COVID-19 infection dynamics in care homes in the East of England: a retrospective genomic epidemiology study

William L. Hamilton*[1,2,+], Gerry Tonkin-Hill*[3], Emily Smith[4], Charlotte J. Houldcroft[5], Ben Warne[1,2], Luke W. Meredith[5], Myra Hosmillo[5], Aminu S. Jahun[5], Martin D. Curran[5], Surendra Parmar[5], Laura G. Caller[5,6], Sarah L. Caddy[5], Fahad A. Khokhar[1], Anna Yakovleva[5], Grant Hall[5], Theresa Feltwell[5], Malte L. Pinckert[5], Iliana Georgana[5], Yasmin Chaudhry[5], Nicholas M. Brown[7,2], Sónia Gonçalves[3], Roberto Amato[3], Ewan M. Harrison[3], Mathew A. Beale[3], Michael Spencer Chapman[3,8], 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[5], M. Estée Török[1,2,+], COVID-19 Genomics Consortium UK[9]

* 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. University of Cambridge, Department of Pathology, Division of Virology, Cambridge, UK
6. The Francis Crick Institute, London, UK
7. Public Health England Clinical Microbiology and Public Health Laboratory, Cambridge UK
8. Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, UK
9. www.cogconsortium.uk

Correspondence

William Hamilton will.l.hamilton@gmail.com

M. Estée Török et317@cam.ac.uk

Word count

Abstract = 246

Research in context = 307

Main text (excluding references) = 3428

References: 30

Figures = 4

Tables = 1

(Max display items = 5)

Key words

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

Abstract

Background

COVID-19 poses a major challenge to infection control in care homes. SARS-CoV-2 is readily transmitted between people in close contact and causes disproportionately severe disease in older people.

Methods

Data and SARS-CoV-2 samples were collected from patients in the East of England (EoE) between 26th February and 10th May 2020. Care home residents were identified using address search terms and Care Quality Commission registration information. Samples were sequenced at the University of Cambridge or the Wellcome Sanger Institute and viral clusters defined based on genomic and time differences between cases.

Findings

7,406 SARS-CoV-2 positive samples from 6,600 patients were identified, of which 1,167 (18.2%) were residents from 337 care homes. 30/71 (42.3%) care home residents tested at Cambridge University Hospitals NHS Foundation Trust (CUH) died. Genomes were available for 700/1,167 (60%) residents from 292 care homes, and 409 distinct viral clusters were defined. We identified several probable transmissions between care home residents and healthcare workers (HCW).

Interpretation

Care home residents had a significant burden of COVID-19 infections and high mortality. Larger viral clusters were consistent with within-care home transmission, while multiple clusters per care home suggested independent acquisitions.

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.

Research in context

Evidence before this study

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. One study in a skilled nursing facility in Washington State, USA, found that 57/89 (64%) residents were SARS-CoV-2 positive and that 27/34 (79%) had genetically similar viral sequences. Another study in four nursing homes in London, UK, found that 126/313 (40%) of residents were SARS-CoV-2 positive; sequencing of 19 samples identified clusters within care homes. Screening and sequencing of staff at five long-term care facilities in Colorado, USA over a two-week period identified a number of clusters among staff working at the same facility, but did not include sequences from residents. A single sequence reported from a care home resident in Hungary was found to differ from other Hungarian SARS-CoV-2 sequences. Finally, an epidemiological study of COVID-19 in 189 care homes in Scotland did not include any genomic data.

Added value of this study

This study includes care home residents tested during the course of the first phase of the COVID-19 pandemic in a large geographical region in the UK. It is more comprehensive and representative than previous studies and includes detailed metadata and genomic data, which are being made openly available as a resource for other researchers. We used a clustering algorithm and demonstrated how genomic and epidemiological data can be integrated to define possible transmission networks.

Implications of all the available evidence

Detailed combined epidemiological and genomic studies are essential to improve our understanding of the transmission and impact of SARS-CoV-2 in long term nursing and residential facilities. Our study has identified two patterns of transmission (outbreaks within care homes and multiple, distinct clusters among care home residents from the same care homes) that will require tailored infection control measures to prevent and mitigate them.

Introduction

Care homes are at high risk of experiencing outbreaks of COVID-19, the disease caused by SARS-CoV-2. COVID-19 is associated with higher mortality in older people and those with comorbidities including cardiovascular and respiratory disease,[1] 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.[2] 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.[3] Despite this observation, 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". [4] 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.[5] 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. [6,7] 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.[8] 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.[9–13] Here, we apply genomic epidemiology to investigate viral transmission dynamics in care home residents across the East of England (EoE). We aimed to address questions of key public health concern: What is the burden of care home associated COVID-19 tested in the region? 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 transmission networks? Are healthcare workers (HCW) involved in these transmission networks?

