Classifying and Differentiating Individuals with Respiratory Syncytial Virus, Influenza, and COVID-19 Cases in OpenSAFELY Between 2016 and 2024

Em Prestige[1*], Charlotte Warren-Gash[1], Jennifer K Quint[2], Dave Evans[3], Ruth E Costello[1], Amir Mehrkar[3], Seb Bacon[3], Ben Goldacre[3], Sarah Barley-McMullen[4], Farheen Yameen[4], Michael Natt[4], Yvonne Alder[4], William J. Hulme[3+], Edward P. K. Parker[1+], Rosalind M. Eggo[1+]

1. Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel St, London. WC1E 7HT, United Kingdom
2. Faculty of Medicine, School of Public Health, Imperial College London, London, United Kingdom
3. Bennett Institute for Applied Data Science, Oxford University, Oxford, United Kingdom
4. Winter Pressures PPIE Advisory Board, London School of Hygiene & Tropical Medicine, Keppel St, London. WC1E 7HT, United Kingdom

*corresponding author: em.prestige@lshtm.ac.uk
+ indicates equal contribution

Keywords: electronic health records, phenotype, identification, classification, respiratory viruses, Covid-19, influenza, RSV

Abstract

Electronic health records (EHRs) are a rich source of data which can be used to analyse health outcomes using computable phenotypes. With the approval of NHS England we used the OpenSAFELY secure analytics platform to design and assess phenotypes to classify three key respiratory viruses – respiratory syncytial virus (RSV), influenza, and COVID-19 – in English coded health data between September 2016 and August 2024. We compared specific and sensitive phenotypes to one another and to publicly available surveillance data. Cases from both phenotypes showed similar seasonal patterns to surveillance data. Sensitive phenotypes led to increased risk of misclassification than specific phenotypes for mild cases. For severe cases the risk of misclassification was higher in infants than for older adults, irrespective of the phenotype used. The phenotypes presented here offer a solution to classifying respiratory viruses from coded health records in the absence of testing information.

Background

Respiratory Syncytial Virus (RSV), influenza, and COVID-19 are major respiratory viruses and key drivers of healthcare utilisation (1–8), but often present with a similar upper respiratory tract infection (URTI) syndrome. Electronic Health Records (EHRs) provide a rich source of information on these viruses, their burden on healthcare systems, and the risk factors for disease. However, the viruses are rarely tested for outside of secondary care - excepting extraordinary circumstances, such as the widespread testing during the COVID-19 state of emergency. Additionally, while molecular testing is regularly conducted in hospital settings, test results are rarely coded in EHRs, making them unavailable for research. The absence of testing data presents a key challenge to accurately classifying and differentiating RSV, influenza, and COVID-19 in EHRs. The ability to detect and distinguish these viruses could broaden research capabilities and improve outbreak detection.

When individuals present to healthcare providers, such as general practitioners (GPs) and hospitals, their consultations are recorded with clinical codes (9). Clinical codes can correspond to specific conditions, symptoms or prescriptions. In secondary care settings in the UK, hospital episodes are classified using ICD-10 codes (10), which can denote pathogen-specific diagnoses (e.g. RSV), diseases (e.g. bronchiolitis), or pathogen-associated presentations (e.g. unspecified lower respiratory tract infections). In primary care settings, clinical events are recorded using a much more diverse system of SNOMED-CT codes (11), again ranging from highly specific diagnoses to broader symptoms.

Current methods of investigating respiratory viruses often rely on using testing data from surveillance systems, for which reporting varies both over time and between viruses (12). Many surveillance systems are reported at an aggregate level and as such cannot be used to investigate individual-level risk factors for disease; those that are individual-level use a smaller sample of practices when compared to broader EHR systems. We aimed to develop a series of computable phenotypes, comprising codelists and corresponding sets of logical rules (13), to classify individuals with RSV, influenza, and COVID-19 in EHRs. For each virus, we developed: (i) phenotypes designed to be specific to each virus; (ii) phenotypes intended to capture cases relating to each virus, with greater sensitivity (but potentially lower specificity); and (iii) phenotypes devised to capture overall respiratory burden which may be attributed to any of

the three viruses. We then evaluated the ability of these phenotypes to classify and differentiate cases in primary and secondary care settings. These phenotypes have scope to be used for: (i) real-time surveillance purposes; (ii) regular reporting on trends for service planning and monitoring; and (iii) epidemiologic research into the impact of new interventions or service disruptions.

