreaks. However, the successful application of the non-test dependent hospital prediction model in England shows great promise for future efforts in syndromic surveillance, where models can be trained in one country with dense test data and adjusted to local trends of hospitalizations in a second country without the need for additional test data.

More than 166,000 participants (2.4% of the adult population in Sweden) joined CSSS in the first five weeks after launch, supporting the feasibility of large-scale app-based syndromic surveillance and the power of citizen science, which can be rapidly scaled-up without needing additional staff or large financial resources. However, although the use of a smart device app was intended to minimize barriers to enrolment[16], a lesser proportion of CSSS participants were male, ≥65 years, or smokers, or

obese, as compared to the general population, indicating an overrepresentation of healthy individuals. Further, owing to limited resources the CSSS app was in Sweden only available in Swedish, which precluded participation of non-Swedish speakers, a group that may be at higher risk of COVID-19 infection[17, 18]. We also observed that the CSSS participants resided in postal code areas with on average less deprivation than the general public. Together, these factors may have limited the ability of the CSSS to detect local outbreaks in vulnerable neighbourhoods, areas where lower community testing rates were also observed during the pandemic[12]. In future epidemic surveillance efforts, combining syndromic surveillance with non-participatory data sources, such as number of calls regarding specific complaints to nurse telephone consultation services[19], measurements of virus occurrence in wastewater[20, 21], monitoring of mobility patters in the population[22], and aggregate data on vaccination rates across neighbourhoods, may constitute a cost-efficient way to characterize community infection trends and predict increased demands on healthcare resources.

Additional potential limitations to the data collection also apply. Firstly, participants may have been more likely to join the study and report daily if they experienced symptoms perceived to be linked with COVID-19 than if they had been healthy, potentially inflating COVID-19 prevalence estimates. We sought to reduce this risk by excluding the first seven days of data collected for each participant, but we cannot exclude the risk of residual participation bias. Secondly, all data collected in the COVID Symptom Study app are self-reported. We had no means of linking the self-reported data to national population or health registers, so we could not validate information on COVID-19 tests or baseline health survey questions. We were able to overcome this limitation by validating the model predicting COVID-19 PCR test positivity in the independent CRUSH Covid dataset. Lastly, some questions in the COVID Symptom Study app were updated during the study period, potentially influencing data collection and analyses. The baseline health survey questions that were modified and/or updated during this time were, however, not included in our analyses and did not affect the prevalence estimates or the accuracy of the hospitalization prediction model. The symptom questions were updated on November 4, 2020. This update was implemented at the same time as incidence was

increasing sharply, but the national prevalence estimate curve and the regional prevalence estimate curves remained smooth, indicating that these updates did not materially influence our findings.

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 Public Health Agency of Sweden also noted high occurrence of symptoms of acute respiratory infections at this time[23]. Laboratory analyses of respiratory viruses later indicated a high incidence of common colds caused by rhinoviruses in September 2020[24]. Hence, the specificity of the CSSS data was compromised when prevalence of other pathogens with similar symptomatology to COVID-19 was elevated. We therefore added time as a variable in the model developed in Step 1, allowing estimated probabilities to vary depending on the PCR test positivity rate during a given period. This second model yielded results more consistent with the national COVID-19 incidence data but is strongly influenced by the proportion of positive COVID-19 tests. The model is thus rather insensitive to a sudden increase or decrease in the reporting of symptoms, and change in prevalence will not be fully captured until this trend is also reflected in the test results reported in the app. The more static main model will, conversely, react more quickly to an increase of reported symptoms, raising the sensitivity of the model but also the risk of false positive healthcare alerts. 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 regions, unless regions are modelled separately. An intermediate solution would be to retrain the model at known events that affect the relationship between symptoms and positive PCR tests, such as when vaccinations are introduced, new variants or other concurrent epidemics emerge. Very few cases of seasonal influenza were confirmed in Sweden in the winter season of 2020/2021 compared with previous years[25], which rendered the lower specificity during this period less problematic.

As the study period ended in February 2021, this study encompassed pandemic waves characterized in Sweden mainly by the early SARS-CoV-2 strains, with the variant of concern Alpha being detected in late December 2020[26]. Moreover, less than 150,000 inhabitants (<2% of the total population) had been inoculated with two doses of vaccine by early February 2021[27]. Further studies on prediction of COVID-19 hospital admissions, including subsequent variants of concern as well as higher vaccination rates in the general population, are therefore warranted.

Conclusion

App-based syndromic surveillance and citizen science may represent a powerful and rapid asset when assessing the early spread of a pandemic virus. The findings from COVID Symptom Study Sweden and validation in the English setting suggest that app-based technologies could contribute to national and regional disease surveillance and early warnings to healthcare systems.

METHODS

COVID Symptom Study Sweden (CSSS)

COVID Symptom Study Sweden (CSSS) was launched in Sweden on April 29, 2020 to provide COVID-19 syndromic surveillance data and to build a large-scale repeated measures database for COVID-19 research. 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[7, 8], 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 disease burden and to predict consequences of COVID-19[9, 13, 28].

All individuals ≥18 years living in Sweden with access to a smart device were eligible to participate in the CSSS after downloading the app and providing informed consent. Participants are asked to report year of birth, sex, height, weight, postal code, whether they work in the healthcare sector, and to complete a health survey including pre-existing health conditions. Subsequently, participants were asked daily (with voluntary response frequency) if they felt "healthy as normal" or not, and to report the date and result of any COVID-19 PCR or serology test. If they did not feel healthy, they were asked about an array of symptoms potentially associated with COVID-19. The symptoms that participants could report included, but were not limited to, loss of smell and/or taste, fever, persistent cough, fatigue, abdominal pain, chest pain, hoarse voice, shortness of breath, headache, muscle pains, skin rashes, nausea, chills, eye soreness, diarrhoea, and confusion. Unspecified symptoms could be added as free text.

The symptom questions were updated on November 4, 2020. The original question related to loss of smell and/or taste was then branched into a) loss of smell and/or taste, and b) altered smell and/or taste. This update was made to improve the specificity with which symptom severity was assessed. Several other new symptom questions were also introduced at this time, including for example

unusual hair loss and earache. Questions pertaining to COVID-19 vaccinations were added on March 27, 2021, and are not included in the current analysis. Information on all questionnaire variables both from the baseline health survey and the daily report, including information on whether the questions were available at launch, or the date when the questions were added and/or removed, are presented in Supplementary Table 6.

Study population and data management

In this study, we included data from April 29, 2020 to February 10, 2021. Participants were in this study excluded from the analyses if they: 1) never submitted a daily report (n=5,931), 2) had missing age or reported an age <18 years or >99 years (n=801), or 3) stated their sex as other or intersex (n=236) as this sample size was insufficient for a further analysis (Supplementary Table 7). Participants 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 reported a PCR test and 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 6). We labelled self-reported height, weight, and body mass index (BMI) as "missing" if <130 or >210 cm, <35 or >300 kg, and <15 or >70kg/m^2, respectively. "Obesity" was defined as BMI ≥30kg/m^2.