Methods

Data were collected on SARS-CoV-2 positive samples from the EoE, tested at the Public Health England (PHE) Clinical Microbiology and Public Health Laboratory (CMPHL) in Cambridge, between 26th February and 10th May 2020 (Appendix p 4). The UK government launched a national care home testing portal on 11th May 2020,[14] in which all care home staff and residents were eligible for testing with priority for homes caring for people aged 65 years or older. Prior to this, systematic screening of all residents within care homes was much less common; testing primarily occurred where there was a suspicion of an outbreak, and hence there is reduced risk of bias introduced by systematic screening. During the study period the scope of testing in hospital and community settings, including care homes, changed several times, as eligibility criteria were modified (Appendix p 15).

Patients were initially identified as potential care home residents if search terms including "care home" or "nursing home" were identified in the address fields of their electronic healthcare records. Next, the names of care homes registered in the CQC database (which aims to include all care homes in England) were matched with patient addresses to identify further care home residents (Appendix pp 4-7, 16). The resulting dataset was manually inspected, linked to CQC registered care homes, and matching care home addresses were assigned anonymised care home codes. We refer to care homes recorded by the CQC as having nursing care available as "nursing homes" and care homes without nursing care as "residential homes".[4]

As part of the COVID-19 Genomics Consortium UK (COG-UK), samples from Cambridge University Hospitals NHS Foundation Trust (CUH) and a random selection of EoE samples were sequenced in the Division of Virology, Department of Pathology at the University of Cambridge, using nanopore sequencing (Oxford Nanopore Technologies, Oxford, UK). The remaining samples submitted from EoE laboratories to PHE CMPHL for diagnostic testing were sent to the Wellcome Sanger Institute (WSI) for sequencing using short-read sequencing (Illumina Inc, Great Chesterford, UK; details in Appendix p 4). Available genomes from care home residents and a randomised subset of non-care home residents were passed through quality control filtering (QC). Genomes were aligned using MAFFT, phylogenetic trees produced using IQ-TREE[15,16] and visualised initially in Microreact[17] and then using the R package *ggtree*[18], and SNP difference matrices produced using *snp-dists* (Appendix pp 9-10). Care home "clusters" were defined using an implementation of the *transcluster* package,[19,20] assuming a viral mutation rate of 1e-3 substitutions/site/year,[21] and serial interval of 5 days,[22,23] using a threshold of $\geq$15% probability of $\leq$2 intermediate hosts between linked cases (Appendix pp 10-11). Date of first positive test was used as a proxy for serial interval (if symptomatology data not

available). Tests for statistical significance were performed in R; non-parametric population comparisons were made using the Wilcoxon rank sum test. The lowest *P*-values reported are <0.0001.

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).

Results

7,406 SARS-CoV-2 positive samples from 6,600 patients were identified in the study period, and care home residency status was determined in 6,413 (Figure 1, Appendix p 16). Positive cases came from 37 submitting organisations including regional hospital laboratories and community-based testing services (Appendix pp 17-19). The study population included almost half of the COVID-19 cases diagnosed in the EoE at this time,[24] 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 (Appendix p 20). 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.[25] 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 (Appendix p 21). There was a slight trend for nursing homes to have more cases per home than residential homes (median 3, IQR 2-4 versus median 2, IQR 1-3, respectively (*P*=0.03, Wilcoxon rank sum test)) (Appendix p 22). The number of cases per care home per week increased over the study period (Appendix p 23), likely reflecting increased testing. While non-care home cases declined during April 2020, care home numbers were initially maintained and then declined more slowly; the proportion of cases coming from care homes relative to non-care homes increased, from <10% in March to >40% in the first week of May (Figure 2, Appendix pp 24-25). While this may partially reflect the changing profile of samples submitted to the CMPHL, a similar trend was observed for cases tested acutely at CUH, 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 (Appendix p 26).

71 / 464 (15.3%) COVID-19 patients diagnosed by the acute medical services at CUH were identified as care home residents (Table 1, Figure 2B), of which <7% were admitted to the intensive care unit (ICU) and 30/71 (42.3%) died (not showing precise values where the number of individuals is equal to or below five, to protect patient anonymity). In comparison, amongst non-care home residents 84 / 393 (21.4%) were admitted to the ICU and 68 / 393 (17.3%) died. The difference in unadjusted mortality rate was significant (*P*=0.0018, Pearson's Chi-squared test). The mortality rates for CUH patients who were residents at nursing homes and residential homes were similar: 20 / 49 (41%) versus 10 / 22 (45%), respectively.