Methods

Data Source

Through the OpenSAFELY analytics platform we extracted linked GP, hospital admissions, and A&E data for 42% of the population of England. All practices running SystmOne software made by the health technology company TPP were included, which includes over 2500 practices. The patient population was representative of the general population with some regional variations (14). We used consultations within primary care, secondary care (Admitted Patient Care Spells - APCS), and emergency care (Emergency Care Data Set - ECDS). Emergency care attendances were available from late 2017, but were incomplete until late 2019 (15). We extracted information on prescriptions from primary care EHRs. For infants, ethnicity from primary care records was supplemented with ethnicity recorded in secondary care due to missingness in ethnicity for infants (~70%), which was reduced to ~4% after supplementation. We did not supplement ethnicity for other cohorts since individuals presenting to secondary care are likely to differ from the general population. OpenSAFELY requires all outputs to undergo statistical disclosure control. As such, all results have been rounded to the nearest midpoint-10; all true zeroes have been retained.

Patient and Public Involvement

A panel of six patient and public members were involved during development of the research and presentation of results. The accessibility of these results has been improved through the input of the panel. Panel members vary by age, gender, ethnicity, socioeconomic status, chronic illness/disability, and lived experience of respiratory virus infections personally or as a carer.

Study Population

We split the study population into four age cohorts (older adults, 65 years or over; adults, 18-64 years; children and young people, 2-17 years; and infants, 0-2 years). We extracted eight annual cohorts from 2016 to 2024, with each cohort including data from 1 September to 31 August of the subsequent year. Individuals were included in any annual cohort for which they met the inclusion criteria of 3 months of continuous registration prior to follow-up; infants were included from birth if they were registered within 3 months of birth. We also explored a subgroup of infants who could be linked to their birthing parent; linked infants were included if their birthing parent had at least 1 year of continuous registration prior to birth.

Follow-up continued until the first of: death, practice deregistration, aging out of age cohort, or end of annual cohort (31 August). Individuals were excluded from the study if they had missing information on age, sex, or IMD quintile, due to potential poor data quality. Care

residents (using a care home flag within TPP) were also excluded as this setting is not representative of the epidemiological burden in the general population. Infants were excluded if they were eligible for Palivizumab due to high risk of complications due to RSV (16) (via the presence of ventilation codes for those under 12 months or those 12-23 months with cardiovascular-related diagnosis codes).

Respiratory Virus Phenotypes

We defined virus phenotypes which applied across all cohorts with specific criteria for infants and older adults. We developed the phenotypes in collaboration with medical professionals in respiratory and primary care. To classify and differentiate cases of each respiratory virus, we created two phenotypes: one specific which aimed to capture cases more likely to be attributable to specific viruses; and one more sensitive, aiming to capture as many relevant cases as possible. In addition, we separated cases into mild or severe, where mild was primary or emergency care (where relevant) and severe was in secondary care (Hospital Episode Statistics - HES). To enforce temporal ordering of case severity, classification as a mild case could not occur after classification as a severe case - with respect to each virus. We also created an overall respiratory virus definition, which aimed to capture the overall burden of the three respiratory viruses of interest, without distinguishing the pathogen driving it. All outcomes used a pre-specified episode length definition of 14 days.

For both mild and severe phenotypes, the specific phenotypes were a subset of the sensitive phenotypes for each virus and severity level (supplementary Table S1, Figure 1A). As such, all cases classified using specific phenotypes were also classified using the sensitive phenotypes.

For mild specific phenotypes, each virus included at least one pathogen-specific SNOMED CT code in the primary care record. Bronchiolitis was used as a virus-specific feature for RSV (17,18). We used emergency attendances for bronchiolitis (from ECDS) in infants as an age-specific feature, to capture mild RSV cases out-of-hours.

For mild sensitive phenotypes, all phenotypes: (i) met the specific phenotype definition; (ii) had two pathogen-associated clinical event SNOMED-CT codes, e.g. respiratory symptoms within an episode (19-22); (iii) or one prescription code (in dm+d) within the same episode as one pathogen-associated clinical event code. To improve the accuracy of sensitive phenotypes, we excluded individuals with codes indicative of non-target infection a month before or after a relevant event or prescription (supplementary Table S1). Age-specific features included exacerbations of COPD (23) and asthma (24) for overall respiratory virus cases in older adults. Virus-specific features for influenza were: (i) the inclusion of codes indicating the 'suspicion' of influenza; or (ii) the presence of influenza-like-illness (ILI), adapted from the WHO definition (25), which required the presence of at least one code indicating an acute respiratory illness (ARI) and at least one code indicating a fever, occurring within a 14-day period. We compared our definition of ILI (see supplementary Table S1) to the definition developed, and validated, by the Royal College of General Practitioners (RCGP) (26). As both phenotypes were developed from the WHO definition of ILI (25), they were closely aligned. Minor differences were present, such as the RCGP definition using a 10-day window for ascertainment (versus our 14 days) and our exclusion of a code for `cough` in our definition. We excluded this requirement due to clinician input which suggested that clinical coding of `cough` would be