Symptom trajectories in study participants with negative and positive COVID-19 PCR tests

We described symptom prevalence and trajectories in all CSSS participants from the perspective of their positive (n=5,178) or negative (n=32,089) PCR test. We included all symptoms reported in the 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.

Representativeness of the study population

We assessed the representativeness of the CSSS study population as a sample of the Swedish general population. We obtained data from the Living Condition Surveys in Sweden from 2018–2019 on prevalence of current smoking and BMI in the general adult population[29]. We further extracted aggregate adult population sociodemographic information for all postal code areas in Sweden from Statistics Sweden, which included proportion of women, proportion of inhabitants ≥65 years, highest achieved education level, proportion engaged in work or studies in the population 20–64 years, median yearly net income for the population ≥20 years, and proportion of population ≥18 years with foreign background. Foreign background is here in defined as birth in a country other than Sweden, and/or birth in Sweden but with both parents born in a country other than Sweden. We then calculated a neighbourhood deprivation index (NDI, lowest corresponds to most disadvantaged) for each postal code area based on the proportion of adult inhabitants employed or studying, the proportion with a university education, and the median yearly net income[30]. We also calculated population density for all postal code areas in Sweden as the total number of inhabitants per square kilometres. The surface area of each postal code was estimated using shapefiles originating from an external company, Postnummerservice Norden AB, that contained detailed information on five-digit postal code area boundaries.

NOVUS

We compared overall daily symptom prevalence in CSSS during the study period with symptom data collected by NOVUS, a private company that conducts opinion polls and other surveys using panels recruited by random sampling of the Swedish population[31]. 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%[32]. 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 (for details, see Supplementary Material).

Statistical analyses strategy

Our analysis strategy consisted of five steps, as illustrated in Figure 2. Each step is detailed below.

Step 1. Training of model for estimation of individual probability of COVID-19

We trained a model to estimate the individual probability of a symptomatic COVID-19, defined as having symptoms and a positive PCR test result. The model training set consisted of data from 19,161 participants who reported at least one PCR test result (of which 2,588 were positive) between April 29 and December 31, 2020, and who reported at least one reported candidate symptom within seven days before or on the test date. As we observed that a higher proportion of study participants reported symptoms within the first week of joining the study than thereafter (Supplementary Figure 8), we excluded all reports submitted during the first week to reduce participation bias from increased motivation among symptomatic individuals in the general population. For participants who had not submitted reports every day, we assumed the last reported observation to be valid 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 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[9]. 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/or taste, as well as age and sex (see Supplementary Table 8 for model coefficients). After the loss of smell and/or taste symptom question was branched (November 4, 2020) into loss of smell and/or

taste and altered smell and/or taste, we combined these two new questions into a single variable, which we subsequently used interchangeably with the original loss of smell and/or taste variable.

As the primary aim was to provide individual probabilities of symptomatic COVID-19, rather than creating a classification rule, we followed the TRIPOD explanation and elaboration document (https://www.equator-network.org/reporting-guidelines/tripod-statement/) guidelines to report discrimination and calibration for the model. Discrimination and calibration were internally evaluated by applying nested tenfold cross-validation within the dataset from April 29 to December 31, 2020. We further assessed discrimination and calibration in CSSS data (1,753 participants, of which 339 tested positive) from January 1 to February 10, 2021, the period not included in model training. The nested tenfold cross-validation is described in detail in the Supplementary Material. Discrimination was quantified using the ROC area under the curve (AUC) as this metric is robust to differences in prevalence. Model calibration was assessed by plots with estimated probabilities divided into deciles.

To externally validate the model, we used data from the CRUSH Covid study, which invited all individuals (≥18 years) to complete a symptom survey when they did a COVID-19 PCR test in Region Uppsala, the fifth largest healthcare region in Sweden. Using data from October 18, 2020 to February 10, 2021, the classification ability of the model for individual probability of symptomatic COVID-19 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 9).

Step 2. Estimation of daily individual probability of COVID-19 in CSSS

We used the model from Step 1 to estimate the daily individual probability of symptomatic COVID-19 in the full CSSS study population, including individuals not reporting any PCR test, across the entire study period, from May 10, 2020, to February 10, 2021. 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 post-acute sequelae of COVID-19. As in Step 1, we a) excluded symptoms reported during the first seven days after a participant had joined the study, b) assumed the last report to be current for no more than seven days, and c) combined all reports on a given day into a single report.

Step 3. Daily regional age- and sex-weighted COVID-19 prevalence estimates

We estimated the daily regional prevalence of symptomatic COVID-19 infection in real-time using a weighted mean of individual predicted probabilities for each of the 21 administrative healthcare regions in Sweden re-weighted by age (<50 and ≥50 years) and sex in the total adult population using direct standardization. Demographic data from 2021 was available from Statistics Sweden. The 95% confidence intervals (95% CI) for predictions were generated using the function *ageadjust.direct* from the epitools package in R[33], using the method of Fay and Feuer[34]. This function accommodates the sum of the model-generated probabilities, number of participants for each of the four age- and sex strata on a given day, and the total population of Sweden. The method assumes that the sum of the model-generated probabilities is Poisson-distributed using an approximation based on the gamma distribution.

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 intercept of the model generated in Step 1 until the nationwide app-based predictions for May 27, 2020, matched the estimated nationwide prevalence of 0.3% (95% CI 0.1–0.5%) from a survey performed by the Public Health Agency of Sweden between May 25–28, 2020[11]. In that survey, 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 Public Health Agency of Sweden survey to be 70%[35, 36]. Although the method of Fay and Feuer[34] used to calculate the confidence intervals may be

regarded as conservative, we assumed the Public Health Agency of Sweden point estimate for May 25–28, 2020, PCR sensitivity, and proportion of asymptomatic individuals to be known quantities.

To compare CSSS prevalence estimates with reported confirmed cases in Sweden, we extracted a pseudonymised dataset of all COVID-19 cases ≥18 in individuals ≥18 years from SmiNet. SmiNet is an electronic notification system of communicable diseases maintained by the Public Health Agency of Sweden, to which all PCR-confirmed cases of COVID-19 are by law reported (details provided in the Supplementary Material). We also acquired data on all COVID-19 hospital admissions in Sweden from the National Patient Register from January 1, 2020 to January 4, 2021, including all individuals ≥18 years hospitalized with a first diagnosis of COVID-19: International Statistical Classification of Diseases and Related Health Problems Tenth Revision [ICD-10] codes U07.1 and U07.2 (for details see Supplementary Material). The delay in registering COVID-19 hospital admissions in the National Patient Register was up to one month. We therefore utilized data from the register until December 4, 2020. We evaluated the agreement of a) CSSS-estimated prevalence (main model) b) CSSS-estimated prevalence (time-dependent model) with official case notification rates from SmiNet and daily new hospital admissions per 100,000 inhabitants ≥18 years (seven-day moving average) on national and regional levels by inspecting trend plots.