Genome sequence data were available for 700 / 1,167 (60.0%) care home residents from 292 care homes in the EoE (Appendix p 27). 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[8] (Appendix p 28). The proportion of viral lineage B.1.1 increased over the study period in both care home residents and non-care home residents (Appendix pp 29-30), consistent with European trends.[26] 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. 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 outbreak is defined as two or more cases linked in time or place[27]) (Figure 3). 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 adjusted for SARS-CoV-2 (Appendix pp 10-11, 31-32).[20] 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, 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) (Appendix pp 33-34). Networks for the ten care homes with the largest number of genomes are shown in Figure 4, indicating linked transmission clusters based on our model assumptions and probability threshold. Consistent with the phylogeny shown in Figure 3, some care homes contained a single transmission cluster involving multiple cases (e.g. CARE0314), while others comprised multiple independent clusters (e.g. CARE0061).

We investigated transmission networks involving care home residents and healthcare workers (HCW) for people tested at CUH (HCW data were not available outside of CUH) (Appendix pp 10-11). 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,[8] 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 determine precise 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 (Appendix p 35); these links are in the top 1.1% of all pairwise transmission probabilities inferred using

the *transcluster* algorithm. However, we accept that without confirmatory epidemiological data this interpretation remains speculative.

Discussion

We have investigated the genomic epidemiology of SARS-CoV-2 in care home residents from the East of England, whose samples were tested at the PHE Clinical Microbiology and Public Health Laboratory in Cambridge. Around half of confirmed COVID-19 cases from the EoE were included with study dates encompassing the majority of cases from the first phase of the pandemic in the EoE. The remainder of cases were tested at other laboratories. 18% of COVID-19 positive cases were care home residents, with a median age of 86 years (21 years older than non-care home residents). 42% of care home residents tested acutely in CUH died, compared with 17% for non-care home residents (without adjusting for age or other confounders). The absolute number of diagnosed COVID-19 cases from care home residents declined more slowly in April than for non-care home residents, increasing the proportion of cases from care homes and contributing to the slow rate of decline in total case numbers during April and early May 2020. This suggests that care home transmission was more recalcitrant to lockdown measures (implemented on 23rd March 2020 in the UK) than in non-care home settings. This may reflect the underlying vulnerability of the care home population, and the infection control challenges of nursing multiple residents who may also share communal living spaces. The largest viral clusters we identified among care home residents comprised >10 cases. In contrast, the UK as a whole had an average of 2.37 people per household in 2019 [28] and in the East region only 2.2% of households were made up of two or more unrelated adults (6.2% in London) [29]. No new viral lineages from outside the UK were seen in our dataset; this may reflect the success of travel restrictions in limiting new viral introductions into the general population. It is important to note that laboratory confirmed case numbers may not reflect the "true" case burden owing to the heterogeneous testing strategy over the study period. Thus whilst the number of laboratory-confirmed cases in care homes was low in our dataset until mid to late March, the actual number of care home cases was likely much higher.

Our transmission modelling suggested that some care homes had experienced "outbreaks" with all or a large proportion of cases linked in a single transmission network. We also observed care homes containing multiple distinct transmission clusters. Some of these may represent hospital-acquired infections for care home residents that were admitted to hospital at or shortly before the time of their positive test result, rather than independent transmission events within the care homes. However, we note that only 7/71 (10%) care home residents tested at CUH had suspected or definite hospital-

acquired COVID-19 infections, so some of the identified instances of multiple transmission clusters from a single care home may represent independent introductions of the virus into the care home. These findings suggest that preventing the introduction of new infections into care homes should be a key priority to limit outbreaks, alongside infection control efforts to reduce transmission within care homes, including once an outbreak has been identified. We also found transmission networks involving care home residents and HCW such as paramedics, care home workers, and hospital-based healthcare staff, suggesting a potential link between care home infections and healthcare-associated COVID-19 cases [8].

We acknowledge several limitations to our study. First, 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. However, we linked every care home included in the analysis to CQC registered care homes, so the care homes included should be accurate. 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. Second, we did not have viral sequence data available for 40% of care home residents; this was due to a combination of missing samples, mismatches between sequences and metadata, genomes not passing QC filtering using a stringent threshold (<10% missing calls), or sequences being unavailable at the time of data extraction. Third, early in the pandemic diagnostic testing was limited to patients admitted to hospital with suspected COVID-19. Thus not every care home resident with COVID-19 will have been tested and our genomic cluster sizes per care home will almost certainly be an underestimate. The introduction of systematic care home screening creates a potential bias in population analyses that assume random sampling. For example, care home screening could inflate the number and size of clusters of related virus relative to non-care home residents that were not screened. This could affect population genetic metrics such as viral genetic diversity. However, we believe 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. Fourth, the cut-off we have used for classifying transmission networks is an arbitrary discontinuity in a continuous probability distribution, and different thresholds would yield different clusters per care home. However, such cut-offs can be helpful for producing understandable outputs from complex genomic epidemiological data for biological and public health interpretation,[20,30] and for focusing investigations with limited public health resources. 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, such as paramedic calls and carers working across multiple care homes. Without this epidemiological information, the low viral genetic diversity makes it difficult to infer transmission. Even within care homes, it is possible some genetically similar viruses are from distinct introduction events, though incorporating genomic data will be more accurate for excluding linked transmission than if only temporal data were available. Finally, we did not have data available on who was a HCW or had hospital-acquired infection for patients tested outside of CUH. This means that some of the care home clusters we identified could represent transmission events occurring in hospitals and other healthcare settings, rather than in the care homes themselves. However, large clusters involving multiple residents from the same care home, who often have reduced mobility, and in the context of a lockdown, are suggestive of transmission taking place between residents within the care home.

