



Estimating the household secondary attack rate and serial interval of COVID-19 using social media



Aarzo Dhiman^{1,2}✉, Elad Yom-Tov^{3,4}, Lorenzo Pellis⁵, Michael Edelstein⁶, Richard Pebody⁷, Andrew Hayward⁸, Thomas House⁵, Thomas Finnie⁷, David Guzman¹, Vasileios Lampos¹✉, Virus Watch Consortium* & Ingemar J. Cox^{1,9}✉

We propose a method to estimate the household secondary attack rate (hSAR) of COVID-19 in the United Kingdom based on activity on the social media platform X, formerly known as Twitter. Conventional methods of hSAR estimation are resource intensive, requiring regular contact tracing of COVID-19 cases. Our proposed framework provides a complementary method that does not rely on conventional contact tracing or laboratory involvement, including the collection, processing, and analysis of biological samples. We use a text classifier to identify reports of people tweeting about themselves and/or members of their household having COVID-19 infections. A probabilistic analysis is then performed to estimate the hSAR based on the number of self or household, and self and household tweets of COVID-19 infection. The analysis includes adjustments for a reluctance of Twitter users to tweet about household members, and the possibility that the secondary infection was not acquired within the household. Experimental results for the UK, both monthly and weekly, are reported for the period from January 2020 to February 2022. Our results agree with previously reported hSAR estimates, varying with the primary variants of concern, e.g. delta and omicron. The serial interval (SI) is based on the time between the two tweets that indicate a primary and secondary infection. Experimental results, though larger than the consensus, are qualitatively similar. The estimation of hSAR and SI using social media data constitutes a new tool that may help in characterizing, forecasting and managing outbreaks and pandemics in a faster, affordable, and more efficient manner.

The household secondary attack rate (hSAR) of a disease measures its potential for spread in the context of repeated close contacts as seen in households, in contrast to measures of overall infectiousness such as the basic reproduction number R_0 . The hSAR can be defined in different ways, but here we use the common definition as the probability of a household member acquiring the disease within an incubation period given another household member is infected. The investigation of hSAR, in addition to reproduction rate, is important to understanding the risk of transmission¹, and also to inform interventions such as the decision to vaccinate close contacts of immunocompromised individuals². The household SAR is

affected by a number of parameters including the disease (e.g. its reproduction number, incubation period, variants of concern)³, the number of household members⁴, the size of the dwelling, contact environment⁵, preventive measures of household members (e.g. vaccination, masks, social distancing)⁶, comorbidities of the contacts⁷, and demographic features of the population such as age structure, sex ratio, and ethnicity⁸. Accurately estimating the SAR of a disease is difficult. The conventional epidemiological methodology requires a repeated cycle of case investigation and contact tracing⁹ for a sample population. Infection is usually determined through laboratory-confirmed and self-reported cases.

¹Department of Computer Science, University College London, London, UK. ²Centre of Excellence for Data Science, AI and Modelling, University of Hull, Hull, UK.

³Microsoft Research, Herzliya, Israel. ⁴Department of Computer Science, Bar Ilan University, Ramat Gan, Israel. ⁵Department of Mathematics, University of Manchester, Manchester, UK. ⁶Azieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. ⁷UK Health Security Agency, 61 Collingdate Avenue, NW9 5EQ London, UK. ⁸UCL Collaborative Centre for Inclusion Health, UCL, London, UK. ⁹Department of Computer Science, University of Copenhagen,

Copenhagen, Denmark. *A list of authors and their affiliations appears at the end of the paper. ✉ e-mail: aarzo.dhiman@ucl.ac.uk; v.lampos@ucl.ac.uk; ingemar@ieee.org

The value of digital footprints, i.e. data people knowingly or unknowingly generate when using electronic services, to infer information about the health of populations or individuals is now well established. This is the basis for digital epidemiology, i.e. epidemiology that uses data generated outside the public health system and not primarily generated for health purposes. There are a wide variety of digital footprints including social media posts, microblogging (X, formerly known as Twitter (For reasons of clarity, we refer to X as Twitter for the remainder of this paper.)), Web search data, and over-the-counter (OTC) sales of medicines. The use of digital trails for syndromic surveillance dates back to at least 1977 when Welliver et al.¹⁰ demonstrated a strong correlation between sales of OTC remedies and influenza-like illness rates. Interest in syndromic surveillance increased in the early 2000's with the US Defence Advanced Research Projects Agency initiative called ENCOMPASS (ENhanced COncurrence Management Planning And Support System) to improve early warning systems to protect against bioterrorism. Contemporaneously, a number of papers demonstrated the utility of Web search data for estimating influenza prevalence^{11–14}. Subsequently, various researchers showed that Twitter data could also be used for the same purpose^{15–18}. The use of digital trails has not been limited to disease prevalence. Other research has shown its utility for a variety of tasks including identifying adverse drug reactions^{19,20}, automatically identifying individuals at higher risk of health events²¹, performing disease diagnosis^{22,23}, health behaviour analysis e.g., addictions²⁴, and mental health prediction²⁵.

The use of Twitter data to infer the hSAR and serial interval (SI) of influenza was described in²⁶. At the risk of over-simplifying, the fundamental idea is to identify tweets indicating that a user and/or a household member have influenza. When users tweet first about both themselves and then about a household member having influenza, or vice versa, within a predetermine interval, it is assumed that this is an instance of household secondary infection. The hSAR is then estimated as the ratio of the number of such pairs to the total number of users tweeting that they or a household member has influenza. The SI is determined by the time between the two tweets that indicate a primary and secondary infection. Our work significantly extends this approach through demonstrating its utility to COVID-19, and extending the analytical framework. We created a labelled set of tweets indicating whether a tweet was about a user having COVID-19, a household member having COVID-19, or simply about a household member. This data was then used to train three corresponding classifiers. A probabilistic approach is then used to estimate hSAR for each month or week. Monthly and weekly hSAR for COVID-19 are estimated for the UK using Twitter data from January 2020 to February 2022. Not all secondary infections originate from within the household, and, as the prevalence of an infectious disease increases, the probability of acquiring the disease from outside of the household increases. We therefore introduce an adjustment factor to partially correct for secondary infections originating from outside of the household. It has been previously noted²⁶ that Twitter users may have varying reluctance to tweet about a household member. This reluctance can lead to under-reporting of household infection and consequently lower our hSAR estimates. We confirmed that this reluctance exists and adjusted for variations in the likelihood of tweeting about a household member. A further contribution is an analysis of the sensitivity to variations in the size of the monthly cohort, and providing weekly as well as monthly hSAR estimates.

Our methodology assumes all households are of size two, based on the UK average of 2.3, and that a Twitter user is only infected once. The rationale and consequences of these assumptions are described in the Discussion.

Results

Twitter cohort classification

We first identified a seed group of 1,226,509 Twitter users that issued experiential tweets that included statements of the form “I have COVID” or “My spouse has COVID”. We then performed geo-filtering, described in Section 4.1 to only include users based in the UK, resulting in 77,016 users.

Three classifiers were trained to independently identify whether a tweet was (i) about the user/Tweeter having COVID-19, (ii) about a household member having COVID-19, or (iii) about a household member (irrespective of topic), see Supplementary material (Supplementary Methods S 5) for details. The classifiers, denoted C1, C2, and C3, respectively, were trained on a labelled set of 7894 tweets and achieved AUC scores of 0.907, 0.934, and 0.768, respectively. The F1 scores for the three classifiers were 0.8538, 0.8733, and 0.7202 respectively. We denote the probability of tweeting about a household member as p_h . The p_h for each user is estimated as the number of tweets classified as positive by C3, i.e. a tweet about a household member, irrespective of topic, divided by the total number of tweets downloaded for this user. We removed users who never tweeted about a household member, i.e. none of a user's tweets was positively classified by C3, reducing the total number of users (cohort size) to $n = 58,555$. We report results for $p_h > 0$ unless otherwise stated. The most recent (up to) 3200 Tweets of these 58,555 users were downloaded. This resulted in 107,921,029 tweets.

