Genomic epidemiology of COVID-19 in care homes in the East of England

William L. Hamilton*[1,2,+], Gerry Tonkin-Hill*[3], Emily Smith[4], Dinesh Aggarwal[1,5], Charlotte J. Houldcroft[6], Ben Warne[1,2], Colin S. Brown[5], Luke W. Meredith[6], Myra Hosmillo[6], Aminu S. Jahun[6], Martin D. Curran[7], Surendra Parmar[7], Laura G. Caller[6,8], Sarah L. Caddy[6], Fahad A. Khokhar[1], Anna Yakovleva[6], Grant Hall[6], Theresa Feltwell[6], Malte L. Pinckert[6], Iliana Georgana[6], Yasmin Chaudhry[6], Nicholas M. Brown[7,2], Sónia Gonçalves[3], Roberto Amato[3], Ewan M. Harrison[3], Mathew A. Beale[3], Michael Spencer Chapman[3,9], David K. Jackson[3], Ian Johnston[3], Alex Alderton[3], John Sillitoe[3], Cordelia Langford[3], Gordon Dougan[1], Sharon J. Peacock[1], Dominic P. Kwiatkowski[3], Ian Goodfellow[6], M. Estée Török[1,2,+], COVID-19 Genomics Consortium UK[10]

* contributed equally

+ Corresponding authors

Affiliations

1. University of Cambridge, Department of Medicine, Cambridge, UK
2. Cambridge University Hospitals NHS Foundation Trust, Departments of Infectious Diseases and Microbiology, Cambridge UK
3. Wellcome Sanger Institute, Hinxton, UK
4. Cambridgeshire County Council, UK
5. Public Health England, Colindale, UK
6. University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK
7. Public Health England Clinical Microbiology and Public Health Laboratory, Cambridge UK
8. The Francis Crick Institute, London, UK
9. Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK
10. www.cogconsortium.uk

Correspondence

William Hamilton will.l.hamilton@gmail.com

M. Estée Török et317@cam.ac.uk

Word count

Abstract = 150 (max 150)

Main text (excluding references or Methods) = 3428

Key words

COVID-19, SARS-CoV-2, genomics, epidemiology, care home; nursing home; residential home

Abstract

COVID-19 poses a major challenge to care homes, as SARS-CoV-2 is readily transmitted and causes disproportionately severe disease in older people. Here, we report on 6,600 COVID-19 cases from the East of England, 1,167 of which were identified as residents from 337 care homes. Older age and being a care home resident were associated with increased mortality. SARS-CoV-2 genomes were available for 700 residents from 292 care homes. By integrating genomic and temporal data we defined 409 viral clusters within the 292 homes, indicating two different patterns - outbreaks among care home residents and independent introductions with limited onward transmission. Approximately 70% of residents in the genomic analysis were admitted to hospital during the study period, providing extensive opportunities for transmission between care homes and hospitals. Limiting viral transmission between care home residents should be a key target for infection control to reduce COVID-19 mortality in this population.

Introduction

Care homes are at high risk of experiencing outbreaks of SARS-CoV-2. COVID-19 is associated with higher mortality in older people and those with comorbidities including cardiovascular and respiratory disease (Williamson et al., 2020), making the care home population especially vulnerable. As of week ending 30th June 2020, the United Kingdom (UK) Office for National Statistics (ONS) estimated that 30.2% of all deaths due to COVID-19 (13,417 deaths) in England occurred in care homes, and 63.9% (28,390 deaths) occurred in hospital (ONS, 2020a). Most of the COVID-19 deaths in hospital were in persons aged 65 years and over (86.1%). This likely represents an underestimate resulting from diagnostic testing limitations; the ONS estimates that from 28 December 2019 to 12 June 2020, there were 29,393 excess deaths in care homes compared to the expected number based on previous years, of which only two thirds are explained by recorded COVID-19 (ONS, 2020b). To date, SARS-CoV-2 transmission in care homes has not been systematically studied with linkage of epidemiological and genomic data on a large scale.

Care homes are defined by the Care Quality Commission (CQC), the independent regulator of adult health and social care in England, as "places where personal care and accommodation are provided together" (CQC, 2020a). In 2011, 291,000 people aged 65 or older were living in care homes in England and Wales, representing 3.2% of the total population at this age; 82.5% of the care home population was aged 65 years or older (ONS, 2014). Care homes are known to be high risk settings for infectious diseases, owing to a combination of the underlying vulnerability of residents who are often frail and elderly with multiple comorbidities, the shared living environment with multiple communal spaces, and the high number of interpersonal contacts between residents, staff and visitors in an enclosed space (Curran, 2017; Lansbury et al., 2017; Strausbaugh et al., 2003). Understanding the transmission dynamics of SARS-CoV-2 within care homes is therefore an urgent public health priority.

Rapid SARS-CoV-2 sequencing combined with detailed epidemiological analysis has been used to trace viral transmission networks in hospital and community-based healthcare settings (Meredith et al., 2020). Previous epidemiological studies of COVID-19 in care homes have been limited in population size, temporal scale and/or the amount of genomic data included (Arons et al., 2020; Burton et al., 2020; Graham et al., 2020; Kemenesi et al., 2020; Quicke et al., 2020). Here, we apply genomic epidemiology to investigate viral transmission dynamics in care home residents across the East of England (EoE). We aimed to address several key questions of public health concern: What is the burden of care home-associated COVID-19 tested in the region? What are the outcomes for care home residents admitted to hospital with COVID-19? Does SARS-CoV-2 spread between care home residents

from the same care home via a single introduction and subsequent transmission, or through multiple independent acquisitions of the virus among residents? Finally, is there evidence of viral transmission between care homes and hospitals?

Results

COVID-19 case numbers from care home and non-care home residents included in the study

7,406 SARS-CoV-2 positive samples from 6,600 individuals were identified in the study period (26th February to 10th May 2020) (Figure 1), and care home residency status was determined in 6,413 (Figure 1, supplement 1) – the remaining 187 cases had missing address data and care home status could not be determined. The samples were tested at the Public Health England (PHE) Clinical Microbiology and Public Health Laboratory (CMPHL) in Cambridge, which receives samples from across the East of England. Positive cases came from 37 submitting organisations including regional hospital laboratories and community-based testing services (Supplementary Table 1). The proportion of samples coming from different sources changed over the study period (Figure 1, supplement 2). This likely reflects a combination of regional hospitals establishing their own testing facilities, increasing availability of community testing in the UK, and the implementation of national policies that increased the scope of care home testing (Figure 1, supplement 3). Overall, the study population included almost half of the COVID-19 cases diagnosed in the EoE at this time (Public Health England, 2020a), with the remainder being tested at other laboratory sites.

1,167 / 6,413 (18.2%) of the study population were identified as care home residents from 337 care homes. 193 / 337 (57.3%) care homes were residential homes and 144 / 337 (42.7%) were nursing homes, with the majority located in five counties across EoE: Essex, Hertfordshire, Bedfordshire, Suffolk and Cambridgeshire (Figure 2). This represents around half of the care homes in the East of England which had reported suspected or confirmed outbreaks to PHE as of 11th May 2020 (UK government, 2020a). As expected, care home residents were older than non-care home residents (median age 86 years (interquartile range (IQR) 79-90 years) versus 65 years (IQR 48-80 years), respectively (P<0.0001, Wilcoxon rank sum test)) (Table 1). There was a median of 2 cases per care home (IQR 1-5, range 1-22), with a highly skewed distribution: the 10 care homes (top 3%) with the largest number of cases contained 164 / 1167 (14.1%) of all care home cases (Figure 2, supplement 1).

The epidemic curve for all cases tested at the Cambridge CMPHL peaked in the end of March and early April (Figure 3). Care home residents comprised a greater proportion of cases in late April and May than in March (Figure 3A, Table 2). This may reflect the changing profile of samples submitted to the CMPHL, as more regional hospitals had their own testing capacity and a greater number of samples were submitted from community testing organisations in later weeks. However, a similar trend was observed for patients tested at Cambridge University Hospitals, with the proportion of community-onset care home-associated cases increasing from <5% in March to a peak of 14/49 (28.6%) in mid-April (Figure 3B, Table 3). This may suggest that transmission involving care home residents took longer to decline following national lockdown (implemented on 23rd March 2020 in the UK) than transmission in the non-care home general community.

Mortality of COVID-19 infections for care home and non-care home residents tested in hospital

464 / 6,600 (7%) individuals with positive COVID-19 tests were patients tested at Cambridge University Hospitals, a secondary care provider and tertiary referral centre in the East of England. We had access to richer metadata for this subset of patients via the hospital electronic records system. 72 / 464 (15.5%) COVID-19 patients diagnosed at CUH were identified as care home residents (Table 1, Figure 3B), of which <7% were admitted to the intensive care unit (ICU) and 34/72 (47.2%) died within 30 days of their first positive test (precise values not shown where the number of individuals is equal to or below five, to protect patient anonymity). In comparison, amongst non-care home residents, 84 / 392 (21.4%) were admitted to the ICU and 78 / 392 (19.9%) died within 30 days of diagnosis. In a logistic regression analysis, older age, care home residency, ICU admission, and lower diagnostic cycle threshold (Ct) values were associated with increased odds of mortality at 30 days from diagnosis (Figure 4, Table 4). In this model, the Odds Ratio (OR) for 30-day mortality of being a care home resident and being ≥80 years old was 3.0 (95% confidence interval (CI) 1.56 – 5.69) and 6.6 (95% CI 3.7 – 12), respectively. The odds of mortality within 30 days of diagnosis did not differ between residents at nursing homes versus residential homes in a separate logistic regression analysis.

Identifying viral clusters within care homes using genomic and epidemiological data

Genome sequence data were available for 700 / 1,167 (60.0%) care home residents from 292 care homes (Figure 2, supplement 2). There was a median of 8 single nucleotide polymorphisms (SNPs) separating care home genomes (IQR 6-12, range 0-29), compared to 9 (IQR 5-13, range 0-28) for randomly selected non-care home samples (*P*=0.95, Wilcoxon rank sum test), similar to the EoE region described previously (Meredith et al., 2020). The proportion of viral lineage B.1.1 increased over the study period in both care home residents and non-care home residents (Figure 5), consistent with

European trends (Alm et al., 2020). With ongoing viral evolution, descendent lineages of B.1 and B.1.1 also rose in frequency and were commonly found in England during the relevant time period. Overall, this suggests that the SARS-CoV-2 lineages circulating in care homes were similar to those found across the EoE. No new viral lineages from outside the UK were observed, which may reflect the success of travel restrictions in limiting introductions of new lineages into the general population.

The ten care homes with the largest number of genomes (top ~3%) contained 154 / 700 (22%) samples (range 7-18). For several of these ten care homes, all cases clustered closely together on a phylogenetic tree with zero or 1 pairwise SNP differences, consistent with a single "outbreak" spreading within the care home (where an outbreak is defined as two or more cases linked in time or place (McAuslane and Morgan, 2014)) (Figure 6). By contrast, several care homes were "polyphyletic", with cases distributed across the phylogenetic tree and higher pairwise SNP difference counts between samples, consistent with multiple independent transmission events (Figure 3).

The probability of two cases having linked transmission in an epidemiologically meaningful timeframe (for example direct transmission or within one or two intermediate hosts) is a function of several factors, including the pairwise genetic differences between viruses and their phylogenetic relatedness, the time difference between cases, and the opportunities for infection between people (for example, the frequency, duration and extent of close contact). For this continuous probability distribution, we used a pragmatic cut-off of $\geq$15% likelihood that samples were connected by $\leq$2 intermediate hosts, using a previously published algorithm called *transcluster* (Stimson et al., 2019), adjusted for SARS-CoV-2 (Methods). We considered each care home as a separate microcosm of transmission and estimated the number of viral clusters per care home, with separate clusters implying distinct acquisition events among residents. This method identified 409 transmission clusters from 292 care homes (median 1 cluster per care home, range 1-4). Within each cluster, 673 / 775 (86.8%) of pairwise links had zero or 1 pairwise SNP differences (maximum 4), and 756 / 775 (97.5%) were sampled <14 days apart (maximum 22 days). Networks for the ten care homes with the largest number of genomes are shown in Figure 7, indicating linked transmission clusters based on our model assumptions and probability threshold. Consistent with the phylogeny shown in Figure 6, some care homes contained a single transmission cluster involving multiple cases (e.g. CARE0314), while others comprised multiple independent clusters (e.g. CARE0061). 133 care home clusters from 124 / 292 (42.5%) care homes comprised 2 or more individuals, consistent with care home outbreaks.

We investigated the role of genomics in defining care home clusters by repeating the *transcluster* algorithm using the same parameters as for the main analysis but assuming all genomes were identical to each other. This yielded 316 clusters from the 700 residents across 292 care homes – 23% fewer than the 409 clusters yielded when incorporating genomics. This suggests that genomics makes a significant contribution to defining viral clusters; without genomic data, separate transmission events occurring around the same time in a care home cannot be readily distinguished (so cluster sizes are over-estimated and the number of separate viral introductions is under-estimated).

Links between care homes and hospitals

We investigated links between care homes and hospitals. 694 / 700 (99%) of the care home residents with genomes included in the cluster analysis had NHS numbers available, which were linked to national hospital admissions data (Methods). 470 / 694 (67.7%) care home residents had at least one hospital admission within the study period, and 398 / 694 (57.3%) were deemed to have been admitted to hospital with COVID-19 (i.e. their first positive sample was taken within 2 days prior to admission up to 7 days post-admission). 40 / 694 (5.8%) cases were categorised as suspected hospital-acquired COVID-19 infections, defined as first positive test being 7 days or more after their hospital admission date and prior to their discharge date (N=13) or within 7 days following their hospital discharge (N=27). Of the latter group, 10 individuals were admitted to hospital and discharged on the same day prior to their positive test, 9 were admitted for 1 to 7 days, and 8 had been admitted for greater than 7 days. 230 / 694 (33.1%) individuals were discharged from hospital within 7 days of their first positive test, and thus could potentially have been infectious at the time of hospital discharge (Byrne et al., 2020).

Viral clusters linking care home residents and healthcare workers

We investigated potential transmission networks involving care home residents and healthcare workers (HCW) for people tested at CUH (HCW data were not available outside of CUH). We defined clusters using the same method as for the care home resident analysis, but allowed HCW to belong to clusters from multiple care homes, allowing for multiple care home residents to be linked to the same HCW. 38 / 54 (70.4%) care home residents had possible links with HCW using our relaxed threshold. However, on review of the medical records we could only identify strong epidemiological links for 14 / 54 (26.0%) residents from two care home clusters, CARE0063 and CARE0114. The CARE0063 cluster has been described previously (Meredith et al., 2020) and includes care home residents, a carer from that same care home and another from an unknown care home, paramedics and people living with the above. The CARE0114 cluster comprises several care home residents and acute medical staff at

CUH who cared for at least one of the residents. Our clustering method does not assign probabilities for directionality of transmission and cannot therefore precisely determine transmission chains. While all residents from a care home cluster may link to a given HCW, in reality the resident-HCW transmission event may have only involved one of the residents from that cluster, so the proportion of residents with links to HCW may be inflated. Nonetheless, these data show that two care home clusters involved HCW, one based mainly in the community and the other with hospital-based staff at CUH.

We also observed cases from a third care home, CARE0273, with possible transmission links to the same paramedics and carers involved in the CARE0063 cluster. These two care homes are within 1 kilometre of each-other and the cases cluster together on the phylogenetic tree, raising the possibility of shared transmission between them. A plausible transmission network connecting the residents at these care homes and the shared HCWs could be made with at most zero SNPs and three days between sampled cases (Figure 7B); these links are in the top 1.1% of all pairwise transmission probabilities inferred using the *transcluster* algorithm. We accept, however, that without confirmatory epidemiological data this interpretation remains speculative.

Discussion

We have investigated the genomic epidemiology of SARS-CoV-2 in care homes in the East of England. We found that care home residents comprised a large fraction of COVID-19 diagnoses in the "first wave" of the pandemic in this region: up to a quarter of patients in the peak weeks of late March and early April tested at CUH were admitted from care homes. Older age and being from a care home were correlated with each other and were both associated with significantly increased odds of mortality within 30 days of diagnosis. Care home residents thus bore a high burden of COVID-19 infections and mortality.

A smaller proportion of care home residents were admitted to ICU compared with people who were not from care homes. What treatments a patient receives, including the invasive treatments provided in intensive care, are complex and individualised decisions based on risk-benefit assessments involving patients, their families and carers, and healthcare professionals (ICS, 2020; NICE, 2020). Of note, non-invasive respiratory support (such as continuous positive airway pressure, high-flow nasal oxygen therapy and non-invasive ventilation) are routinely provided outside ICU in many UK centres.

We defined viral clusters within each care home by integrating temporal and genetic differences between cases. This provides a "high resolution" picture of viral transmission; without genomic data, separate introductions of the virus occurring around the same time would be impossible to distinguish. We found that care homes frequently experienced "outbreaks" of multiple cases within clusters (the largest of which had >10 residents), consistent with substantial person-to-person transmission taking place within care homes. We also observed care homes containing multiple distinct clusters (up to 4), consistent with independent acquisitions of COVID-19 among residents. The majority of care home residents in the genomic analysis did not acquire COVID-19 in hospital. In the context of a national lockdown, the most likely location they acquired their infection was the care home. The high frequency of care home outbreaks may reflect the underlying vulnerability of this population to COVID-19 and the challenges of infection control in care homes. In contrast, the UK as a whole had an average of 2.37 people per household in 2019 (ONS, 2019a) and in the East region only 2.2% of households were made up of two or more unrelated adults (6.2% in London) (ONS, 2019b).