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. Further research is needed to refine our understanding of care home transmission mechanisms, both within- and between-care homes, and to elucidate the interplay between infections in care homes and healthcare-associated infections. Future work can build on population-level analyses using integrated genomic and epidemiological data to elucidate transmission networks probabilistically for targeting public health interventions most effectively.

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.

References

1 Williamson EJ, Walker AJ, Bhaskaran K, *et al.* OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature* 2020. DOI:10.1038/s41586-020-2521-4.

2 Deaths involving COVID-19, England and Wales - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsinvolvingcovid19englandandwales/deathsoccurringinjune2020 (accessed July 30, 2020).

3 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 (accessed July 30, 2020).

4 Service types | Care Quality Commission. https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types#care-homes-without-nursing (accessed July 30, 2020).

5 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 (accessed July 30, 2020).

6 Lansbury LE, Brown CS, Nguyen-Van-Tam JS. Influenza in long-term care facilities. *Influenza Other Respi Viruses* 2017; **11**: 356–66.

7 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/ (accessed July 30, 2020).

8 Meredith LW, Hamilton WL, Warne B, *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 2020; **0**. DOI:10.1016/S1473-3099(20)30562-4.

9 Arons MM, Hatfield KM, Reddy SC, *et al.* Presymptomatic SARS-CoV-2 infections and transmission in a skilled nursing facility. *N Engl J Med* 2020; **382**: 2081–90.

10 Graham NSN, Junghans C, Downes R, *et al.* SARS-CoV-2 infection, clinical features and outcome of COVID-19 in United Kingdom nursing homes. *J Infect* 2020. DOI:10.1016/j.jinf.2020.05.073.

11 Quicke K, Gallichote E, Sexton N, *et al.* Longitudinal Surveillance for SARS-CoV-2 RNA Among Asymptomatic Staff in Five Colorado Skilled Nursing Facilities: Epidemiologic, Virologic and Sequence Analysis. *medRxiv* 2020; : 2020.06.08.20125989.

12 Kemenesi G, Kornya L, Tóth GE, *et al.* Nursing homes and the elderly regarding the COVID-19 pandemic: situation report from Hungary. *GeroScience* 2020; **2**. DOI:10.1007/s11357-020-00195-z.

13 Burton JK, Bayne G, Evans C, *et al.* Evolution and impact of COVID-19 outbreaks in care homes: population analysis in 189 care homes in one geographic region. *medRxiv* 2020; :

2020.07.09.20149583.

14 Get coronavirus tests for a care home - GOV.UK. https://www.gov.uk/apply-coronavirus-test-care-home (accessed July 30, 2020).

15 Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–9.

16 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. *Mol Biol Evol* 2018; **35**: 518–22.

17 Argimón S, Abudahab K, Goater RJE, *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb genomics* 2016; **2**: e000093.

18 Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods Ecol Evol* 2017; **8**: 28–36.

19 gtonkinhill/fasttranscluster. https://github.com/gtonkinhill/fasttranscluster (accessed July 30, 2020).

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

21 Fauver JR, Petrone ME, Hodcroft EB, *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* 2020; **181**: 990–6.

22 Zhang J, Litvinova M, Wang W, *et al.* Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. *Lancet Infect Dis* 2020; **20**: 793–802.

23 He X, Lau EHY, Wu P, *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* 2020; **26**: 672–5.

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

25 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 (accessed July 30, 2020).