For each monthly or weekly period, we then identified the subset of users with one or more tweets originating in this time interval and being positively classified by C1 and/or C2, i.e. who tweeted that they and/or a household member had COVID-19. Figure 1 depicts the monthly and weekly cohort size for the period from January 2020 to February 2022, inclusive.

Household secondary attack rate estimation

Based on this cohort, Fig. 1 also shows the inferred monthly and weekly hSAR estimates (hSAR_{br}), where it is assumed that household secondary infection occurs within 14 days of the primary infection. This period is the assumed maximum serial interval. Experiments with a longer assumed serial interval produced very similar results. The time of onset of the primary and secondary infections is assumed to be the dates of the corresponding tweets. UKHSA estimates for hSAR for the period from January 2021 to February 2022 are also reported for comparison purposes. Note that UKHSA did not report hSAR estimates prior to January 2021. See Supplementary Methods S 3 for further details.

Sensitivity analysis

Figure 2 examines the sensitivity of the hSAR_{br} estimates to the cohort size. We note that for the period from January 2020 to February 2022 the three smallest monthly cohort sizes were 48 (January 2020), 173 (February 2020), and 947 (August 2020). The average (median) monthly cohort size was 3143.19 (2497).

Adjustment for infection outside of the household

Our estimate, hSAR_{br} , incorporates two adjustments. The first adjustment attempts to account for the probability that the source of the secondary infection was outside of the household. This is of particular concern as our cohort size is determined by the number of index cases tweeting about COVID-19. As such, it is strongly related to the community incidence of COVID-19. When the community incidence is high it is more likely that a second case in the household will in fact have been acquired outside the household, leading to higher levels of estimated hSAR when the cohort size is large. The correlation between our hSAR estimates, denoted as hSAR_{br} , and the corresponding monthly and weekly cohort sizes are 0.504 ($p = 0.0085$) and 0.389 ($p = 2.01e-05$), respectively. For comparison, UKHSA SAR estimates have a correlation of 0.5304 ($p = 0.0509$) with their corresponding cohort sizes. Figure 3A illustrates the secondary attack rate, hSAR_b , before the removal of the probability of non-household secondary infection, the latter probability denoted as $r\text{SAR}$. Without this adjustment, the correlation between the estimated secondary attack rate, hSAR_b , and the corresponding cohort size is significantly stronger at 0.7220.

Adjustment for reluctance to tweet about household members

The second adjustment attempts to correct for people's reluctance to tweet about a household member. Previous work²⁶ has noted that the probability of tweeting about a household member having influenza-

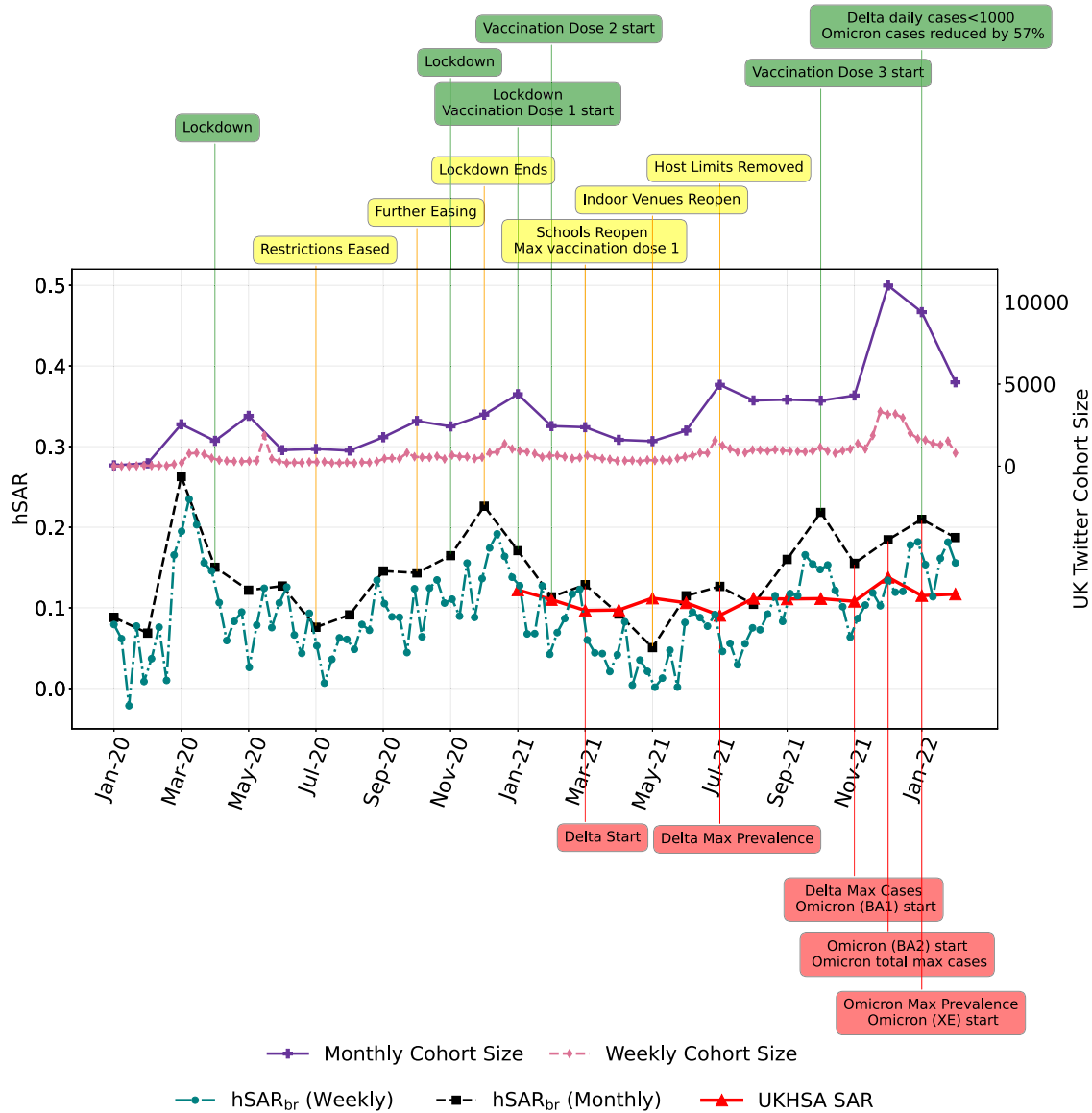


Fig. 1 | Monthly/weekly household secondary attack rates with key events for the period from January 2020 to February 2022. Monthly and weekly cohort sizes and household SAR ($hSAR_{br}$) marked with key events for the period from January 2020 to February 2022, inclusive, for users with $p_h > 0$ and an assumed maximum serial

interval of 14 days. The UK Health Security Agency (UKHSA) (formerly Public Health England) SAR scores are the weighted average of SAR scores for different variants published by UKHSA.

like illness (ILI) after tweeting that the Twitter user had ILI, is likely to be less than the converse, i.e. the probability of tweeting that the Twitter user has ILI after tweeting that a household member is infected. The former probability is expected to be less due to the reluctance to tweet about a household member. To examine this we computed the $hSAR$ independently for the two groups, as depicted in Fig. 3B. The two curves, denoted α_1 and α_2 , represent the $hSAR$ estimates from (1) the users that first tweeted about being infected and then tweeted about a household member being infected, and (2) users that first tweeted about household members being infected and then tweeted about themselves being infected. If there was no reluctance to tweet about household members, we would expect the two curves to be very similar. Instead α_1 is consistently less than α_2 , indicating that household infections are being under-reported. Figure 3A depicts the $hSAR$ estimates before and after adjustment i.e. $hSAR$ and $hSAR_b$. A plot of α_1 against α_2 (see Supplementary Fig. 1 of the supplementary material) gives a best fit line with gradient 0.56 indicating that household infection is under-reported by about 44%.