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. However, this peak coincided with regional reports of rhinovirus surges. To better estimate the time course of symptomatic COVID-19 retrospectively from the app data, we constructed a time-dependent model for individual probability of symptomatic COVID-19, based on the variables utilized in Step 1 and with data from the same time period (from April 29 to December 31, 2020) with the addition of restricted cubic splines for calendar time with six knots placed according to Harrell's recommendations[37] (coefficients in Supplementary Table 10). Discrimination and calibration were assessed in CSSS data from January 1 to February 10, 2021, which is a period that was not included in model training.

Consistent with the main model, we also performed nested tenfold cross-validation, described in detail in the Supplementary Material, and independently validated the model in the CRUSH Covid dataset.

Step 4. Predictions of regional COVID-19 hospital admissions the following week

We assessed the ability of the CSSS prevalence estimates to predict the following week's COVID-19 hospitalizations during the first and second pandemic waves in Sweden. We defined the end of the first pandemic wave as when the rate of daily new hospital admissions in Sweden dropped below 0.5 individuals per 100,000 inhabitants ≥18 years (July 3, 2020), and the beginning of the second wave as when the hospital admission rate again rose above that threshold (October 19, 2020).

We utilized data from the entire study period in Swedish data to assess which variables to include in the model predicting hospital admission seven days ahead. The candidate variables were: daily regional CSSS prevalence estimate on day 0 or -1; current regional rate of COVID-19 hospitalizations per 100,000 inhabitants ≥18 years on day 0 or -1; weekday of hospitalization (Monday through Sunday); mean age in region; and mean regional Neighbourhood Deprivation Index (NDI). The final model was a weighted linear regression model, including daily regional CSSS prevalence estimate on day 0 and daily regional rate of COVID-19 hospitalizations per 100,000 inhabitants ≥18 years on day 0, assuming both variables to have a linear relationship with the outcome. Mean age and mean NDI did not notably influence the predictions and were therefore not included. To create weights for the linear regression we multiplied the number of inhabitants ≥18 years in each region with a density assigned to each day. The density was created using a Gaussian kernel function, with the highest density for the most recent observation. The standard deviation for the kernel function was set to be two days by trial-and-error. Weights for days more recent than seven days prior to the observation being forecasted was set to zero.

The model can be portrayed with the following equation, where β_0 represents the intercept, "prevalence" the CSSS prevalence and "hospital" the admissions/100,000 inhabitants ≥18 years. The subscript t_0 represents day 0 and the subscript t_7 day 7.

Equation 1.

$$hospital_t_7 = \beta_0 + \beta_{prevalence} \times prevalence_t_0 + \beta_{hospital} \times hospital_t_0$$

We developed an iterative time-updated prediction model process by first training the model in available outcome data up to the first day 0 (June 1, 2020) to derive the coefficients β0, $\beta_{prevalence}$ and $\beta_{hospital}$ for Eq. 1. We inserted the CSSS prevalence estimate and daily regional rate of COVID-19 hospitalizations per 100,000 inhabitants ≥18 years on June 1, 2020 into the equation, applying the derived coefficients to predict hospitalization rates on June 8, 2020. We then repeated the model fit and prediction from June 2 to November 29, 2020, with past data influencing the daily new intercept as well as the daily new two betas.

To evaluate the prediction model performance, we transformed both observed and predicted rates to number of hospitalizations per day and region. We then calculated the median absolute percentage error (MdAPE) between predicted and observed number of hospitalizations across the first wave (May 11 to July 3, 2020), the summer period (July 4 to October 18, 2020), and the second wave (October 19 to November 29, 2020), for each of the 21 healthcare regions as well as for the five most populated regions combined (Stockholm, Västra Götaland, Skåne, Östergötland and Uppsala). If a region had zero new hospitalizations on any day, that data point was excluded when calculating the absolute percentage error.

For comparison, we further created a second regression model, where we instead of the regional CSSS prevalence estimates inserted the daily regional case notification rates as registered in SmiNet (on day 0) in the equation. We evaluated both date of PCR test and date of registration in SmiNet in the model. When using date of PCR test, we only counted tests that were registered before or on day 0 in each time-updated iteration. Using the date of registration provided better predictions. We

found the same density and weights suitable for the SmiNet model as for the CSSS-based model. Model fit was repeated daily across the same period. We further plotted the relative error in % per day and region in hospital admission predictions over the range of observed admissions.

Step 5. Validation of hospital prediction model using external data from England by repeating Steps 2–4

We repeated Step 2 in English data by applying the Swedish model for estimation of individual probabilities from Step 1, applying the same symptom coefficients as in Sweden, and thereby assessed the daily individual probability of symptomatic COVID-19 in ZOE COVID Study (CSS UK) English participants ≥18 years from March 30, 2020 to January 31, 2021. We included participants residing in all postal code areas (n=2,261) within any of the seven English healthcare regions (South East, London, North West, East of England, South West, West and East Midlands, Yorkshire and the Humber and North East; total population ≥18 years n=43.2 million). If postal code areas overlapped two or more regions the postal code area was randomly assigned to one of them in our analyses.

We further repeated Step 3 and assessed daily age- and sex-weighted averages of the individual probabilities to estimate daily COVID-19 prevalence across the seven English healthcare regions. Demographic data was extracted from the UK Office for National Statistics[38].

We concluded by seeking replication of Step 4 by applying Eq. 1 in the English dataset using the equivalent variables used in the analysis of CSSS data. We trained the model in available outcome data up to the first day 0 (April 27, 2020) to derive the coefficients β_0, $\beta_{prevalence}$, and $\beta_{hospital}$. We inserted the ZOE COVID Study prevalence estimate and daily regional rate of COVID-19 hospitalizations per 100,000 inhabitants ≥18 years on April 27, 2020 into the equation, applying the derived coefficients to predict hospitalization rates on May 4, 2020. We then repeated the model fit and prediction from May 5, 2020 to February 11, 2021, with past data influencing the daily new intercept as well as the daily new two betas.

Information on COVID-19 hospital admissions from April 6, 2020 to February 7, 2021, was obtained from the UK government COVID-19 dashboard[39]. We defined the end of the first wave (June 19, 2020), and the start of the second (September 20, 2020), using the same threshold as in Swedish data. We calculated the MdAPEs between predicted and observed number of hospitalizations for each of the seven English healthcare regions during the first wave (April 6 to June 19, 2020), the summer period (June 20 to September 19, 2020), and the second wave (September 20, 2020 to February 7, 2021).

Ethical approvals

The Swedish Ethical Review Authority has approved both the CSSS (DNR reference number 2020-01803 with addendums 2020-04006, 2020-04145, 2020-04451, 2020-07080, and 2021-02316) and CRUSH Covid (DNR reference number 2020-07080, and DNR 2020-04210 with addendum 2020-06315). In the United Kingdom, the CSS UK has been approved by King's College London (KCL) ethics committee REMAS ID 18210, review reference LRS-19/20-18210.

