PLOS PATHOGENS

RESEARCH ARTICLE

Viral burden is associated with age, vaccination, and viral variant in a population-representative study of SARS-CoV-2 that accounts for time-since-infection-related sampling bias

Helen R. Fryer[1]*, Tanya Golubchik[1,2], Matthew Hall[1], Christophe Fraser[1,3], Robert Hinch[1], Luca Ferretti[1,3], Laura Thomson[1], Anel Nurtay[1], Lorenzo Pellis[4,5], Thomas House[4], George MacIntyre-Cockett[6], Amy Trebes[6], David Buck[6], Paolo Piazza[6], Angie Green[6], Lorne J Lonie[6], Darren Smith[7], Matthew Bashton[7], Matthew Crown[7], Andrew Nelson[8], Clare M. McCann[8], Mohammed Adnan Tariq[8], Claire J. Elstob[8], Rui Nunes Dos Santos[8], Zack Richards[8], Xin Xhang[8], Joseph Hawley[8], Mark R. Lee[8], Priscilla Carrillo-Barragan[8], Isobel Chapman[8], Sarah Harthern-Flint[8], The COVID-19 Genomics UK (COG-UK) consortium¶, David Bonsall[3,6], Katrina A. Lythgoe[1,3,9]

1 Big Data Institute, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford, United Kingdom, **2** Sydney Infectious Diseases Institute (Sydney ID), School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Sydney, Australia, **3** Pandemic Sciences Institute, University of Oxford, Old Road Campus, Oxford, United Kingdom, **4** Department of Mathematics, University of Manchester, Manchester, United Kingdom, **5** The Alan Turing Institute, London, United Kingdom, **6** Wellcome Centre for Human Genetics, Oxford, United Kingdom, **7** The Hub for Biotechnology in the Built Environment, Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom, **8** Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Newcastle upon Tyne, United Kingdom, **9** Department of Biology, University of Oxford, Oxford, United Kingdom

¶ Membership of The COVID-19 Genomics UK (COG-UK) consortium is provided in Supporting Information file S1 Text.
* helen.fryer@bdi.ox.ac.uk

Abstract

In this study, we evaluated the impact of viral variant, in addition to other variables, on within-host viral burden, by analysing cycle threshold (Ct) values derived from nose and throat swabs, collected as part of the UK COVID-19 Infection Survey. Because viral burden distributions determined from community survey data can be biased due to the impact of variant epidemiology on the time-since-infection of samples, we developed a method to explicitly adjust observed Ct value distributions to account for the expected bias. By analysing the adjusted Ct values using partial least squares regression, we found that among unvaccinated individuals with no known prior exposure, viral burden was 44% lower among Alpha variant infections, compared to those with the predecessor strain, B.1.177. Vaccination reduced viral burden by 67%, and among vaccinated individuals, viral burden was 286% higher among Delta variant, compared to Alpha variant, infections. In addition, viral burden increased by 17% for every 10-year age increment of the infected individual. In summary, within-host viral burden increases with age, is reduced by vaccination, and is influenced by the interplay of vaccination status and viral variant.

OPEN ACCESS

Citation: Fryer HR, Golubchik T, Hall M, Fraser C, Hinch R, Ferretti L, et al. (2023) Viral burden is associated with age, vaccination, and viral variant in a population-representative study of SARS-CoV-2 that accounts for time-since-infection-related sampling bias. PLoS Pathog 19(8): e1011461. https://doi.org/10.1371/journal.ppat.1011461

Editor: Adi Stern, Tel Aviv University, ISRAEL

Received: December 16, 2022

Accepted: June 5, 2023

Published: August 14, 2023

Data Availability Statement: Genomes included in this manuscript can be accessed through the European Nucleotide Archive (ENA) at https://www.ebi.ac.uk/ena/browser/text-search/. A list of accession numbers are provided in S1 Data. All analyses were implemented in Matlab and the code is available at https://github.com/helenfryer1000000/epidemiologically-adjusted-viral-load. Quantile median regression was implemented using the function qr_standard, provided at: https://github.com/zjph602xtc/Quantile_reg.

Funding: KAL and HF were supported by The Wellcome Trust and The Royal Society (107652/Z/15/Z to KAL) and by the Li Ka Shing Foundation funding awarded to KAL David Bonsall was supported by the Bill and Melinda Gates Foundation (INV-003680) and by COG-UK (which is funded by the Medical Research Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health Research (NIHR) (grant code: MC_PC_19027), and Genome Research Limited, operating as the Wellcome Sanger Institute.TG is funded by a strategic grant from the Sydney Institute for Infectious Diseases (Sydney ID), University of Sydney. RH, MH, LF, LT, AN and CF were funded by a Li Ka Shing Foundation grant to CF. LP was funded by the Wellcome Trust and Royal Society (grant number 202562/Z/16/Z), the UKRI through the JUNIPER modelling consortium (grant number MR/V038613/1 to LP) and the Alan Turing Institute under the EPSRC grant (EP/N510129/1 to LP). TH was funded by the Royal Society (grant code: INF\R2\180067). AT, David Buck, PP, LL, AG and GM-C were funded by The Wellcome Trust (203141/Z/16/Z awarded to David Bonsall). DS, MB, AN, CM were funded by UKHSA. DS was funded by COG-UK. DS, MB, GY were funded by Research England E3: HBBE. DS, AN, CM were supported by Northumbria University. CE, XX, JH, ML, PC-B, IC, SH-F and MAT were contract researchers at Northumbria University, funded by UKHSA as a DNA sequencing resilience site. ZR, MC and RNDS were supported by HBBE/COG-UK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

The SARS-CoV-2 epidemic in the United Kingdom (UK) has been characterized by the successive transmission of distinct viral variants. Viral variation can impact viral properties, including infectiousness and pathogenicity. These properties may be linked to the amount of virus present in an infected individual. In this study, we examined the association between intra-patient viral burden and a range of factors, including viral variant, using a large, population-representative SARS-CoV-2 survey conducted in the UK. As part of our investigation, we developed a novel method to account for bias in sampled viral burden resulting from the study's sampling methodology. Our findings indicate that viral burden within the host increases with age, is reduced by vaccination, and is influenced by the interplay between vaccination status and viral variant.

Introduction

The SARS-CoV-2 epidemic in the United Kingdom (UK) has been characterised by the appearance of a series of distinct viral variants that, in order of emergence, include the B.1.177 lineage, and the Alpha (B.1.1.7 lineage), Delta (B.1.617.2 lineage) and Omicron (BA.1, BA.2, BA.4 and BA.5 lineages) variants. Explaining their successive abilities to spread, the Alpha, Delta and Omicron variants have been estimated to have a transmission advantage of 43–100% [1–3], 60–70% [4] and 52% [5] compared to their preceding variant. The underlying causes of these differences are unclear, but could include differences in within-host viral burden [6], infectious period, or the per-virion probability of between-host transmission. In turn, these could be influenced by many factors [7], including changes in virus attachment to human cells and the continuous interplay of population acquisition of immunity and the emergence of immune escape variants [8, 9]. In this study, we investigate the association between within-host viral burden and viral variant by analysing nose and throat swabs collected as part of the UK's nationally representative SARS-CoV-2 surveillance study [10, 11].

A number of studies have compared viral burden amongst infections with the Alpha variant and predecessor variants (S1 Table)[12–18] with mixed findings. For example, two detailed longitudinal surveys of a small number of infected individuals have suggested that viral burden is similar among infections with the Alpha variant and predecessor strains [16, 17]. However, a much larger, but less intensive study of viral burden at symptom onset has identified that viral burden is higher among infections with the Alpha-variant, compared to those with a predecessor lineage [15]. The impact of later variants on viral burden has also been studied [11, 15, 16, 19], indicating that viral burden is higher among Delta-variant, compared to the Alpha-variant infections, in vaccinated individuals [11] in one survey, but that there is no difference between infections with these variants in another [16]. The study design and cohorts used to investigate viral burden have varied and this may explain the different findings. In addition to the differences in sample sizes and sampling frequency, the study populations have varied. Some have been based upon testing symptomatic individuals or their close contacts [12, 14, 15] and have thereby excluded some asymptomatically infected individuals, who make up an estimated 40% [20] of infections. Others have focussed on a specific group of people, with examples being hospitalized individuals [12] and persons associated with a professional sporting league [16]. Methods to identify variants have also varied, with some surveys using Spike gene target failure (SGTF) [12, 14, 15] during PCR testing or sample date [11] to classify the viral variants, whereas other have used whole genome sequencing [13, 16, 17].

The Office for National Statistics (ONS) COVID-19 Infection Survey (CIS) is a large household-based surveillance study based in the United Kingdom [10, 11]. We analysed data from the CIS to investigate the impact of viral variant on viral burden. The survey randomly selects private households on a continuous basis from address lists and previous surveys to provide a representative UK sample. Individuals were asked to provide information that included demographics, symptoms, and vaccination details. As part of the survey, nose and throat swabs were collected and tested for SARS-CoV-2 using RT-PCR, and, if positive, individuals with a cycle threshold (Ct) less than 30 were sequenced using whole genome sequencing. Since the Ct value of a sample is inversely correlated with $\log_{10}$(viral burden) of that sample [21], this study design enables viral burden to be investigated. Although the accuracy with which the viral burden sampled from a nose and throat swab informs the viral burden throughout the body is unclear [22], this study does allow for investigation into viral burden in a manner that avoids biases associated with samples from symptomatic individuals or small studies of particular demographic groups.

The survey is simultaneously a cross-sectional survey of the population through time and a longitudinal survey of individuals, with individuals sampled approximately weekly during the first month following enrolment and then monthly thereafter, regardless of symptoms. This weekly or monthly sampling leads to uncertainty in the time-since-infection of positive samples. In addition, the different epidemiological trajectories of the variants mean that the distribution of time-since-infection for each variant at any given time can be skewed depending on when the samples were collected. For example, if a variant is increasing in prevalence, a cross sectional sample will contain more individuals with that variant who are earlier on in their infection, compared to those who are later on in their infection [23]. Because within-host viral burden trajectories are asymmetric, with the peak in viral load closer to the start of infection than to the end [16], this can affect the sampled distribution of viral burden and complicate comparisons between viral variants. The impact of SARS-CoV-2 epidemiology on sampled Ct values is sufficiently strong for its shifts to be inferred from changes in Ct values measured over time [23, 24].

We are unaware of any published studies comparing viral burden associated with viral variants from a large population-representative surveillance survey that directly estimates the impact of variant-specific epidemiological trajectories. Here, we address this gap by developing a methodology that directly estimates the combined impact of variant-specific within-host viral burden and epidemiological trajectories on randomly sampled viral burden. We apply this methodology to data from the CIS to investigate the impact of a range of factors, including variant, vaccination status, and age, on viral burden, as measured by Ct values. As many countries move towards implementing SARS-CoV-2 surveillance surveys, the concepts and methodologies described here will be valuable for informing public health decisions. Moreover, the concepts and methods will be applicable to other pathogens for which sampled viral burden is affected by epidemiological dynamics.