Serial interval estimation

Finally, Fig. 4 depicts the distribution of the serial (time) interval between the primary and secondary reports of household infection, aggregated over the entire period from Jan 2020 to Feb 2022 and the periods where the Alpha, Delta, and Omicron variants were dominant. The mean (median) serial intervals are 6.49 (6), 6.67 (6), 6.61 (6), and 6.10 (5), respectively.

Discussion

Figure 1 shows the monthly/weekly cohort size gradually increasing over time, with a significant jump in November and December of 2021. This increase might be partially attributable to an increase in willingness to publicly reveal/discuss personal and household COVID-19 status. However, we hypothesise that the increase is primarily driven by increases in the incidence of COVID-19 during this period. Based on UK government statistics²⁷, the number of reported cases on 2nd November 2021 was 31,328, on 1st December 2021 it was 47,263, and on 1st January 2022 it was 99,304.

As expected, Fig. 2 shows that estimates of $hSAR_{br}$ become noisier as cohort size decreases. However, for cohort sizes greater than 1000, and

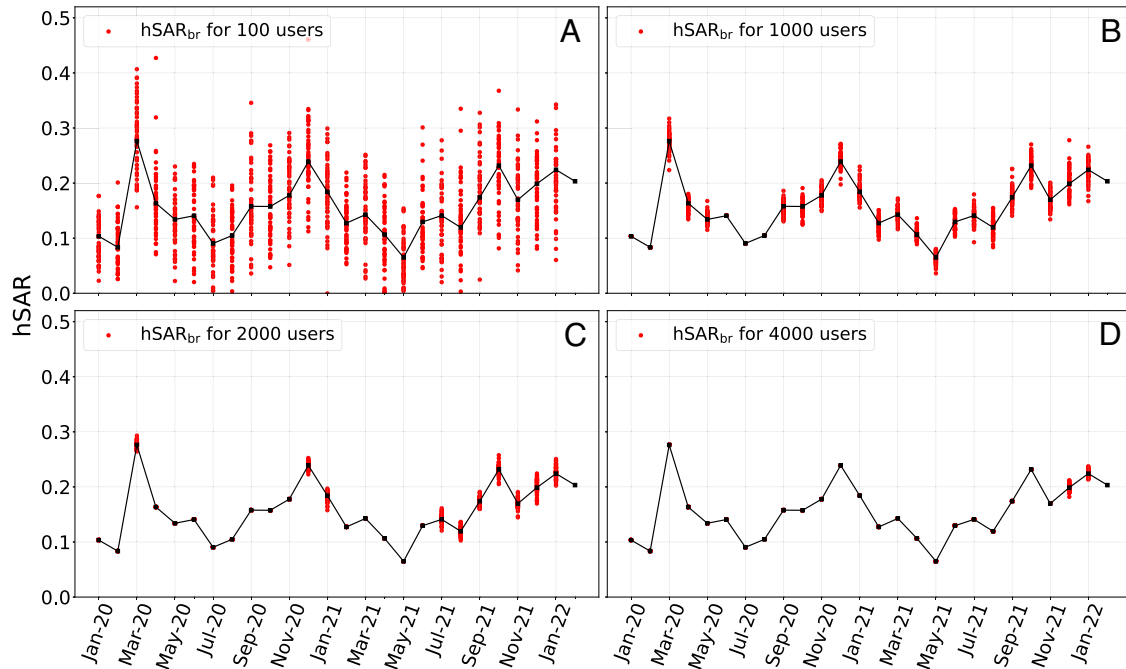


Fig. 2 | Comparison of $hSAR_{br}$ across for various cohort sizes. The solid line depicts $hSAR_{br}$ values calculated using all the users in our Twitter cohort. The dotted values show $hSAR_{br}$ for fixed cohort sizes of **A.** 100, **B.** 1000, **C.** 2000 **D.**4000 users.

For a given cohort size, we create 50 cohorts, uniformly sampled from the available data, total cohort size permitting.

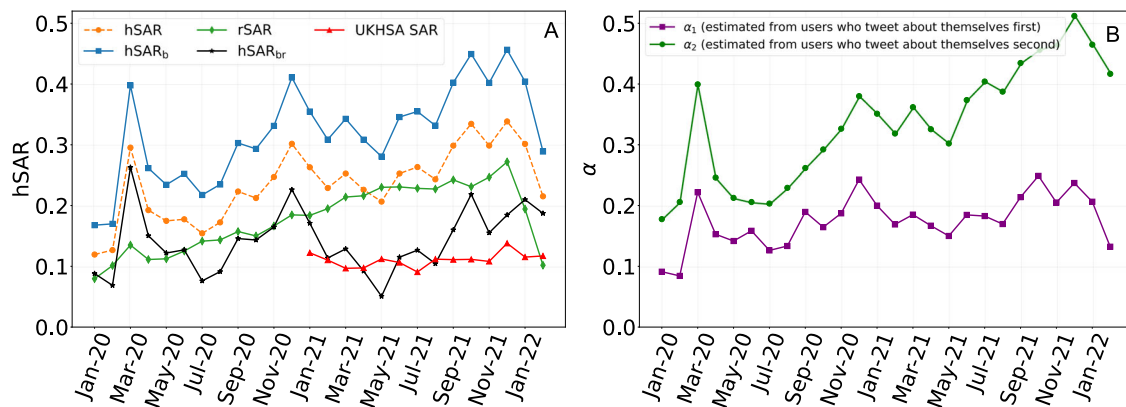


Fig. 3 | Adjustments to the household secondary attack rate estimates. **A** $hSAR$ depicts the monthly household SAR assuming no reluctance to tweet about a household member. $hSAR_b$ depicts the household SAR adjusted for reluctance. $rSAR$ is an estimate of second infections from outside of the household. Our final estimate of household SAR is $hSAR_{br} = hSAR_b - rSAR$. **B** The monthly values for α_1 (the

$hSAR$ estimate calculated from the subset of users who tweeted about being infected and subsequently tweeted that a household member was infected) and α_2 (the estimate calculated from the subset of users who tweeted about a household member being infected and subsequently tweeted that they were infected) for users with $p_h > 0$ assuming a maximum serial interval of 14 days.

certainly 2000, the variation in estimates is small. The monthly (weekly) cohort size is primarily determined by two factors, namely, the incidence of COVID-19 in the general population, and the threshold p_h , i.e. the probability of tweeting about a household member. The later factor is under our control. However the monthly cohort size decreases rapidly with increasing threshold, as discussed in Supplementary Methods S 2 of the Supplementary material. For example, when $p_h \geq 0.05$ the total cohort sizes drops from 58,555 to only 8,244.