These findings emphasise the importance of limiting viral transmission within care homes in order to prevent outbreaks. Given there is increasing evidence for asymptomatic and presymptomatic transmission of SARS-CoV-2 (Arons et al., 2020; Goldberg et al., 2020; He et al., 2020), isolating residents or staff when they develop symptoms is not sufficient to prevent within-care home spread once the virus has entered the care home. Certain measures may be required on an ongoing basis within care homes when there is sustained community transmission, even when no outbreak is suspected (at least until the morbidity and mortality of the virus in older people has been reduced substantially through vaccination or treatments). These may include use of appropriate Personal Protective Equipment (PPE) for staff and visitors (including visiting healthcare professionals and friends and family), rigorous hand hygiene, social distancing, and making use of larger, well-ventilated rooms for social interactions or socialising outdoors, providing that this is practical and safe (N. R. Jones et al., 2020). This is consistent with current national guidance for care homes in England (Public Health England, 2020b; UK government, 2020b). Face coverings for residents themselves when interacting socially in communal indoor areas could be considered, if acceptable to residents.

The majority of residents had hospital contact during the study period, indicating substantial opportunity for infections to pass between care homes and hospitals in either direction. A third of patients were discharged from hospital within 7 days of their first positive test, and thus were potentially infectious at discharge. We identified transmission clusters that would be consistent with COVID-19 spread between care home residents and HCW, based both in the community and in

hospitals. A previous study found that working across different homes was associated with higher SARS-CoV-2 positivity among staff (Ladhani et al., 2020). Limiting the spread of COVID-19 between care home residents, HCW and hospitals is a therefore another key target for infection control and prevention.

We acknowledge several limitations to this study. First, we have not captured all of the COVID-19 cases from the East of England. Serology data indicate that 10.5% of all residents in care homes for people aged 65 and older in England had been infected with SARS-CoV-2 by early June, the majority of whom were asymptomatic (UK government, 2020c). The Cambridge CMPHL did not receive all of the samples from the region, though based on national data we estimate we have captured around half of the COVID-19 cases reported from EoE during the study period. We did not have viral sequence data available for 40% of care home residents, as a result of missing samples, mismatches between sequences and metadata, genomes not passing quality control filtering using a stringent threshold (<10% missing calls), or sequences being unavailable at the time of data extraction. We may therefore have underestimated viral cluster sizes.

Second, the nature of diagnostic testing sites changed during the study period as regional hospitals developed their own in-house testing capacity and community testing laboratories were set up. “Pillar 2” testing in the UK was outsourced to high-throughput laboratories during April 2020 and performed an increasing proportion of community testing. It is possible that some care home residents from the same care home could have been tested through different routes, with symptomatic cases more likely to be tested in “Pillar 1” via the CMPHL (and included in this dataset), and asymptomatic screening occurring more via the Pillar 2 laboratories. However, most care homes in EoE only began systematic screening after the end of our study following the introduction of the UK care home testing portal on 11th May 2020. Moreover, the *transcluster* algorithm allows for “missing links” within a cluster (our thresholds assumed a $\geq$15% probability of infections being linked within $\leq$2 intermediate hosts), reducing the impact of missing care home cases on defined clusters. The changing profile of COVID-19 testing in the UK between March and May 2020 should therefore be factored into all interpretations of COVID-19 epidemiology from that period.

Third, defining who is a care home resident from large electronic healthcare records is challenging and, despite our best efforts, we may not have identified all care home residents. Using pre-defined coding such as care home CQC registration numbers when patients are booked into hospital systems, rather than free-text data entry, would help considerably with care home surveillance. We undertook

multiple rounds of electronic searches and manual inspection to identify as many care home residents as possible, cross-referencing every patient included in the analysis with a CQC database. The care homes included in this study should therefore be accurate.

Fourth, low viral sequence diversity limits the power of genomics to infer transmission clusters. We have mostly avoided commenting on between-care home transmission because, unlike within-care home cases, opportunities for transfer of SARS-CoV-2 between care homes cannot be assumed or inferred from our data. This could be assessed in a dedicated prospective study gathering epidemiological data on between-care home contacts. Even within care homes, it is possible some genetically similar viruses are from unconnected introduction events. However, incorporating genomic data is more accurate for excluding linked transmission than if only temporal data are available. Genomics can thus be used to "rule out" cases as being part of a linked cluster if the genetic difference is greater than would be expected given the viral mutation rate. This could be practically informative for care homes (along with other organisations at risk of COVID-19 outbreaks like factories (Middleton et al., 2020)), with implications for infection control procedures.

In conclusion, care homes represent a major burden of COVID-19 morbidity and mortality, with transmission events introducing SARS-CoV-2 into care homes and subsequent transmission within them. Genomic data can be used in outbreak investigations to define viral clusters; this is critically dependent on integration with epidemiological data. The cut-offs we used for defining care home clusters were pragmatic but plausible given our current understanding of the biology and epidemiology of SARS-CoV-2. Such cut-offs can be helpful for producing understandable outputs for biological and public health interpretation (MacFadden et al., 2018; Stimson et al., 2019), and for focusing investigations with limited public health resources. Future work will need to prospectively integrate genomic and epidemiological data to rapidly identify viral clusters, thus enabling deployment of infection control and public health interventions in real time.

Methods

Study overview

Data were collected on SARS-CoV-2 positive samples from the East of England, tested at the PHE CMPHL in Cambridge, between 26th February and 10th May 2020. The UK government launched a national care home testing portal on 11th May 2020 (UK government, 2020d), in which all care home staff and residents were eligible for testing with priority for homes caring for people aged 65 years or older. The 10th May was selected as a study end-date because it encompassed the bulk of the "first

wave" of the epidemic in the East of England. Furthermore, prior to the 11th May 2020, systematic screening of all residents within care homes was much less common and testing primarily occurred where there was a suspicion of an outbreak; our strategy reduced risk of bias which would have been introduced had we included systematic screening. During the study period the scope of testing in hospital, community and care home settings changed several times, as eligibility criteria were modified (Figure 1 Supplement 1). When interpreting trends in COVID-19 cases in the UK during this period it is essential to consider the changing capacity and policies surrounding testing.

Diagnostic testing, metadata collection and genome sequencing

For details on diagnostic testing, patient metadata collection and nanopore genome sequencing see (Meredith et al., 2020). Briefly, CMPHL used an in-house generated and validated one-step RT q-PCR assay detecting a 222-bp region of the RdRp genes, along with an MS2 bacteriophage internal extraction control, using the Rotorgene™ PCR instrument. Samples that generated a Ct value $\leq$36 were considered positive. Samples from Cambridge University Hospitals NHS Foundation Trust (CUH) and a selection of East of England (EoE) samples were sequenced on site to provide rapid information on hospital-acquired infections (Meredith et al., 2020). Nanopore sequencing (Oxford Nanopore Technologies) took place in the Division of Virology, Department of Pathology, University of Cambridge, following the ARTICnetwork V3 protocol and assembled using the ARTICnetwork assembly pipeline. The remaining EoE samples, where available, were sent to the Wellcome Sanger Institute (WSI) for sequencing.

Sequencing at WSI used Illumina technology. cDNA was generated from SARS-CoV-2 viral nucleic acid extracts and subsequently amplified to produce 400nt amplicons tiling the viral genome using V3 nCov-2019 primers (ARTIC). This was followed by Illumina library generation using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs Inc, Cat. No. E7645L). Libraries were amplified with KAPA HiFi Ready Mix (Kapa Biosystems, Cat. No. 07958927001) and uniquely indexed with a 100 µM i5 and i7 primer mix (50 µM each) (Integrated DNA Technologies) to allow multiplexing of up to 384 SARS-CoV-2 viral extracts into one sequencing pool. The PCR products were pooled in equal volume and purified with an AMPure XP workflow (Beckman Coulter, Cat. No. A63880). The purified pool was quantified by qPCR (Illumina Library Quantitation Complete kit, Cat. No. KK4824) and sequenced on one lane of an Illumina NovaSeq SP flow cell (Illumina Inc, NovaSeq 6000 SP Reagent Kit v1.5 (500 cycles), Cat. No. 20028402), with XP workflow (Illumina Inc, NovaSeq XP 2 lane kit v1.5, Cat. No. 20043130). Genomes were generated for each library's sequencing data using bwa mem (Li, 2013) for alignment with MN908947.3 (Wu et al., 2020) as reference, samtools (Li et al., 2009) for pileup and

ivar (Grubaugh et al., 2019) for trimming and consensus generation, all orchestrated by the ncov2019-artic-nf pipeline (Bull, 2020).

Sequences were available from both Illumina and Nanopore platforms for eight care home residents included in the study (in all cases the Illumina data were used for the study analysis). In 7/8 cases the sequence pairs were identical; in one case there were two SNP differences between the consensus fasta sequences (both C in the Illumina data and T in the Nanopore data). The sequence pairs are shown below:

Illumina sample - COG-UK ID	Illumina sample - date	Nanopore sample - COG-UK ID	Nanopore sample - date	Pairwise SNP difference
CAMB-761D5	30/03/2020	CAMB-7B088	11/04/2020	zero
CAMB-1AF1F0	30/04/2020	CAMB-1AD8A2	30/04/2020	zero
CAMB-1AE7C2	30/04/2020	CAMB-1AC269	30/04/2020	2
CAMB-80590	09/04/2020	CAMB-789BD	06/04/2020	zero
CAMB-1AB23D	20/04/2020	CAMB-840B9	26/04/2020	zero
CAMB-83AAD	15/04/2020	CAMB-8416B	25/04/2020	zero
CAMB-1ABE2A	21/04/2020	CAMB-8468A	27/04/2020	zero
CAMB-1AB631	21/04/2020	CAMB-1ABF18	27/04/2020	zero

As with all the sample dates used in this study, the above dates are based on sample collection date where available, with missing data substituted with the date of receipt in the laboratory. SNP differences were identified from a vcf file produced from the alignments using the package *snp-sites* v 2.5.1 (Page et al., 2016), command:

```
snp-sites -v alignment_file.aln
```

COG-UK IDs and GISAID accession numbers for genomes analysed in this study are included in Supplementary Materials, along with a complete author list for the COG-UK consortium.

Sample selection

As described in (Meredith et al., 2020), patient metadata were downloaded daily from the electronic medical record system (Epic Systems, Verona, WI, USA) and metadata manipulations were performed in R (v 3.6.2) using the *tidyverse* packages (v 1.3.0) installed on CUH computers. Positive samples were collected and assigned either for nanopore sequencing on site (focusing on CUH samples and a randomised selection of EoE samples), or sent to WSI for Illumina sequencing. Metadata were uploaded weekly to the MRC CLIMB system as part of the COG-UK Consortium. Samples included

healthcare workers (HCW) tested in the CUH HCW screening programme (N. K. Jones et al., 2020; Rivett et al., 2020), all of which were nanopore sequenced on site.

Identifying care home residents

We identified care home residents using a two-stage data mining approach followed by manual inspection and linking of putative care home addresses to care homes registered to the Care Quality Commission (CQC).

Step 1: Search terms in patient address fields

We searched patient address lines 1 and 2 for the following list of key phrases (not case sensitive) in their electronic healthcare records; if any phrases were present the patient was labelled as being from a care home:

- "residential home"
- "care home"
- "nursing home"
- "care centre"
- "care hom"
- "nursing hom"
- "residential hom"
- "carehome"

This identified 765 patients as being care home residents.

Step 2: Matching location names to CQC registered care facilities

Many care homes do not have the above list of phrases in their address names. To capture these facilities, we used the publicly available database of care homes registered to the CQC, the independent regulator of health and adult social care in England. All organisations providing accommodation for persons who require nursing or personal care must be registered with the CQC, including care homes with or without nursing care (CQC, 2020b). Details of the CQC registration scope can be found in "The scope of registration (Registration under the Health and Social Care Act 2008)", March 2015, available at this link as of 24th June 2020: (CQC, 2015).

We accessed the file "CQC care directory – with filters (1 June 2020)" on 23rd June 2020 from the CQC website: (CQC, 2020c), and applied the following filters:

- Total facilities in CQC database: N = 49,516
- "Carehome?" column filtered to "Y": N = 15,507*

- We only included care homes for which "Location Postal Code" column matched at least 1 postcode from the dataset of 6,600 patients included in our study, yielding N = 444 care homes. Following manual review and consistifying postcodes with the sample metadata, we used a set of 469 CQC care homes.**

*Filtering using the "carehome?" column was based on advice given after correspondence with the CQC.

** Requiring CQC registered care homes to match postcodes from the patient dataset minimised the number of "false positives" – patients whose address name matched a CQC registered care home name by coincidence. 25 CQC-registered care homes were added following manual review of the identified putative care home residents, who had a different postcode documented in the electronic healthcare records for the same care home, yielding the final "CQC EoE care home search set" of 469 care homes.

We then used the values from the "Location name" column of the filtered CQC dataset (i.e. the care home facility names) as search phrases for address line 1 in the patient database. Any patients with exactly matching phrases were labelled as care home residents. This increased the number of care home residents identified by a further 382 to 1,147, i.e. around one third of care home residents were identified using CQC facility names and would have been missed by relying on generic care home-related search phrases alone.

Step 3: Manual inspection and data clean up

Address lines for the non-care home patients were manually inspected; this identified a further 89 care home residents. Most of these had not been detected in steps 1 and 2 due to spelling or formatting issues with the patient addresses (e.g. short-hand abbreviations used for care home names, or inclusion of extra details like flat number meaning the string did not match a CQC care home name exactly).

Next, address lines for the care home residents were manually inspected and 14 were deemed not to be care home residents. Most of these were due to unrelated locations sharing the same address name as a CQC registered care home. The manual filtering steps thus yielded a care home resident count of 1,147 + 89 – 14 = 1,222. Address line 1 for all 1,222 care home residents was manually inspected and formatted to ensure residents from the same care home had matching terms in this column. This was necessary due to discrepant address entrance formats for identical care homes;

without this step, residents from the same care home would be incorrectly assigned to different anonymised care home codes.

Step 4: Linking care home addresses to CQC registered care homes

First line of patient address and postcodes were matched to care home names and postcodes from the CQC EoE care home search set (described above). Any discrepancies (care homes not matching the CQC data) were manually inspected and in the majority of cases the discrepancy could be reconciled (e.g. alternative name or postcode used for the same care home). In 55 cases a "care home" was reclassified to non-care home, either because the address was independent housing with a matching name to a care home by coincidence, or because a care facility was determined by CQC definitions to not be a care home - e.g. several mental health community hospitals, drug rehabilitation centres, and supported living environments were excluded. Ultimately all 337 care homes included in the study were linked to CQC data; in two cases the care home had been previously registered but had since been "archived", and the most recent CQC data for defining whether residential or nursing care was being provided was used.

Care home location IDs assigned by the CQC were turned into anonymised codes (format: CARE followed by a 4-digit numeric code). We determined whether care homes were "residential homes" or "nursing homes" using the CQC data column "Service type - Care home service with nursing" filtered to "Y" for care homes with nursing, and column "Service type - Care home service without nursing" = "Y" for care homes without nursing ("residential homes"). If both fields were "Y" then the care home was coded as being a nursing home.

Linking care home data to CUH acute medical testing data

The dataset of 7,407 PCR-positive samples with metadata were collected prospectively as part of the COG-UK study. Data on CUH acute care testing, including categorisations of whether infections were community- or hospital-acquired (definitions provided in (Meredith et al., 2020)) and data on patient outcomes (mortality at 30 days and ICU admissions), were collected separately as part of CUH and national monitoring. During the study period, 464 patients tested positive for COVID-19 at CUH.

When merging the metadata collected for COG-UK (including the above care home categorisations) with CUH acute testing data we identified 71 care home residents tested at CUH. However, there were 23 samples that had tested positive in CUH that were not in the COG-UK dataset. 21/23 of these were tested on the SAMBA platform at CUH (Collier et al., 2020), which is not PCR-based; sequencing was

not possible for these samples owing to rapid RNA degradation. For technical reasons, SAMBA results were therefore not included in the data collected prospectively in the Cambridge COG-UK study. The remaining two discrepancies were not captured in our electronic patient record downloads, which likely reflects periods where the download processes and coding methods were being established. Of the 23 missing samples, 20 were community-onset community-associated, two were hospital-onset indeterminate healthcare-associated, and one was a healthcare worker. These are counted as such and depicted with the above categorisations in the CUH epidemic curve shown in Figure 3B. Of the 23 CUH samples missing from the Cambridge COG-UK dataset, one was determined to be a care home resident, bringing the total CUH care home residents analysed to 72.

Statistics

All statistical analyses were performed in R. The logistic regression model used to estimate odds of 30-day mortality was coded as follows:

```
glm.fit <- glm(mortality_30_days ~ age + sex + care_status + ICU_admission + diagnostic_ct_value, data=data, family=binomial)
summary(glm.fit)
```

Odds ratios and 95% confidence intervals were derived by exponentiating the model coefficients:

```
exp(cbind(coef(glm.fit), confint(glm.fit)))
```

To produce the plot of odds ratios shown in Figure 4, the age and diagnostic Ct value continuous variables were transformed into binary categoricals using cut-offs of age $\geq$80 years and Ct value <20.

Selecting randomised sample of non-care home residents as comparison group

We selected a randomised sample of non-care home residents to use as a control group for comparison of viral lineage composition against the care home residents. Because this group was intended to be representative of non-care home community-acquired transmission, we applied the following inclusion criteria prior to randomisation:

- Patient address available
- Not one of the identified care home residents
- Not a healthcare worker (information only available for people tested at CUH)
- Not a CUH case of indeterminate, suspected or definite hospital acquired infection
- Not living in a long-term care facility other than a care home (e.g. mental health hospital, rehabilitation unit, *etc*)
- Not living in a prison

We attempted to have a roughly equivalent representation of nanopore and WSI sequenced samples as present in the care home database. Samples were selected using the R randomisation command *sample_n()* from available genomes in the CLIMB database passing QC filters. Having identified 698 samples, any cases with matching addresses that had been excluded were added to yield the final set of 700 non-care home genomes for comparison. Of the 700 non-care home samples included, we note that there were five instances of pairs of samples sharing the same address; in all five cases the pairwise SNP difference was zero or 1, and in 4/5 cases the people shared the same surname. This non-care home comparison set is only used for the viral lineage comparison shown in Figure 5, and is not part of the main care home viral cluster analysis.