minimal, leading to underascertainment of ILI cases. However, SNOMED-CT usage suggests that clinical coding for 'cough' symptoms may be more common than expected (27).

For severe specific phenotypes, all viruses were defined as requiring a pathogen-specific ICD-10 code in first or second position of the hospital record (APCS). For severe sensitive phenotypes, common features across viruses included: (i) meeting the specific phenotype definition; or (ii) the presence of a relevant pathogen-associated ICD-10 code in first or second position. To improve the accuracy of the sensitive phenotypes, we excluded individuals with codes indicative of non-target infection within a month prior to or proceeding the case (supplementary Table S1). Virus-specific features included the use of COPD (23) or asthma (24) exacerbations as a proxy for severe overall respiratory virus cases.

To define the phenotypes above, we used a total of 37 codelists (20 in SNOMED CT format, 14 in ICD-10 format, and 3 in dm+d format). Eight of these codelists comprised single SNOMED CT or ICD-10 codes. Six were existing codelists available on OpenCodelists (28). Two codelists were adapted by converting codelists used in CPRD (29) and supplementing the conversion by re-running the searches used to build the original codelist. Another two codelists were developed using validated search terms (30), with codes then reviewed by a relevant clinician. The remaining 23 codelists were newly developed with clinician input. This involved working collaboratively with clinicians to develop a list of relevant search terms, and verifying the subsequent codelist. Codelists are available for re-use on opencodelists.org, with corresponding search terms, for details on codelists used see the relevant references included within supplementary Table S1.

Reinfections

We explored decisions made when designing the phenotypes, such as episode length and reinfections, using the specific phenotypes. We quantified the proportion of cases who went on to be classified a second time with the same case-type within the same season. We then quantified the proportion of those reinfections occurring within 28 days over time and by severity. A high proportion of reinfections within 28 days would suggest that the defined episode length of 14 days may be too short.

Case Classification Types

In order to evaluate whether the phenotypes were achieving the expected differences in sensitivity we compared phenotypes to one another by calculating the change in the number of classified cases using specific versus sensitive phenotypes (Figure 1A). We also quantified the change in overlap in case classification – cases simultaneously attributable to multiple viruses and therefore more prone to case misclassification (Figure 1B). In addition, we reviewed the proportion of cases classified as overall respiratory virus that were also classified as either RSV, influenza, or COVID-19. Below we highlight results for the groups with the highest burden of disease in the main text (older adults and infants) for annual cohorts of highest burden (2017-18, 2018-19, 2020-21) and the most recent year (2023-24) (Figure 1C).