While “recruitment” of the Twitter cohort is straightforward, cohort size is also determined by the number of index cases tweeting about COVID-19, and thus is strongly related to the community incidence of COVID-19. As noted earlier, when the community incidence is high it is more likely that a second case in the household

may have been acquired outside the household. This would have the effect of increasing the estimated $hSAR$ when the cohort size is large. This is observed in Fig. 3A where the curve of $hSAR_b$ (no adjustment for probability that source of the secondary infection was outside the household) has a strong correlation (0.7220) with the size of the cohort. We adjust for this by estimating the probability that the source of the second infection was from outside of the household. This is accomplished by randomly pairing Twitter users, assuming homogeneous mixing, considering one as the index and the other as the secondary, and calculating the probability, denoted $rSAR$, that a random pair will be infected within a 14-day interval. Further details are provided in Section 4.2. As expected, $rSAR$ generally increases with the size of the cohort. After applying this adjustment, the

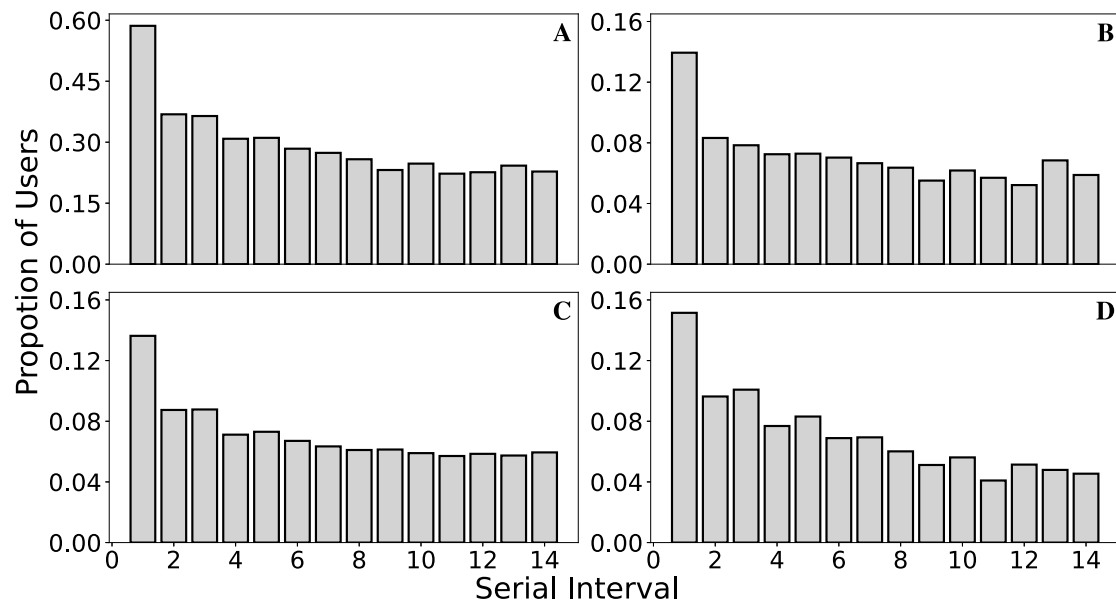


Fig. 4 | Serial interval for Alpha, Delta, and Omicron dominant periods. Histogram of serial interval for daily bins with an assumed maximum serial interval of 14 days for the periods **A** All the months from Jan 2020 to Feb 2022. **B** Alpha

dominant period (18 December 2020 to 15 May 2021). **C** Delta dominant period (22 May 2021 to 19 December 2021). **D** Omicron dominant period (23 December 2021 to 28 February 2022).

hSAR_{br} estimates have a correlation of 0.504 which is similar to the correlation (0.5304) of UKHSA hSAR estimates and the sizes of their corresponding monthly cohorts.

The average hSAR_{br} value of the period from January 2020 to February 2022 is 0.1598 (CI, 0.1412 to 0.1784). While our hSAR_{br} estimates are usually higher than those of UKHSA, our average is very close to the hSAR estimate of 0.166 (CI 0.140, 0.193) (throughout confidence intervals are 95%) of²⁸ based on a meta analysis of 54 relevant studies published until October 2020. A follow-up meta analysis based on 87 studies, published between October 2020 and June 2021, reported an overall SAR of 0.189 (CI 0.162, 0.220)²⁹. A further meta-analysis of 63 studies from January 2020 to January 2022 with midpoints through April 2020 reported SAR of 0.155 (CI 0.132-0.182)³⁰. A review of studies in the early pandemic phase, dominated by the ancestral strain only, was carried out by³¹ with estimated SAR values ranging from about 0.1 to about 0.45.

During the period under study, there were three primary variants of concern, namely, Alpha, Delta, and Omicron. According to the UK Office of National Statistics³², the Alpha variant was dominant from the week ending 18 December 2020 to the week ending 15 May 2021, Delta from the week ending 22 May 2021 to the week ending 19 December 2021 and Omicron from the week ending 23 December 2021 to the week ending 5 September 2022. The average hSAR_{br} during these periods were 0.148 (95% CI, 0.125 to 0.17), 0.154 (95% CI, 0.139 to 0.168), and 0.223 (95% CI, 0.202 to 0.245), respectively. A meta-analysis of hSAR by variant³⁰ reported corresponding values of 0.364 (95% CI, 0.334 to 0.395), 0.297 (95% CI, 0.230 to 0.373), and 0.427 (95% CI, 0.354 to 0.504). There is generally good qualitative agreement, with small differences in hSAR between Alpha and Delta, and a much larger hSAR for Omicron. Lyngse et al.³³ analysed 87,677 individuals in 26,675 households in Denmark (restricted to sizes between 2–6 and an average household size of 3.28) during the period 9–22 December 2021 (after contact tracing was stopped, but before the Christmas holidays commenced). During this period Omicron was replacing Delta as the dominant variant. The SAR was estimated as 0.21 for households with a Delta primary case, and 0.29 for households with an Omicron primary case.

Figure 1 provides the dates of commencement (and termination) of primary interventions during the COVID-19 pandemic in England. We observe that the hSAR_{br} declines significantly in the month before the first lockdown in March 2020. It continues to decline until restrictions are eased in June 2020. As expected, the hSAR_{br} then progressively increases. We observe a

steady decline in hSAR_{br} beginning in January 2021, when COVID-19 vaccinations were introduced, until May 2021 when hSAR_{br} estimates begin to increase. We note that May 2021 marks the time when the Delta variant becomes dominant. The hSAR_{br} continues to increase until October 2022 which marks the introduction of the third vaccine dose. However, the hSAR_{br} increases after November 2022 as the Omicron variant becomes dominant.

Figure 3B clearly demonstrates that some Twitter users are much more reluctant to disclose the health status of household members. In fact, we estimate that approximately 44% of household COVID-19 infections are not reported. Nevertheless, it is straightforward to estimate the under-reporting and correct for it. An alternative solution might have been to select a cohort that was less reluctant, i.e. to select users for which p_h was significantly greater than 0. However, as discussed earlier, this leads to a very large decline in the size of the monthly cohorts.

We assume in “Methods” that the household size is 2, since the average household size in the UK is 2.3³⁴. Our estimates of hSAR are overestimated for household sizes greater than 2 and, conversely, are underestimated for household sizes of 1. On average, we assume that the two effects negate one another as we do not have knowledge of a Tweeter’s household size. If such knowledge is available, or the average household size is higher, e.g. 3, it is straightforward to adjust the Methods accordingly. A further assumption is that a user is only infected once. This is enforced by only taking the single most probable output or output pair from the classifiers. Relaxing this constraint is also straightforward and effectively increases the cohort size, as a user and/or household member will be counted multiple times, once for each period of infection. However, the classifiers are noisy and the increased cohort size comes with the increased risk that positively classified infections are erroneous, since we are essentially lowering the classifier thresholds. To reduce this risk, and because the cohort sizes for each month/week were adequate, we chose to be conservative and only consider the single most likely infection.

We assumed a maximum serial interval of two weeks, i.e. if the two cases are separated by more than two weeks, the cases were treated as independent. The distribution of the serial interval, depicted in Fig. 4A, is almost monotonically decreasing with mean and median values of 6.49 and 6 respectively. While the distribution is similar to that in³⁵, the mean/median values are larger than generally reported. A systematic review³⁶ of research articles studying the serial interval estimated that the weighted pooled mean serial interval of COVID-19 was 5.2, and a serial interval of 4 days was reported in a study in Spain³⁷. Figure 4B–D depict the distribution during the

periods when the Alpha, Delta, and Omicron variants were dominant, respectively. The serial intervals for Alpha and Delta are similar with mean values of 6.67 and 6.61. However the distribution for the Omicron variant is clearly different with a mean value of 6.10. While these serial intervals are longer than previously reported elsewhere, they qualitatively support the evidence that the SI for Omicron was shorter than those for the Alpha and Delta variants. The pooled mean serial interval for Delta was estimated to be 3.9 days and Omicron 3.2 days³⁸. The UKHSA³⁹ has estimated Delta and Omicron serial interval distributions from UK contact tracing data with mean serial intervals of 3.87 days and 3.64 days respectively. Our analysis of the serial interval has at least two potential sources of error. First, we assume that the timestamps associated with the pair of tweets represents the dates of the index and secondary infection. This may not be true - there may be random lags between infection and associated tweet. Further, a tweet of the form "I had COVID 3 days ago" will be classified as the user having COVID-19 at the time the tweet was posted, not 3 days earlier. Further natural language processing could detect and correct for this but remains an avenue for future work.