Care home viral phylogenetics and cluster analysis

Consensus fasta sequences were downloaded from the MRC-CLIMB website (https://www.climb.ac.uk/) (Connor et al., 2016). Genomes were de-duplicated (1 genome per person) and passed through quality control (QC) filtering using the same criteria as in (Meredith et al., 2020): genome size >29Kb, N count <2990 (i.e. >90% coverage). Where there were multiple sequences from the same patient, we selected the sequence passing QC filters that was collected first for genomic analysis (closest to the onset of symptoms).

The 700 de-duplicated viral genomes from care home residents passing QC were aligned using MAFFT (v 7.458) (Katoh and Standley, 2013) with default settings. Command:

```
"/PATH/mafft"  --retree 2 --inputorder "multi_fasta_filename.fasta" > "alignment_filename"
```

A SNP difference matrix was produced from the alignment using *snp-dists* v 0.7.0 (Seemann, 2020) installed in a conda environment, run with the following command:

```
snp-dists -c alignment_filename.aln > snp_diff_matrix_filename.csv
```

The SNP difference matrix was manipulated in R using the *Matrix* and *tidyverse* packages to generate the SNP difference histogram and boxplots.

Phylogenetic trees were generated using IQ-TREE (v 1.6.12 built Aug 15 2019). An alignment was generated as above including a reference genome from Wuhan, China, collected December 2019 and used to root the tree (GISAID ID: EPI_ISL_402123). The IQ-TREE Model Finder Plus option was used (Kalyaanamoorthy et al., 2017) which searches from a database of available nucleotide substitution models and selects the best fit to the analysis, command line:

```
~/PATH/iqtree -s alignment_filename -m MFP
```

The best-fit nucleotide substitution model according to BIC was GTR+F+R2. The tree shown in this manuscript was produced using the GTR+F+R2 model with the ultrafast bootstrap option (Hoang et al., 2018) run through 1,000 iterations to estimate branch support values, using command:

```
~/PATH/iqtree -s alignment_filename -m GTR+F+R2 -bb 1000
```

Newick trees were manipulated in *FigTree* (v 1.4.4) to root on the Wuhan sample and put in increasing node order. Trees were visualised initially using the microreact online tool (Argimón et al., 2016), and Figure 6 was produced in R using *ggtree* (v 2.0.4) (Yu et al., 2017).

Lineage assignment

Viral lineages were assigned using the Pangolin COVID-19 Lineage Assigner web utility (COG-UK, 2020). Analysis was performed with Pangolin (Andrew Rambaut et al., 2020) version 1.1.14, lineages version 2020-05-19-2. Contextual information about lineages was taken from (A Rambaut et al., 2020), accessed 24/07/2020.

Clustering

Clusters were produced using an adaptation of the *transcluster* package (Stimson et al., 2019; Tonkin-Hill, 2020), assuming a viral mutation rate of 1e-3 substitutions/site/year (Fauver et al., 2020) and serial interval of five days (He et al., 2020; Zhang et al., 2020). We used days between first positive sampling date as a proxy for serial interval (which is the date of symptom onset, rather than swab positivity). As above, where collection date was missing, we used the date the sample was received in the Cambridge PHE laboratory. The resulting pairwise transmission probabilities were used to generate a pairwise distance matrix and clustering was performed using single linkage hierarchical clustering with the R *hclust* function. Links were only considered if they involved residents from the same care home; thus, the largest theoretical number of clusters in this analysis would be 700 (every individual is their own distinct cluster), and the smallest was 292 (one cluster for each care home).

We explored the relationship between the probability of infections being linked by $\leq$2 intermediate hosts and the resulting number of care home clusters. A higher threshold leads to more care home clusters, with greater likelihood of linked transmission within each cluster than when using a lower threshold. We pragmatically chose a cut-off of $\leq$15% probability, yielding 409 clusters. The majority of pairwise comparisons within clusters were zero or 1 SNP different and <14 days apart.

For 16/700 (2.3%) genomes, the sample that produced the analysed sequence was not the first positive test for that individual in the dataset. This could have occurred if the first positive test was not sequenced, or the sequencing failed or did not pass QC filters. This could theoretically lead to different clustering outcomes, if two cases were counted as further apart temporally than they really were from the date of first positive swab. To ensure this had not biased our findings, we re-ran the *transcluster* analysis with identical thresholds using the date of first positive test for each individual (keeping the same genomes). There was no change in the number of clusters identified (n=409).

Investigating hospital admissions for care home residents

Hospital Episode Statistics (HES) data from 26th February to 10th May 2020 were linked to cases from this study using matching NHS numbers. The data were accessed by the Public Health England Healthcare Associated Infections (HCAI) division via the PHE Data Lake. This was possible for 694/700 (99%) of the care home residents with genomes available (used in the cluster analysis); six cases could not be linked to admission data due to missing NHS numbers in the study metadata.

Hospital admission coding included transfer of care between medical units as separate admissions. We condensed these into single admissions if the time interval between the preceding discharge and the following admission was less than or equal to 1 day. i.e. an admission had to occur 2 days or more after the preceding discharge to be counted as a new admission.

Hospital admission data were parsed to yield the following outputs:

- COVID-19 related hospital admission: first positive test date was -2 to +7 days inclusive from a hospital admission date
- Suspected hospital acquired: first positive test date was +7 days from a hospital admission to +7 days from a hospital discharge, inclusive. The people testing positive in the community within 7 days of discharge from hospital are categorised as, “community onset, suspected hospital acquired”; the people testing positive after 7 days from admission but before their discharge are categorised as, “hospital onset, suspected hospital acquired”.
- For the 6 individuals with no NHS number, we assumed they were not discharged within 7 days of a positive test.

For the residents with community-onset, suspected hospital-acquired infections, we calculated the number of days the patient had been admitted to hospital prior to their positive test.

CUH HCW-Care home resident cluster analysis

To investigate transmission between healthcare workers (HCW) and care home residents we focused on CUH cases, where we had richest metadata including HCW status.

Of 6,600 PCR-positive patients, 91 had been identified as HCW. 74 of these were from the CUH HCW screening programme (which includes symptomatic, asymptomatic and household contact arms) (N. K. Jones et al., 2020; Rivett et al., 2020) and 17 had presented acutely to CUH medical services, and been identified as HCW during their initial medical clerking and subsequent note reviews. Of the 91 HCW, 76 had genomes available for analysis (breakdown: 56 samples identified through the CUH HCW screening programme, 9 CUH HCW who presented to acute medical services at CUH, and 11 HCW from community settings (paramedics and care home workers) that had been flagged as HCW through admission clerkings). Of 464 CUH cases in the study period, 71 were care home residents and 54 of these had available genomes for analysis. The total combined analysis set of CUH HCW and care home residents was therefore 76+54 = 130.

The 130 genomes were aligned using MAFFT and underwent the same cluster analysis using the *transcluster* algorithm as described above. We focused on clusters containing at least one of both a HCW and a care home resident. Transmission links between care homes were excluded as were links between HCWs. We allowed for HCWs to belong to multiple clusters from different care homes to allow for the possibility of a HCW seeding multiple care home infections. We identified 21 clusters involving both care home residents and HCWs. Of the 54 care home residents, 38 had links with HCWs within the 0.15 probability threshold. Medical notes for potential care home resident-HCW transmission pairs were reviewed by author WLH as described in (Meredith et al., 2020), with cases being categorised as strongly linked epidemiologically (e.g. the HCW documented in the care home residents' medical notes); possibly linked (e.g. both working in the hospital at the same time but not in the same wards); or no evidence of an epidemiological link.

Ethics

This study was conducted as part of surveillance for COVID-19 infections under the auspices of Section 251 of the NHS Act 2006. It therefore did not require individual patient consent or ethical approval.

The COG-UK study protocol was approved by the Public Health England Research Ethics Governance Group (reference: R&D NR0195).

Funding

This work was funded by COG-UK (supported by the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute); the Wellcome Trust; the Academy of Medical Sciences; the Health Foundation; and the Cambridge NIHR Biomedical Research Centre.

Acknowledgements

We gratefully acknowledge the invaluable contributions of all members of the Wellcome Sanger Institute Covid-19 Surveillance Team (www.sanger.ac.uk/covid-team) who have supported this project. We would also like to thank Nick Donnelly for advice with statistical analyses, and the Public Health England Hospital Acquired Infection (HCAI) division, in particular Rebecca Guy and Mehdi Minaji, for assistance accessing hospital admission data for this study.

Author contributions

WLH and MET designed and supervised the study and wrote the first draft of the manuscript. WLH, DA, CJH, BW, LWM, MH, ASJ, MDC, SP, LGC, SLC, FAK, AY, GH, TF, MLP, IGe, YC, NMB, SG, RA, EH, MAB, MS, DKJ, IJ, AA, JS, CL, GD, SJP, DPK, IG and MET collected the data. WLH, GTH, ES, DA, CJH, BW, CSB, IG and MET analysed and interpreted the data. WLH and GTH produced the figures and tables. WLH, GTH, ES, DA, CJH and MET reviewed the literature. WLH, GTH, ES, DA, CJH, BW, CSB, GD, SJP, DPK, IG and MET contributed to writing the manuscript. All authors reviewed and approved the final manuscript. Author contributions for the COVID-19 Genomics Consortium UK are shown in supplementary materials.

References

Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-stroh S. 2020. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region , January to June 2020 1–8.

Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA,

Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb genomics* **2**:e000093. doi:10.1099/mgen.0.000093

Arons MM, Hatfield KM, Reddy SC, Kimball A, James A, Jacobs JR, Taylor J, Spicer K, Bardossy AC, Oakley LP, Tanwar S, Dyal JW, Harney J, Chisty Z, Bell M, Methner M, Paul P, Carlson CM, McLaughlin HP, Thornburg N, Tong S, Tamin A, Tao Y, Uehara A, Harcourt J, Clark S, Brostrom-Smith C, Page LC, Kay M, Lewis J, Montgomery P, Stone ND, Clark TA, Honein MA, Duchin JS, Jernigan JA. 2020. Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *N Engl J Med* **382**:2081–2090. doi:10.1056/NEJMoa2008457

Bull M. 2020. A Nextflow pipeline for running the ARTIC network's fieldbioinformatics tools with a focus on ncov2019. https://github.com/connor-lab/ncov2019-artic-nf

Burton JK, Bayne G, Evans C, Garbe F, Gorman D, Honhold N, McCormick D, Othieno R, Stevenson J, Swietlik S, Templeton K, Tranter M, Willocks L, Guthrie B. 2020. Evolution and impact of COVID-19 outbreaks in care homes: population analysis in 189 care homes in one geographic region. *medRxiv* 2020.07.09.20149583. doi:10.1101/2020.07.09.20149583

Byrne AW, McEvoy D, Collins AB, Hunt K, Casey M, Barber A, Butler F, Griffin J, Lane EA, McAloon C, O'Brien K, Wall P, Walsh KA, More SJ. 2020. Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open* **10**:e039856. doi:10.1136/bmjopen-2020-039856

COG-UK. 2020. Pangolin COVID-19 Lineage Assigner. https://pangolin.cog-uk.io/

Collier D, Assennato S, Sithole N, Sharrocks K, Ritchie A, Ravji P, Routledge M, Sparkes D, Skittrall J, Warne B, smielewska A, RAMSEY I, Goel N, CURRAN M, ENOCH D, TASSELL R, LINEHAM M, VAGHELA D, LEONG C, Gupta R. 2020. Rapid point of care nucleic acid testing for SARS-CoV-2 in hospitalised patients: a clinical trial and implementation study. *Cell Reports Med* 100062. doi:10.1101/2020.05.31.20114520

Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Thompson SE, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ. 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb genomics* **2**:e000086. doi:10.1099/mgen.0.000086

Coronavirus testing - GOV.UK. 2020. https://www.gov.uk/government/news/coronavirus-testing

COVID-19 policy tracker | The Health Foundation. 2020. https://www.health.org.uk/news-and-comment/charts-and-infographics/covid-19-policy-tracker

CQC. 2020a. Service types | Care Quality Commission. https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types#care-homes-without-nursing

CQC. 2020b. What is registration? | Care Quality Commission. https://www.cqc.org.uk/guidance-providers/registration/what-registration

CQC. 2020c. Using CQC data | Care Quality Commission. https://www.cqc.org.uk/about-us/transparency/using-cqc-data

CQC. 2015. The scope of registration.

Curran ET. 2017. Infection outbreaks in care homes: prevention and management, Nursing Times. https://www.nursingtimes.net/clinical-archive/infection-control/infection-outbreaks-in-care-homes-prevention-and-management-14-08-2017/

Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, Vogels CBF, Brito AF, Alpert T, Muyombwe A, Razeq J, Downing R, Cheemarla NR, Wyllie AL, Kalinich CC, Ott IM, Quick J, Loman NJ, Neugebauer KM, Greninger AL, Jerome KR, Roychoudhury P, Xie H, Shrestha L, Huang ML, Pitzer VE, Iwasaki A, Omer SB, Khan K, Bogoch II, Martinello RA, Foxman EF, Landry ML, Neher RA, Ko AI, Grubaugh ND. 2020. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**:990–996. doi:10.1016/j.cell.2020.04.021

Goldberg SA, Lennerz J, Klompas M, Mark E, Pierce VM, Thompson RW, Pu CT, Ritterhouse LL, Dighe A, Rosenberg ES, Grabowski DC. 2020. Presymptomatic Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 Among Residents and Staff at a Skilled Nursing Facility: Results of Real-time Polymerase Chain Reaction and Serologic Testing. *Clin Infect Dis*. doi:10.1093/cid/ciaa991

Graham NSN, Junghans C, Downes R, Sendall C, Lai H, McKirdy A, Elliott P, Howard R, Wingfield D, Priestman M, Ciechonska M, Cameron L, Storch M, Crone MA, Freemont PS, Randell P, McLaren R, Lang N, Ladhani S, Sanderson F, Sharp DJ. 2020. SARS-CoV-2 infection, clinical features and outcome of COVID-19 in United Kingdom nursing homes. *J Infect*. doi:10.1016/j.jinf.2020.05.073

Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM, Brackney DE, Grewal S, Gurfield N, Van Rompay KKA, Isern S, Michael SF, Coffey LL, Loman NJ, Andersen KG. 2019. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* **20**:1–19. doi:10.1186/s13059-018-1618-7

He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, Mo X, Chen Y, Liao B, Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM. 2020. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* **26**:672–675. doi:10.1038/s41591-020-0869-5

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast

Bootstrap Approximation. Molecular biology and evolution. *Mol Biol Evol* **35**:518–522. doi:10.5281/zenodo.854445

ICS. 2020. Assessing whether COVID-19 patients will benefit from critical care, and an objective approach to capacity challenges. https://www.rcplondon.ac.uk/file/20726/download

Jones NK, Rivett L, Sparkes D, Forrest S, Sridhar S, Young J, Pereira-Dias J, Cormie C, Gill H, Reynolds N, Wantoch M, Routledge M, Warne B, Levy J, Jiménez WDC, Samad FNB, McNicholas C, Ferris M, Gray J, Gill M, Curran MD, Fuller S, Chaudhry A, Shaw A, Bradley JR, Hannon GJ, Goodfellow IG, Dougan G, Smith KGC, Lehner PJ, Wright G, Matheson NJ, Baker S, Weekes MP. 2020. Effective control of sars-cov-2 transmission between healthcare workers during a period of diminished community prevalence of covid-19. *Elife* **9**:1–10. doi:10.7554/eLife.59391

Jones NR, Qureshi ZU, Temple RJ, Larwood JPJ, Greenhalgh T, Bourouiba L. 2020. Two metres or one: what is the evidence for physical distancing in covid-19? *BMJ* **370 :m3223**. doi:https://doi.org/10.1136/bmj.m3223

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**:587–589. doi:10.1038/nmeth.4285

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Kemenesi G, Kornya L, Tóth GE, Kurucz K, Zeghbib S, Somogyi BA, Zöldi V, Urbán P, Herczeg R, Jakab F. 2020. Nursing homes and the elderly regarding the COVID-19 pandemic: situation report from Hungary. *GeroScience* **2**. doi:10.1007/s11357-020-00195-z

Ladhani SN, Chow JY, Janarthanan R, Fok J, Crawley-Boevey E, Vusirikala A, Fernandez E, Perez MS, Tang S, Dun-Campbell K, Wynne-Evans E, Bell A, Patel B, Amin-Chowdhury Z, Aiano F, Paranthaman K, Ma T, Saavedra-Campos M, Myers R, Ellis J, Lackenby A, Gopal R, Patel M, Chand M, Brown K, Hopkins S, Consortium CG, Shetty N, Zambon M, Ramsay ME. 2020. Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced CoVID-19 outbreak investigations in London care Homes. *J Infect* **81**:621–624. doi:10.1016/j.jinf.2020.07.027

Lansbury LE, Brown CS, Nguyen-Van-Tam JS. 2017. Influenza in long-term care facilities. *Influenza Other Respi Viruses* **11**:356–366. doi:10.1111/irv.12464

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM **00**:1–3.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. doi:10.1093/bioinformatics/btp352

MacFadden DR, McGeer A, Athey T, Perusini S, Olsha R, Li A, Eshaghi A, Gubbay JB, Hanage WP. 2018. Use of genome sequencing to define institutional influenza outbreaks, Toronto, Ontario, Canada, 2014-15. *Emerg Infect Dis* **24**:492–497. doi:10.3201/eid2403.171499

McAuslane H, Morgan D. 2014. Communicable Disease Outbreak Management. Operational guidance. *Public Heal Engl* 1–66.

Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, Curran MD, Parmar S, Caller LG, Caddy SL, Khokhar FA, Yakovleva A, Hall G, Feltwell T, Forrest S, Sridhar S, Weekes MP, Baker S, Brown N, Moore E, Popay A, Roddick I, Reacher M, Gouliouris T, Peacock SJ, Dougan G, Török ME, Goodfellow I. 2020. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* **0**. doi:10.1016/S1473-3099(20)30562-4

Middleton J, Reintjes R, Lopes H. 2020. Meat plants-a new front line in the covid-19 pandemic. *BMJ* **370**:1–2. doi:10.1136/bmj.m2716

NICE. 2020. NICE guideline NG159. COVID-19 rapid guideline: critical care in adults. 2. Admission to critical care.