Data availability

Primary data in this study were collected by ZOE Limited and provided to CSSS under a data-sharing agreement. Additional anonymized data originated from the National Board of Health and Welfare, the Public Health Agency of Sweden, Statistics Sweden, and NOVUS. Restrictions apply to the availability of these data, which were used under license and ethical approval and are not publicly available. Pseudonymized individual-level data are, however, available from the authors upon reasonable request and with written permission from the Swedish Ethical Review Authority. Group-level prevalence estimates calculated from individual-level data, hospitalization data, and cases of COVID positivity are available as the Source Data (Figures.zip), which were used to generate Figures

2–5. We provide mock CSSS individual level datasets, that can be used together with the code. These data do not represent real observations. The Source Data that support the findings of this study are available on GitHub at https://github.com/ulfha881/App-based-COVID-19-syndromic-surveillance-and-prediction-of-hospital-admissions-The-COVID-Symptom-S, or, in archived form, at [40].

Code availability

All code necessary for the replication of our results, including reproducibility instructions, is available on GitHub at https://github.com/ulfha881/App-based-COVID-19-syndromic-surveillance-and-prediction-of-hospital-admissions-The-COVID-Symptom-S, or, in archived form, at [40]. R and Stata codes are provided that will allow readers to generate Figures 2–5 from Source Data (Figures.zip). Codes to calculate prevalence estimates and the prediction models are provided, but without unmodified individual level datasets, they can't be executed in full.

ACKNOWLEDGEMENTS

We thank all COVID Symptom Study Sweden 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, and Riia Sustarsic for project management support; Sara Liedholm, Johanna Sandahl, Lars Uhlin, Caroline Runéus, and Katrin Ståhl, along with Lund University and Uppsala University communication teams, for dissemination and outreach. Jacqueline Postma and Lund University and Uppsala University 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.

Funding

Swedish Heart-Lung Foundation (20190470, 20140776), Swedish Research Council (EXODIAB, 2009-1039; 2014-03529), European Commission (ERC-2015-CoG - 681742 NASCENT), eSSENCE@LU 8:8 (eSSENCE - The e-Science Collaboration), Swedish Foundation for Strategic Research (LUDC-IRC, 15-0067), Crafoord Foundation (20211011), and EUGLOHRIA (101017752) to MG and/or PF, and 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). JB was financially supported by Swedish Research Council (2019-00198 and 2021-04665) and by Sweden's Innovation Agency (Vinnova; 2021-02648). 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 also 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.

CONTRIBUTORS

MG, PF, TF, BK, and UH conceived the study and designed the analysis. UH created and trained the prediction model, and together with HF, MM, NT, and NO performed the data analyses. BK, TF, and UH wrote the first draft of the manuscript. All authors together interpreted the findings, and reviewed, edited, and approved the final article. MG, PF, and 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.

MG, PF, and TF share senior authorship.

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. World Health Organization. Public health surveillance for COVID-19: Interrim Guidance. 16 December 2020. Preprint at https://apps.who.int/iris/bitstream/handle/10665/337897/WHO-2019-nCoV-SurveillanceGuidance-2020.8-eng.pdf (2020).

2. European Centre for Disease Prevention and Control. COVID-19 surveillance guidance. October 2021. Preprint at https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-surveillance-guidance.pdf (2021).

3. Denis F*, et al.* Epidemiological Observations on the Association Between Anosmia and COVID-19 Infection: Analysis of Data From a Self-Assessment Web Application. *J Med Internet Res* **22**, e19855 (2020).

4. Timmers T, Janssen L, Stohr J, Murk JL, Berrevoets MAH. Using eHealth to Support COVID-19 Education, Self-Assessment, and Symptom Monitoring in the Netherlands: Observational Study. *JMIR Mhealth Uhealth* **8**, e19822 (2020).

5. Yoneoka D*, et al.* Early SNS-Based Monitoring System for the COVID-19 Outbreak in Japan: A Population-Level Observational Study. *J Epidemiol* **30**, 362-370 (2020).

6. Rossman H*, et al.* A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. *Nat Med* **26**, 634-638 (2020).

7. Chan AT*, et al.* The COronavirus Pandemic Epidemiology (COPE) Consortium: A Call to Action. *Cancer Epidemiol Biomarkers Prev* **29**, 1283-1289 (2020).

8. Drew DA*, et al.* Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* **368**, 1362-1367 (2020).

9. Menni C*, et al.* Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nat Med* **26**, 1037-1040 (2020).

10. Ludvigsson JF. The first eight months of Sweden's COVID-19 strategy and the key actions and actors that were involved. *Acta Paediatr* **109**, 2459-2471 (2020).

11. Public Health Agency of Sweden. The prevelance of COVID-19 in Sweden April 21-24 and May 25-28, 2020 (Förekomsten av covid-19 i Sverige 21–24 april och 25–28 maj 2020). https://www.folkhalsomyndigheten.se/contentassets/fb47e03453554372ba75ca3d3a6ba1e7/forekomsten-covid-19-sverige-21-24-april-25-28-maj-2020_2.pdf Accessed: April 29, 2021.

12. Almgren M, Björk J. Mapping of differences in the regions' efforts for sampling and contact tracing during the COVID-19 pandemic (Kartläggning av skillnader i regionernas insatser för provtagning och smittspårning under coronapandemin). https://coronakommissionen.com/wp-content/uploads/2021/10/underlagsrapport-m-almgren-kartlaggning-av-skillnader-i-regionernas-insatser-for-provtagning-och-smittsparning-under-coronapandemin.pdf Accessed: November 19, 2021.

13. Varsavsky T*, 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* **6**, e21-e29 (2021).

14. COVID Symptom Study Sverige. COVID Symptom Study Sverige. https://csss-resultat.shinyapps.io/csss_dashboard/ Accessed: June 2021.

15. Public Health Agency of Sweden. Confirmed cases in Sweden (Bekräftade fall i Sverige). ttps://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/bekraftade-fall-i-sverige/ Accessed: July 1, 2021.

16. Pilemalm S, Timpka T. Third generation participatory design in health informatics--making user participation applicable to large-scale information system projects. *J Biomed Inform* **41**, 327-339 (2008).

17. Drefahl S*, et al.* A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden. *Nature Communications* **11**, 5097 (2020).

18. Rostila M, Cederstrom A, Wallace M, Branden M, Malmberg B, Andersson G. Disparities in Coronavirus Disease 2019 Mortality by Country of Birth in Stockholm, Sweden: A Total-Population-Based Cohort Study. *Am J Epidemiol* **190**, 1510-1518 (2021).

19. Spreco A*, et al.* Nowcasting (Short-Term Forecasting) of COVID-19 Hospitalizations Using Syndromic Healthcare Data, Sweden, 2020. *Emerging Infectious Diseases* **28**, (2022).