There are several limitations to this study. At a practical level, Twitter's new restrictions are an impediment to replication and extensions to our study. However, the impediment is not technological but financial. Twitter's new policies no longer permit free access to Tweets. Based on Twitter's current terms and conditions, we estimate that the data collection would now cost between US\$700K to US\$1.3M. Of course, this is for a period of 26 months. To estimate the hSAR for a single month would cost between US\$27K-US\$50K.

The study only considered the UK and it may be the case that Twitter behaviour is significantly different in other geographic regions. However, there is evidence⁴⁰ that there is considerable correlation in behaviour across countries, both English speaking (Australia, USA) and non-English speaking. It may be that the classifiers need to be re-trained to identify relevant tweets. However, we note that we obtained sufficient accuracy with a training set of 7894 tweets that were quickly and inexpensively labelled using a crowdsourcing platform. The proposed method may be more useful in low and middle income countries where the conventional health infrastructure needed to support standard epidemiological studies of hSAR is poor or absent, but access to the Web via mobile devices is prevalent. However, verifying the utility of the method is difficult when ground truth data is absent. We further note that the demographics of Twitter users is unlikely to be fully representative of the UK population. It is reported⁴¹ that there are 25.60 million users in the United Kingdom in early 2024, of which 38.1% are female and 61.9% are male. Data from 2018⁴² states that 33% or users were between the ages of 15 to 24, and that more than half of all users were above the age of 34.

Estimating household secondary attack using classic approaches requires demographic data (household composition) epidemiological data (dates of onset, etc.) and microbiological data (test results). This makes such studies logistically complex, time consuming, and potentially expensive. This new approach opens the potential for hSAR estimates that are cheaper, faster and do not require the collection of data or biological specimens from individuals. With the right calibration and adjustments, our results suggest estimations are within similar ranges as classic methods.

The method is generalizable to other studies. In fact, our work builds on previous work on estimating the hSAR for influenza, as noted earlier. Its appropriateness to other infectious diseases is primarily determined by (i) whether users are inclined/reluctant to tweet about the disease, and (ii) the prevalence of the disease in a population. Thus, for example, it is unlikely to be useful for sexually transmitted diseases where the associated stigma strongly discourages public acknowledgement. Conversely, there have been several studies estimating the prevalence of a variety of infectious diseases, including dengue fever⁴³, Zika^{44,45}, and Monkeypox⁴⁶ from Twitter, and if prevalence can be estimated, it is likely that the corresponding hSAR can also be estimated.

Methods

We first describe the data collection and pre-processing steps and then describe the data analysis.

Data collection

Data collection consists of the following steps:

- Step 1: *Identify a seed group.* For the period from January 2020 until March 2022 inclusive, we queried the Twitter API for all tweets that contained keywords or phrase that implied that the tweeter or a household member had COVID-19. The full set of keywords can be found in the supplementary material in Supplementary Methods S4. Examples of keywords or phrases are "I have covid", "I have been tested positive for corona", "husband got covid" and "kid has coronavirus". Note that the keywords do not, by themselves, define tweets as being about family. For example, a Tweet reading "Joe Biden's wife's first name is Jill" contains the keyword "wife" but is not about a family member. Similarly, the absence of a keyword, e.g. "grandmother", does not imply that there are no tweets containing the word grandmother. For example, a tweet of "My wife is now a grandmother" might be included in the training set since the tweet contains the keyword "wife". Note further, that the keywords are only used to construct the training set, which is subsequently manually labelled. During training, the classifiers are free to select and weight any words to optimise performance. This query resulted in a total of 2,001,896 tweets from 1,226,509 unique seed users.
- Step 2: *Geolocation.* The 1,226,509 unique seed users can be located anywhere in the world. We therefore applied geolocation filtering to retain only users whose tweets originate from the UK. To identify if a user tweeted from the UK, we collect the user information of each user. This is publicly available information associated with a Twitter user account, and consists of several fields, including 'user creation timestamp', 'user description', and 'location'. If a Twitter user's description is unavailable, we delete the user. This reduced the number of unique users from 1,226,509 to 1,145,503. We perform a keyword lookup in the *user description* and the *location* fields to identify if the user belongs to the UK. We use a list of the top 20 most populated cities in the UK⁴⁷ as well as additional keywords. The additional keywords used are {'England', 'Scotland', 'Wales', 'Northern Ireland', 'United Kingdom', 'UK', 'Newport', 'Belfast', 'Derry'}. This list includes four regions of the UK and the top two most populated cities in these regions. Many of the 23 cities in the United Kingdom have the same name as cities in the United States or elsewhere. To disambiguate city names, a user is considered to be tweeting from UK only if the city name (that is same in the UK and any other country) is accompanied by one of the keywords from the list {'England', 'United Kingdom', 'UK'}. Note that 15 cities (Bristol, Leeds, Preston, Liverpool, London, England, Islington, Reading, Sheffield, Birmingham, Leicester, Manchester, Coventry, Nottingham, Sunderland) required disambiguation. For example, 'Liverpool' is a city in both the UK and the US. We mark it as UK location only if it is in one of the following forms {'Liverpool, England', 'Liverpool, United Kingdom', 'Liverpool, UK'}. Even though 'London' is located in the UK, US and Canada, we always mark it as a UK location because of its high population in the UK. After geofiltering, the number of unique users is reduced from 1,145,503 to 77,016.
- Step 3: *Download users' timelines.* After geofiltering, we downloaded the most recent 3200 tweets for each of the 75,440 users. Note that this is less than the 77,016 remaining in the previous step, and reflects that some users (1576) were no longer available. Note that 3200 is an arbitrary number determined by the Twitter API. Also note, that some users will have less than the maximum number of 3200 tweets.
- Step 4: *Filtering Twitter Cohort.* Collecting the set of seed tweets and corresponding seed users occurred some months before we collected users' timelines. (This delay was in part due to staff interruptions and the COVID-19 pandemic.) We observed that almost 19% of users' timelines did *not* include the original seed tweet. There are at least two possible reasons for the absence of a seed tweet. First, the user may have deleted the tweet. Second, for prolific tweeters, the

duration of the 3200 timeline may not encompass the seed. We decided to remove all users whose seed tweet was not found in their time line. There were 13,010 prolific users (97.24%) out of 13,379 users not containing the seed tweets. This reduced the number of unique users from 75,440 to 62,061.

- Step 5: *Filtering user’s timeline.* We apply the Household classifier (C3) to the timelines of each of the 62,061 users to determine each user’s probability of tweeting about a household member. We remove all users who never tweet about a household member. This reduces the number of unique users from 62,061 to 58,555. More generally, we can set a threshold on the probability, p_h , of tweeting about a household member and remove all users whose probabilities are less than p_h .
- Step 6: *Identifying a user’s COVID-19 tweet(s)* We assume that a user is only infected once. However, usually there are multiple references to infection in a user’s timeline. The following process was used to filter a user’s timeline to a single pair of tweets, one indicating that the user had COVID-19 and the other that a household member was infected, or a single tweet indicating that the user or household member was infected.