Results

Covid-19-Infection survey

We used data from the Office for National Statistics Covid infection survey (ISRCTN21086382CT, https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey). The survey has been described in detail elsewhere [24] and the survey protocol is provided online (https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/protocol-and-information-sheets). However, in brief, private households were randomly selected on a continuing basis in order to provide a representative sample of inhabitants of the UK. Following agreement to participate, self-collected nose and throat swabs were taken by participants–or their parents/guardians if they were aged 12 or under–as instructed by a study worker. The intended schedule of swabbing was weekly for the first month of participation and monthly thereafter, for up to a year. However, there was variability among participants due to missed or late swabs, and participants could also choose to participate only once, or only for the first month, rather than on an ongoing basis, and were also free to leave the study at any time. Overall however, attrition rates for the survey were low (typically less than 1% of participants per month), as detailed in a report by the ONS [25], providing confidence in the accuracy of the survey's findings.

Individuals were asked about demographics, symptoms, contacts, and relevant behaviours (https://www.ndm.ox.ac.uk/covid-19/covid-19-infection-survey/case-record-forms). An analysis of how representative the Covid-19 Infection Survey is of the UK population has previously been described in detail [26]. In a random 10–20% of households, participants who were 16 years or older were invited to provide monthly venous blood samples for assays of anti-trimeric spike protein IgG.

RT-PCR Covid-19 testing

In this study we focused on swabs and blood samples that were part of the CIS and were sent (at ambient temperatures) to the UK's national Lighthouse Laboratory at Glasgow. RT-PCR for three SARS-CoV-2 genes (N protein, S protein, and ORF1ab) used the Thermo Fisher TaqPath RT-PCR COVID-19 Kit, analysed using UgenTec Fast Finder 3.300.5 (TaqMan 2019-nCoV Assay Kit V2 UK NHS ABI 7500 v2.1). The Assay Plugin was used to convert the qualitative amplification Assay PCR raw data from the ABI 7500 Fast into test results with minimal manual intervention, using an assay-specific algorithm and decision mechanism. Samples were considered positive if at least one N gene and/or ORF1ab gene was present. The lighthouse laboratory was assessed by the external quality assessment laboratories, NEQAS. We analysed RT-qPCR SARS-CoV-2 positive samples that were sequenced at Oxford University (sampled between 27/09/20 and 17/06/21) or Northumbria University (sampled between 20/09/21 and 19/01/22) and had a Ct$\leq$30. These samples cover the period of the epidemic that includes sections of the B.1.177, Alpha, Delta, and BA.1 Omicron waves.

Sequencing and lineage identification

All swabs were tested for SARS-CoV-2 using RT-QPCR, and the cycle threshold (Ct) values of positive samples were recorded. For samples collected before December 2020, real-time sequencing of samples was undertaken where possible, with some additional retrospective sequencing. From December 2020 onwards, sequencing was attempted on all positive samples with Ct$\leq$30. Sequenced samples collected between 27th Sep 2020 and 17th July 2021 were sequenced at the University of Oxford using veSEQ. This employs an RNASeq protocol based on a quantitative targeted enrichment strategy [27] and sequencing on the Illumina Novaseq platform. For a full description of the sequencing protocol, see [27, 28]. Samples collected between 20th Sep 2021 and 19th Jan 2022 were sequenced at the University of Northumbria using the CoronaHiT [29] variant of the ARTIC protocol and Illumina Novaseq 550.

All samples sequenced in Oxford with Ct$\leq$30 were retained for analysis, with the added restriction of $\geq$50% genome coverage required for samples sequenced in Northumbria. For the periods that these two datasets span, the number of positive samples tested in Glasgow was 51101. Amongst these samples 32852 (64%) had a Ct$\leq$30. Consensus sequences were produced using the *shiver* pipeline [30] and lineages were assigned using PangoLEARN [31], with samples assigned as B.1.177 (and sublineages), Alpha (B.1.1.7 and sublineages), Delta (B.1.617.2 and sublineages) and BA.1 Omicron (BA.1 and sublineages) used for this analysis.

For Oxford sequenced samples with <50% coverage, and which could not be reliably assigned using PangoLEARN, we assigned one of the four major lineages if a consensus base was called at three or more lineage defining sites, and with more than two-thirds of these calls consistent with the lineage. Of a total of 5562 and 27290 samples with a Ct≤30 from the time periods coinciding with when samples were sent for sequencing at Oxford and Northumbria, a lineage could be assigned for 5315 (96%) and 21805 (80%) samples, respectively. Among these samples 4829 (91%) and 20191 (93%) could reliably be assigned to one of the four lineages described above, and the remainder were excluded from further analysis. Additional data processing steps are detailed in the Methods section. Because the sequencing protocols and data availability criteria used at Oxford and Northumbria were different, data from the two sequencing labs were incorporated into separate regression analyses.

Serology testing

Blood samples were analyzed for antibodies at the University of Oxford. The immunoassay used for this purpose was developed by the University of Oxford's National SARS-CoV-2 Serology Assay Evaluation Group in conjunction with the study team. The study protocol was approved by the South Central Berkshire B Research Ethics Committee (20/SC/0195). A positive result was recorded for antibody titres above 8 million units (National SARS-CoV-2 Serology Assay Evaluation Group, 2020) on the original fluorometric version of the assay, and above 42 units on the colorimetric version used from March 1, 2021.

A new framework to infer epidemiologically adjusted Ct values

To enable us to investigate the association between viral burden and viral variant, we developed a framework that adjusts observed Ct values to account for the different epidemiological trajectories of different viral variants (see methods). In brief, variant-specific incidence rates for each of the major variants in the sample data (B.1.177, Alpha, Delta and BA.1 Omicron) (Fig 1A) were inferred by combining estimates of total SARS-CoV-2 incidence rates in England (www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/coronaviruscovid19infectionsurveydata) with estimates of the proportion of incident infections with each variant, as inferred from the COVID-19 infection consortium data repository (www.cogconsortium.uk/). These data were used rather than the equivalent estimates available directly from the CIS to prevent the introduction of a time lag between incidence and prevalence into our study. The variant-specific incidence rates were combined with normally distributed infection periods to estimate how the expected distribution of time since infection from randomly sampled individuals changes over calendar time for each of the variants. For each PCR positive sequenced sample in our analysis, the expected distribution of the time since infection corresponding to its variant and sample date was identified and truncated to account for expected bounds, where these could be determined by previous positive or negative samples from the same individual. For each sample, we next estimated an expected distribution of Ct values. This was achieved by assuming that within-host Ct values are described by a piecewise, valley-shaped trajectory (Fig 1B) with depth (viral burden peak) and width (infected period) taken from normal distributions. The timing of the valley trough (peak viral burden) was fixed at a chosen fraction across the width. The parameters describing these metrics were estimated from an alternative data source [16]. However, the mean maximum valley depth (mean peak viral burden) was iteratively inferred, and other parameters–including the timing of the peak viral burden–were varied during sensitivity analyses. For each sample, an adjusted Ct value was then inferred as follows, and as demonstrated graphically in Fig 1C. Firstly, we identified what percentile the observed Ct value lies at among

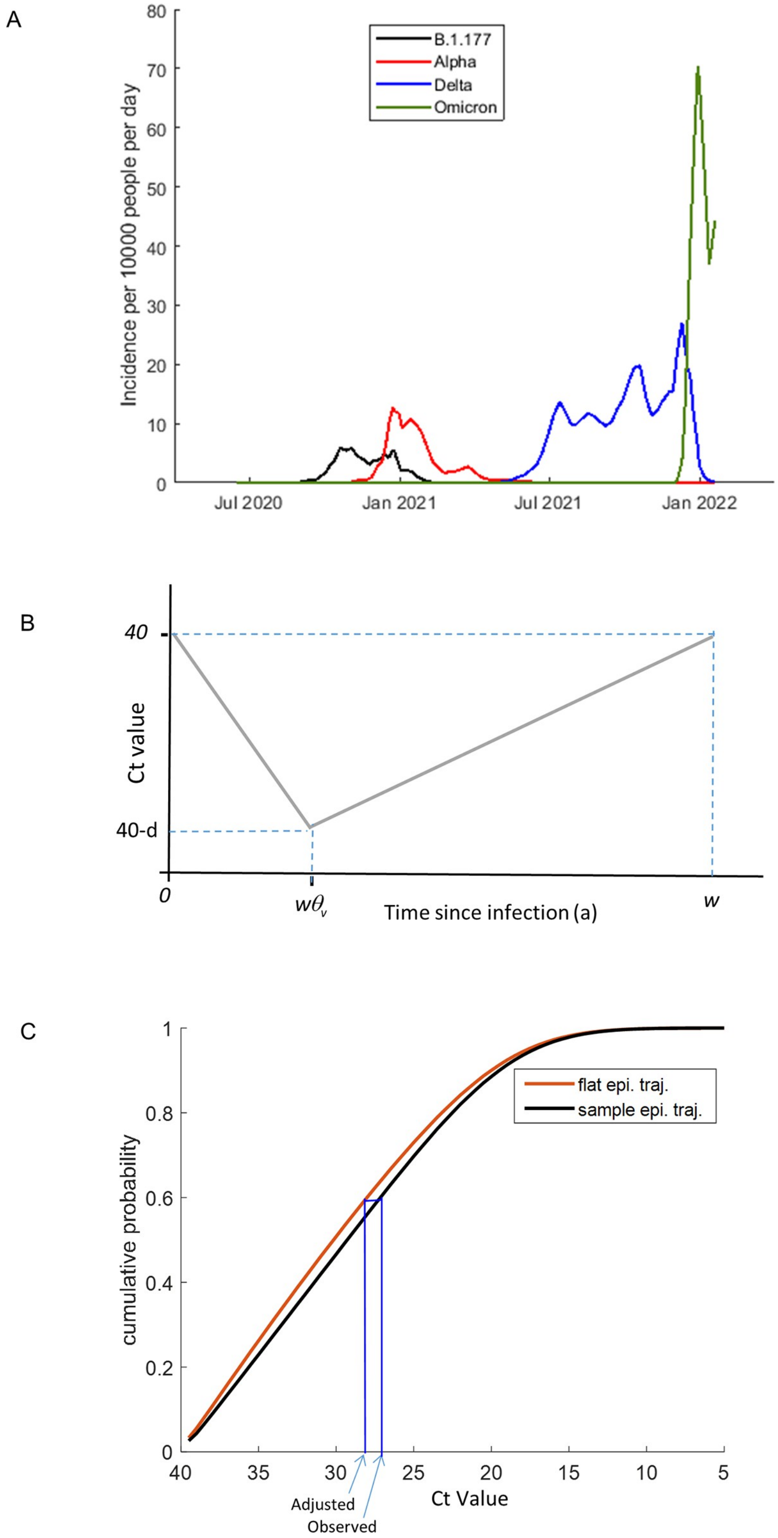

Fig 1. A method for estimating epidemiologically adjusted Ct values. A) Inferred daily incidence with the B.1.177 lineage and the Alpha, Delta and BA.1 Omicron variants between July 2020 and January 2022 in the UK. These were estimated to equal the product of the daily incidence of SARS-CoV-2 and the fraction of incident infections of that variant. B) Within-host Ct trajectories were assumed to be valley shaped, with infected period (width) w, and depth d. The valley trough was estimated to be a fraction θ_v across the width. C) Adjusted Ct values were inferred by first

estimating the cumulative probability distribution of Ct values based upon the sample date and the known epidemiological trajectory of the sample variant and identifying the percentile at which the observed Ct value falls within this distribution. Second, the cumulative probability distribution of Ct values under an assumption of a flat epidemiological trajectory was estimated and the Ct value at the selected percentile was identified.

https://doi.org/10.1371/journal.ppat.1011461.g001

the Ct distribution that is expected, based upon the sample variant (see the black line in Fig 1C). Secondly, we selected the Ct value at the corresponding percentile, from the Ct distribution that is expected, based upon a flat (constant incidence) epidemic trajectory (see the red line in Fig 1C).