20. Galani A*, et al.* SARS-CoV-2 wastewater surveillance data can predict hospitalizations and ICU admissions. *Science of The Total Environment* **804**, 150151 (2022).

21. Saguti F*, et al.* Surveillance of wastewater revealed peaks of SARS-CoV-2 preceding those of hospitalized patients with COVID-19. *Water Research* **189**, 116620 (2021).

22. Gerlee P*, et al.* Predicting regional COVID-19 hospital admissions in Sweden using mobility data. *Scientific Reports* **11**, 24171 (2021).

23. Public Health Agency of Sweden. Weekly report about COVID-19, week 39 (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.

24. Karolinska Universitetslaboratoriet. Respiratory pathogens (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.

25. Public Health Agency of Sweden. The influenza season 2020-2021 (Influensasäsongen 2020–2021). https://www.folkhalsomyndigheten.se/globalassets/statistik-uppfoljning/smittsamma-sjukdomar/veckorapporter-influensa/2020-2021/influensasasongen-2020-2021-sasongssammanfattning-final-v2_23juli.pdf Accessed: July 30, 2021.

26. Public Health Agency of Sweden. Statistics about SARS-CoV-2 Variants of Concern (Statistik om SARS-CoV-2 virusvarianter av särskild betydelse) https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/sars-cov-2-virusvarianter-av-sarskild-betydelse/ Accessed: January 31, 2022.

27. Public Health Agency of Sweden. Statistics for COVID-19 vaccination (Statistik för vaccination mot covid-19). https://www.folkhalsomyndigheten.se/smittskydd-beredskap/vaccinationer/nationella-vaccinationsregistret/rapportering-covid-19-vaccinationer/ Accessed: January 31, 2022.

28. Sudre CH*, et al.* Attributes and predictors of long COVID. *Nat Med* **27**, 626-631 (2021).

29. Statistics Sweden. Statistics on Income and Living Conditions (SILC) (Undersökningarna av levnadsförhållanden). https://www.scb.se/hitta-statistik/statistik-efter-amne/levnadsforhallanden/levnadsforhallanden/undersokningarna-av-levnadsforhallanden-ulf-silc/ Accessed: April 20, 2021.

30. Messer LC*, et al.* The development of a standardized neighborhood deprivation index. *J Urban Health* **83**, 1041-1062 (2006).

31. Novus. Novus Sweden Panel (Novus Sverigepanel). https://novus.se/metoder/sverigepanel/.

32. Novus. Novus Corona Status (Coronastatus 210226). https://novus.se/coronastatus-210226/.

33. R Core Team. R: A Language and Environment for Statistical Computing.). R Foundation for Statistical Computing (2021).

34. Fay MP, Feuer EJ. Confidence intervals for directly standardized rates: a method based on the gamma distribution. *Stat Med* **16**, 791-801 (1997).

35. 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.

36. Woloshin S, Patel N, Kesselheim AS. False Negative Tests for SARS-CoV-2 Infection - Challenges and Implications. *N Engl J Med* **383**, e38 (2020).

37. Harrell Jr FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer (2015).

38. UK Office for National Statistics. Office for National Statistics. https://www.ons.gov.uk/.

39. UK government. The official UK government website for data and insights on coronavirus (COVID-19). https://coronavirus.data.gov.uk/.

40. Kennedy B*, et al.* App-based COVID-19 syndromic surveillance and prediction of hospital admissions The COVID Symptom Study Sweden. https://doi.org/10.5281/zenodo.6069218.

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[2]		48 (37, 59)	47 (35, 57)	50 (38, 61)
	≥65 (%)	22,272 (15.5)	11,641 (13.0)	10,631 (19.7)
Pregnant (%)		..	1251 (1.4)	..
BMI, kg/m^2 [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 (%)		1,107 (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)	1,444 (1.6)	612 (1.1)
Current cancer (%)		1,476 (1.0)	644 (0.7)	832 (1.5)
Immunosuppressive medication[3] (%)		5,817 (4.1)	3,926 (4.4)	1 891 (3.5)
Healthcare 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	2,229 (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[2]		43 (13, 119)	43 (14, 116)	43 (13, 124)
Duration of study participation, days[2,4]		151 (52, 252)	154 (53, 253)	147 (50, 252)
PCR test[5] (%)		43,501 (30.3)	30,702 (34.3)	12,799 (23.7)
Antibody test[5] (%)		29,208 (20.3)	19,216 (21.5)	9,992 (18.5)

NDI[6,7]	0.37 (0.95)	0.36 (0.94)	0.41 (0.96)
Foreign background,[1,7] %	19 (13, 27)	19 (12, 27)	19 (13, 27)
Population density, inhabitants/km[1,7]	1,706 (357, 5,244)	1,686 (334, 5,190)	1,729 (389, 5,340)

[1]Row percentage, [2]Median (first and third quartile), [3]Corticosteroids, methotrexate and/or biological agents (treatment of cancer and/or rheumatic disease), [4]From first to last daily report, [5]At any time during follow-up, [6]Aggregate data on postal code area level, [7]Mean (standard deviation)
BMI: Body Mass Index; NDI: Neighbourhood Deprivation Index; Foreign background is defined as born outside of Sweden and/or with both parents born outside Sweden.

Table 2. Median absolute percentage errors (MdAPEs) for prediction of new daily hospitalizations, across the first pandemic wave (May 11 to July 3, 2020), the summer period (July 4 to October 18, 2020), and the second pandemic wave (October 19 to November 29, 2020). The iterative prediction model used current regional COVID Symptom Study Sweden prevalence estimates and hospital data to predict hospital admissions seven days ahead.

	Median absolute percentage errors (%)		
	First wave May 11 to July 3, 2020	Summer period July 4 to October 18, 2020	Second wave October 19 to November 29, 2020
All 21 regions combined	37.0	48.2	42.4
Top 5 most populated regions *	25.9	38.6	26.8
Blekinge	41.4	67.6	55.8
Dalarna	58.5	49.4	47.7
Gotland	48.3	83.5	49.2
Gävleborg	29.4	54.1	44.0
Halland	39.4	49.8	73.4
Jämtland	31.9	74.5	51.6
Jönköping	32.4	45.5	28.5
Kalmar	52.5	53.7	44.7
Kronoberg	53.4	69.4	52.5
Norrbotten	40.3	50.3	68.6
Skåne	24.3	46.8	18.9
Stockholm	12.2	31.6	16.6
Södermanland	54.2	37.6	44.8
Uppsala	40.5	60.1	31.1
Värmland	41.0	55.1	43.7
Västerbotten	60.9	48.2	34.2
Västernorrland	36.3	54.1	42.9
Västmanland	39.3	49.1	30.2
Västra Götaland	20.4	35.3	22.4
Örebro	27.0	46.1	67.9
Östergötland	55.2	40.2	43.9

* Stockholm, Västra Götaland, Skåne, Östergötland, Uppsala

Figure 1. The prevalence of symptoms reported by participants in COVID Symptom Study Sweden with a) a positive PCR test for COVID-19 (n=5,178), and b) a negative PCR test for COVID-19 (n=32,089), across the study period April 29, 2020 to February 10, 2021.