ONS. 2020a. Deaths involving COVID-19, England and Wales - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020

ONS. 2020b. Deaths involving COVID-19 in the care sector, England and Wales - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional

ONS. 2019a. Families and households in the UK - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2019

ONS. 2019b. Families and households - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholdsfamiliesandhouseholds

ONS. 2014. Changes in the Older Resident Care Home Population between 2001 and 2011 - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/changesintheolderresidentcarehomepopulationbetween2001and2011/2014-08-01

Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid

efficient extraction of SNPs from multi-FASTA alignments. *Microb genomics* **2**:e000056. doi:10.1099/mgen.0.000056

Public Health England. 2020a. Weekly Coronavirus Disease 2019 (COVID-19) Surveillance Report Confirmed cases in England. Year: 2020, Week: 20. *Summ COVID-19 Surveill Syst*.

Public Health England. 2020b. COVID-19 Personal protective equipment (PPE) – resource for care workers working in care homes during sustained COVID-19 transmission in England.

Quicke K, Gallichote E, Sexton N, Young M, Janich A, Gahm G, Carlton EJ, Ehrhart N, Ebel GD. 2020. Longitudinal Surveillance for SARS-CoV-2 RNA Among Asymptomatic Staff in Five Colorado Skilled Nursing Facilities: Epidemiologic, Virologic and Sequence Analysis. *medRxiv* 2020.06.08.20125989. doi:10.1101/2020.06.08.20125989

Rambaut A, Holmes E, O'Toole Á, Hill V, McCrone J, Ruis C, du Plessis L, Pybus O. 2020. SARS-CoV-2 lineages. https://cov-lineages.org/descriptions.html

Rambaut Andrew, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. doi:10.1038/s41564-020-0770-5

Rivett L, Sridhar S, Sparkes D, Routledge M, Jones NK, Forrest S, Young J, Pereira-Dias J, Hamilton WL, Ferris M, Torok ME, Meredith L, Curran M, Fuller S, Chaudhry A, Shaw A, Samworth RJ, Bradley JR, Dougan G, Smith KGC, Lehner PJ, Matheson NJ, Wright G, Goodfellow I, Baker S, Weekes MP. 2020. Screening of healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-19 transmission. *Elife* **9**:1–20. doi:10.7554/eLife.58728

Seemann T. 2020. tseemann/snp-dists: Pairwise SNP distance matrix from a FASTA sequence alignment. https://github.com/tseemann/snp-dists

Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. 2019. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Mol Biol Evol* **36**:587–603. doi:10.1093/molbev/msy242

Strausbaugh LJ, Sukumar SR, Joseph CL. 2003. Infectious disease outbreaks in nursing homes: An unappreciated hazard for frail elderly persons. *Clin Infect Dis* **36**:870–876. doi:10.1086/368197

Tonkin-Hill G. 2020. gtonkinhill/fasttranscluster. https://github.com/gtonkinhill/fasttranscluster

UK government. 2020a. COVID-19: number of outbreaks in care homes - management information - GOV.UK. https://www.gov.uk/government/statistical-data-sets/covid-19-number-of-outbreaks-in-care-homes-management-information

UK government. 2020b. Update on policies for visiting arrangements in care homes. https://www.gov.uk/government/publications/visiting-care-homes-during-coronavirus/update-on-policies-for-visiting-arrangements-in-care-homes#section-4

UK government. 2020c. Vivaldi 1: COVID-19 care homes study report - GOV.UK. https://www.gov.uk/government/publications/vivaldi-1-coronavirus-covid-19-care-homes-study-report/vivaldi-1-covid-19-care-homes-study-report

UK government. 2020d. Get coronavirus tests for a care home - GOV.UK. https://www.gov.uk/apply-coronavirus-test-care-home

Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, Curtis HJ, Mehrkar A, Evans D, Inglesby P, Cockburn J, McDonald HI, MacKenna B, Tomlinson L, Douglas IJ, Rentsch CT, Mathur R, Wong AYS, Grieve R, Harrison D, Forbes H, Schultze A, Croker R, Parry J, Hester F, Harper S, Perera R, Evans SJW, Smeeth L, Goldacre B. 2020. OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature*. doi:10.1038/s41586-020-2521-4

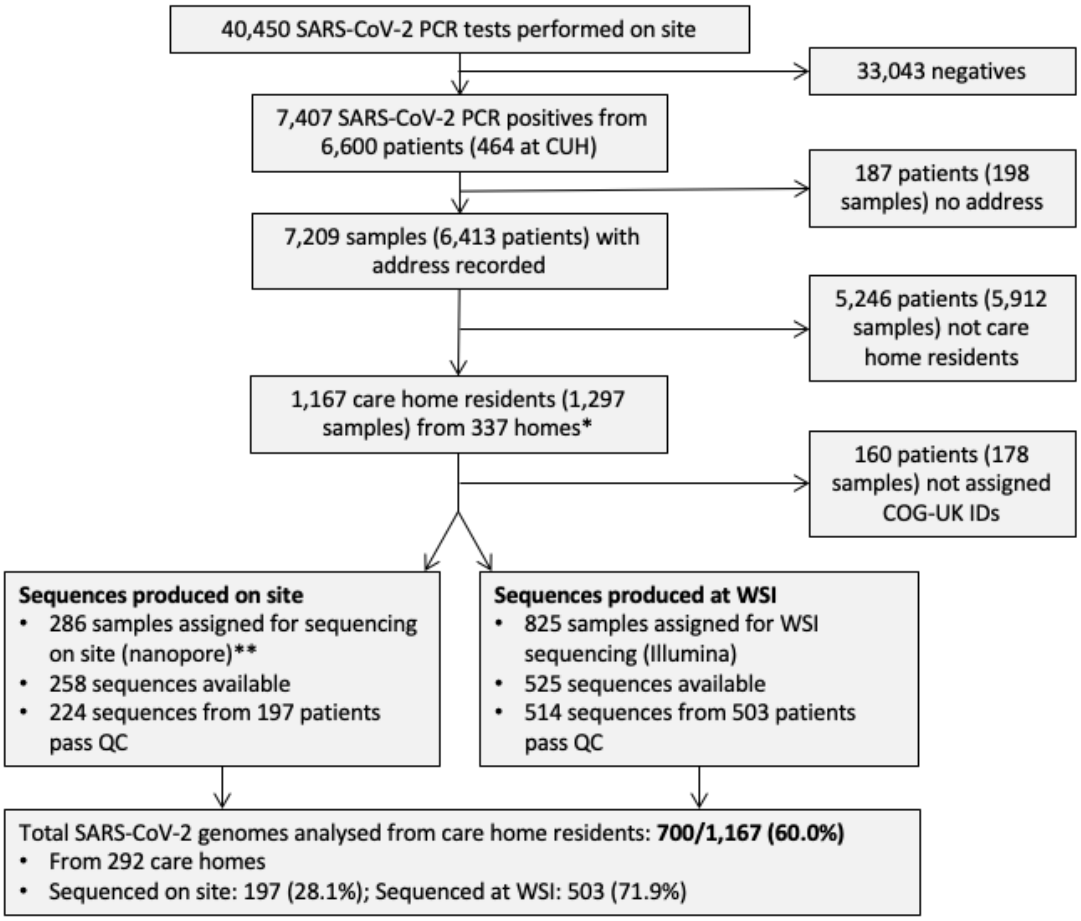
Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* **579**:265–269. doi:10.1038/s41586-020-2008-3

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods Ecol Evol* **8**:28–36. doi:10.1111/2041-210X.12628

Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen Xinghui, Li M, Zheng W, Yi L, Chen Xinhua, Wu Q, Liang Y, Wang X, Yang J, Sun K, Longini IM, Halloran ME, Wu P, Cowling BJ, Merler S, Viboud C, Vespignani A, Ajelli M, Yu H. 2020. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect Dis* **20**:793–802. doi:10.1016/S1473-3099(20)30230-9

Main display items

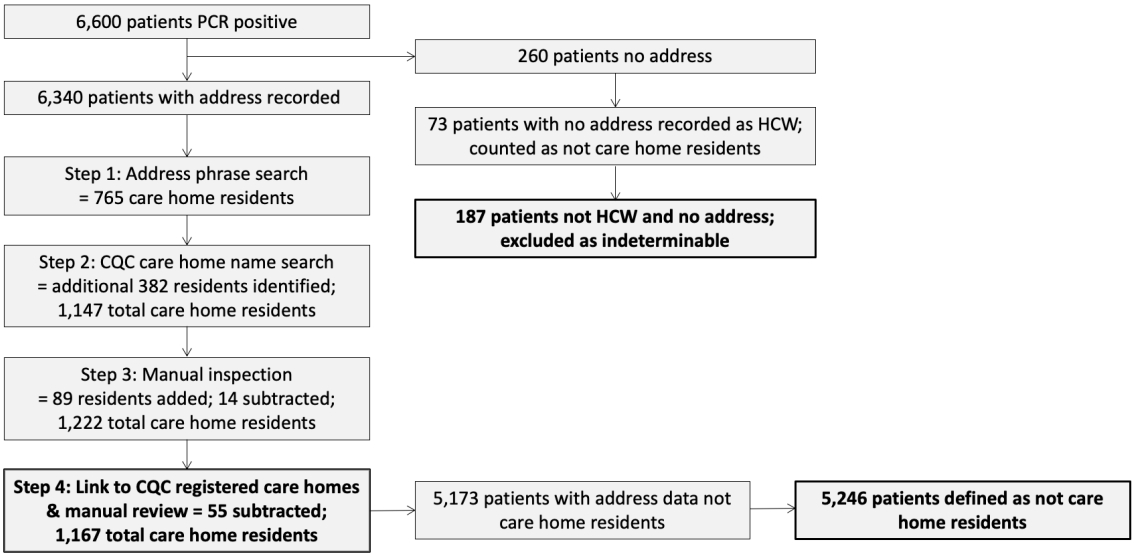
Figure 1



Study flow diagram

765 care home residents were identified by having care-related search terms in their addresses; a further 382 had CQC registered care home names in their addresses; 89 were identified manually and 69 were removed after linking the addresses to CQC registered care home names and manual review. * 193 care homes were registered with the CQC as being residential homes without nursing care and 144 had nursing care available. ** Samples were selected for sequencing on site if they were CUH inpatients or CUH HCW, where we prioritised rapid turnaround time to investigate hospital-acquired infections, plus a randomised selection of other EoE samples to provide genomic context. The remainder, where available, were sent to WSI for sequencing.

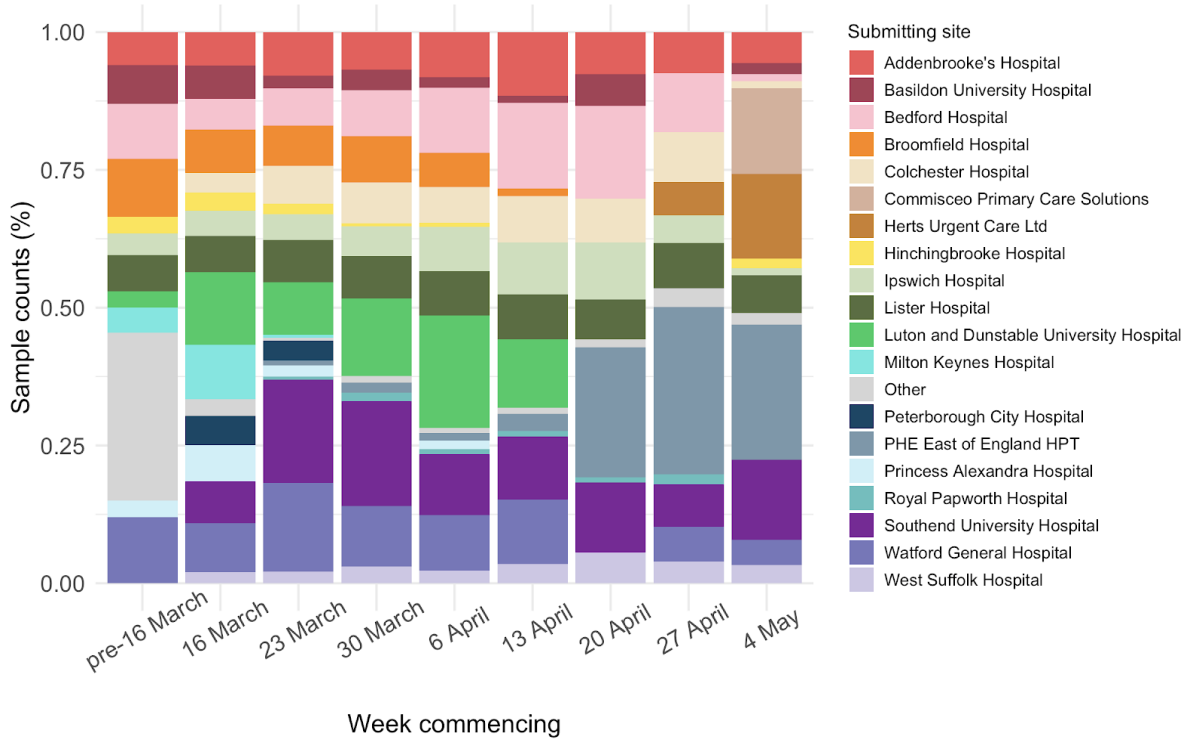
Figure 1, supplement 1



Flow diagram for identifying care homes from Cambridge-COGUK metadata

Steps for identifying care home residents. First, we searched for phrase matches in the address field of the patient electronic healthcare records ("care home", "nursing home", *etc*). Second, we searched for care home names in the patient address field using a dataset of care homes from the Care Quality Commission (CQC). The resulting list was manually inspected and every care home included in the study was linked to a registered CQC care home, allowing us to use CQC coding of whether nursing care was or was not available at each care home (referred to as "nursing homes" and "residential homes" in this study, respectively). If the address information was incomplete (no postcode and/or no address line) then the case was excluded as impossible to determine whether or not the patient was from a care home, unless the person was known to be a healthcare worker (HCW), in which case it was assumed they were not a care home resident. This process yielded the final result of 1,167 care home residents from 337 care homes; 5,246 individuals that were not care home residents, and 187 individuals that were indeterminable.

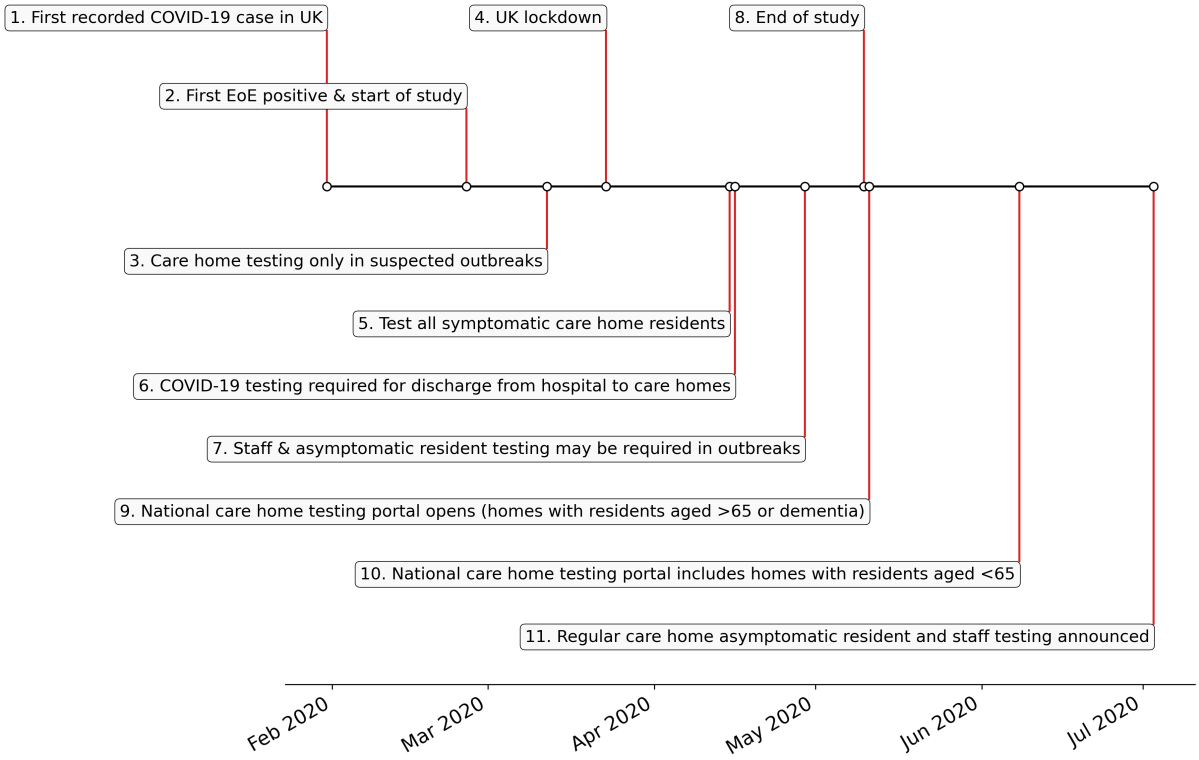
Figure 1, supplement 2



Breakdown of main organisations submitting samples to Cambridge PHE Laboratory over study period per week

Only showing sites that submitted samples from >50 people with positive test results over study period, otherwise counted as "Other". To maintain patient anonymity, per time interval only showing sites that submitted samples from >5 people with positive test results (otherwise counted as "Other"). Data prior to 16 March is amalgamated here due to low sample numbers. Note that over the course of the study, some sites changed testing provider from CMPHL as further testing sites became available around the region. This explains some of the variation in the relative proportion of cases submitted from each site. The numbers reported here do not necessarily reflect total case numbers for each hospital or submitting organisation, as tests may have been performed elsewhere or metadata not collected in this study; the numbers are included purely to indicate where the samples included in this study originated from.