Ct values from early and late during the Alpha wave are more closely aligned after epidemiological adjustment

Since we had data spanning a large portion of the epidemiological trajectory of the Alpha wave in the UK, we determined the impact of our method when applied to data collected at different stages during its trajectory. We applied the adjustment to Alpha-variant samples collected from unvaccinated individuals who had no known prior exposure (i.e. no recorded prior infection nor a positive spike-antibody test) (n = 2465; approximately 15% of individuals prior to the start of the Alpha wave were tested for antibodies). By splitting the samples according to sample date into sets representing the early-phase (growth) and late-phase (decline) of the Alpha wave, we visualised how the timing of sampling during the epidemiological trajectory impacted observed Ct values (Fig 2).

The median of the unadjusted Ct values was lower for early-phase samples than for late-phase samples (Fig 2A), consistent with the expected impact of the epidemiological effect. For each sample, our method estimates a probability distribution for the time since infection for that sample, based upon the sample variant, sample collection date, and, where available, the dates of recent positive and negative samples within the same infection. The mean time since infection derived from each of these distributions is plotted in Fig 2B. On average the mean time since infection is longer among the late-phase compared to early-phase samples. Because Ct values are, on average, lower in early infection compared to late infection (Fig 1B), the adjustment acted in the opposite direction and increased the Ct values of early-phase samples, but decreased the Ct values of late-phase samples (Fig 2C).

When the epidemiological adjustment was applied to the Ct values, the adjusted distribution of Ct values for the early-phase and late-phase were more closely aligned compared to the unadjusted values (Fig 2D). For comparison, the application of the method to data from the whole Alpha wave is also shown (middle column in Fig 2), revealing that the net adjustment applied to the full set of samples is negligible. This emphasises the value of using the epidemiological adjustment when samples are only available for part of the epidemiological trajectory of a variant, such as during the emergence phase of a new variant.

The asymmetry of the within-host viral trajectory impacts comparisons

Our framework highlights that the combined impact of the shape of the within-host viral trajectory and the epidemiological stage of a variant can affect viral burden measured at the population level. Plausible changes to our assumption of the mean infected period have only a small impact upon the adjusted values (Fig 3A), whereas plausible changes to the fractional position of the viral burden peak across this period have a bigger effect on the adjusted values (Fig 3B) (although absolute changes are still modest compared with variability between individuals). The closer the peak viral burden is to the start of infection, the greater the epidemiological correction applied to samples selected from just early on or just late on during the Alpha wave.

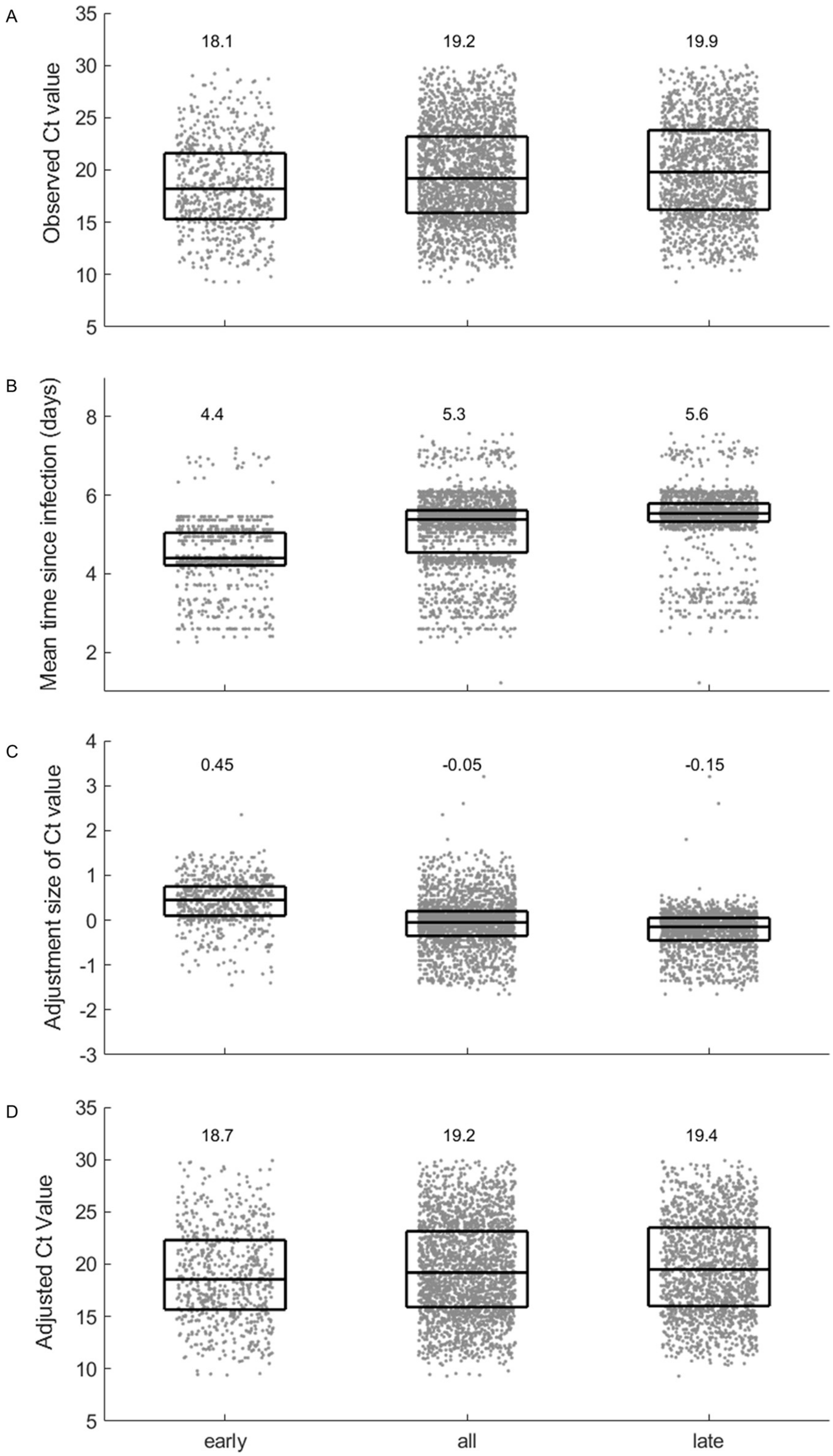

Fig 2. Epidemiological adjustment results in more closely aligned estimates of mean viral burden from samples taken early and late during the Alpha wave. Samples that correspond to Alpha-variant infections in individuals who were unvaccinated and had not been identified as having had a prior exposure were split according to sample date. Four metrics were applied to data from the early phase, all phases and the late phase. In each panel, median and interquartile ranges are overlaid onto individual data points (equivalent distribution plots are provided in S2 Fig). A) The observed Ct values are, on average, higher for late phase, compared to early phase samples. B) The estimated mean

time since infection is, on average, longer for late-phase, compared to early-phase samples. C) The Ct adjustment size is, on average, positive for early phase samples, negative for late phase samples and negligible when all data are considered. D) On average, the adjusted Ct values relating to the early and late phase are more closely aligned than the observed Ct values. However, adjusted values remain, on average, higher in late-phase, compared to early-phase samples.

https://doi.org/10.1371/journal.ppat.1011461.g002

This can be understood by noting that in a random sample, early-phase samples have, on-average, shorter times since infection than late-phase samples and the greater the asymmetry of the within-host viral burden, the greater the difference in expected viral burden between infections in the earlier or later phases of infection (Fig 3C).

In calculating the adjusted Ct values for samples with the Alpha-variant (Fig 2D) we assumed that peak viral burden occurs at a fraction 0.3 across the infected period, based upon prior data from 103 individuals [16]. It is noteworthy that using this parameter estimate, the median adjusted Ct value remains higher for late-phase, compared to early-phase samples. This can be visualised by comparing the red and blues lines shown in Fig 3B at a value $\theta_{\text{Alpha}} = 0.3$ along the x-axis (grey dashed vertical line). For a mean infected period of 10 days [16] (classified according to having a Ct value above 40), this would correspond to the peak infection occurring at day 3. The median adjusted Ct values of the early-phase and late-phase samples are closer when the asymmetry of the within-host trajectory is increased. Arguably, changing this parameter estimate so that the peak is closer to the start of infection than we have assumed, may therefore provide a better estimate of its true value compared to the one that we derived from published work.

However, there are other explanations for a higher viral burden (lower Ct values) in the early-phase samples. Because the CIS conducted a large round of recruitment in September-October 2020, many participants at the start of the Alpha wave were still undergoing more regular–approximately weekly–follow-up, meaning they may genuinely have been sampled closer to the start of infection in the early phase than the later phase. Second, CIS tested antibodies in only ~15% of participants prior to the Alpha wave, so we cannot rule out that some samples come from individuals who had an undetected prior infection and that the number of such individuals increased over the duration of the Alpha wave. For a description of the relationship between sero-status and test positivity see Walker et al. [24]. It is thus credible that more intensive sampling and lower population levels of immunity present earlier on in the Alpha wave could contribute to the pattern of lower adjusted Ct values in early-phase compared to late-phase samples.

Investigating variables associated with within-host viral burden

We investigated whether variables, including viral variant, are associated with adjusted Ct values sampled in the CIS and sequenced at Oxford or Northumbria Universities using partial least squares regression (PLS). Samples sequenced at Oxford were collected between 27th September 2020 and 17th July 2021, and cover the period of the epidemic that includes parts of the B.1.177, Alpha and Delta waves. Samples sequenced at Northumbria were collected between 20th September 2021 and 19th January 2022 and cover parts of the Delta and BA.1 Omicron waves. We have analysed the samples sequenced from the two centres separately so that differences in sequencing protocols and genomic coverage inclusion criteria do not affect our results.