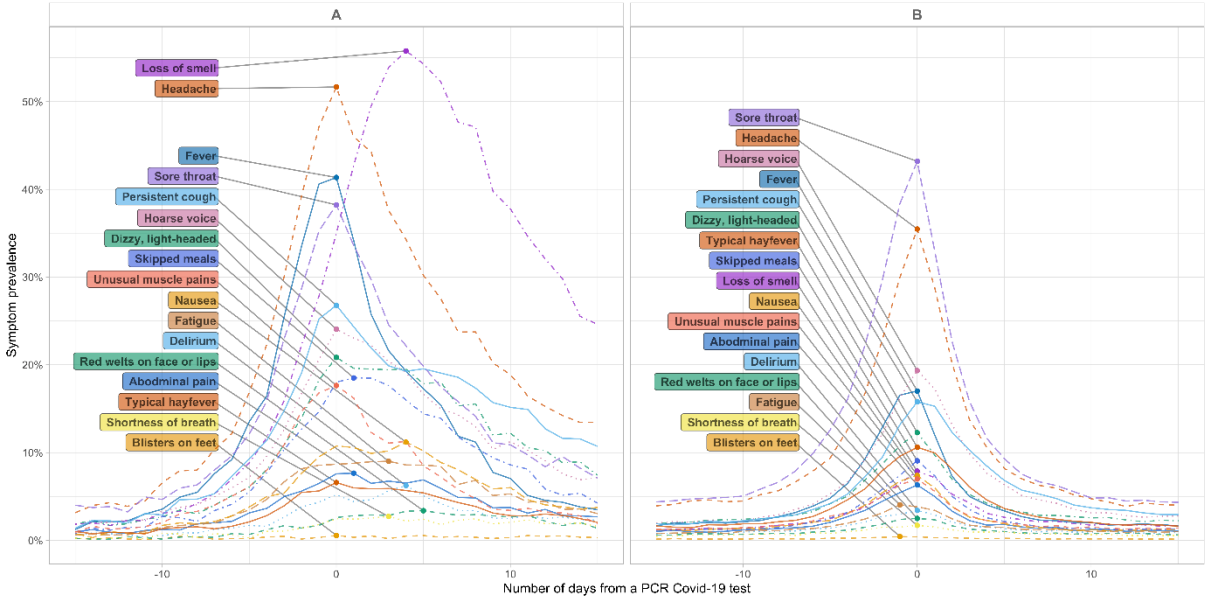


Figure 2. Analysis strategy and data sources

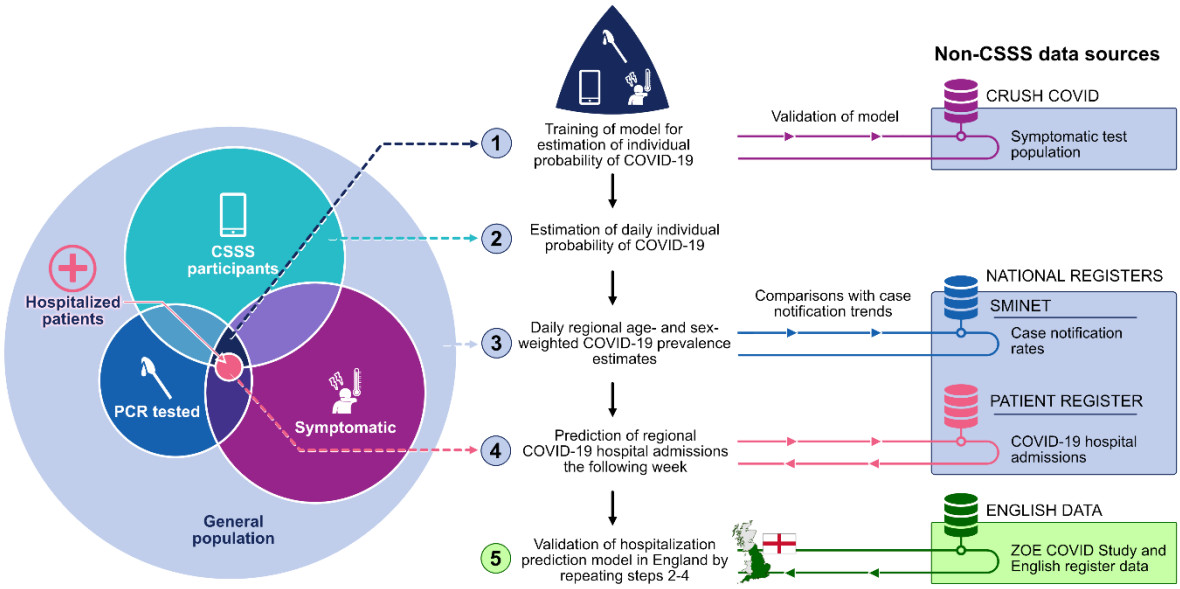
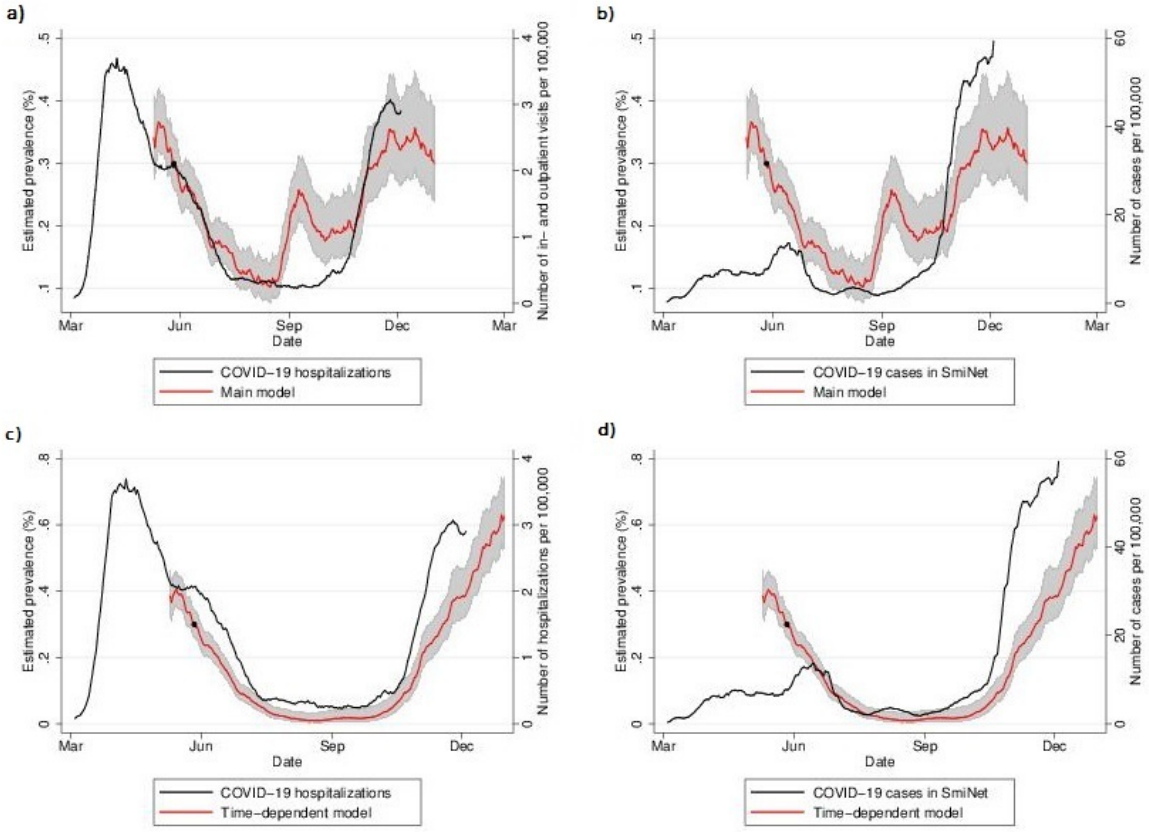