Figure 1, supplement 3



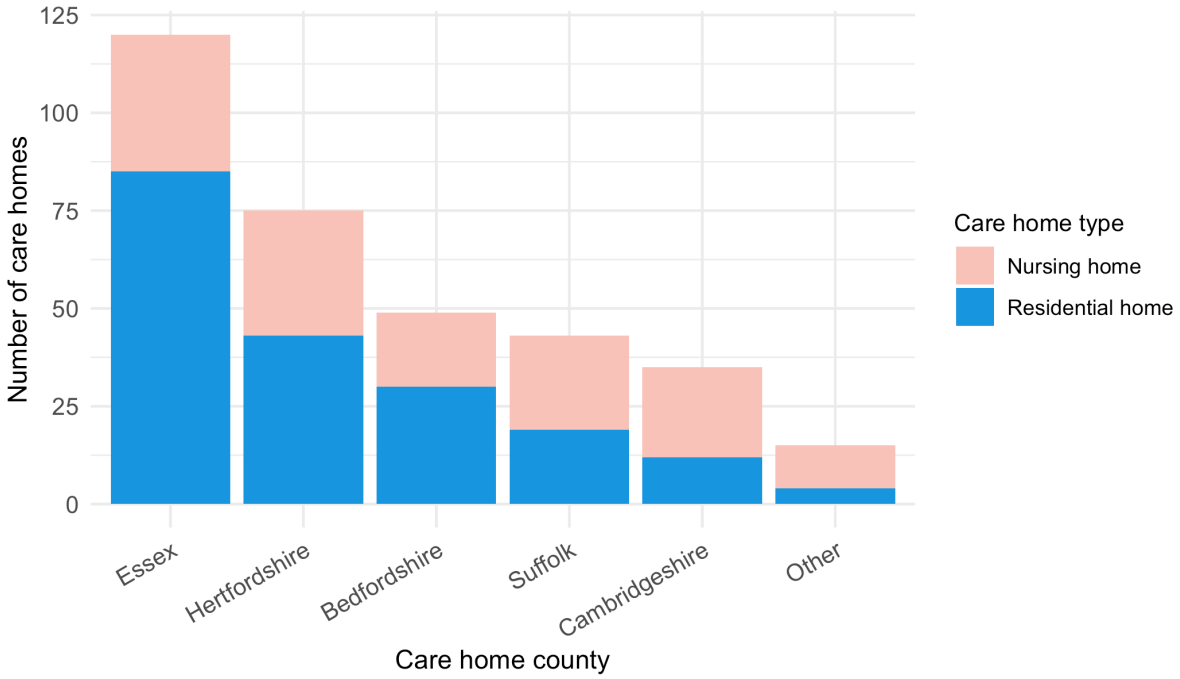
UK care home testing policy timeline

1. 31st January - first recorded case of covid-19 in the UK
2. 26th February - first case of COVID-19 in the East of England; start date of this study
3. 12th March – individuals in the community advised to self-isolate for 7 days, without testing. Testing only offered to care homes in the context of a suspected outbreak
4. 23rd March - UK lockdown officially begins
5. 15th April – action plan announced to test all symptomatic residents in care homes, plus testing of all residents prior to admission to care home from hospital
6. 29th April – testing guidance amended to reflect that asymptomatic as well symptomatic residents and staff in care homes may need to be tested as part of an outbreak
7. Policy for COVID-19 testing prior to discharge to care homes instigated 16th April: https://www.gov.uk/government/publications/coronavirus-covid-19-adult-social-care-action-plan/covid-19-our-action-plan-for-adult-social-care
8. 10th May - end date of this study
9. 11th May – national whole care home testing portal (offering a single test to all staff and residents) goes live for care homes with residents aged 65 years and over or dementia patients
10. 8th June – national whole care home testing portal extends eligibility to care homes with residents aged under 65 years

11. 3rd July – announcement that regular asymptomatic testing for care home staff and residents will be rolled out through the national whole care home testing portal in July for homes with residents aged over 65 years or dementia patients

References: ("Coronavirus testing - GOV.UK," 2020, "COVID-19 policy tracker | The Health Foundation," 2020)

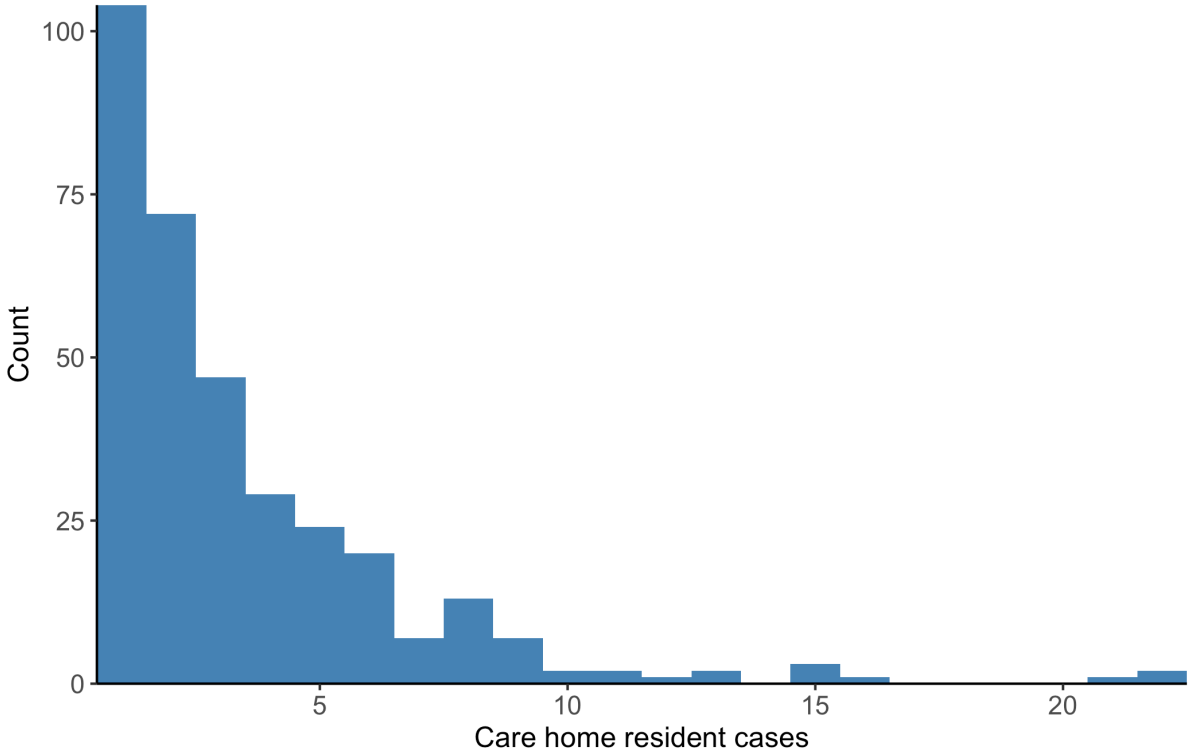
Figure 2



Care home locations by county, showing nursing and residential homes

Only showing the five counties with the largest number of cases (all >25) to preserve patient anonymity. Definitions of "nursing home" and "residential home" are based on Care Quality Commission (CQC) information on whether nursing care is or is not present. If no nursing care is available the home is classified as a residential home. If the same care home offers both nursing and residential care then the home is classified as a nursing home.

Figure 2, supplement 1

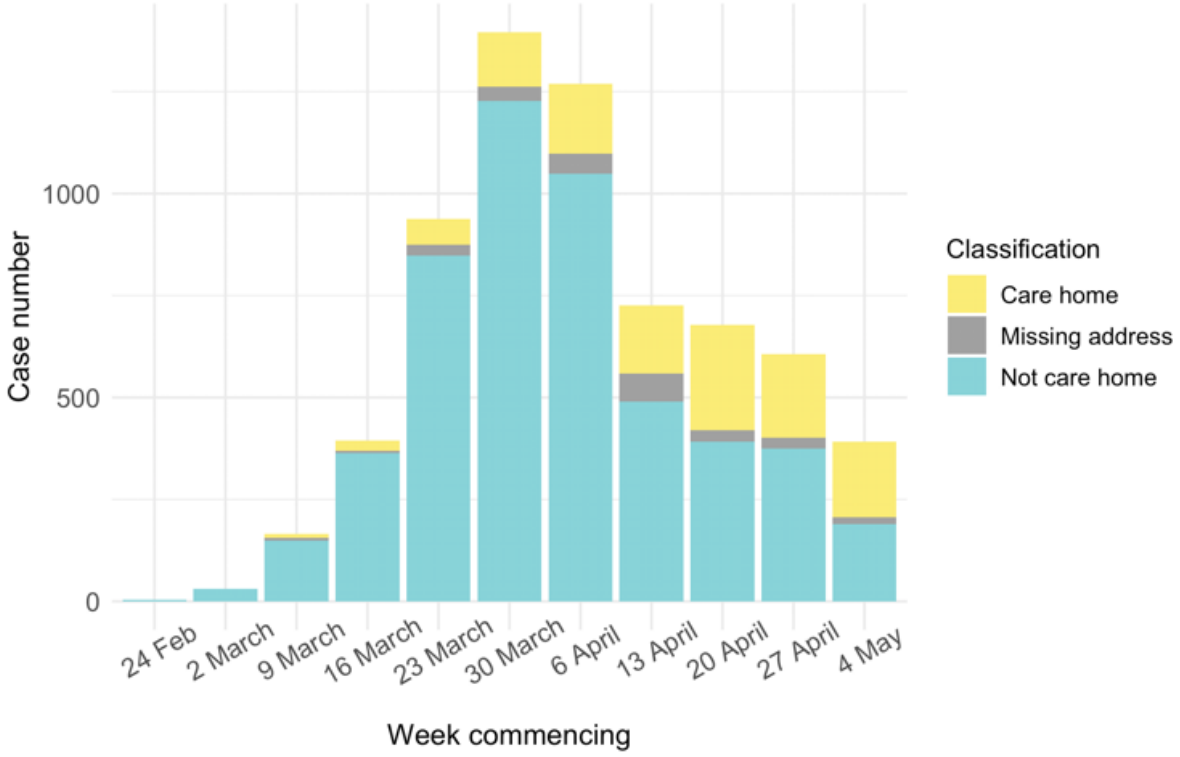


Distribution of cases per care home

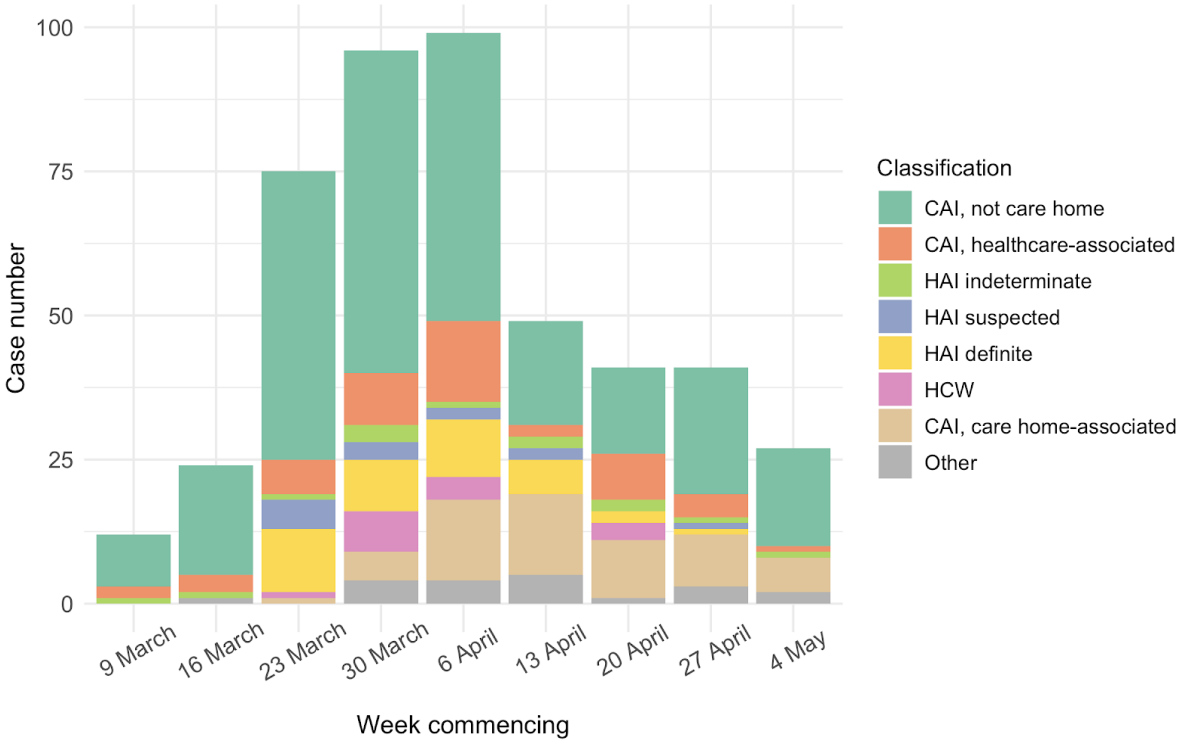
The number of positive cases per care home was highly skewed, such that a relatively small number of care homes contributed a large proportion of cases (right-hand side of the plot). Plot produced with R package *ggplot2* using geom_histogram with binwidth=1.

Figure 3

A


Case number
1000
500
0
24 Feb
2 March
9 March
16 March
23 March
30 March
6 April
13 April
20 April
27 April
4 May
Week commencing
Classification
Care home
Missing address
Not care home

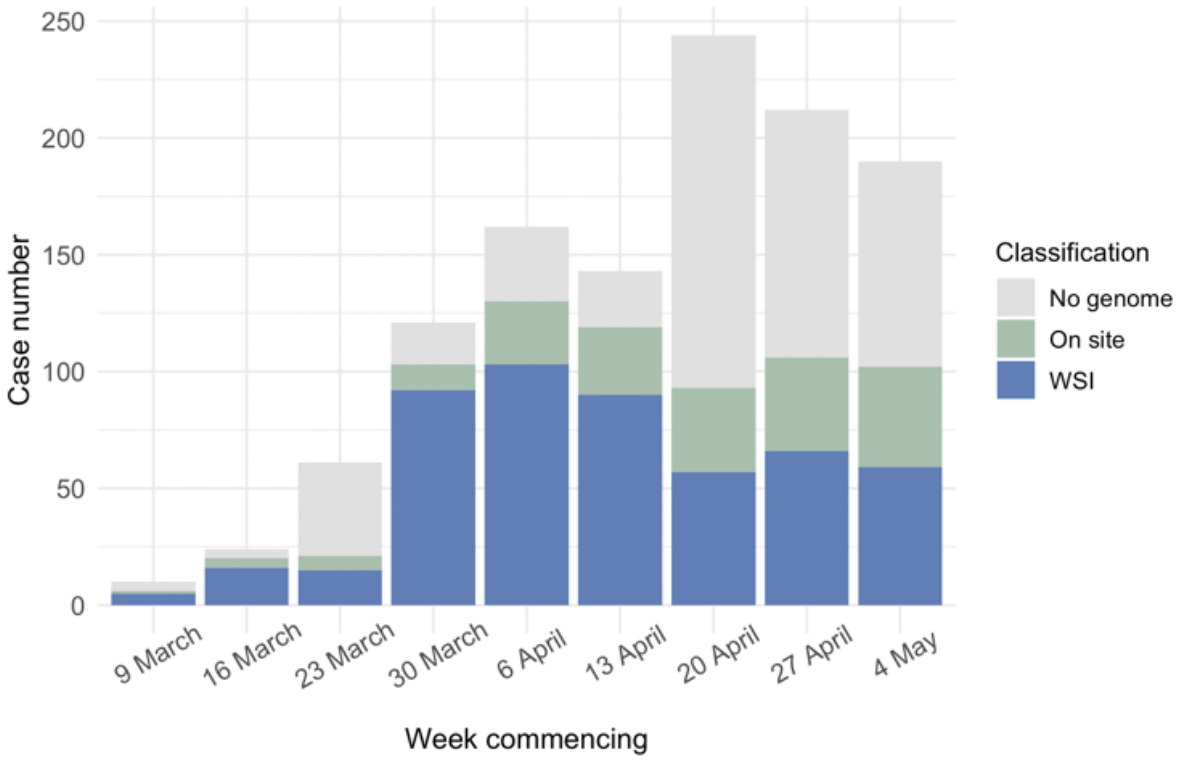
B


Case number
100
75
50
25
0
9 March
16 March
23 March
30 March
6 April
13 April
20 April
27 April
4 May
Week commencing
Classification
CAI, not care home
CAI, healthcare-associated
HAI indeterminate
HAI suspected
HAI definite
HCW
CAI, care home-associated
Other

Epidemic curves for EoE and CUH showing care home residents

Number of positive cases per week over the study period for different infection sources, for all samples tested from EoE at the Cambridge PHE laboratory (A), or those tested at CUH acute medical services (B). Peak of the epidemic for samples tested at the Cambridge PHE laboratory and CUH acute medical services were weeks commencing 30th March and 6th April, respectively. UK lockdown started 23rd March 2020. In both settings a prolonged right-hand "tail" was observed as case numbers gradually fell. The relative proportion of cases admitted from care homes increased over this period for both sample sets, while the contribution of general community cases fell more quickly. However, interpreting these trends is confounded by the changing profile of COVID-19 testing nationally and regionally. Note that if the patient address was missing, and they were not a HCW, then the care home status was undetermined. CAI = Community Acquired Infection; HAI = hospital acquired infection; HCW = healthcare worker; "Other" mainly comprise inpatient transfers from other hospitals to CUH for which metadata was lacking to determine the infection category. CAI was considered "healthcare-associated" if there had been healthcare contact within 14 days of first positive swab. The three categories of HAI were defined based on the difference in days between admission and first positive swab, reflecting increasing likelihood of hospital acquisition: indeterminate = 3-6 days; suspected 7-14 days; definite >14 days (as used in (Meredith et al., 2020), with further detail in Methods).