We constructed three classifiers, see Supplementary Methods S 6 and 7 for details, to independently identify whether a tweet was (i) about the user/ Tweeter having COVID-19, (ii) about a household member having COVID-19, and/or (iii) about a household member (irrespective of topic), denoted C1, C2, and C3, respectively. The classifiers were constructed based on a labelled dataset of 7894 tweets created using a crowdsourcing platform. Each tweet was labelled by three labellers. For all cases where there was disagreement between labellers, the labellers were required to resolve the discrepancies. The best performing classifiers, based on 10-fold cross validation, used CT-BERT⁴⁸, and resulted in accuracies of 0.8886, 0.9324 and 0.7839 for the classifier C1, C2, and C3, respectively. The AUCs for the three classifiers are 0.907, 0.934, and 0.768, respectively. Note that for a tweet to be classified as about a household member having COVID-19, it must also be positively classified by both C2 and C3, i.e. the tweet must be classified as both “about a household member having COVID-19” and “about a household member”. Clearly, if this is not the case, one of the classifiers is wrong, but which classifier is incorrect is unknown. In such cases, we conservatively chose to ignore the positive classification.

For each user, u , we determine the sets of tweets $\{C1_u\}$ and $\{C2_u\}$ that exceed the classifiers’ thresholds. $\{C1_u\}$ is the set of Tweets from user u that were positively classified as being about the user, u , having COVID-19. Similarly, $\{C2_u\}$ is the set of Tweets from user u that were positively classified as being about the user, u ’s, household members having COVID-19. Each set may have 0, 1 or more entries.

We observed that each set may contain almost identical tweets, e.g. a retweet of a previous tweet. To remove these copies we performed the following steps:

1. We removed (i) all user-mentions starting with ‘@’, (ii) hyperlinks or website links starting with ‘https’ or ‘www’, and (iii) replaced all emojis with their textual equivalent using Python package *emoji.demojize*. All that remains is alphanumeric text.
2. Tweets with identical text were then identified and only the earliest tweet is retained.

This results in two filtered sets $\{C1'_u\}$ and $\{C2'_u\}$. The classifier score for each tweet, i , in these sets is then converted to a “probability” with values between 0 and 1 using the equation

$$P1(u, i) = \frac{1}{2} + \frac{1}{2} \cdot \frac{C1(u, i) - \tau_1}{\max(C1) - \tau_1} \tag{1}$$

$$P2(u, j) = \frac{1}{2} + \frac{1}{2} \cdot \frac{C2(u, j) - \tau_2}{\max(C2) - \tau_2} \tag{2}$$

where, $C1(u, i)$ and $C2(u, j)$ are the classification scores for C1 and C2 for tweet i or j of user u . The classifier thresholds are denoted by τ_1 and τ_2 . The function $\max()$ returns the maximum score across all users from the classifier passed to it.

Next, we consider all $\{C1'_u\} \times \{C2'_u\}$ pairs of tweets. All tweets include their corresponding date of publication (Timestamp). We remove pairs where the time between them exceeds a specified threshold (two weeks). The remaining pairs, if any, are ranked according to the product of their probabilities. Only the top-ranked pair is retained. This pair, if it exists, is given the date of the earliest of the two tweets.

In the case where there is no pair, there are two possible scenarios:

- Either the set $\{C1'_u\}$ or $\{C2'_u\}$ is empty. In this case, the user is assigned to the category of the non-empty set.
 - If $\{C1'_u\}$ is not empty, the tweet with the maximum $P1$ is selected
 - If $\{C2'_u\}$ is not empty, the tweet with the maximum $P2$ is selected
- All pairs in $\{C1'_u\} \times \{C2'_u\}$ have a time difference of more than two weeks. In this case, the two tweets with the maximum $P1$ and $P2$ are retained.

Analysis

For a given month (week), we determine the set of users, A , who tweeted that they and/or a household member had COVID-19 in this month. The set A is the union of $A_{1,0}$ (those that only tweet about themselves), $A_{2,0}$ (only about household members), $A_{1,2}$ (those that tweet about themselves and then household members), and $A_{2,1}$ (household members then themselves).

Consider the two cases $A_1 = A_{1,0} \cup A_{1,2}$, i.e. the subset of users who only tweet about themselves having COVID-19 or who tweet that they have COVID-19 before tweeting about a household member having COVID-19, and $A_2 = A_{2,0} \cup A_{2,1}$, i.e. the subset of users who only tweet about a household member having COVID-19 or who tweet that a household member has COVID-19 before tweeting about themselves having COVID-19. Let a_1 and a_2 denote the actions of tweeting about the user or a household member having COVID-19, respectively. Then the probability that individual, i , will tweet about a household member having COVID-19 after tweeting that they have COVID-19 is

$$P_i(a_2|a_1) = (1 - (1 - \alpha_1)^{n-1})P_i(a_2|E_2, a_1) \tag{3}$$

where E_2 represents the event that a household member has COVID-19, and $(1 - (1 - \alpha_1)^{n-1})$ represents the probability that a household member will get infected given α_1 is the secondary attack rate and n is the size of the household. The average household size for UK is $n = 2.3$, so $(n - 1) = 1.3 \approx 1$, which gives us

$$P_i(a_2|a_1) = \alpha_1 P_i(a_2|E_2, a_1) \tag{4}$$

As in²⁶, we assume that $P_i(a_2|E_2, a_1)$ is equal to $P_i(a_3)$, where, a_3 denotes the action of user i tweeting about a household member, irrespective of the subject.

Summing over all m_1 users in A_1 , and dividing both sides by m_1 , we have

$$\frac{1}{m_1} \sum_{i=1}^{m_1} P_i(a_2|a_1) = \frac{\alpha_1}{m_1} \sum_{i=1}^{m_1} P_i(a_3) \tag{5}$$

The LHS is simply the average probability, $P(a_2|a_1)$ of tweeting about a household member with COVID-19 after tweeting that the user has COVID-19, and can be empirically estimated as,

$$P(a_2|a_1) = \frac{|A_{1,2}|}{|A_{1,2}| + |A_{1,0}|} \tag{6}$$

The SAR, α_1 , estimated from the cohort subset A_1 is then

$$\alpha_1 = \frac{|A_{1,2}|}{P(a_3) \times (|A_{1,2}| + |A_{1,0}|)} \tag{7}$$

where

$$P(a_3) = \frac{1}{m_1} \sum_{i=1}^{m_1} P_i(a_3) \tag{8}$$

Similarly for the subset A_2 , since we assume that the household size $n = 2$, the probability that user, i , tweets about having COVID-19 after tweeting that a household member has COVID-19 is simply

$$P_i(a_1|a_2) = \alpha_2 P_i(a_1|E_1, a_2) \tag{9}$$

where α_2 is the secondary attack rate, and E_1 denotes the event that the users has COVID-19. Since user i has already tweeted that a householder has COVID-19, we assume that $P_i(a_1|E_1, a_2) = 1$. Note that this probability is likely to be less than 1, so this is a lower bound on the secondary attack rate. Summing Equation (9) over all users, m_2 , and rearranging, we get

$$\alpha_2 = \frac{1}{m_2} \sum_{i=1}^{m_2} P_i(a_1|a_2) \tag{10}$$

The right hand side of Equation (10) is the average probability, $P(a_1|a_2)$, which can be empirically estimated as

$$\alpha_2 = \frac{|A_{2,1}|}{|A_{2,1}| + |A_{2,0}|} \tag{11}$$

The overall SAR estimate, α , is a weighted average of α_1 and α_2 , i.e. $\alpha = \alpha_1 w_1 + \alpha_2 w_2$ where w_1 and w_2 are given by Equation (12), respectively.

$$w_1 = \frac{|A_{1,2}| + |A_{1,0}|}{|A_{1,2}| + |A_{1,0}| + |A_{2,1}| + |A_{2,0}|} = \frac{|A_{1,2}| + |A_{1,0}|}{|A|} \tag{12}$$

$$w_2 = \frac{|A_{2,1}| + |A_{2,0}|}{|A_{1,2}| + |A_{1,0}| + |A_{2,1}| + |A_{2,0}|} = \frac{|A_{2,1}| + |A_{2,0}|}{|A|}$$

This results in Equation (13) for SAR approximation.

$$hSAR_b = \frac{P(a_3) \cdot |A_{2,1}| + |A_{1,2}|}{P(a_3) \cdot (|A_{2,1}| + |A_{2,0}| + |A_{1,2}| + |A_{1,0}|)} \tag{13}$$

When $P(a_3) = 1$, Equation (13) provides a lower bound for the SAR scores.