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

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

28 Families and households in the UK - Office for National Statistics.

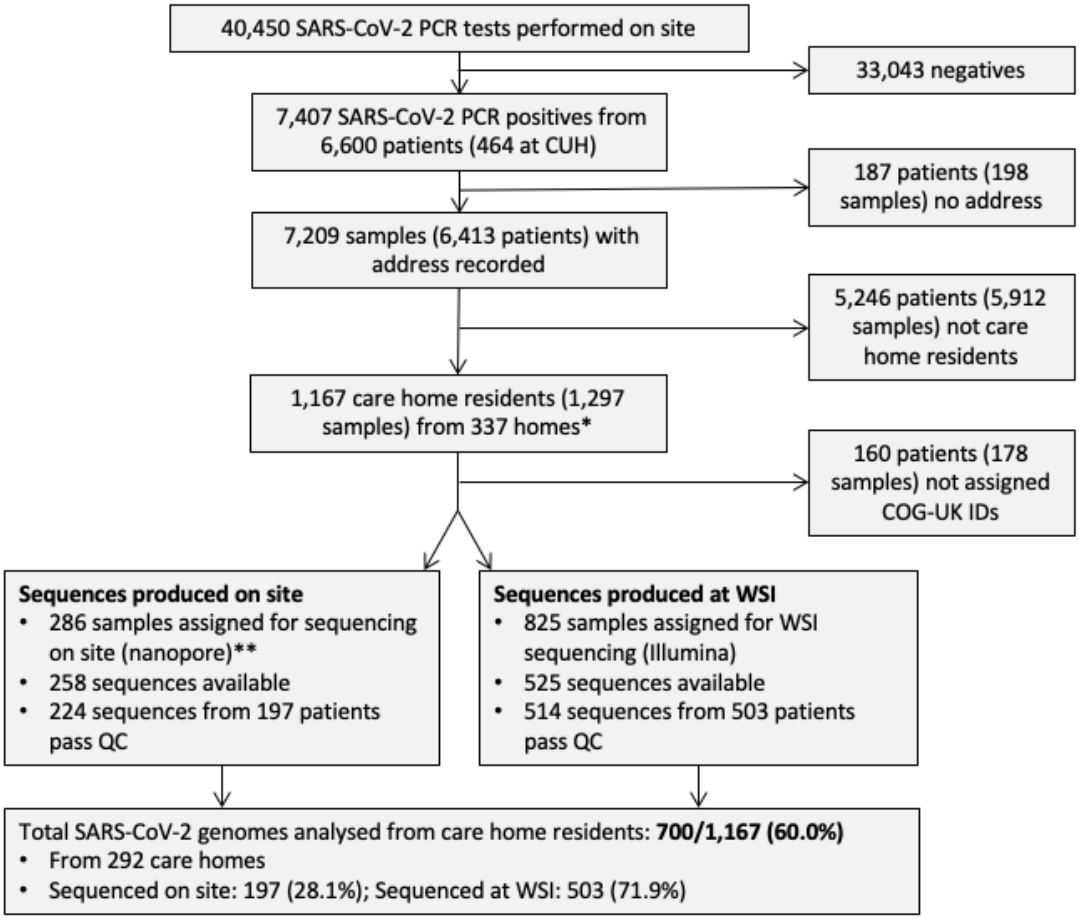
https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/bulletins/familiesandhouseholds/2019 (accessed July 30, 2020).

29 Families and households - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/families/datasets/familiesandhouseholdsfamiliesandhouseholds (accessed July 30, 2020).

30 MacFadden DR, McGeer A, Athey T, *et al.* Use of genome sequencing to define institutional influenza outbreaks, Toronto, Ontario, Canada, 2014-15. *Emerg Infect Dis* 2018; **24**: 492–7.