1 **Figure 3, supplement 1**



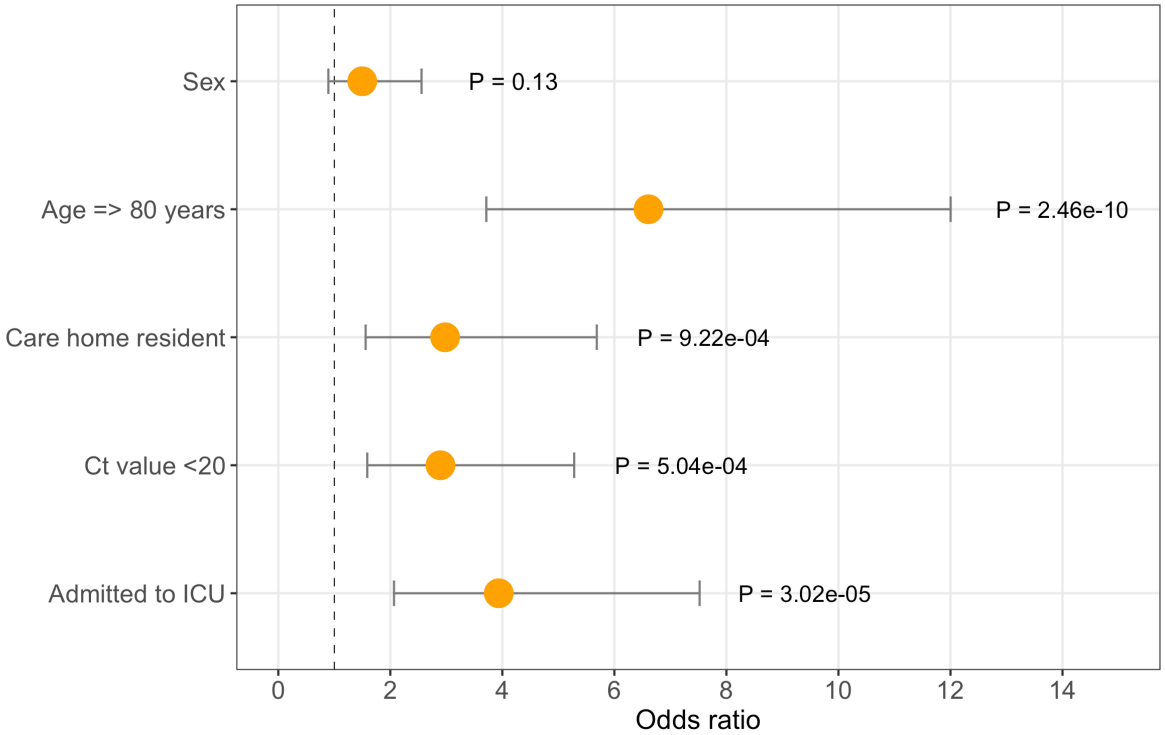
2

3 **Care home residents per week showing genome sequencing site**

4 Plot shows total care home residents testing positive per week over the study period, showing number
5 of care home residents with genomes included in the study broken down by sequencing location (on
6 site in the Department of Pathology, Division of Virology or at the Wellcome Sanger Institute).

7
8

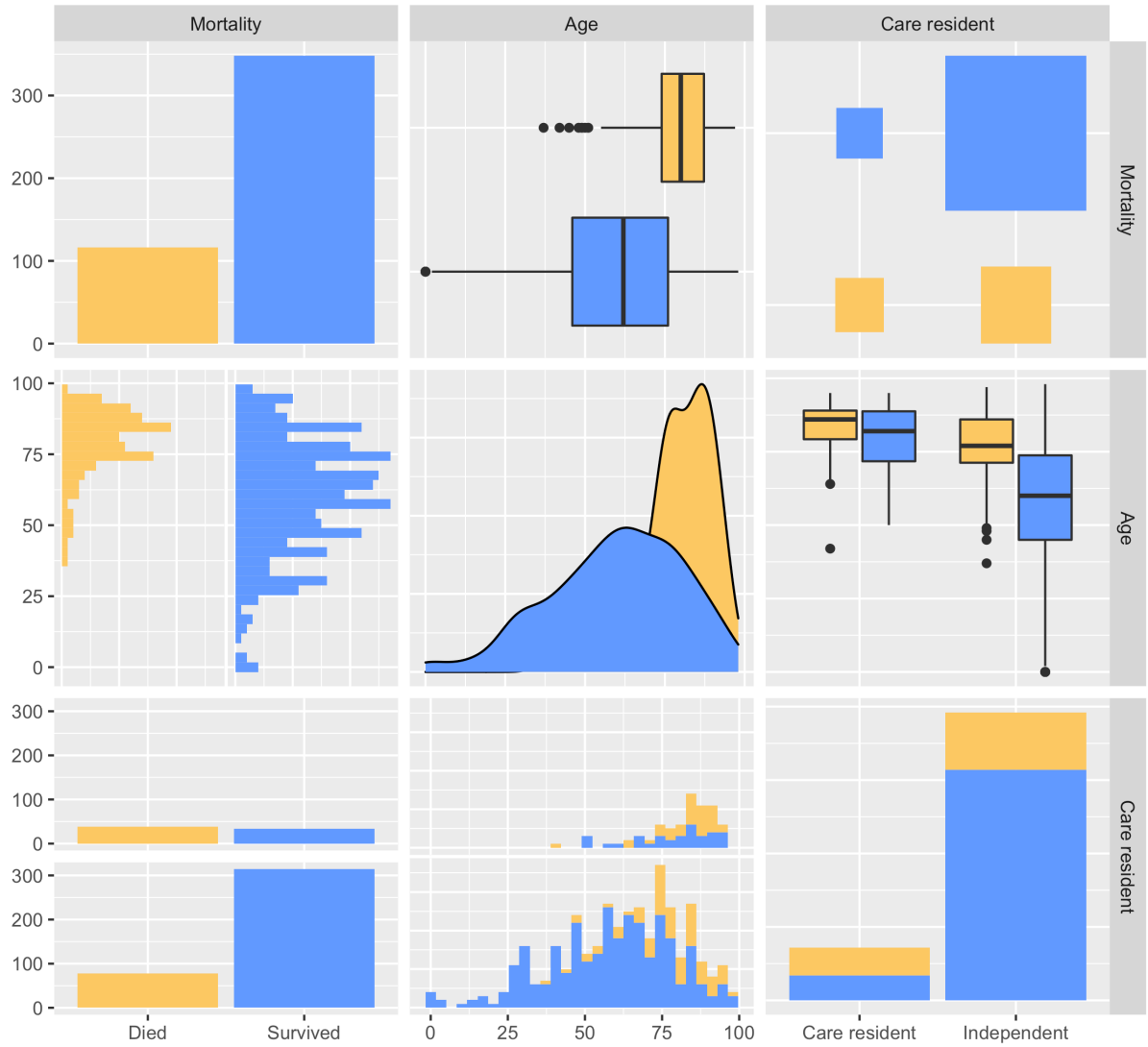
Figure 4



Odds ratios for mortality at 30 days

Logistic regression analysis showing odds of death at 30 days (with 95% confidence intervals) for five available metadata variables: patient sex, age (here categorised as ≥80 years), whether they were a care home resident, the diagnostic Ct value (here categorised as <20), and whether they were admitted to the intensive care unit. Overall there were 116 deaths within 30 days of diagnosis (out of 464 CUH patients). ICU = intensive care unit. Ct = Cycle threshold for diagnostic PCR.

Figure 4, supplement 1

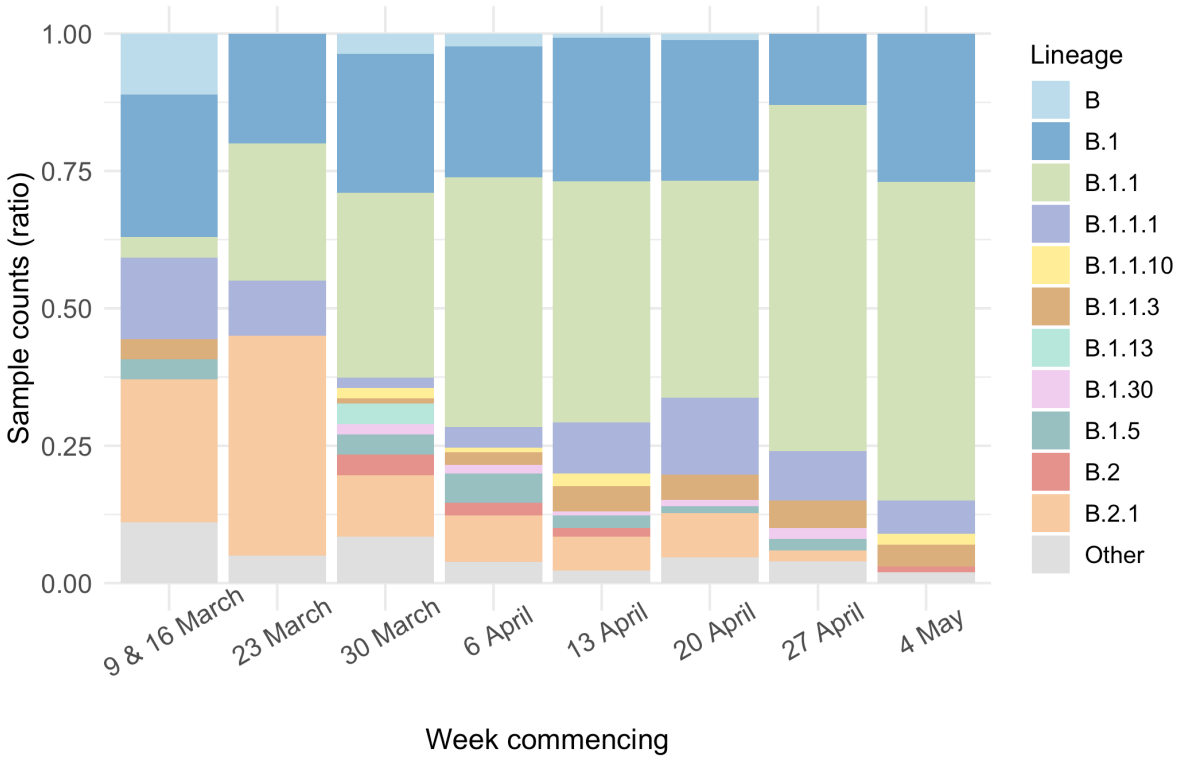


Pairwise comparisons of mortality at 30 days, age and whether the person was a care home resident

Each plot compares two of these three variables to visualise cross-associations, and the data are divided in each case into individuals that died (yellow) or survived (blue). The plot was produced using *GGally::ggpairs().*

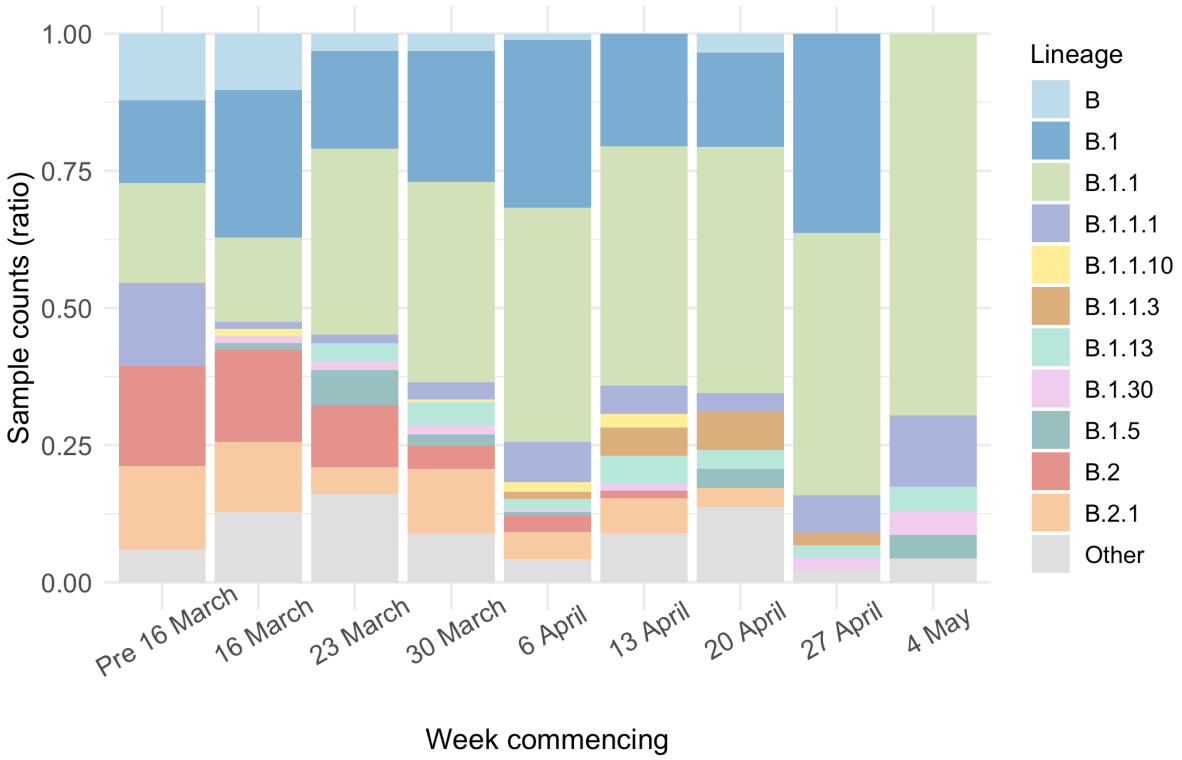
1 **Figure 5**

2 **A**



3

4 **B**



5

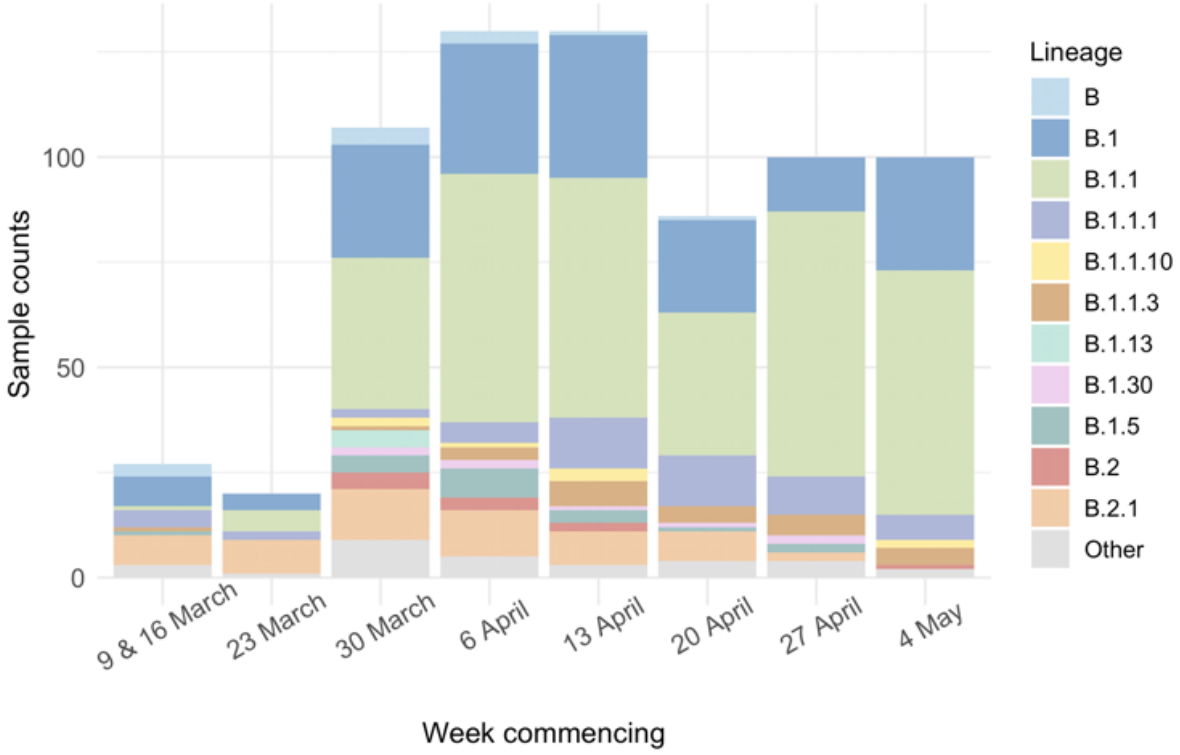
6

7 **Viral lineage compositions in care home and non-care home samples**

Plots showing the ratios of SARS-CoV-2 viral lineages for 700 care home resident genomes (A) and a randomly selected subset of 700 non-care home residents (B). The proportion of lineage B.1.1 increased over the study period in both care home and non-care home residents. Lineages defined using *pangolin*.