Adjusted Ct values for samples from these two centres are shown in Fig 4, categorised according to sample date (Fig 4A and 4B), participant age (Fig 4C and 4D), and a combination of prior exposure category and variant (Fig 4C and 4F). Using partial least squares regression

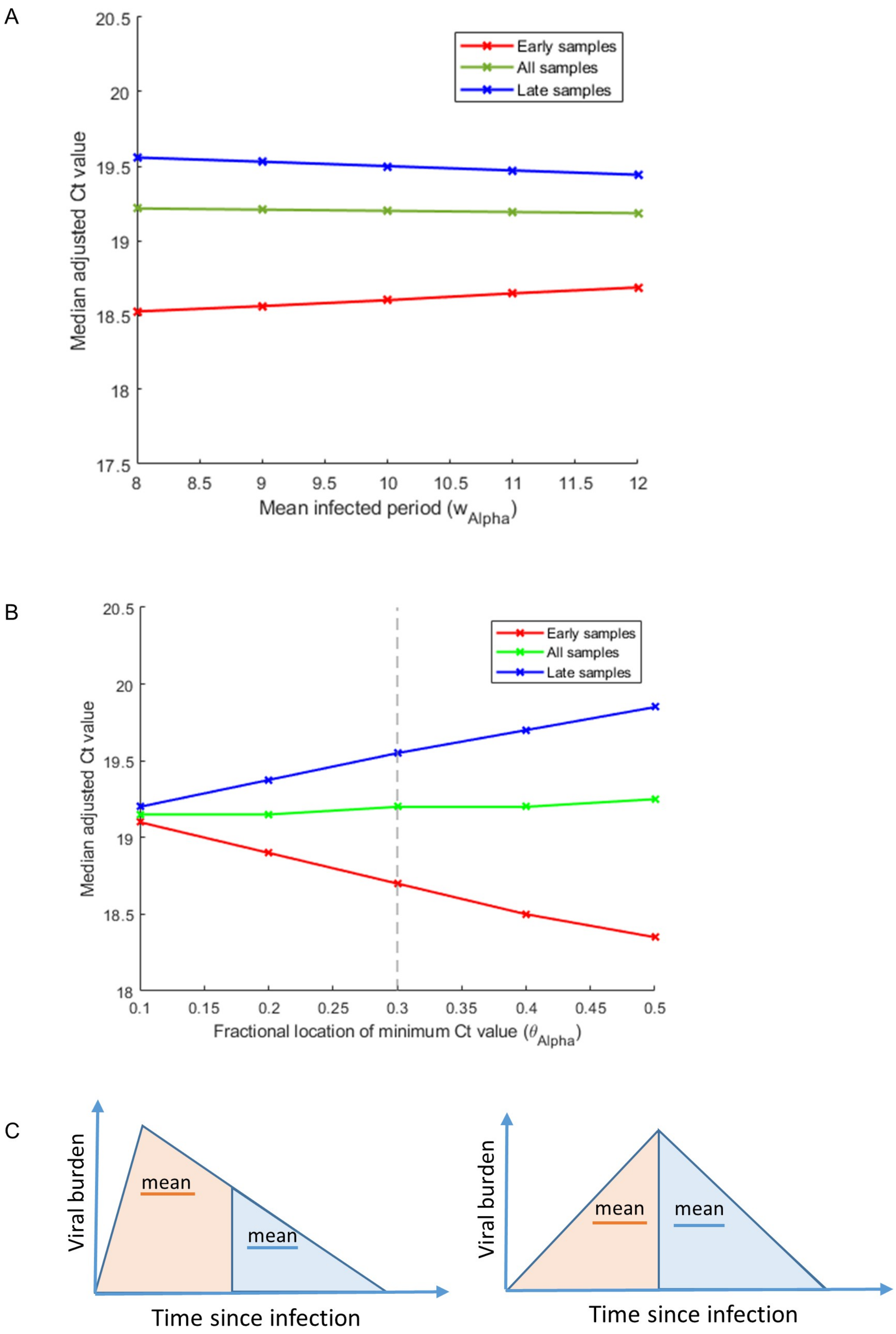
A
Early samples
All samples
Late samples
Median adjusted Ct value
20.5
20
19.5
19
18.5
18
17.5
8
8.5
9
9.5
10
10.5
11
11.5
12
Mean infected period (w_{Alpha})
B
Early samples
All samples
Late samples
Median adjusted Ct value
20.5
20
19.5
19
18.5
18
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Fractional location of minimum Ct value (θ_{Alpha})
C
Viral burden
mean
mean
Time since infection
Viral burden
mean
mean
Time since infection

Fig 3. The epidemiological stage and the asymmetry of the within-host viral trajectory impact the Ct adjustment size. In panels A) and B) samples that correspond to Alpha-variant infections in individuals who were unvaccinated and had not been identified as having had a prior exposure are split according to sample date. The medians of the adjusted Ct values are plotted for early samples (red), late samples (blue) and all samples (green) under different assumptions about the asymmetry and the mean width of the within-host viral burden trajectory. In panel A) the infected period is varied under an assumption that the viral burden trajectory is skewed towards the start of infection ($\theta_{\mathrm{Alpha}} = 0.3$). This shows that Ct values are lower (viral burden is higher) amongst samples taken earlier on during infection, but vary to only a limited degree with changes in the mean infected period (w_{Alpha}). In panel B) the fractional location of the peak viral burden, θ_{Alpha}, is varied under the assumption that the mean infected period (with Ct value ≤40) is 10 days ($w_{\mathrm{Alpha}} = 10$) [16]. This shows that the asymmetry of the within-host viral burden trajectory measurably impacts the adjusted Ct values and that the early- and late-phase Alpha-variant samples are most closely aligned when θ_{Alpha} is smaller. Panel C) highlights how, when the within-host trajectory is skewed towards earlier during infection, viral burden sampled during early infection will on average be higher than that sampled later on in infection.

https://doi.org/10.1371/journal.ppat.1011461.g003

analysis, we assessed the impact of sample date, sex, ethnicity, health care worker status, first vaccine dose product (AstraZeneca ChAdOx1 nCoV-19, Pfizer/BioNTech BNT162b2) and prior exposure category (unvaccinated with no known prior exposure, unvaccinated with a known prior exposure, 1 vaccine dose, 2 vaccine doses, 3 vaccine doses) on adjusted Ct values. It is noteworthy that because not all individuals were tested for spike antibodies prior to infection that individuals with no known prior exposure could have experienced an unidentified prior exposure. In addition, within each prior exposure category we assessed the impact of variant (B.1.177, Alpha, Delta and BA.1 Omicron).

PLS was used for this study instead of standard multivariate regression analysis, because it explicitly acknowledges multicollinearity, which is measurable in our data. This was evident from an investigation of variance inflation factors (discussed below) and from the fact that standard regression analyses gave varying results with different selection procedures. Unlike standard regression analysis, PLS does not explicitly test effects for individual explanatory variables. Instead it acknowledges the issue of multicollinearity by analysing latent (principle) components. Whilst the results are not as intuitive as for multivariate regression, PLS does produce Beta scores (which can be considered equivalent to regression coefficients), and variance in projection (VIP) scores, which can be used to assess the magnitude and importance of the contribution of the different variables to the response, respectively. Other methods, such as Bayesian model averaging (BMA) or principle component analysis (PCA) could alternatively have been employed for this problem. However, the results from BMA can be influenced by the choice of prior and the dimension reduction technique in PCA only acknowledges the relationships between the explanatory variables. PLS, however, is independent of user choices and additionally acknowledge the relationships between the explanatory variables and the predictor variable.

Prior to application of the PLS regression model, we investigated multicollinearity among predictor variables by calculating variance inflation factor (VIF) values (S2 Table). Although no clear consensus on VIF threshold criteria exist, a review of relevant literature (summarized in [32]) suggests that a VIF> 3.3 can be considered an indicator of moderate multicollinearity and a VIF>5 an indicator of strong multicollinearity. Because several of the VIF values among both sample sets were greater than 3.3, we analysed our data using PLS regression to acknowledge the difficulties in disentangling the relative roles of different factors in explaining viral burden.

Viral burden is higher among older individuals

For samples sequenced in Oxford, six components (linear combinations of the predictors that are orthogonal to each other) describe the data (S1A Fig), as determined by the number that minimises the mean squared prediction error. Although these components only explain a

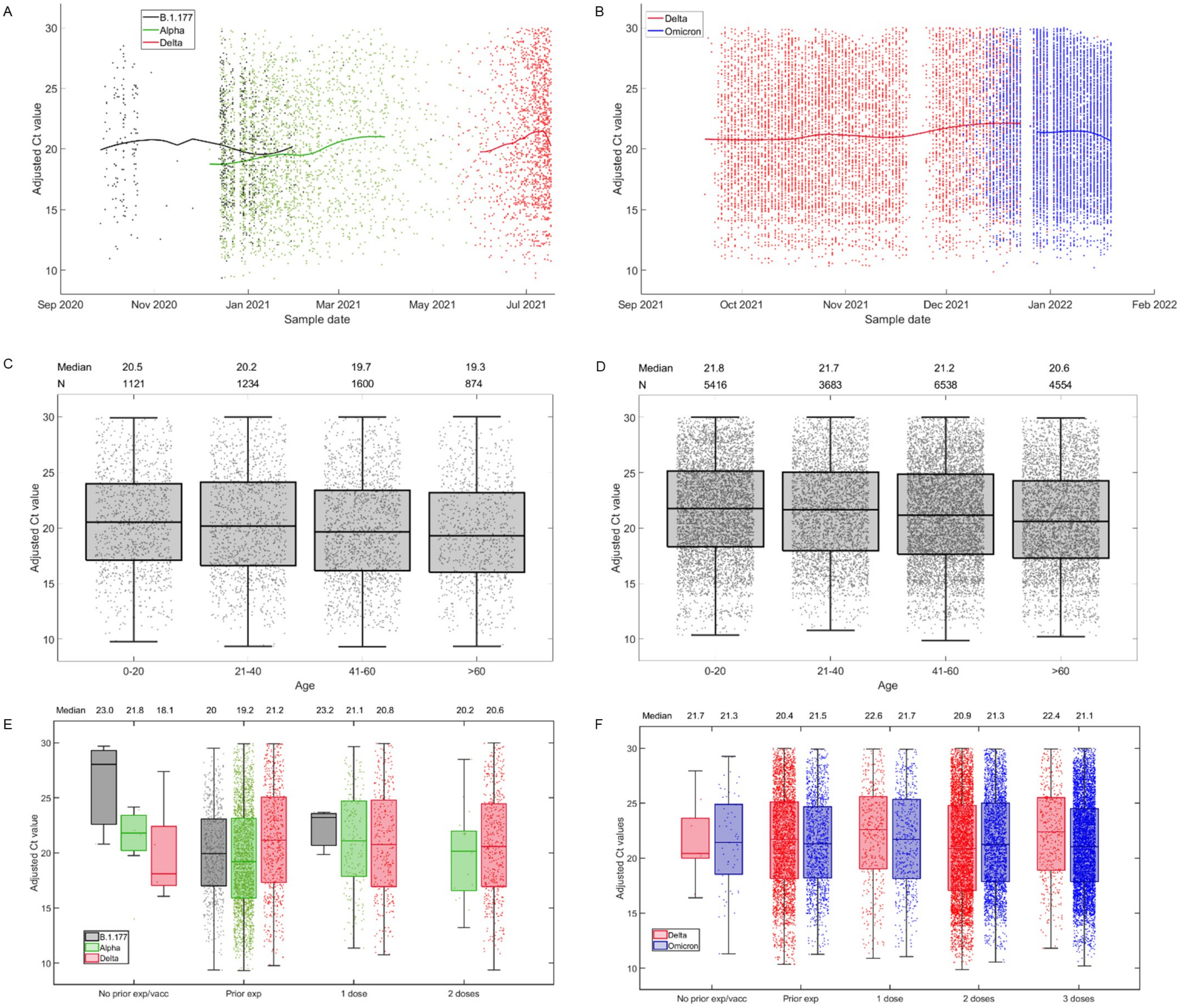

Fig 4. Adjusted Ct values plotted against different factors. For samples sequenced at Oxford (A, C and E) and at Northumbria (d, e and f), adjusted Ct values are plotted against different variabless. Panel A) and B) show a LOESS fit (smoothing parameter = 0.55) of adjusted Ct values over sample date, categorised by variant. Panels C) and D) show box and whisker plots of adjusted Ct values by age category. Panels E) and F) show box and whisker plots of adjusted Ct values by prior vaccination and/or infection, by variant. Horizontal lines represent the median and interquartile range. Parameter values used in these calculations are listed in Table 3.

https://doi.org/10.1371/journal.ppat.1011461.g004

small amount of variance in the adjusted Ct values (2.1%), the first two are both significant in predicting the values in a quantile median regression model (p<0.0001 and p = 0.0004) (used to acknowledge non-normality in the residuals). For the Northumbria samples, six latent components also minimise the mean squared prediction error, the first three of which significantly predict (p<0.0001) the adjusted Ct values (S1A and S1B Fig). This analysis highlights that, taken together, factors included in our model significantly impact viral burden. For reference, loading plots for the first two latent components of each sample are shown in S1C and S1D Fig.