Figure 3. National prevalence estimate, with 95% confidence interval, of symptomatic COVID-19 in COVID Symptom Study Sweden (main model utilized for real-time prediction estimates, and retrospective time-dependent model), combined in a) and c) with retrospective data on daily number of new hospital admissions registered in the National Patient Register per 100,000 inhabitants ≥18 years, and in b) and d) with daily number of new COVID-19 cases registered in SmiNet, per 100,000 inhabitants ≥18 years.



***** Time-point for recalibration of CSSS national COVID-19 prevalence estimate using national point prevalence survey findings from the Public Health Agency of Sweden

Figure 4. Predicted number of daily hospital admissions seven days ahead across the five most populated regions in Sweden ordered by population size. The median absolute percentage errors (MdAPEs) of the predictions are denoted for the first pandemic wave (May 11 to July 3, 2020), the summer period (July 4 to October 18, 2020), and the second pandemic wave (October 19 to November 29, 2020).

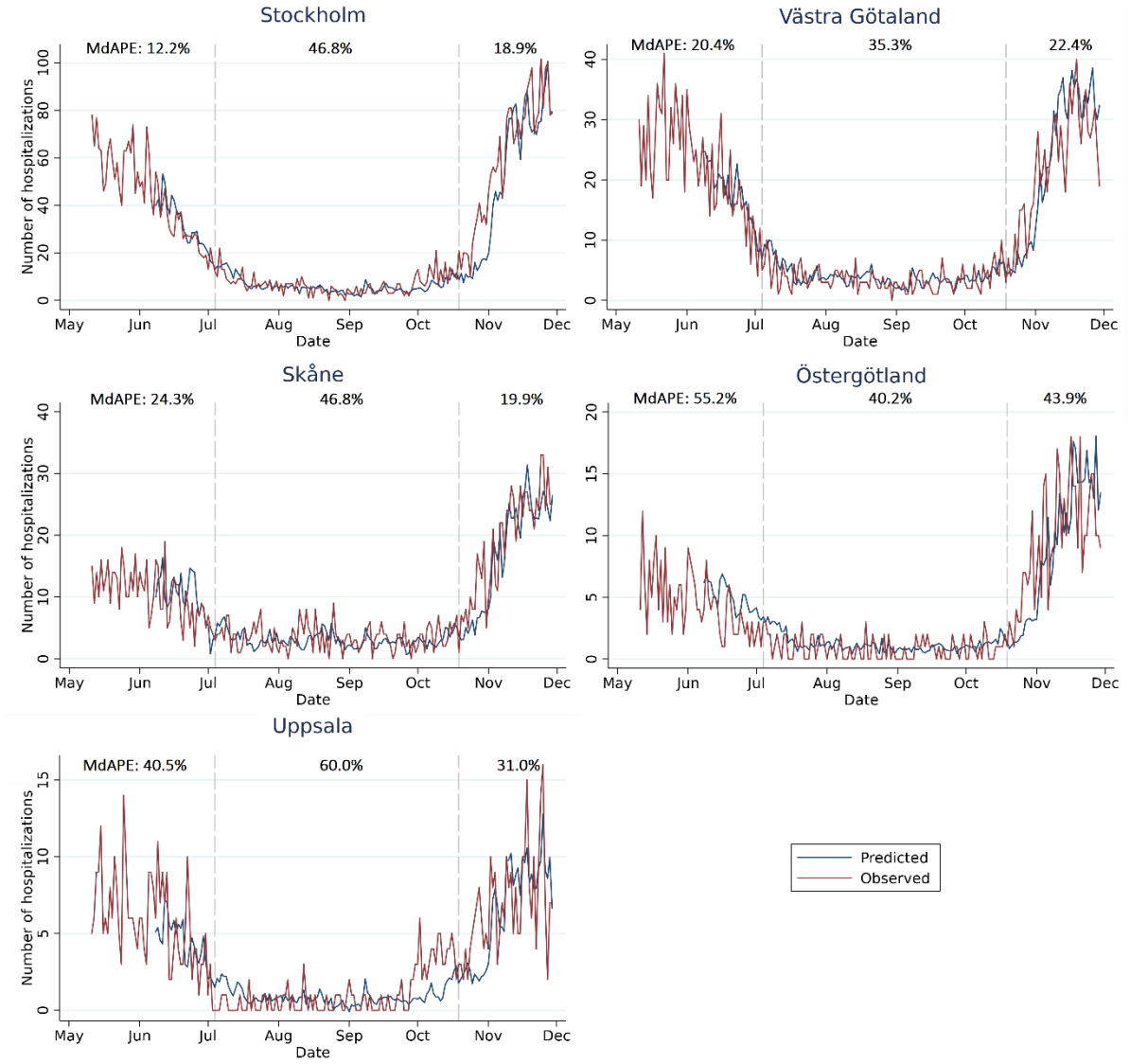


Figure 5. Predicted number of daily hospital admissions seven days ahead across the seven English healthcare regions. The median absolute percentage errors (MdAPEs) of the predictions are denoted for the first pandemic wave (April 6 to June 19, 2020), the summer period (June 20 to September 19, 2020), and the second pandemic wave (September 20, 2020 to February 7, 2021).