1 Main display items

2 **Figure 1. Study flow diagram**



3

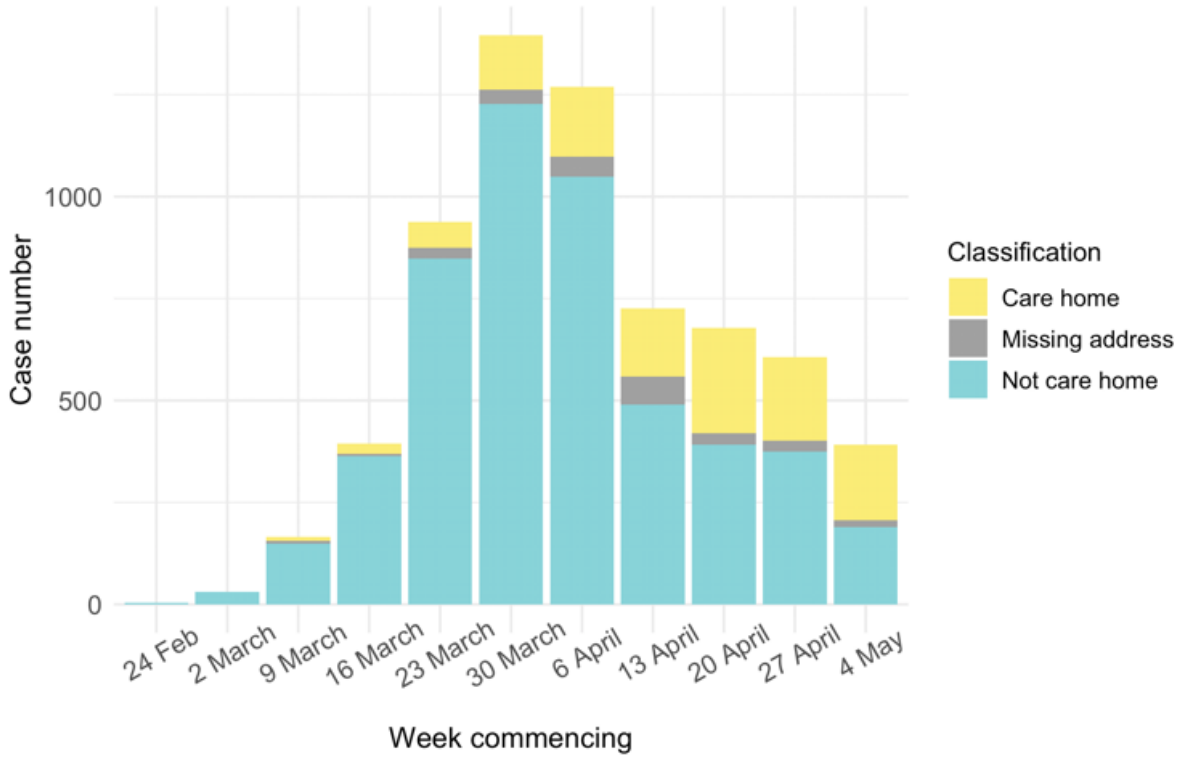
4 765 care home residents were identified by having care-related search terms in their addresses; a
5 further 382 had CQC registered care home names in their addresses; 89 were identified manually and
6 69 were removed after linking the addresses to CQC registered care home names and manual review
7 (Appendix p 15). * 193 care homes were registered with the CQC as being residential homes without
8 nursing care and 144 had nursing care available. ** Samples were selected for sequencing on site if
9 they were CUH inpatients or CUH HCW, where we prioritised rapid turnaround time to investigate
10 hospital-acquired infections, plus a randomised selection of other EoE samples to provide genomic
11 context. The remainder, where available, were sent to WSI for sequencing.