Adjustment for reluctance to tweet about household members. Equation (13) includes an adjustment, $P(a_3)$ to account for user reluctance to tweet about a household member. This probability should be empirically estimated via Equation (8). However, in practice, the individual probabilities, $P_i(a_3)$ are quite small and the resulting SAR estimates were unstable with values sometimes exceeding 1. To resolve this issue, we considered the two independent estimates of hSAR given by α_2 and α_1 . We assume a linear relationship, i.e.

$$\alpha_{1,t} = g\alpha_{2,t} + c \tag{14}$$

where t denotes time in months or weeks. If there was no reluctance to tweet about a household member, we expect $g = 1$ and $c = 0$. In practice, the empirical gradient is 0.56 and $c = 0.06$. We ignore the bias, c , and set $P(a_3) = g$. See Supplementary Methods S 1 of the Supplementary material for further details.

Adjustment for infection outside of the household. The fact that a user and a household member both acquire COVID-19 within a given serial

interval does not preclude the probability that the sources of both infections may be outside the home. This probability increases as the prevalence of the disease in the community increases. We accounted and enumerated for this as follows.

For the given cohort ($n = 58,555$) we created $n/2$ random pairs assuming homogeneous mixing. The first member of the pair is considered the user (primary infection) and the second member the household member (secondary infection). The monthly estimation of SAR proceeds as before where the first member's tweets about themselves having COVID-19 are paired with the second member's tweets about themselves having COVID-19, i.e. the output from the classifier C2 is *not* used. The resulting values, denoted rSAR, are an estimate of the probability that the secondary infection was acquired outside of the household. These monthly (weekly) estimates are subtracted from our estimate hSAR_b to produce our final adjusted estimate, hSAR_{br}.

$$hSAR_{br} = hSAR_b - rSAR \tag{15}$$

Inclusion and ethics statement

This research has been conducted in accordance with ethical standards and principles. Approval for the study protocol, including the collection, analysis, and publication of data, was obtained from the UCL Research Ethics Committee (REC), (i) UCL REC 16621/003 “Estimating the secondary attack rate and serial interval of COVID-19 using Twitter” and (ii) the UCL Computer Science REC /CSREC/R/30 “Estimating the secondary attack rate and serial interval of COVID-19 using X, formerly known as Twitter”. Given the large size of the Twitter cohort, obtaining informed consent from all participants involved in the study was not feasible. This exception was approved by the ethics committees at UCL REC 16621/003 and UCL/CSREC/R/30.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Twitter only permits up to a maximum of 1,500,000 Tweet IDs to be distributed. Due to this limitation, we only provide the Tweet ID's of each of the tweets that are positively classified by the three classifiers. We also provide the labelled data used to train the three classifiers. This is a total of 460,979 ID's. Data is available at <https://figshare.com/s/375a15bbcca69af95822>.

Code availability

The code for this study is publicly available at https://github.com/AarzoDhiman/COVID_SAR. The code is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Received: 11 October 2023; Accepted: 10 June 2024;

Published online: 20 July 2024

References

1. Liu, Y., Eggo, R. M. & Kucharski, A. J. Secondary attack rate and superspreading events for SARS-CoV-2. *Lancet* **395**, e47 (2020).
2. NHS England. Nhs flu and covid vaccine programmes brought forward due to risk of new covid variant (2023). <https://www.england.nhs.uk/2023/08/nhs-flu-and-covid-vaccine-programmes-brought-forward-due-to-risk-of-new-covid-variant/>.
3. Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage b. 1.1. 7 in England. *Science* **372**, eabg3055 (2021).
4. Jing, Q.-L. et al. Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 1141–1150 (2020).
5. Zhao, X. et al. A network meta-analysis of secondary attack rates of COVID-19 in different contact environments. *Epidemiol. Infect.* **149**, e219 (2021).