Beta scores and variance in projection (VIP) scores, used to assess the magnitude and importance of the contribution of the different variables to the response, respectively, are provided in (Table 1). Variables with VIP values greater than 1 are typically considered to be important and those with VIP values greater than 0.8 are considered to be borderline important (and were retained in our model). Using this approach, we identified sample collection

Table 1. Beta scores and variance in projection (VIP) values for the partial least squares analysis of samples sequenced in Oxford and Northumbria. A breakdown of sample sizes, by category is also provided. *based upon a Ct value decrease of 3 being equivalent to a 10-fold increase in viral load [34].

Samples	Oxford			Northumbria		
Result	N (4829)	Beta score (viral load factor change*)	VIP	N (20191)	Beta score (viral load factor change)	VIP
Included in model						
Sample Date		0.008 per year (0.99 per year)	1.83		0.013 per year (0.99 per year)	1.22
Age		-0.016 per year (1.13 per 10 years)	1.11		-0.021 per year (1.17 per 10 years)	1.84
Prior exposure						
Ref = Unvaccinated with no known prior exposure	3768			4637		
Known prior exposure	15	1.56 (0.30)	0.30	75	-0.12 (1.10)	0.18
Vaccinated	1046	1.44 (0.33)	1.16	15479	0.34 (0.77)	0.80
Variant amongst unvaccinated individuals with no known prior exposure						
Ref = Alpha (Oxford), Delta (Northumbria)	3526					
B.1.177	658	1.00 (2.15)	1.05			
Delta	645	-0.01 (1.01)	1.22	3310		
BA.1 Omicron	0			1327	-0.95 (2.07)	0.87
Variant amongst vaccinated individuals						
Ref = Alpha (Oxford), *Delta (Northumbria)*	233					
B.1.177	3	1.82 (0.25)	0.22	0		
Delta	820	-1.37 (2.86)	1.13	6118		
BA.1 Omicron	0			9361	-0.59 (1.57)	0.83
Variant amongst individuals with a known prior exposure						
Ref = Alpha (Oxford), Delta (Northumbria)	7					
B.1.177	3	5.17 (0.02)	0.57	0		
Delta	5	-3.05 (10.39)	0.20	10		
BA.1 Omicron	0			65	-0.35 (1.31)	0.18
Vaccine product						
Ref = Pfizer/BioNTech BNT162b2	2812			11506		
AstraZeneca ChAdOx1 nCoV-19	2017	-0.19 (1.16)	0.89	8685	-0.10 (1.08)	0.98
Not included in model						
Vaccine dose amongst vaccinated individuals						
Ref = 1 dose	527			1022		
≥2 doses	519	-0.09 (1.07)	0.73	14457	0.07 (0.95)	0.78
Ethnicity						
Ref = White	4290			18349		
All other ethnicities	539	0.24 (0.83)	0.32	1842	0.46 (0.70)	0.79
Sex						
Ref = Female	2445			10293		
Male	2384	-0.11 (1.09)	0.21	9898	0.50 (0.68)	0.50
Health care worker						
Ref = no	4646			19511		
Yes	183	0.44 (0.71)	0.31	680	0.28 (0.81)	0.31

https://doi.org/10.1371/journal.ppat.1011461.t001 date as an important predictor of Ct values among the samples sequenced at both Oxford (Beta score = 0.008 per year, VIP = 1.83) and Northumbria (Beta score = 0.013 per year, VIP = 1.22). Infection-acquired immunity increased in the population over this period ([33]), which likely contributed to this effect. Age was also an important predictor in both sample sets (Oxford: Beta score = -0.016 per year, VIP = 1.11; Northumbria: Beta score = -0.021 per year, VIP = 1.84). Based upon a 3-point decrease in Ct value being equivalent to a 10-fold increase in viral load (see [34]), these age effects equate to an approximate 14% and 17% increase in viral load for every 10 year age increase, respectively.

There was no strong evidence of an association between viral burden and either sex, ethnicity or being a health care worker (see Table 1 data).

Among unvaccinated individuals with no known prior exposure, viral burden was higher among Alpha compared to B.1.177 infections

We defined unvaccinated individuals with no known prior exposure as those individuals who had neither a previous recorded infection, a previous positive test for spike antibodies, nor a vaccine at least 14 days prior. For samples sequenced at Oxford, Ct values in this group were higher for B.1.177 samples (Beta score = 1.00, VIP = 1.05) compared to Alpha. This corresponds to a 44% reduction in viral loads with Alpha. Among unvaccinated individuals with no known prior exposure, Ct values were similar among Alpha and Delta infections (Oxford: Beta score = -0.01, VIP = 1.22) and lower with BA.1 Omicron compared to Delta with borderline importance (Beta score = -0.95, VIP = 0.87).

Among vaccinated individuals, viral burden was higher among Delta compared to Alpha infection

For individuals who were vaccinated or had a known prior exposure, we further categorised them according to whether they had either tested positive for spike antibodies prior to the first PCR-positive sample in the infection, or had 1 vaccine dose, or 2 or more vaccine doses. Individuals who had both a known prior exposure and were vaccinated were assigned to the appropriate vaccination group. Among the Oxford samples, Ct values were higher among vaccinated individuals, compared to those with no known prior exposure (Beta score = 1.44, VIP = 1.16), equating to a 67% reduction in viral burden with vaccination. Though the magnitude and importance of the signal was weaker–perhaps due to an increase in unidentified prior infections in the population–a similar pattern was observed among the Northumbria samples (Beta score = 0.34, VIP = 0.80). The impact of two or more vaccine doses over one on Ct values was not shown to be important (see Table 1) but the impact of variant among vaccinated individuals was important. Ct values were lower among Delta compared to Alpha infections (Oxford samples: Beta score = -1.37, VIP = 1.13), perhaps resulting from a greater genetic differences between the Delta variant (relative to the Alpha-variant) and the vaccine sequence. The measured effect corresponds to viral burden being 286% higher among Delta, compared to Alpha, infections. With borderline importance, we observed that Ct values were also lower among BA.1 Omicron infections, compared to Delta infections (Northumbria samples: Beta score = -0.59, VIP = 0.83) and were lower with the AstraZeneca ChAdOx1 nCoV-19 vaccine compared to Pfizer/BioNTech BNT162b2 (Oxford: Beta score = -0.19, VIP = 0.89; Northumbria samples: Beta score = -0.10, VIP = 0.98).

The number of samples with a described prior exposure (defined as an individual with a known prior infection or prior antibody positive result) was small (n = 90 across both data sets) and no clear effect was seen across the two dataset (see Table 1).

The association between novel variants and higher viral burden is robust to assumptions about the shape of the within-host viral trajectory

Given our previous observation that the assumed asymmetry in the within-host viral burden trajectory can have a measurable impact on the adjusted Ct value, we conducted a sensitivity analysis on our PLS regression. We varied the parameter that determines the asymmetry of the within-host viral burden trajectory for each of the variants. Both the Beta score (Fig 5A) and

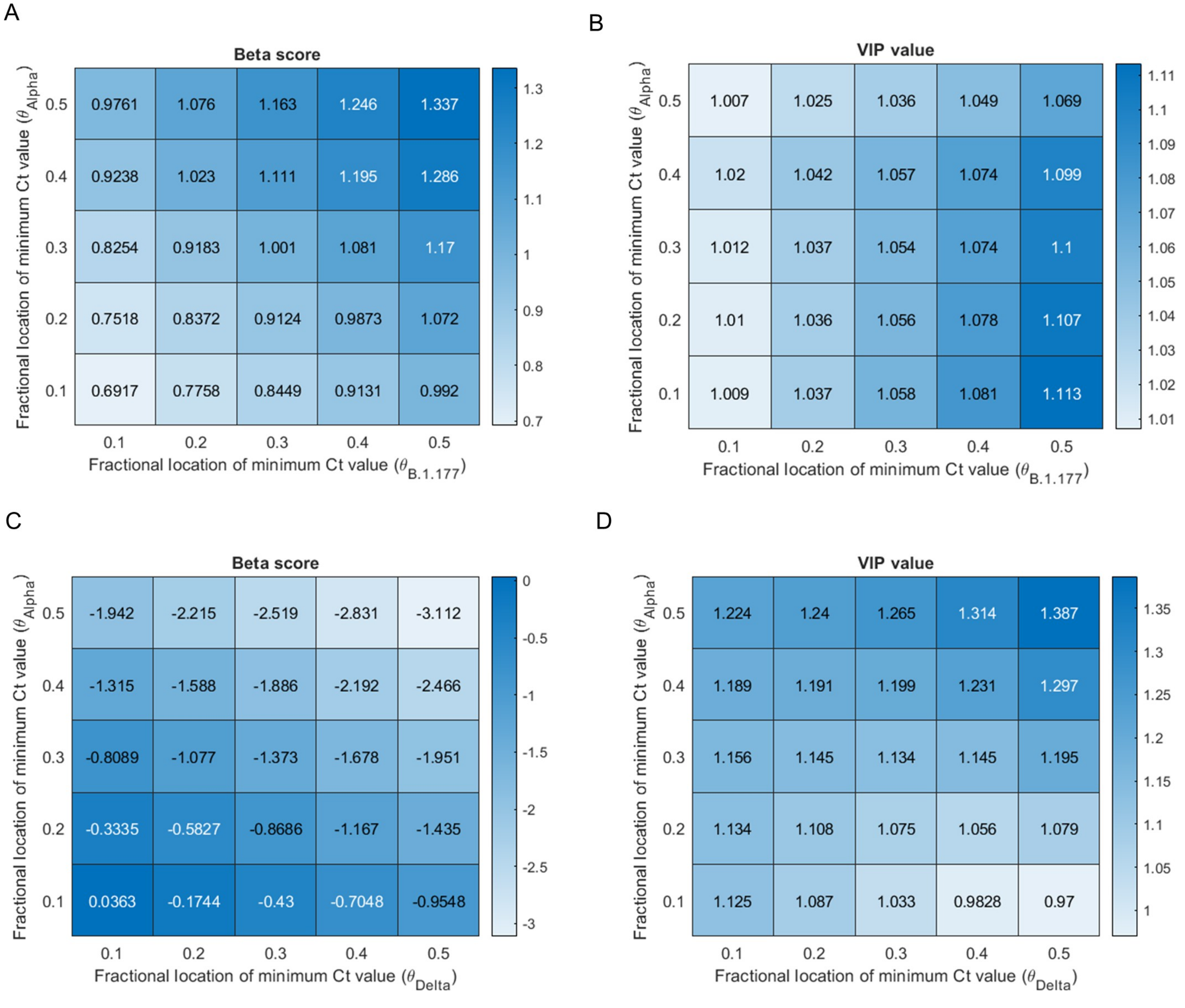

Fig 5. Sensitivity analysis investigating the impact of the shape of within-host viral trajectory on PLS regression analysis into the impact of variant on Ct values. Panels A) and C) show Beta scores, which can be considered to be equivalent to regression coefficients, defining the magnitude of the effect of the variant on the adjusted Ct values. Panels B) and D) show VIP values defining the importance of the association–where values greater than 1 are typically considered to indicate importance. Panels A) and B) investigate the association between the variant being B.1.177 (relative to Alpha) and Ct values among unvaccinated individuals with no known prior exposure. The Beta scores and VIP values vary with changes to the assumed asymmetry of the within-host viral burden trajectory associated with the B.1.177 lineage and the Alpha variant. The asymmetry is determined by changes to the fractional location of the minimum Ct value (peak viral burden) for each variant ($\theta_{B.1.177}$ and θ_{Alpha}, respectively). Data sampled at Oxford. Panels C) and D) investigate the association between the variant being Delta (relative to Alpha) and Ct values among vaccinated individuals and how the Beta scores and VIP values vary with changes to θ_{Alpha} and θ_{Delta}, respectively. Data from samples sequenced at Oxford.