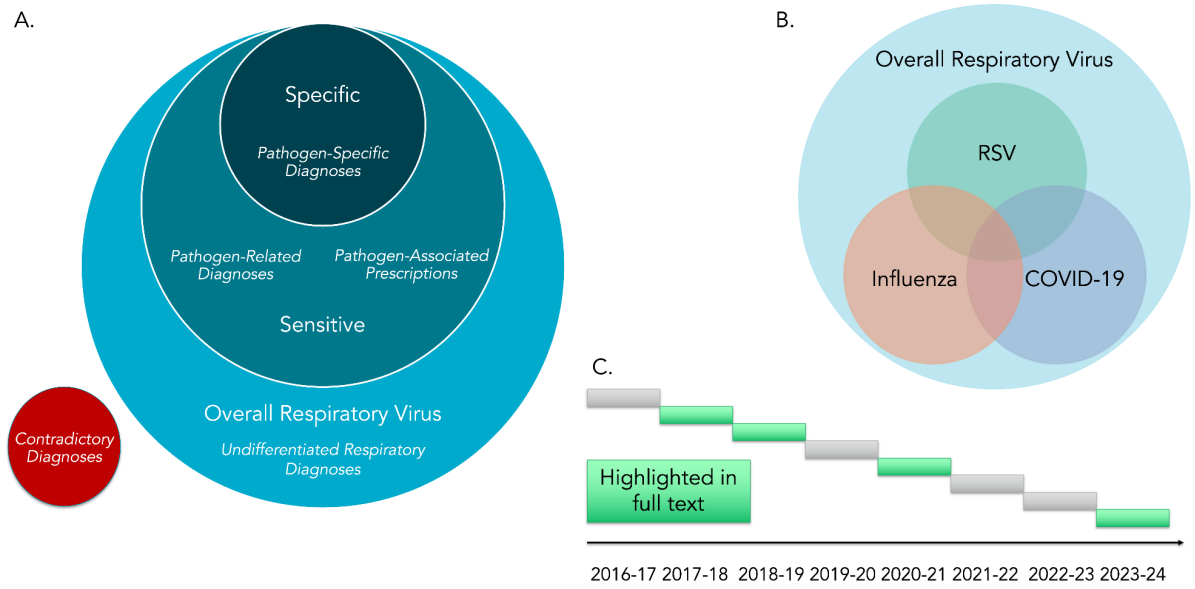

Figure 1: Summary of analysis structure. A: Schematic of relationship between phenotypes, showing how the specific phenotype is a subset of the sensitive and so forth. Contradictory diagnoses (red) only apply to sensitive phenotypes. B: Cases can be assigned to more than one virus, which we term "overlap". All virus-specific outcomes are captured by the overall phenotype (light blue). C: Each of 8 annual cohorts runs from 1st September, and the annual cohorts shown in the main text are highlighted in green.

Comparison to Surveillance Data

We compared case counts for each of our EHR phenotypes to time-series surveillance data from the WHO RespiMart (31). Data submissions to RespiMart come from national reporting systems, in this case the UKHSA, and contain information on both sentinel and non-sentinel surveillance. Countries submit laboratory-confirmed testing volumes and will be included for each virus respectively if at least one specimen per week is tested. The time-series surveillance data were provided as weekly case counts, which we aggregated to monthly counts. We also explored how our estimates compared to UKHSA data from 2016-2020 (32), through the same descriptive comparisons. UKHSA data were not available from 2020 due to changes in reporting procedures, after which only test positivity was publicly available.

We visually compared surveillance data with phenotyped cases over time to explore the alignment of timing and scale of cases classified as RSV, influenza, and COVID-19 when contrasted with testing data. We then used Pearson's correlation coefficient (33) to ascertain levels of correlation between monthly case counts obtained by phenotyped cases and surveillance data. Correlations were calculated using data for all months (combining across the eight seasons) and for each season separately.

Results

Study Population

Each annual cohort included between 21-25 million individuals (supplementary Table S1 and Figure S1), of which 4-5 million were older adults (65+), 12-15 million adults (18-64), 4-5 million children and young people (2-17), and 600-700 thousand infants (<2). Of the infants, between 95-150 thousand infants each year were linked to their birthing parent.

Baseline characteristics for participants remained broadly stable each year (supplementary Tables S2-6), although population-level changes over time led to fluctuations in the distributions of some characteristics such as Index of Multiple Deprivation (IMD) and age (34–36). There were some changes in ethnicity profile over time reflecting changes in the recording of ethnicity data in EHRs (37,38).

Respiratory Virus Phenotypes

Older Adult Cohort

In older adults, when using the mild specific phenotypes, the number of observed cases was highest for influenza until 2020, after which COVID-19 cases dominated (Figure 2A). Around 3,000-4,000 cases of mild influenza were seen per season pre-COVID-19, while over 100,000 cases of COVID-19 were seen in 2020-21. There were much lower numbers of RSV cases (100-1,000 per season), with a larger number in the most recent season (2023-24). For the specific phenotype, overlap of virus classification was minimal for mild cases - at most 1% (supplementary Figure S2).

We found much higher numbers of potential RSV cases when using the mild sensitive phenotypes (Figure 2B), ranging from 100-500 fold higher than the specific phenotype. This meant that depending on phenotype chosen, the "dominant" winter virus in England was different. When compared to the specific phenotypes, the number of cases classified as influenza and COVID-19 by the sensitive phenotypes was slightly higher (~1.5-2 fold higher for influenza and ~2-10 fold higher for COVID-19). This increase in cases classified as RSV and COVID-19 was likely driven by the inclusion of a number of symptom-based codes which were not included for influenza. The level of overlap in case classification was also higher when using sensitive compared with specific phenotypes (between ~5-10 fold pre-COVID-19). Subsequently, we found many cases of overlap between RSV and COVID-19 cases - with over 50% of cases being classified simultaneously as RSV and COVID-19. Compared with the specific phenotype which had no more than 1% overlap across all identified cases, the sensitive phenotype had up to 27% overlap (supplementary Figure S2).