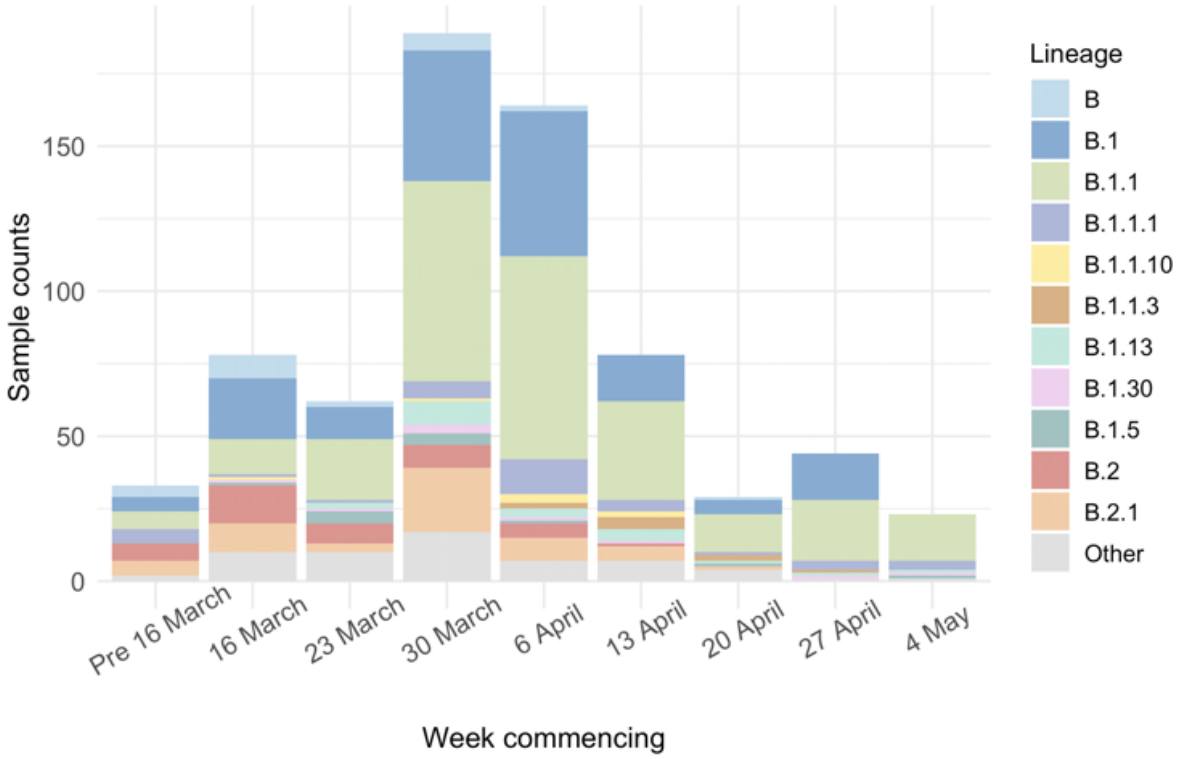
1 **Figure 5, Supplement 1**

2 **A**



3

4 **B**



5

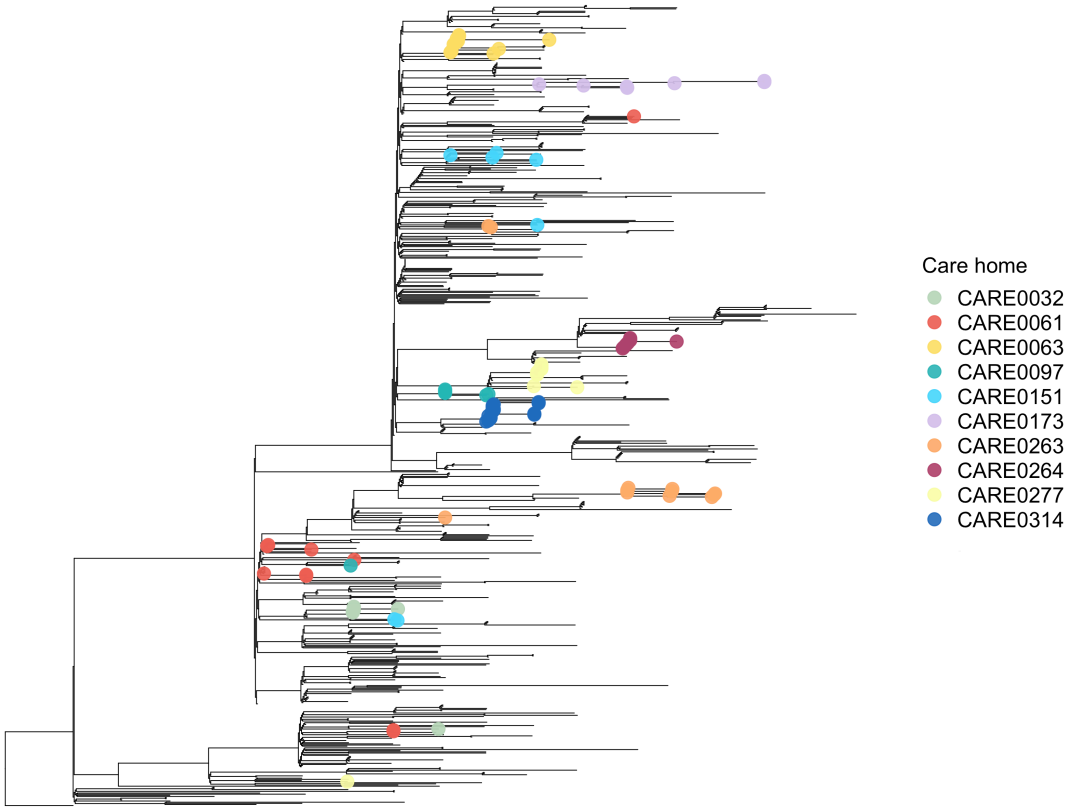
6

7 **Viral lineage compositions in care home and non-care home samples by count**

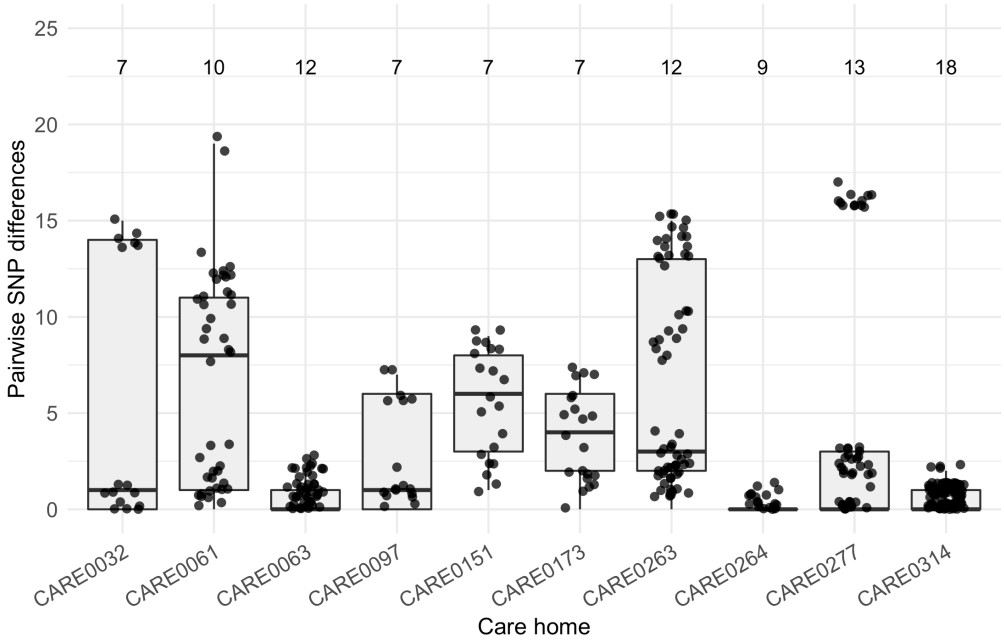
Plots showing the counts of SARS-CoV-2 viral lineages for 700 care home resident genomes (A) and a randomly selected subset of 700 non-care home residents (B). Lineages defined using *pangolin*.

Figure 6

A


Care home
CARE0032
CARE0061
CARE0063
CARE0097
CARE0151
CARE0173
CARE0263
CARE0264
CARE0277
CARE0314

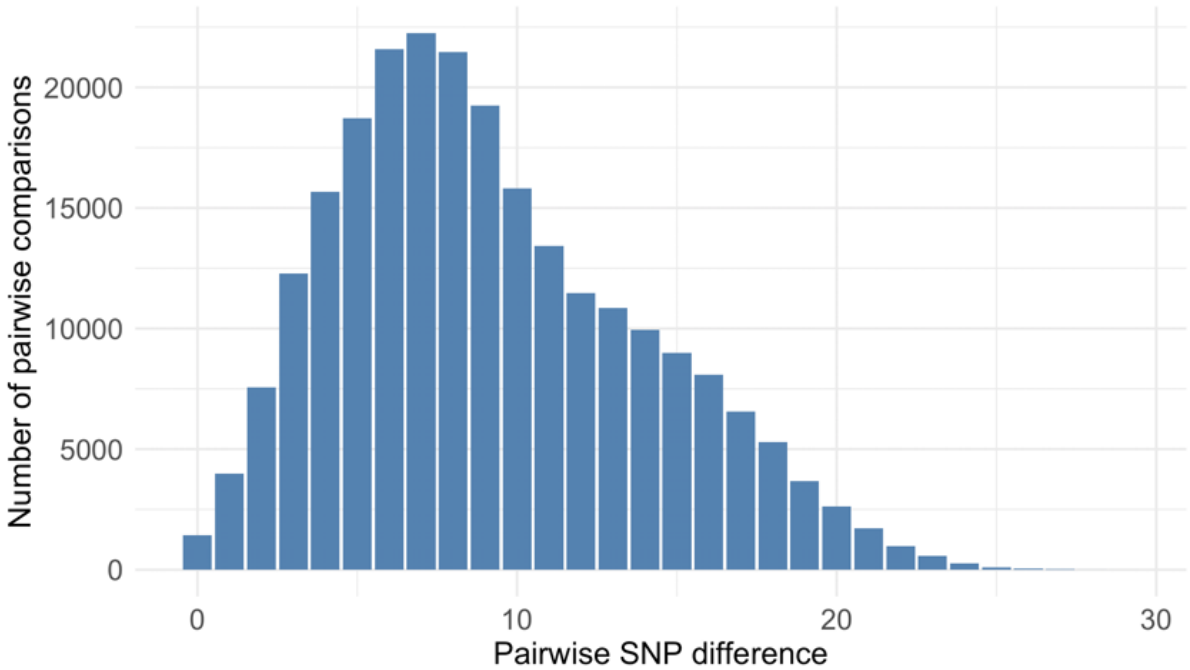
B


Pairwise SNP differences
25
20
15
10
5
0
7
10
12
7
7
7
12
9
13
18
CARE0032
CARE0061
CARE0063
CARE0097
CARE0151
CARE0173
CARE0263
CARE0264
CARE0277
CARE0314
Care home

Care home clustering on viral phylogenetic tree and within-care home pairwise SNP differences

A. Phylogenetic tree of 700 care home SARS-CoV-2 genomes, rooted on a sample from Wuhan, China, collected December 2019. Colour bar indicates the ten care homes with the largest number of genomes. B. Distributions of pairwise SNP differences for the ten care homes with the largest number of genomes (as shown in panel A). Numbers above each box indicate the number of genomes present from that care home. Among the ten care homes with the largest number of genomes, some clustered closely on the phylogenetic tree with low pairwise SNP differences (e.g. CARE0063, CARE0264, CARE0314); in contrast, some care homes were distributed across the tree with higher pairwise SNP differences (e.g. CARE0061, CARE0151, CARE0173, CARE0263).

1 **Figure 6, supplement 1**



2

3 **Distribution of pairwise SNP differences between care home samples**

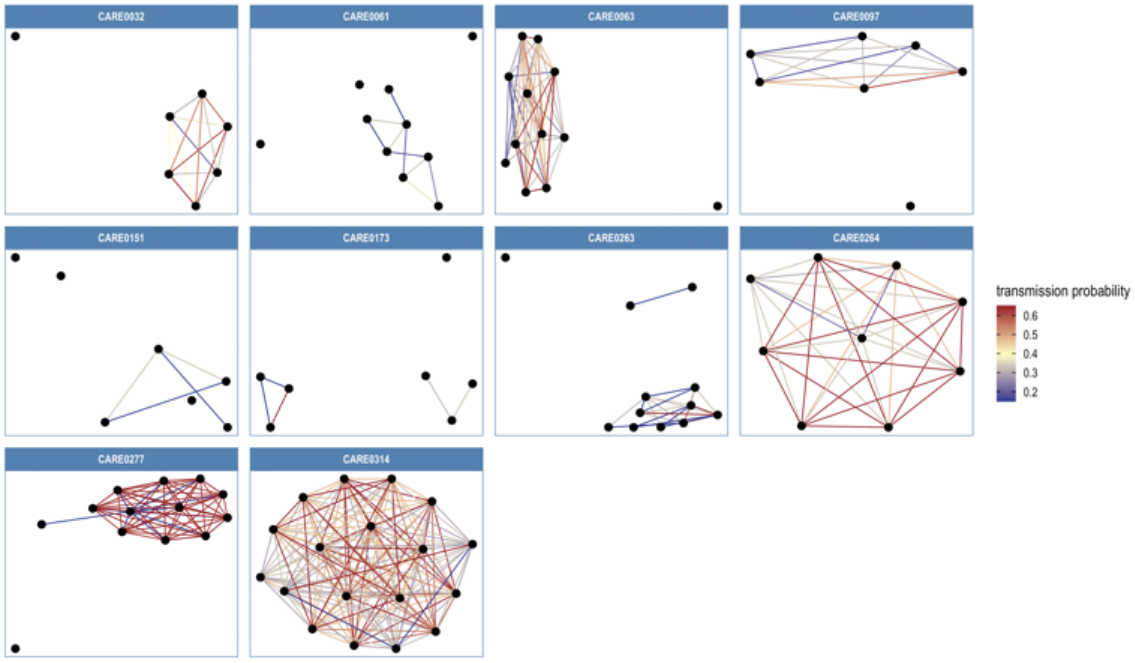
4 Pairwise SNP differences between the 700 care home residents (244,650 comparisons). Median
5 pairwise SNP difference was 8.

6

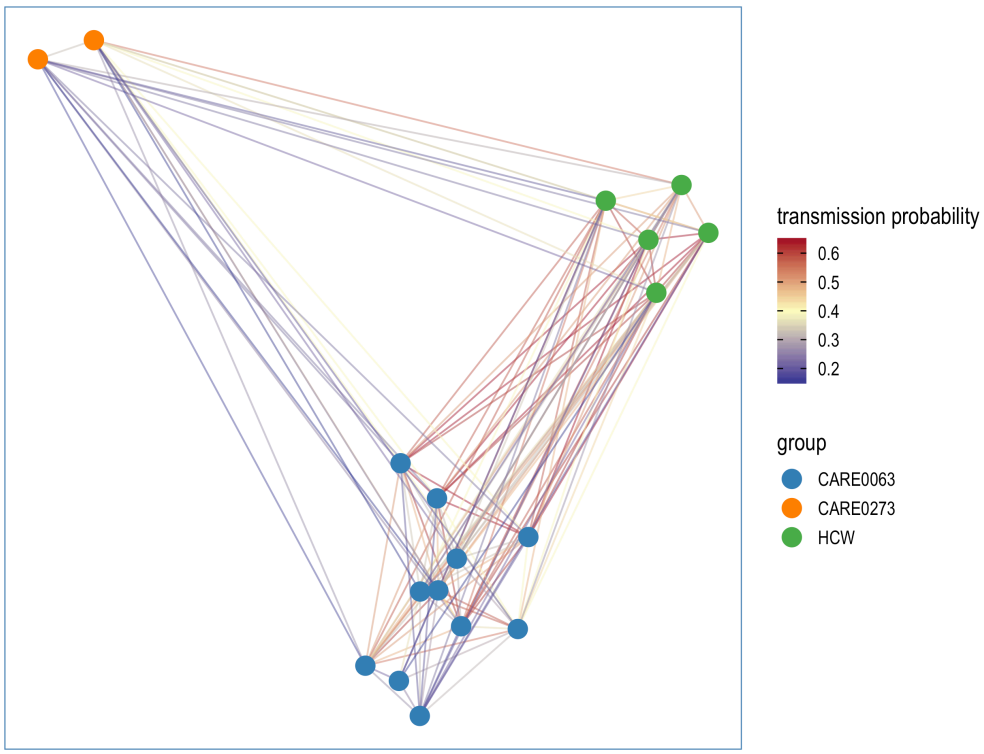
7

Figure 7

A


CARE0032
CARE0061
CARE0063
CARE0097
CARE0151
CARE0173
CARE0263
CARE0264
CARE0277
CARE0314
transmission probability
0.6
0.5
0.4
0.3
0.2

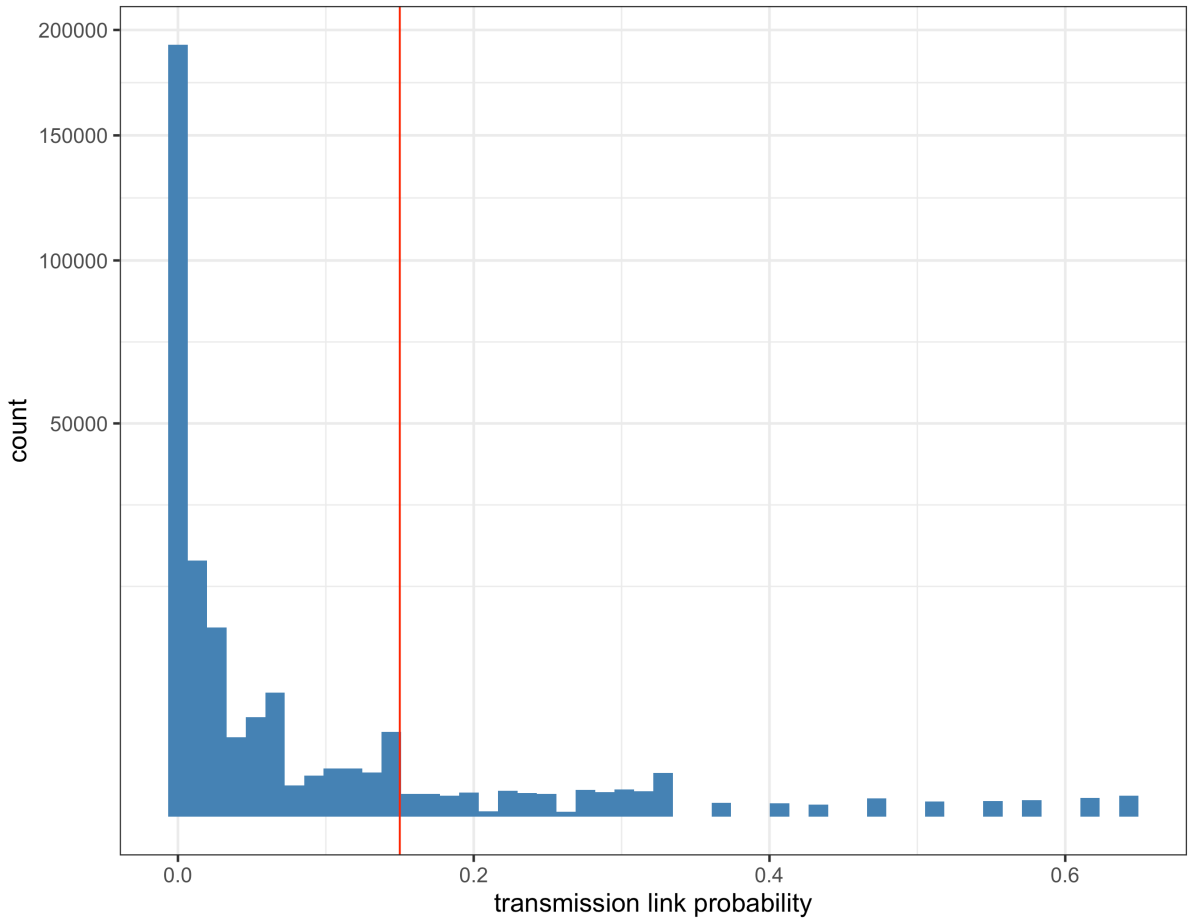
B


transmission probability
0.6
0.5
0.4
0.3
0.2
group
CARE0063
CARE0273
HCW

Visualisations of SARS-CoV-2 clusters among care home residents

Transmission networks were produced using a derivative of the *transcluster* algorithm, which incorporates pairwise date and genetic differences to estimate the probability of cases being connected within a defined number of intermediate hosts; clusters were defined using a probability threshold of ≥15% for cases being linked by ≤2 intermediate hosts (further details in Methods). A. Transmission clusters for the ten care homes with the largest number of care home residents with available genomes. Consistent with Figure 3, several of the ten care homes with the largest number of genomes comprised single transmission clusters (e.g. CARE0264 and CARE0314), while others contained two or more clusters consistent with multiple independent transmission sources among the residents. These data alone do not indicate where the residents acquired their infections, and hospital-acquired infections for some of the clusters is a possibility alongside multiple introductions into the same care homes. B. Visualisation of transmission links between residents of two nearby carehomes and a group of healthcare workers (HCW). Two care homes, CARE0063 (blue) and CARE0273 (orange), each had strong transmission links identified with the *transcluster* algorithm to a group of HCW (green). The HCW comprised paramedics and care home carers – one working at CARE0063 and the other working at an unknown care home. We do not have confirmatory epidemiological data available, but this raises the possibility of the cases sharing a linked transmission network.