https://doi.org/10.1371/journal.ppat.1011461.g005

VIP value (Fig 5B) for the indicator for the variant being B.1.177 rather than Alpha among unvaccinated individuals sampled at Oxford with no known prior exposure (i.e. no known prior infection nor prior antibody positive test) decreased as the assumed viral burden trajectories of B.1.177 were more skewed towards the start of the infection compared to Alpha (Fig 5A and 5B). These relationships are linked to the fact that although the Oxford samples span a large portion of the Alpha wave, they did not span the early part of the B.1.177 wave. It is noteworthy that the VIP value remained greater than unity across plausible parameter combinations, providing support for the observation that viral burden is higher in infections with the Alpha-variant compared to B.1.177, among unvaccinated individuals with no known prior exposure.

When evaluating the impact on Ct values of the variant being Delta (rather than Alpha) among vaccinated individuals (Fig 5C and 5D), both the VIP value and the magnitude of the Beta score increased as the assumed viral burden of Delta was more skewed towards the start of the infection compared to Alpha. These relationships are linked to the fact that the Oxford samples did not span the latter part of the Delta wave. The VIP value remained greater than unity (or very close to for higher discordance between the shape of the within-host trajectories of the two variants) across plausible parameter combinations. This analysis therefore provides support for the finding that samples with the Delta variant had a lower viral burden compared to samples with the Alpha variant among vaccinated individuals. In summary, our sensitivity analysis supports the finding that both the shift from B.1.177 to Alpha and the shift from Alpha to Delta was associated with an increase in viral burden amongst the population (mostly unvaccinated and mostly vaccinated, respectively) that prevailed at each stage.

Discussion

We developed a framework to compare within-host viral burden across different SARS-CoV-2 variants from random survey data, such as the CIS. The method directly estimates the level of uncertainty in the time-since-infection of each sample due to the sparse nature of the sampling and the effect of differing epidemiological trends of SARS-CoV-2 variants. The method highlights how the combination of the within-host viral trajectory and the epidemiological trajectory of a viral variant can influence observed viral burden in survey data.

Using this framework, we inferred epidemiologically adjusted Ct values from samples sequenced as part of the CIS, a large-scale community survey, recruiting randomly selected private residential households and testing participants regardless of symptoms. Using partial least squares regression, we showed that in combination, factors included our analysis are, strongly associated with Ct values. We also assessed the magnitude and importance of the contribution of different variables to the response. Overall the findings indicate that viral burden shifts over time as population levels of immunity (notably from vaccination) change. In addition, for a given population-level immunological background, viral burden is influenced by viral variant. However, the effect sizes that we observed were modest in relation to the overall spread of the data.

We found that viral burden decreased over time and was higher among older individuals (13% and 17% higher for every 10 year age increase, in the samples sequenced in Oxford and Northumbria, respectively). Among individuals with no known prior immunity, viral burden was, on average, 115% higher among Alpha-variant compared to B.1.177 samples. Vaccination reduced viral burden, with the average reduction being 75% and 30% amongst the samples sequenced at Oxford and Northumbria, respectively. Among vaccinated individuals, there was also a pattern of viral burden increasing with the onset of a novel variant. Notably, viral burden was 286% higher among infections with the Delta-variant, compared to the Alpha-variant. We hypothesize that this finding could be linked to a greater genetic difference between the Delta variant (relative to the Alpha variant) and the vaccine strain.

A key question in virology is to understand how, and to what extent, evolution of a virus impacts its ability to transmit amongst a population. The degree to which evolution impacts within-host viral burden is one potential mechanisms. Although our study does not directly address this question, the effects sizes that we observe can be compared to data from a recent study by Lee et al. [35], in which SARS-CoV-2 infection probability among household contacts was investigated. That study found that on average, a Ct reduction of 15 (5 log increase in viral load) corresponded to a 160% increase in SARS-CoV-2 infection probability. This suggests that the effects sizes that we have observed in relation to viral variant and other variables, including age, are small. It is noteworthy, however, that Lee et al. [35], also showed that variables that we found to be associated with viral burden were associated with transmissibility. Notably, infections among children were less transmissible than those among adults, and infections with the Alpha variant were 50% more transmissible than those with the predecessor variant (B.1.177). In summary therefore, evidence suggests that variables such as age and viral variant impact transmissibility. Although viral burden differences may contribute to this effect, our data suggests that any effect is modest and that other factors must be at play. Such factors include differences between variants in the viral shedding rate, infectious period [36], or the per-virion probability of transmission.

For this study we determined viral variant from viral sequence data, which in practice meant excluding samples with a low viral burden. This is because only samples with Ct $\leq$30 are routinely sequenced, and additionally, samples with higher Ct values (a lower viral burden) are less likely to have sufficient genomic coverage to determine the variant. Although these restrictions could impact our qualitative estimates, we do not expect them to bias our main qualitative results. Furthermore, since individuals with a low viral burden contribute little to viral transmission [37], our study reflects the impact of viral variant and other variables on viral burden at levels that are relevant for transmission.

Monitoring the characteristics of novel SARS-CoV-2 variants will continue to be critical to public health decisions in the foreseeable future. As more countries roll out population representative surveys, accounting for epidemiological effects will remain important. More generally, any studies using community surveillance data that aim to consider traits that vary though infection (e.g. Ct values, immune markers), could be impacted by pathogen epidemiology and therefore could benefit from epidemiological adjustment. In summary, our study promotes a new way of critically analysing random survey data to acknowledge the combined impact of pathogen epidemiology and within-host traits that vary over the course of an infection. Although we have applied these methods to SARS-CoV-2 data, the methods are also applicable to the study of data relating to other pathogens.

Methods

Infection data processing

All individuals with at least one positive sample sequenced in Oxford or Northumbria, and with the virus assigned to one of the four major lineages as described above, were included in our analysis, and indexed $i = 1...n$, where n is the number of individuals. If an individual was infected by more than one major lineage during the study period, these were designated with an infection number j, where $j = 1$ represents the first infection, $j = 2$ the second infection, and so on. Positive samples were assumed to be part of the same infection if they were of the same major variant and were in a continuous sequence of positive samples (i.e. no negative intermediate samples). The index k denotes the kth sample of the infection. In the case of a non-

Table 2. Data used in the study.

Variable	Description
t_{ijk}	Sample date of the *k*th sample of the *j*th infection of the *i*th individual
$\widetilde{t}_{ij}$	Sample date of the last negative before the first positive of the *j*th infection of the *i*th individual
c_{ijk}	Observed Ct value of the *k*th sample of the *j*th infection of the *i*th individual.
v_{ij}	Major variant of the *j*th infection of the *i*th individual
φ_i	Sex of the *i*th individual
e_i	Age group of the *i*th individual
f_i	Vaccine product (AstraZeneca or Pfizer) of the first vaccine dose of the *i*th individual
h_i^r	Date of the *r*th vaccine dose of the *i*th individual

https://doi.org/10.1371/journal.ppat.1011461.t002

continuous sequence of positive samples of the same major lineage, any additional positive samples were excluded from our study. Infections which were of the same major lineage but not in a continuous sequence of positive samples were excluded from the analysis. The list of variables used to describe the data are given in Table 2.

Calculating epidemiologically adjusted Ct values

Step 1. Describing the within-host Ct trajectory. We assume that within-host Ct trajectories are piecewise linear and valley-shaped (Fig 1B), defined by the infected period (width, *w*) and the difference between the minimum Ct value and 40 (depth, *d*). Probability distributions for these variables (calculated in a discrete manner, each spaced by value 0.25 and 0.5 respectively) are derived from truncated discretised normal distributions, described by *p(d)* (Eq 1) and *p(w)* (Eq 2), with means W_v^{mean} *and* D_v^{mean} and standard deviations, W^{SD}, D^{SD}, so that

$$p(d) = (\Phi_D(d) - \Phi_D(d - 0.5))/(\Phi_D(5) - \Phi_D(32)) \text{ for d} = [5.5, 6.0, 6.5, \ldots, 32] \quad (1)$$

$$p(w) = \frac{(\Phi_W(w) - \Phi_W(w - 0.25))}{(\Phi_W(35) - \Phi_W(3))} \text{ for } w = [3.25, 3.50, 3.75, .., 35] \quad (2)$$

where

$$\Phi_D(d) = normalCDF_{(D_v^{mean}, D^{SD})}(d) \quad (3)$$

$$\Phi_W(w) = normalCDF_{(W_v^{mean}, W^{SD})}(w) \quad (4)$$

The peak viral burden is assumed to occur at a time since infection equal to a fraction, θ_v<1, of the total infected period. The parameters W_v^{mean}, W^{SD}, D^{SD} and θ_v are derived from previous studies and varied in sensitivity analyses. The parameter D_v^{mean}, is iteratively inferred to a tolerance of 0.1 following implementation of the methodology described–which, for each sample, estimates an adjusted Ct value–and calculated to equal twice the difference between 40 and the mean adjusted Ct value for that variant. For ease of reference, all other variables described here and throughout the following derivation are listed in Table 3.

Step 2. Estimating the distribution of time since infection for different SARS-CoV-2 variants over calendar time. We estimated the distribution of infections in the population stratified by variant and time since infection over calendar time using published estimates of total incidence of SARS-CoV-2 in the UK (www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/coronaviruscovid19infectionsurveydata) and published estimates of the proportion of incident infections with each of the major variants under study (B.1.177, Alpha, Delta and BA.1 Omicron) over time from the COVID-19

Table 3. Description of additional variables and parameters used in calculation of adjusted Ct values.