6. Ng, O. T. et al. Impact of Delta variant and vaccination on SARS-CoV-2 secondary attack rate among household close contacts. *Lancet Regional Health-West. Pac.* **17**, 100299 (2021).
7. Bernal, J. L. et al. Transmission dynamics of COVID-19 in household and community settings in the United Kingdom, January to March 2020. *Eurosurveillance* **27**, 2001551 (2022).
8. Bi, Q. et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *Lancet Infect. Dis.* **20**, 911–919 (2020).
9. National Center for Immunization and Respiratory Diseases (U.S.). Division of Viral Diseases. Contact Tracing workflow in a non-US setting. Available at: <https://stacks.cdc.gov/view/cdc/94172> (2020). [Accessed: April 22, 2024].
10. Welliver, R. C. et al. Sales of Nonprescription Cold Remedies: A Unique Method of Influenza Surveillance. *Pediatr. Res.* **13**, 1015–1017 (1979).
11. Wagner, M. M. et al. The emerging science of very early detection of disease outbreaks. *J. Public Health Manag. Pract.* **7**, 51–59 (2001).
12. Eysenbach, G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA Annu. Symp. Proc.* **2006**, 244 (2006).
13. Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D. & Weinstein, R. A. Using internet searches for influenza surveillance. *Clin. Infect. Dis.* **47**, 1443–1448 (2008).
14. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
15. Lampos, V. & Cristianini, N. Tracking the flu pandemic by monitoring the social web. *2010 2nd International Workshop on Cognitive Information Processing* 411–416 (2010).
16. Culotta, A. Towards detecting influenza epidemics by analyzing Twitter messages. *Proceedings of the first workshop on social media analytics* 115–122 (2010).
17. Signorini, A., Segre, A. M. & Polgreen, P. M. The use of twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS one* **6**, e19467 (2011).
18. Aramaki, E., Maskawa, S. & Morita, M. Twitter catches the flu: detecting influenza epidemics using Twitter. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* 1568–1576 (2011).
19. Leaman, R. et al. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts in health-related social networks. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing* 117–125 (2010).
20. Yom-Tov, E. & Gabrilovich, E. et al. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *J. Med. Internet Res.* **15**, e2614 (2013).
21. Yom-Tov, E. et al. Automatic identification of web-based risk markers for health events. *J. Med. Internet Res.* **17**, e4082 (2015).
22. Sadilek, A. et al. Lymelight: forecasting lyme disease risk using web search data. *NPJ Digital Med.* **3**, 16 (2020).
23. White, R. W., Doraiswamy, P. M. & Horvitz, E. Detecting neurodegenerative disorders from web search signals. *NPJ Digital Med.* **1**, 8 (2018).
24. Myslín, M., Zhu, S.-H., Chapman, W. & Conway, M. et al. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *J. Med. Internet Res.* **15**, e2534 (2013).
25. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. *Proc. Int. AAAI Conf. web Soc. media* **7**, 128–137 (2013).
26. Yom-Tov, E., Cox, I. J., Lampos, V. & Hayward, A. C. Estimating the secondary attack rate and serial interval of influenza-like illnesses using social media. *Influenza other respiratory viruses* **9**, 191–199 (2015).
27. UK Health Security Agency (UKHSA) Data Dashboard. Available at: <https://ukhsa-dashboard.data.gov.uk/>. [Accessed: April 22, 2024].
28. Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E. & Dean, N. E. Household transmission of SARS-CoV-2: a systematic review and meta-analysis. *JAMA Netw. Open* **3**, e2031756–e2031756 (2020).
29. Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E. & Dean, N. E. Factors associated with household transmission of SARS-CoV-2: an updated systematic review and meta-analysis. *JAMA Netw. Open* **4**, e2122240–e2122240 (2021).
30. Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E. & Dean, N. E. Household secondary attack rates of SARS-CoV-2 by variant and vaccination status: an updated systematic review and meta-analysis. *JAMA Netw. Open* **5**, e229317–e229317 (2022).
31. Tsang, T. K. et al. The effect of variation of individual infectiousness on SARS-CoV-2 transmission in households. *Elife* **12**, e82611 (2023).
32. Carbury, R. & Flower, G. Regional and sub-regional estimates of Coronavirus (COVID-19) positivity over time, UK: 12 January 2023. Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/regionalandsubregionalestimatesofcoronaviruscovid19positivityoverimeuk12january2023/previousReleases> (2023). [Online; Accessed: Feb, 2023].
33. Lyngse, F. P. et al. Household transmission of the SARS-CoV-2 Omicron variant in Denmark. *Nat. Commun.* **13**, 5573 (2022).
34. UK average household size by region 2022 | Statista - statista.com. Available at: <https://www.statista.com/statistics/295548/average-household-size-in-the-uk-by-region/>. [Accessed: March 18, 2024].
35. Ali, S. T. et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. *Science* **369**, 1106–1109 (2020).
36. Alene, M. et al. Serial interval and incubation period of COVID-19: a systematic review and meta-analysis. *BMC Infect. Dis.* **21**, 1–9 (2021).
37. Arnedo-Pena, A. et al. COVID-19 secondary attack rate and risk factors in household contacts in Castellon (Spain): Preliminary report. *Enfermedades Emerg.* **19**, 64–70 (2020).
38. Madewell, Z. J. et al. Rapid review and meta-analysis of serial intervals for sars-cov-2 delta and omicron variants. *BMC Infect. Dis.* **23**, 1–9 (2023).
39. Overton, C. & Ward, T. Infectious Disease Modelling Team Omicron and Delta serial interval distributions from UK contact tracing data. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1046481/S1480_UKHSA_Omicron_serial_intervals.pdf (2021). [Accessed: April 22, 2024].
40. Zou, B., Lampos, V. & Cox, I. Transfer learning for unsupervised influenza-like illness models from online search data (2019).
41. Kemp, S. Digital 2024: The United Kingdom – DataReportal - Global Digital Insights — datareportal.com. Available at: <https://datareportal.com/reports/digital-2024-united-kingdom>. [Accessed: April 22, 2024].
42. Strugar, M. Top 25 Surprising Twitter Statistics UK Edition [2024] — cybercrew.uk. Available at: <https://cybercrew.uk/blog/twitter-statistics-uk> (2024). [Accessed: April 22, 2024].
43. Coberly, J. S. et al. Tweeting fever: can twitter be used to monitor the incidence of dengue-like illness in the philippines? *Johns. Hopkins APL Tech. Dig.* **32**, 714–25 (2014).
44. McGough, S. F., Brownstein, J. S., Hawkins, J. B. & Santillana, M. Forecasting zika incidence in the 2016 latin america outbreak combining traditional disease surveillance with search, social media, and news report data. *PLoS Neglected Trop. Dis.* **11**, e0005295 (2017).
45. Masri, S. et al. Use of twitter data to improve zika virus surveillance in the united states during the 2016 epidemic. *BMC Public Health* **19**, 1–14 (2019).

46. Jahanbin, K., Jokar, M. & Rahmanian, V. Using twitter and web news mining to predict the monkeypox outbreak. *Asian Pac. J. Trop. Med.* **15**, 236–238 (2022).
47. Estimates of the population for the UK, England, Wales, Scotland, and Northern Ireland - Office for National Statistics — [ons.gov.uk](https://ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland). Available at: <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland>. [Accessed: April 22, 2024].
48. Müller, M., Salathé, M. & Kummervold, P. E. Covid-twitter-bert: A natural language processing model to analyse COVID-19 content on Twitter. *Front. Artif. Intell.* **6**, 1023281 (2023).

Acknowledgements

We would like to acknowledge all levels of support from the EPSRC projects “EPSRC IRC in Early-Warning Sensing Systems for Infectious Diseases” (EP/K031953/1), “i-sense: EPSRC IRC in Agile Early Warning Sensing Systems for Infectious Diseases and Antimicrobial Resistance” and its COVID-19 plus award “EPSRC i-sense COVID-19: Harnessing digital and diagnostic technologies for COVID-19” (EP/R00529X/1). We also acknowledge a financial gift from Google in support of our work on COVID-19. We would also like to acknowledge the support from the MRC/NIHR project “Virus Watch: Understanding community incidence, symptom profiles, and transmission of COVID-19 in relation to population movement and behaviour” (MC_PC_19070). We also acknowledge the assistance of Tomasz Czernuszenko. L.P. gratefully acknowledges the Wellcome Trust and Royal Society Sir Henry Dale Fellowship (202562/Z/16/Z). L.P. and T.H. are also supported by Wellcome Trust Discovery Award “Harnessing epidemiological and genomic data for understanding of respiratory virus transmission at multiple scales” (227438/Z/23/Z), the UKRI Impact Acceleration Award (IAA 386) and the JUNIPER modelling consortium (MR/V038613/1).

Author contributions

I.J.C., V.L. and E.Y.T. conceived the study. A.D. was responsible for implementation, data collection, data analysis, testing, and algorithm refinement, together with I.J.C. and V.L. D.G. led the ethics approval and

provided technical support with the data collection pipeline and computational servers. M.E., T.F., A.H., T.H., L.P., R.P. provided epidemiological insights, assessments, and critiques. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-024-01160-2>.

Correspondence and requests for materials should be addressed to Aarzo Dhiman, Vasileios Lampos or Ingemar J. Cox.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© Crown 2024

Virus Watch Consortium

Rob Aldridge¹⁰, Sarah Beale^{10,11}, Thomas Byrne¹⁰, Jana Kovar¹⁰, Isobel Braithwaite^{10,12}, Ellen Fragaszy^{11,13}, Wing Lam Erica Fong¹⁰, Cyril Geismar^{10,14}, Susan Hoskins¹¹, Annalan Navaratnam^{10,11}, Vincent Nguyen^{10,11}, Parth Patel¹⁰, Maddie Shrotri¹⁰, Alexei Yavlinsky¹⁰, Pia Hardelid¹⁵, Linda Wijlaars¹⁵, Eleni Nastouli^{15,16,17}, Moira Spyer¹⁶, Anna Aryee¹⁰, Rachel McKendry¹⁸, Tao Cheng¹⁹, Anne Johnson²⁰, Susan Michie²¹, Jo Gibbs²², Richard Gilson²² & Alison Rodger^{22,23}

¹⁰UCL Institute Of Health Informatics, University College London, London, UK. ¹¹UCL Institute of Epidemiology and Health Care, University College London, London, UK. ¹²Extreme Events and Health Protection Team, Centre for Radiation, Chemicals and Environmental Hazards, Public Health England, London, UK. ¹³Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. ¹⁴MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK. ¹⁵Department of Population, Policy and Practice, UCL Great Ormond Street Institute of Child Health, University College London, London, UK. ¹⁶Francis Crick Institute, London, UK. ¹⁷University College London Hospital, London, UK. ¹⁸London Centre for Nanotechnology and Division of Medicine, University College London, London, UK. ¹⁹Department of Civil, Environmental and Geomatic Engineering, University College London, London, UK. ²⁰Centre for Population Research in Sexual Health and HIV, Institute for Global Health, London, UK. ²¹Centre for Behaviour Change, University College London, London, UK. ²²Institute for Global Health, University College London, London, UK. ²³Royal Free London NHS Foundation Trust, London, UK.