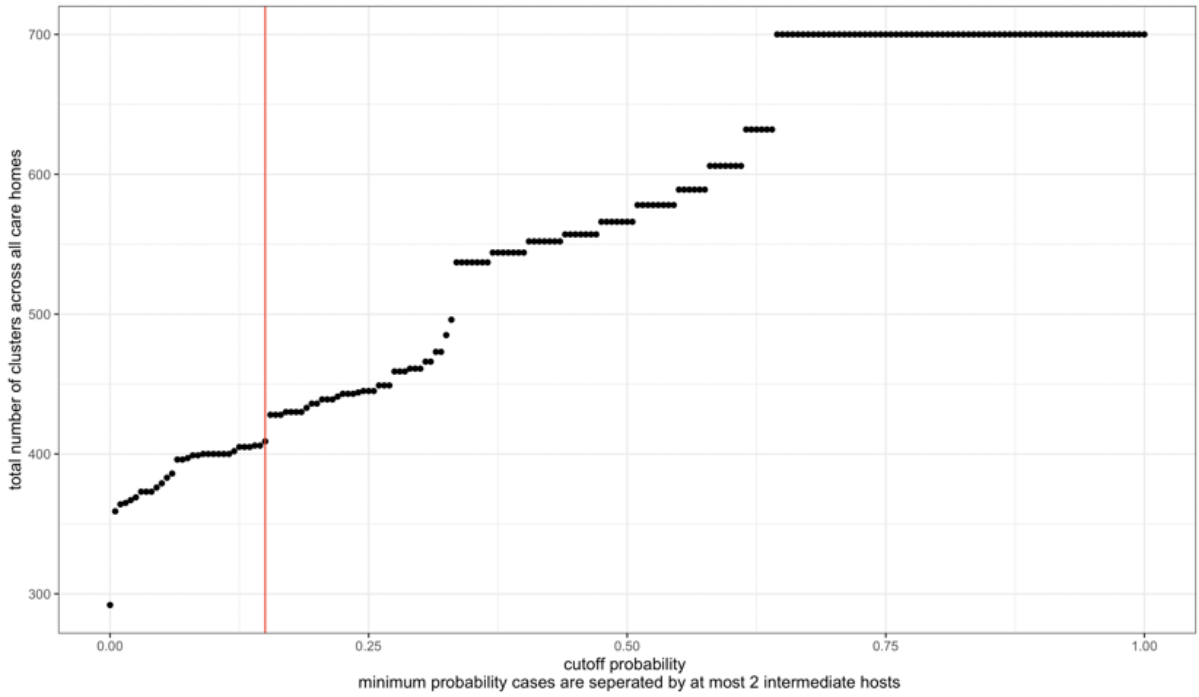
Figure 7, supplement 1



Histogram of pairwise transmission probabilities between care home samples

Histogram of the pairwise probabilities for cases being connected by ≤2 intermediate hosts for all 700 care home residents as inferred by the *transcluster* algorithm, with vertical red line at 0.15 showing the cutoff used to identify care home clusters in our analysis. Note the data gaps along the x-axis reflect the inherent discontinuity of the input datasets, measured in days and SNP differences between cases.

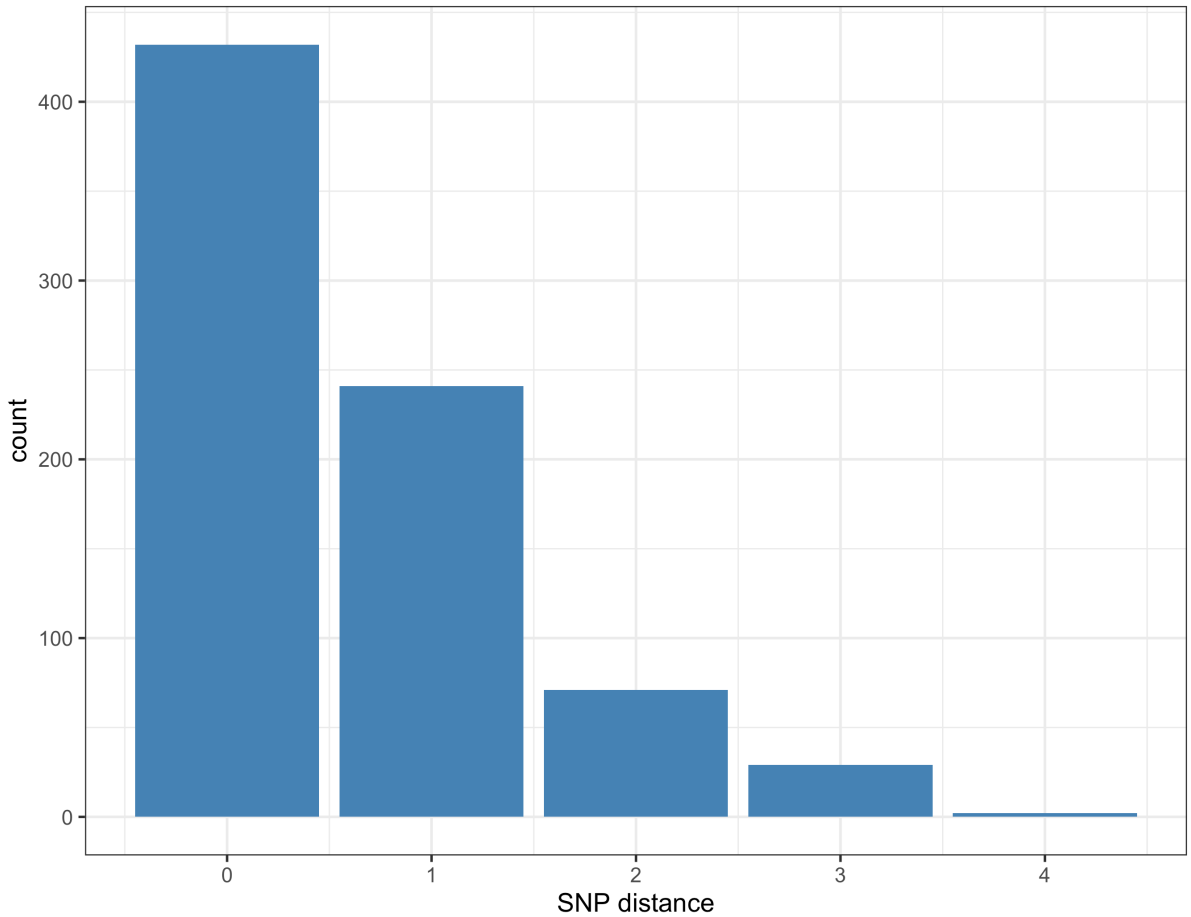
Figure 7, supplement 2



Transmission probability threshold vs number of care home clusters

The *transcluster* algorithm computes the likelihood of two samples being linked within a given number of intermediate hosts, based on the date and genetic differences between samples (assuming a given serial interval and mutation rate, further details in Methods). Changing the probability threshold used to define clusters changes the number of clusters defined, with a higher threshold yielding more clusters (and higher likelihood of transmission within each cluster). Our sample set contained 700 genomes from residents in 292 care homes, and we treated each care home separately as microcosms of potential infection networks. Therefore, the highest theoretical number of clusters is 700, if every genome were its own cluster; and the lowest possible number of clusters is 292, if every person within each care home was part of the same cluster. The cutoff we used ($\geq$15% probability of transmission with $\leq$2 intermediate hosts) is indicated by the red vertical line. This is arbitrary, and was selected 1) because the distribution of pairwise SNP and date differences within resulting clusters appeared reasonable (Figure 7, supplement 3 and 4) and because of a "jump" in the number of clusters occurring at that point.

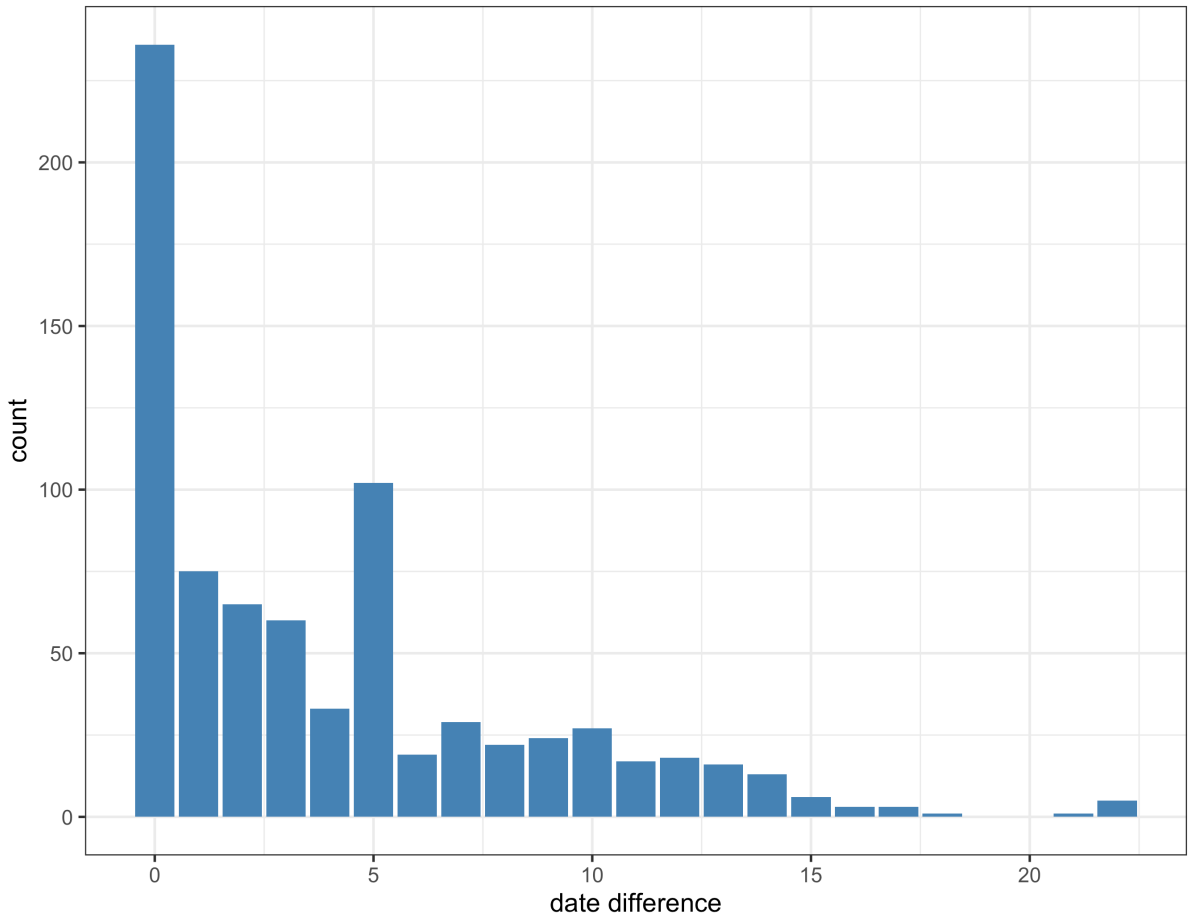
Figure 7, supplement 3



Pairwise SNP difference distribution within clusters

Most genomes within the clusters defined using our model parameters were separated by zero or 1 SNPs, with a small number having higher SNP differences

Figure 7, supplement 4



Pairwise date difference distribution within clusters

Most samples within the clusters defined using our model parameters were separated by <14 days, with a small number separated by longer durations.

1 **Table 1**

2

Variable	Care home residents (all)	Non-care home residents (all)	Care home residents with genomes
Number (%)	1167/6413 (18.2%)	5246/6413 (81.8%)	700/1167 (60%)
Female (%)	624/1167 (53.5%)	2338/5246 (44.6%)	363/700 (51.9%)
Male (%)	543/1167 (46.5%)	2908/5246 (55.4%)	337/700 (48.1%)
Age in years (median, IQR, range)	86 (IQR: 79-90, range: 30-100)	65 (IQR: 48-80, range: 0-100)	86 (IQR: 78-90, range: 42-99)
Diagnostic Ct value	26 (IQR: 22-29)	25 (IQR: 21-29)	24 (IQR: 20-27)
Tested at CUH (%)	72/464 (15.5%)	392/464 (84.5%)	54/72 (75%)
CUH patient admitted to ICU (%)	<5/72 (<7%)	84/392 (21.4%)	<5/54 (<9%)
CUH patient died (%)	34/72 (47.2%)	78/392 (19.9%)	23/54 (42.6%)
Number of care homes	337	-	292
Cases/ care home (median, IQR, range)	2 (IQR: 1-5, range: 1-22)	-	2 (IQR: 1-3, range: 1-18)
Care homes with ≥5 cases	85/337 (25.2%)	-	32/292 (11%)

3

4 **Epidemiological characteristics of care home and non-care home residents with COVID-19 included**
5 **in the study**

6 Not showing precise values where the number of cases is equal to or less than five individuals, to
7 preserve patient anonymity.

8

9

Table 2

Week commencing	Care home resident	Not determined	Not care home resident	Weekly total	Care home resident (%)
24-Feb	0	0	≤5	≤5	0.0%
02-Mar	0	0	31	31	0.0%
09-Mar	10	6	149	165	6.1%
16-Mar	25	6	364	395	6.3%
23-Mar	60	26	852	938	6.4%
30-Mar	126	35	1235	1396	9.0%
06-Apr	162	43	1064	1269	12.8%
13-Apr	154	31	540	725	21.2%
20-Apr	247	16	415	678	36.4%
27-Apr	198	16	393	607	32.6%
04-May	185	8	199	392	47.2%

Case numbers from care homes and non-care home residents per week for full dataset tested at Cambridge CMPHL

Data plotted in Figure 3A of main text, showing case numbers for care homes, non-care homes, and undetermined, for all EoE samples tested at CMPHL. The proportion of COVID-19 cases from care home residents increased in April and May; however, this may reflect the changing profile of samples submitted to the Cambridge CMPHL rather than underlying epidemiological trends.

Table 3

Week	Total weekly COVID-19 cases	Community acquired, care home-associated (%)
09-Mar	12	0 (0%)
16-Mar	24	0 (0%)
23-Mar	75	≤5 (≤7%)
30-Mar	96	≤5 (≤5.2%)
06-Apr	99	14 (14.1%)
13-Apr	49	14 (28.6%)
20-Apr	41	10 (24.4%)
27-Apr	41	9 (22.0%)
04-May	27	6 (22.2%)

Proportion of community acquired, care home-associated COVID-19 infections tested at Cambridge University Hospitals

The proportion of community onset, care home-associated COVID-19 infections tested at Cambridge University Hospitals (CUH) peaked in mid to late April. Total cases shows the total number of new COVID-19 cases diagnosed at CUH that week. "Community acquired" was defined as first positive test ≤48 hours from admission and no healthcare contact within the previous 14 days. Not showing precise values if number of patients is less than or equal to 5 to preserve patient anonymity.

1 **Table 4**

2

Variable	OR	95% CI low	95% CI high	*P* value
Age >= 80	6.6	3.7	12.0	2.46E-10
Sex	1.5	0.9	2.6	1.30E-01
Care resident status	3.0	1.6	5.7	9.22E-04
ICU admission	3.9	2.1	7.5	3.02E-05
Ct value <20	2.9	1.6	5.3	5.04E-04

3

4 **Odds ratios for mortality at 30 days**

5 Logistic regression analysis of odds of mortality at 30 days. Age ≥80 years, being a care home resident,
6 being admitted to ICU and Ct<20 were significantly associated with increased odds of death at 30 days
7 post-diagnosis (*P*<0.05). OR = Odds Ratios. CI = Confidence Interval. ICU = intensive care unit. Ct =
8 Cycle threshold for diagnostic PCR.