Variable	Description	
a	Time since infection (discrete: each unit equivalent to 0.25 days)	
d	Minimum Ct -40 (viral trajectory depth)	
w	Infected period (viral trajectory width) (days)	
v	Variant	
τ	Time step (discrete: each unit equivalent to 0.25 days)	
a_{ijk}	Time since infection of the *k*th sample of the *j*th infection of the *i*th individual (days) (discrete: each unit equivalent to 0.25 days)	
$u_{a,\tau,v}$	Estimated of number of people with time since infection*a* at time step τ, with variant v	
$r_{\tau,v}$	The proportion of incident infections during time step τ that are of variant v	
I_τ	Number of new infections (incidence) during time step τ	
$A_{d,w,\theta}(C)$	Time since infection at Ct value, C, during the down phase of the assumed valley shaped Ct trajectory	
$\widetilde{A}_{d,w,\theta}(C)$	Time since infection at the Ct value, C, during the up phase of the assumed valley shaped Ct trajectory	
$\widetilde{c}_{ijk}$	Adjusted Ct value of the *k*th sample of the *j*th infection of the *i*th individual	
$F_{sample_{ijk}}(C)$	Cumulative probability for the expected Ct value, *C* for sample *ijk*	
$F_{flat}(C)$	Cumulative probability for the expected Ct value, C, assuming a flat trajectory (constant incidence)	
Parameters	**Description**	**Values**
θ_v	Fractional location of the minimal Ct across the infected period, with variant v	0.3
W_v^{mean}	Mean viral trajectory width (infected period, days)	10.1
D_v^{mean}	Mean viral trajectory depth (difference between minimum Ct value and 40)	Iteratively inferred to equal 10 + 2(30 −mean adj Ct) with initial condition: Ct = 20.
W^{SD}	Standard deviation of viral trajectory width	5
D^{SD}	Standard deviation of viral trajectory depth	1.7

https://doi.org/10.1371/journal.ppat.1011461.t003

Genomics UK Consortium (COG-UK: www.cogconsortium.uk). Working in discrete time steps (τ = 1,2,3. . .) that are 0.25 days each, we define I_τ to be the incidence during time step, τ and $r_{\tau,v}$ to be the proportion of incident infections during time step τ that are of variant v (v = 1:4 represent B.1.177, Alpha, Delta and BA.1 Omicron, respectively). We further define $u_{a,\tau,v}$ to be the number of infections with time since infection, a (stratified as discrete time steps of 0.25 days each), during time step τ with variant v. The number of incident infections (i.e. infections with time since infection = 0) during time step τ with each variant v is estimated to be the product of the total incidence during that time step and the fraction of incident infections of that variant ($u_{0,\tau,v} = r_{\tau,v}I_\tau$). To estimate $u_{a,\tau,v}$ for each $a>0$, we assume that the infected periods are taken from a truncated normal distribution with mean, W_v^{mean}, and variance W^{SD}. Therefore, the number of infections of time since infection a, at time step τ is calculated to be the number of incident infections from time step τ-a that are still persisting after a time a, thus:

$$u_{a,\tau,v} = u_{0,\tau-a,v}(1 - normalCDF_{(W_v^{mean},W^{SD})}(a)). \qquad (5)$$

Step 3. For each sample and each infected period, estimate a time since infection distribution. For each sample and for each assumed infected period (w), we inferred the distribution of time since infection. We first selected the distribution (Step 2) that corresponds to the sample date and variant of the sample and adjusted it to account for known bounds on the time since infection for that sample, measured in days. The bounds ($a_{ij}^{\widetilde{max_{ijk\,ij}}}$ and $a_{ij}^{min_{ijk_{ij,k-1}}}$) are derived by considering information on Ct values at previous samples and scaled to account for the transformation to discrete time steps. The time since infection probability distribution for each sample is then given by:

$$p\left(a_{ijk}|w, t_{ijk}, v_{ij}\right) = \begin{cases} 0 \text{ if } a_{ijk} \begin{cases} > w \text{ or} \\ > a_{ij}^{max} \text{ or} \\ < a_{ij}^{min} \{ \end{cases} \end{cases} \left| \frac{u_{a_{ijk},4t_{ijk},v_{ij}}}{\sum_{a=4a_{ij}^{min}}^{4\ min(a_{ij}^{max}} \sum u_{a,4t_{ijk},v_{ij}} \text{ otherwise}} \right. \{ \tag{6}$$

Step 4. Infer a sample-specific expected distribution of Ct values. For each sample, based upon the sample time (t_{ijk}) and variant (v_{ij}), we derived an expected distribution of Ct values (Eq 7). This was done by conditioning on the time since infection (a) and the depth (d) and width (w) of the within host viral trajectory. These conditional probabilities were combined with the time since infection distributions derived in step 3 and the within-host parameter distributions described in step 1.

$$p\left(C - 0.5 \leq c < C|t_{ijk}, v_{ij}\right) = \frac{\sum_{a_{ijk}} \sum_{d} \sum_{w} p(C - 0.5 \leq c < C|a_{ijk}, d, w) p(a_{ijk}, d, w|t_{ijk}, v_{ij})}{\sum_{c=0.5}^{40} \sum_{a_{ijk}} \sum_{d} \sum_{w} p(C - 0.5 \leq c < C|a_{ijk}, d, w) p(a_{ijk}, d, w|t_{ijk}, v_{ij})} \tag{7}$$

where the probability of a particular time since infection (a_{ijk}), trajectory width (w) and trajectory depth (d) is given by:

$$p(a_{ijk}, d, w|t_{ijk}, v_{ij}) = p(a_{ijk}|w, t_{ijk}, v_{ij}) p(d) p(w) \tag{8}$$

and the probability of the Ct value (c) falling within a certain discrete boundary, given the time since infection and the width and depth of the viral trajectory, is defined as 1 or 0 depending upon whether it matches up with the valley-shaped viral trajectory curve (Fig 1B), as shown below:

$$p(C - 0.5 \leq c < C|a_{ijk}, w, d) = \begin{cases} 1 \text{ if } A_{d,w,\theta_v}(C - 0.5) < a_{ijk} < A_{d,w,\theta_v}(C) \text{ and } a_{ijk} \leq \theta w \\ 1 \text{ if } \widetilde{A}_{d,w,\theta_v}(C - 0.5) < a_{ijk} < \widetilde{A}_{d,w,\theta_v}(C) \text{ and } a_{ijk} > \theta w \\ 0 \text{ otherwise} \end{cases} \tag{9}$$

Where C is a dummy variable representing the Ct value, and

$$A_{d,w,\theta_v}(C) = \frac{(40 - C)\theta_v w}{d} \tag{10}$$

and

$$\widetilde{A}_{d,w,\theta_v}(C) = w - \frac{(40 - C)(1 - \theta_v)w}{d} \tag{11}$$

are dummy variables that describe the relationship between the Ct value (C) and the time since infection ($A_{d,w,\theta_v}(C)$ and $\widetilde{A}_{d,w,\theta_v}(C)$), during down the phase and up phase of the valley-shaped trajectory, respectively.

Step 5. Calculate an expected distribution of Ct values for a flat epidemic trajectory. The full process for calculating an expected distribution of Ct values (steps 1–4) was repeated under an assumption of a flat epidemic trajectory (constant incidence) rather than a variant-specific trajectory.

Step 6. For each sample, infer an epidemiologically adjusted Ct value. For each sample, we identified the percentile that the observed Ct (c_{ijk}) falls in, among the sample-specific expected Ct distribution. The adjusted Ct value ($\widetilde{c}_{ijk}$) was then derived by identifying the Ct value at that percentile within the expected distribution of Ct values based upon a flat epidemic trajectory (Fig 1C).

$$\widetilde{c}_{ijk} = F_{flat}^{-1}(F_{sample_{ijk}}(c_{ijk})) \tag{12}$$

where

$$F_{sample_{ijk}}(\hat{C}) = p(\hat{c} < \hat{C}|sample_{ijk}) = \sum\nolimits_{C=5.5,6.0,\ldots}^{\hat{C}} p(C - 0.5 \leq c < C|t_{ijk}, v_{ij}) \tag{13}$$

$$F_{flat}(\hat{C}) = p(\hat{c} < \hat{C}|\text{flat epidemic}) = \sum\nolimits_{C=5.5,6.0,\ldots}^{\hat{C}} p(C - 0.5 \leq c < C|\text{flat epidemic}) \tag{14}$$

Implementation of analysis

All analyses were implemented in Matlab and the code is available at https://github.com/helenfryer1000000/epidemiologically-adjusted-viral-load. Estimation of adjusted Ct values was implemented using a bespoke script. Partial least squares regression was implemented using the PLSregress function, which is part of the Statistics and Machine Learning toolbox in Matlab. Quantile median regression was implemented using the function qr_standard, provided at: https://github.com/zjph602xtc/Quantile_reg.

Supporting information

S1 Table. A review of published studies investigating the impact of viral variant on Ct values.
(DOCX)

S2 Table. Variance inflation factor (VIF) values.
(DOCX)

S1 Fig. Mean squared error and loading plots relating to the partial least squares regression analysis.
(TIF)

S2 Fig. Plots showing the distribution of the data provided in Fig 2.
(TIF)

S1 Data. A list of accession number for samples included in this study. Sequences can be accessed via the the European Nucleotide Archive (ENA) at https://www.ebi.ac.uk/ena/browser/home.
(TXT)

S1 Text. A list of members of the COVID-19 Genomics UK (COG-UK) consortium.
(DOCX)

Acknowledgments

The authors acknowledge use of data generated through the COVID-19 Genomics Programme funded by the Department of Health and Social Care.

Author Contributions

Conceptualization: Helen R. Fryer, Tanya Golubchik, David Bonsall, Katrina A. Lythgoe.

Data curation: Helen R. Fryer, Tanya Golubchik, Matthew Hall, Laura Thomson, Matthew Bashton, Clare M. McCann, Katrina A. Lythgoe.

Formal analysis: Helen R. Fryer.

Funding acquisition: Tanya Golubchik, David Bonsall, Katrina A. Lythgoe.

Investigation: Helen R. Fryer, Tanya Golubchik, Christophe Fraser, George MacIntyre-Cockett, Amy Trebes, David Buck, Paolo Piazza, Angie Green, Lorne J Lonie, Darren Smith, Matthew Crown, Andrew Nelson, Mohammed Adnan Tariq, Claire J. Elstob, Rui Nunes Dos Santos, Zack Richards, Xin Xhang, Joseph Hawley, Mark R. Lee, Priscilla Carrillo-Barragan, Isobel Chapman, Sarah Harthern-Flint, David Bonsall.

Methodology: Helen R. Fryer, Tanya Golubchik, Robert Hinch, Luca Ferretti, Anel Nurtay, Lorenzo Pellis, Thomas House.

Project administration: Helen R. Fryer, Tanya Golubchik, David Bonsall, Katrina A. Lythgoe.

Software: Helen R. Fryer, Laura Thomson.

Supervision: Tanya Golubchik, Christophe Fraser, Lorenzo Pellis, David Bonsall, Katrina A. Lythgoe.

Validation: Helen R. Fryer.

Visualization: Helen R. Fryer.

Writing – original draft: Helen R. Fryer.

Writing – review & editing: Helen R. Fryer, Tanya Golubchik, Matthew Hall, Anel Nurtay, Lorenzo Pellis, David Bonsall, Katrina A. Lythgoe.