12

13

1 **Figure 2. Epidemic curves for EoE and CUH showing care home residents**

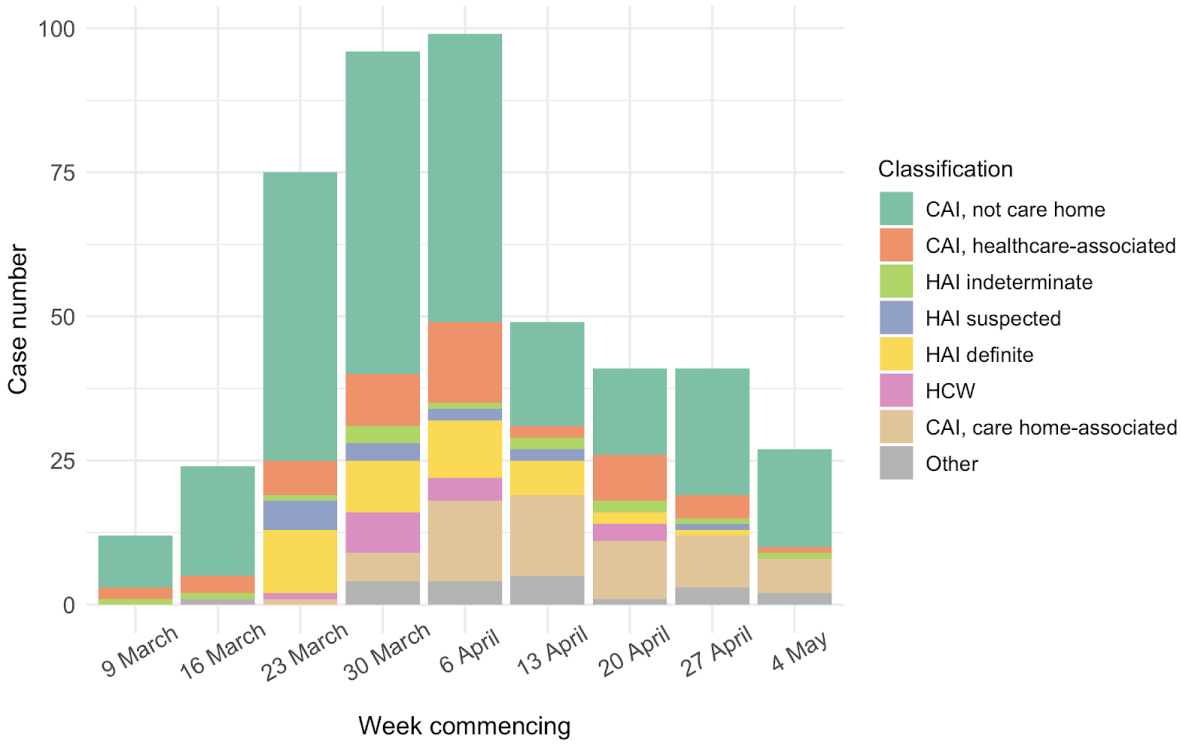
2 **A**



3

4 **B**

5

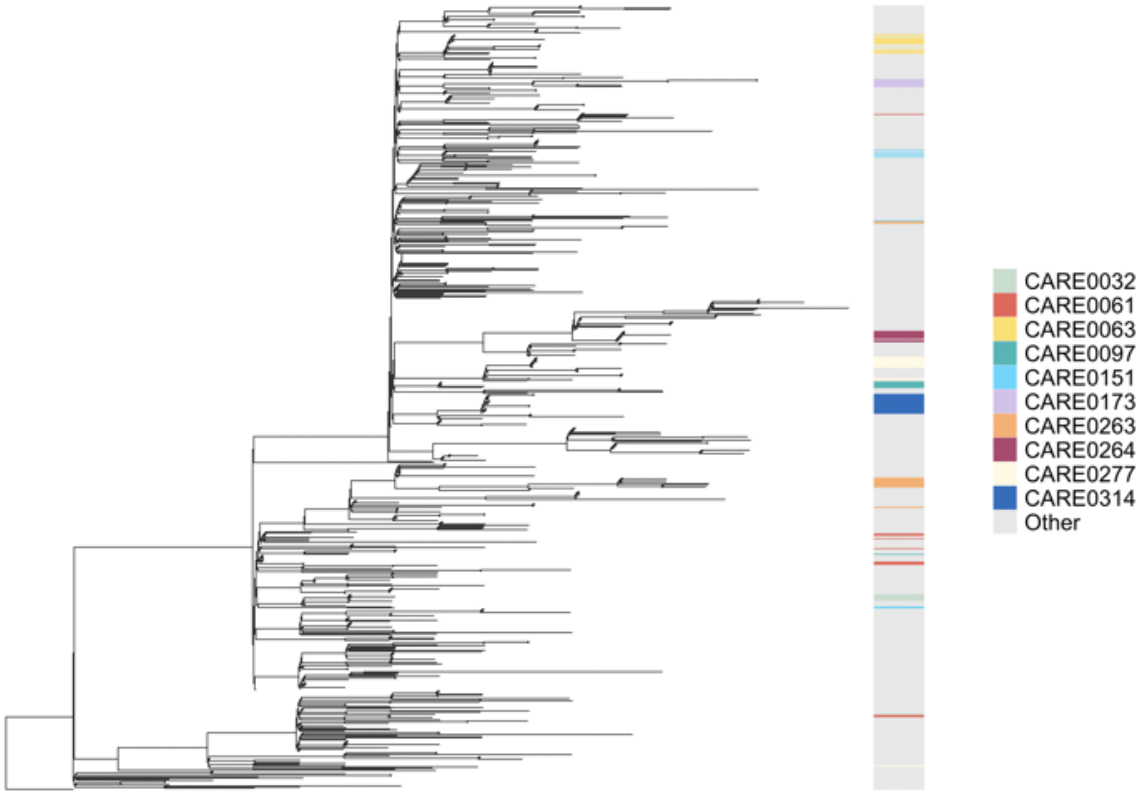


6

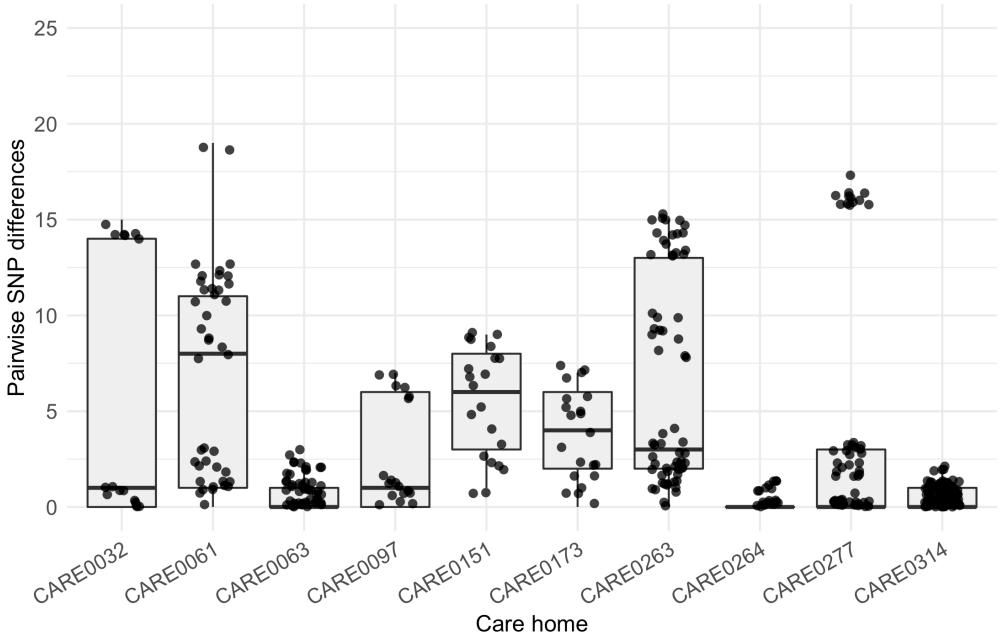
Number of positive cases per week over the study period for different infection sources, for all samples tested from EoE at the Cambridge PHE laboratory (A), or those tested at CUH acute medical services (B). Peak of the epidemic for samples tested at the Cambridge PHE laboratory and CUH acute medical 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 (numbers provided in Appendix), while the contribution of general community cases fell more quickly. 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 = 2-6 days; suspected 7-14 days; definite >14 days (as used in [8], with further detail in Appendix p 26).

Figure 3. Care home clustering on viral phylogenetic tree and within-care home genetic diversity

A

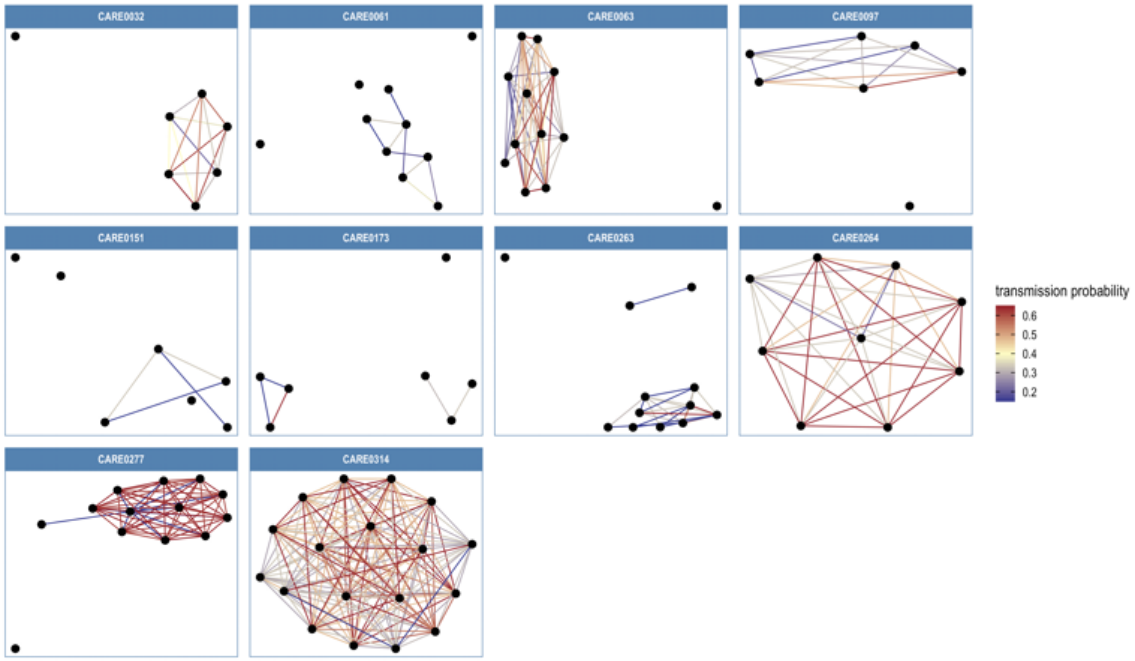


B



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). 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).

Figure 4. Transmission networks for the ten care homes with the largest number of SARS-CoV-2 genomes



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 main text and Appendix). 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.

1 **Table 1. Epidemiological characteristics of care home and non-care home residents**
2 **with COVID-19**

3

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%)	2908 / 5246 (55.4%)	363 / 700 (51.9%)
Male (%)	543 / 1167 (46.5%)	2338 / 5246 (44.6%)	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)
Tested at CUH (%)	71 / 464 (15.3%)	393 / 464 (84.7%)	54 / 71 (76.1%)
CUH patient admitted to ICU (%)	≤5 / 71 (≤7.0%)	84 / 393 (21.4%)	≤5 / 54 (≤9.3%)
CUH patient died (%)	30 / 71 (42.3%)	68 / 393 (17.3%)	24 / 54 (44.4%)
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 (11.0%)

4

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