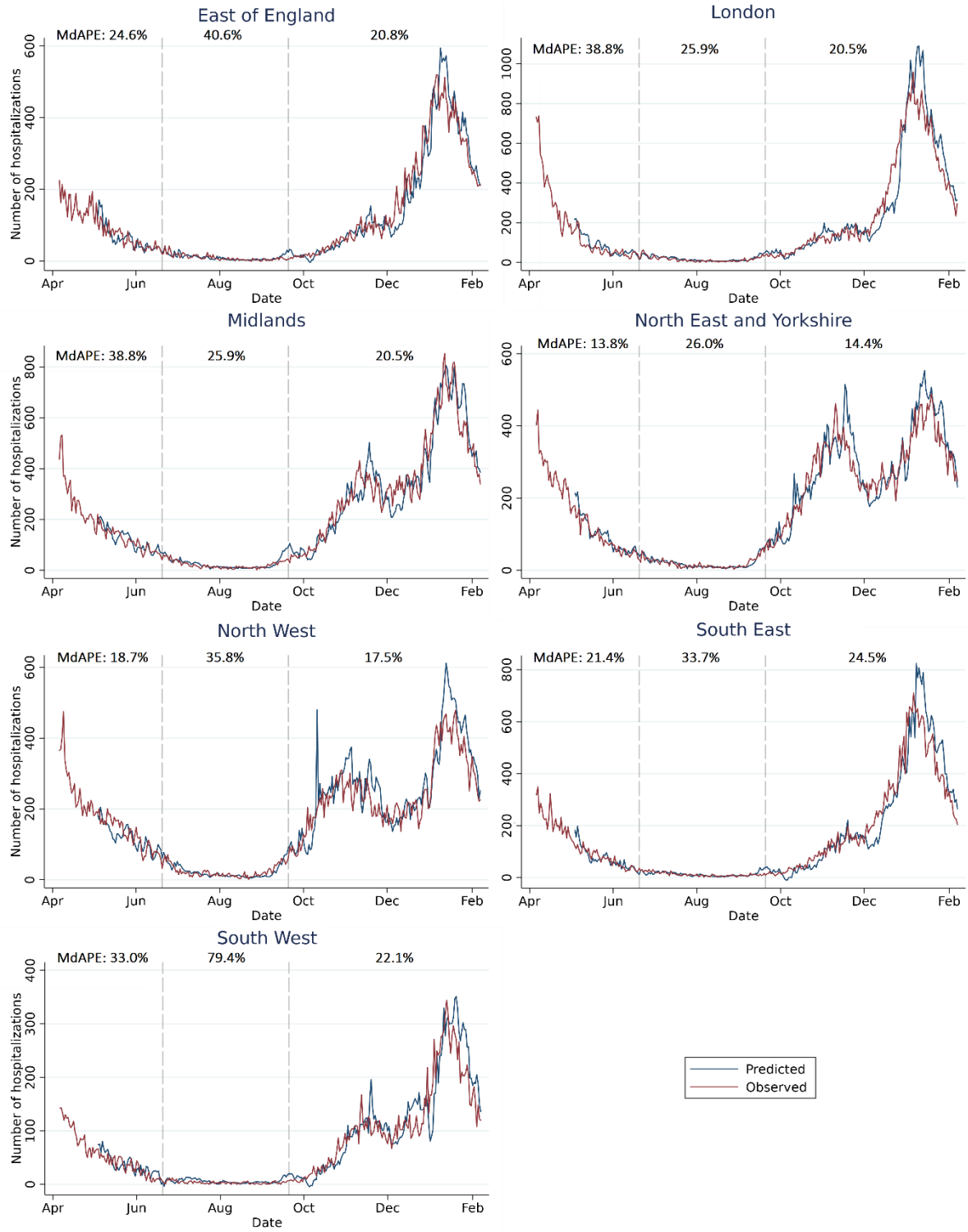
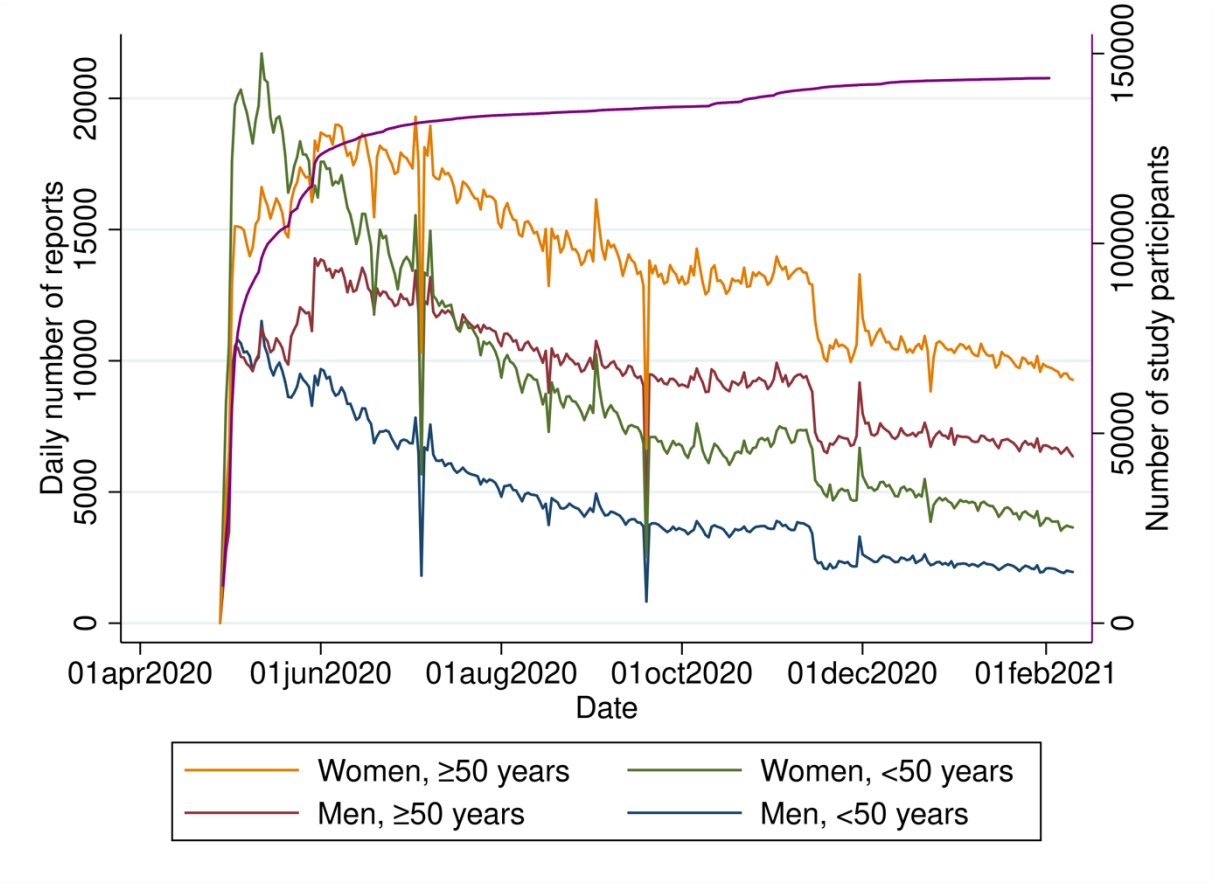



Figure 6. Cumulative number of study participants (n total=143,531), and number of daily reports from these participants stratified by sex and age (<50 and ≥50 years), in COVID Symptom Study Sweden during the study period April 29, 2020 to February 10, 2021.