References

1. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday JD, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science. 2021; 372(6538). Epub 2021/03/05. https://doi.org/10.1126/science.abg3055 PMID: 33658326; PubMed Central PMCID: PMC8128288.
2. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 2021; 593(7858):266–9. Epub 2021/03/27. https://doi.org/10.1038/s41586-021-03470-x PMID: 33767447.
3. Lyngse FP, Molbak K, Skov RL, Christiansen LE, Mortensen LH, Albertsen M, et al. Increased transmissibility of SARS-CoV-2 lineage B.1.1.7 by age and viral load. Nat Commun. 2021; 12(1):7251. Epub 2021/12/15. https://doi.org/10.1038/s41467-021-27202-x PMID: 34903718; PubMed Central PMCID: PMC8669007.
4. Allen H, Vusirikala A, Flannagan J, Twohig KA, Zaidi A, Chudasama D, et al. Household transmission of COVID-19 cases associated with SARS-CoV-2 delta variant (B.1.617.2): national case-control study. Lancet Reg Health Eur. 2022; 12:100252. Epub 2021/11/04. https://doi.org/10.1016/j.lanepe.2021.100252 PMID: 34729548; PubMed Central PMCID: PMC8552812.
5. Jorgensen SB, Nygard K, Kacelnik O, Telle K. Secondary Attack Rates for Omicron and Delta Variants of SARS-CoV-2 in Norwegian Households. JAMA. 2022; 327(16):1610–1. Epub 2022/03/08. https://doi.org/10.1001/jama.2022.3780 PMID: 35254379; PubMed Central PMCID: PMC8902690.

6. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2021; 592(7852):116–21. Epub 2020/10/28. https://doi.org/10.1038/s41586-020-2895-3 PMID: 33106671; PubMed Central PMCID: PMC8158177.

7. Leung NHL. Transmissibility and transmission of respiratory viruses. Nat Rev Microbiol. 2021; 19 (8):528–45. Epub 2021/03/24. https://doi.org/10.1038/s41579-021-00535-6 PMID: 33753932; PubMed Central PMCID: PMC7982882.

8. Planas D, Saunders N, Maes P, Guivel-Benhassine F, Planchais C, Buchrieser J, et al. Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. Nature. 2022; 602(7898):671–5. Epub 2022/01/12. https://doi.org/10.1038/s41586-021-04389-z PMID: 35016199.

9. Planas D, Veyer D, Baidaliuk A, Staropoli I, Guivel-Benhassine F, Rajah MM, et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature. 2021; 596(7871):276–80. Epub 2021/07/09. https://doi.org/10.1038/s41586-021-03777-9 PMID: 34237773.

10. Pouwels KB, House T, Pritchard E, Robotham JV, Birrell PJ, Gelman A, et al. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. Lancet Public Health. 2021; 6(1):e30–e8. Epub 2020/12/15. https://doi.org/10.1016/S2468-2667(20)30282-6 PMID: 33308423; PubMed Central PMCID: PMC7786000.

11. Pouwels KB, Pritchard E, Matthews PC, Stoesser N, Eyre DW, Vihta KD, et al. Effect of Delta variant on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. Nat Med. 2021; 27(12):2127–35. Epub 2021/10/16. https://doi.org/10.1038/s41591-021-01548-7 PMID: 34650248; PubMed Central PMCID: PMC8674129.

12. Frampton D, Rampling T, Cross A, Bailey H, Heaney J, Byott M, et al. Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. Lancet Infect Dis. 2021; 21(9):1246–56. Epub 2021/04/16. https://doi.org/10.1016/S1473-3099(21)00170-5 PMID: 33857406; PubMed Central PMCID: PMC8041359 manufacture of UCL-Ventura continuous positive airway pressure device for patients with COVID-19 from the UK Department of Health and Social Care, during the conduct of the study; grants and advisory board fees paid to institution research fund from NewB; grants from DSTL; advisory board and speaking fees paid into institutional research fund from Amormed, Biotest, General ElectricBaxter, Baxter, Roche, Bayer, and Shionogi; and grants from Critical Pressure and Apollo Therapeutics, outside the submitted work. All other authors declare no competing interests.

13. Calistri P, Amato L, Puglia I, Cito F, Di Giuseppe A, Danzetta ML, et al. Infection sustained by lineage B.1.1.7 of SARS-CoV-2 is characterised by longer persistence and higher viral RNA loads in nasopharyngeal swabs. Int J Infect Dis. 2021; 105:753–5. Epub 2021/03/09. https://doi.org/10.1016/j.ijid.2021.03.005 PMID: 33684558; PubMed Central PMCID: PMC7934691.

14. Kidd M, Richter A, Best A, Cumley N, Mirza J, Percival B, et al. S-Variant SARS-CoV-2 Lineage B1.1.7 Is Associated With Significantly Higher Viral Load in Samples Tested by TaqPath Polymerase Chain Reaction. J Infect Dis. 2021; 223(10):1666–70. Epub 2021/02/14. https://doi.org/10.1093/infdis/jiab082 PMID: 33580259; PubMed Central PMCID: PMC7928763.

15. Cosentino G, Bernard M, Ambroise J, Giannoli JM, Guedj J, Debarre F, et al. SARS-CoV-2 viral dynamics in infections with Alpha and Beta variants of concern in the French community. J Infect. 2022; 84 (1):94–118. Epub 2021/07/31. https://doi.org/10.1016/j.jinf.2021.07.031 PMID: 34329672.

16. Kissler SM, Fauver JR, Mack C, Tai CG, Breban MI, Watkins AE, et al. Viral Dynamics of SARS-CoV-2 Variants in Vaccinated and Unvaccinated Persons. N Engl J Med. 2021; 385(26):2489–91. Epub 2021/12/24. https://doi.org/10.1056/NEJMc2102507 PMID: 34941024; PubMed Central PMCID: PMC8693673.

17. Ke R, Martinez PP, Smith RL, Gibson LL, Mirza A, Conte M, et al. Daily longitudinal sampling of SARS-CoV-2 infection reveals substantial heterogeneity in infectiousness. Nat Microbiol. 2022; 7(5):640–52. Epub 2022/04/29. https://doi.org/10.1038/s41564-022-01105-z PMID: 35484231.

18. Golubchik T, lythgoe KA, Hall M, Ferretti L, Fryer HR, MacIntyre-Cockett G, et al. Early analysis of a potential link between viral load and the N501Y mutation in the SARS-COV-2 spike protein. medRxiv. 2021.

19. Li B, Deng A, Li K, Hu Y, Li Z, Shi Y, et al. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. Nat Commun. 2022; 13(1):460. Epub 2022/01/26. https://doi.org/10.1038/s41467-022-28089-y PMID: 35075154; PubMed Central PMCID: PMC8786931.

20. Ma Q, Liu J, Liu Q, Kang L, Liu R, Jing W, et al. Global Percentage of Asymptomatic SARS-CoV-2 Infections Among the Tested Population and Individuals With Confirmed COVID-19 Diagnosis: A Systematic Review and Meta-analysis. JAMA Netw Open. 2021; 4(12):e2137257. Epub 2021/12/15. https://doi.org/10.1001/jamanetworkopen.2021.37257 PMID: 34905008; PubMed Central PMCID: PMC8672238.

21. Chang MC, Hur J, Park D. Interpreting the COVID-19 Test Results: A Guide for Physiatrists. Am J Phys Med Rehabil. 2020; 99(7):583–5. Epub 2020/05/20. https://doi.org/10.1097/PHM.0000000000001471 PMID: 32427600; PubMed Central PMCID: PMC7268832.

22. Trypsteen W, Van Cleemput J, Snippenberg WV, Gerlo S, Vandekerckhove L. On the whereabouts of SARS-CoV-2 in the human body: A systematic review. PLoS Pathog. 2020; 16(10):e1009037. Epub 2020/10/31. https://doi.org/10.1371/journal.ppat.1009037 PMID: 33125439; PubMed Central PMCID: PMC7679000.

23. Hay JA, Kennedy-Shaffer L, Kanjilal S, Lennon NJ, Gabriel SB, Lipsitch M, et al. Estimating epidemiologic dynamics from cross-sectional viral load distributions. Science. 2021; 373(6552). Epub 2021/06/05. https://doi.org/10.1126/science.abh0635 PMID: 34083451; PubMed Central PMCID: PMC8527857.

24. Walker AS, Pritchard E, House T, Robotham JV, Birrell PJ, Bell I, et al. Ct threshold values, a proxy for viral load in community SARS-CoV-2 cases, demonstrate wide variation across populations and over time. Elife. 2021; 10. Epub 2021/07/13. https://doi.org/10.7554/eLife.64683 PMID: 34250907; PubMed Central PMCID: PMC8282332.

25. Statistics OfN. Coronavirus (COVID-19) Infection Survey: methods and further information. 2023.

26. Statistics OfN. Coronavirus (COVID-19) Infection Survey: quality and methodology information (QMI). 2023.

27. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, et al. SARS-CoV-2 within-host diversity and transmission. Science. 2021; 372(6539). Epub 2021/03/11. https://doi.org/10.1126/science.abg0821 PMID: 33688063; PubMed Central PMCID: PMC8128293.

28. Bonsall D, Golubchik T, de Cesare M, Limbada M, Kosloff B, MacIntyre-Cockett G, et al. A Comprehensive Genomics Solution for HIV Surveillance and Clinical Monitoring in Low-Income Settings. J Clin Microbiol. 2020; 58(10). Epub 2020/07/17. https://doi.org/10.1128/JCM.00382-20 PMID: 32669382; PubMed Central PMCID: PMC7512176.

29. Baker DJ, Aydin A, Le-Viet T, Kay GL, Rudder S, de Oliveira Martins L, et al. CoronaHiT: high-throughput sequencing of SARS-CoV-2 genomes. Genome Med. 2021; 13(1):21. Epub 2021/02/11. https://doi.org/10.1186/s13073-021-00839-5 PMID: 33563320; PubMed Central PMCID: PMC7871948.

30. Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D, et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. Virus Evol. 2018; 4(1):vey007. Epub 2018/06/08. https://doi.org/10.1093/ve/vey007 PMID: 29876136; PubMed Central PMCID: PMC5961307.

31. O'Toole A, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol. 2021; 7(2):veab064. Epub 2021/09/17. https://doi.org/10.1093/ve/veab064 PMID: 34527285; PubMed Central PMCID: PMC8344591.

32. Kock N, Lynn G. Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. Journal of the Association for information Systems. 2012;13(7).

33. Statistics OfN. Coronavirus (COVID-19) Infection Survey, antibody data, UK: 29 March 2023. UK: 2023.

34. England PH. Understanding cycle threshold (Ct) in SARS-CoV-2 RT-PCR: a guide for health protection teams. 2020.

35. Lee LYW, Rozmanowski S, Pang M, Charlett A, Anderson C, Hughes GJ, et al. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Infectivity by Viral Load, S Gene Variants and Demographic Factors, and the Utility of Lateral Flow Devices to Prevent Transmission. Clin Infect Dis. 2022; 74(3):407–15. https://doi.org/10.1093/cid/ciab421 PMID: 33972994; PubMed Central PMCID: PMC8136027.

36. Hart WS, Miller E, Andrews NJ, Waight P, Maini PK, Funk S, et al. Generation time of the alpha and delta SARS-CoV-2 variants: an epidemiological analysis. Lancet Infect Dis. 2022. Epub 2022/02/18. https://doi.org/10.1016/S1473-3099(22)00001-9 PMID: 35176230; PubMed Central PMCID: PMC8843191 report participation in discussions of the UK Government's Scientific Pandemic Influenza Group on Modelling, Operational subgroup.

37. Marc A, Kerioui M, Blanquart F, Bertrand J, Mitja O, Corbacho-Monne M, et al. Quantifying the relationship between SARS-CoV-2 viral load and infectiousness. Elife. 2021; 10. Epub 2021/09/28. https://doi.org/10.7554/eLife.69302 PMID: 34569939; PubMed Central PMCID: PMC8476126.