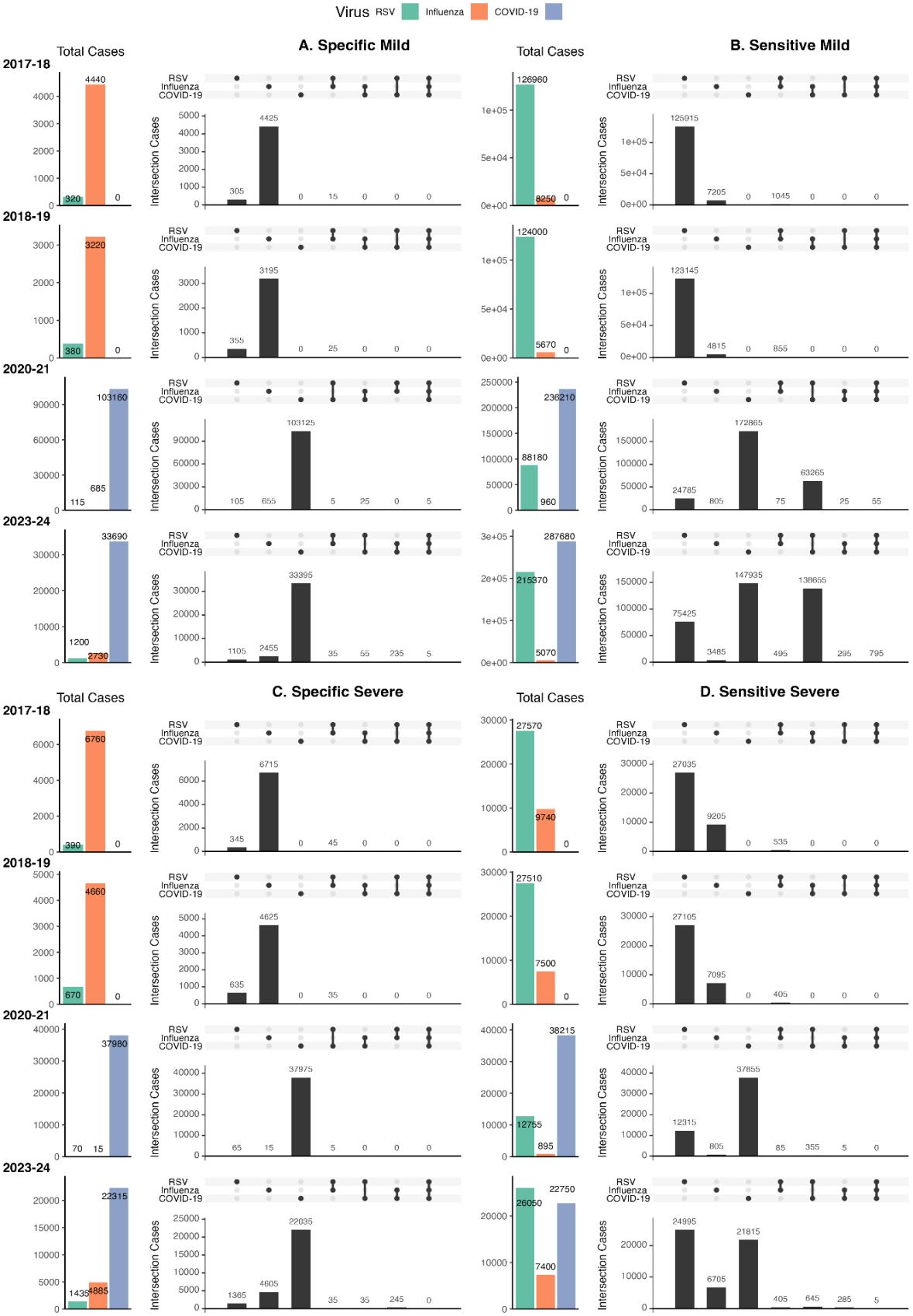

Figure 2: Classification of respiratory viruses in older adults. We compared specific phenotypes (A, C) and sensitive phenotypes (B, D), where cases included RSV (green), influenza (orange) or COVID-19 (purple). Mild cases (A, B) and severe cases (C, D) were classified independently.

For severe cases in older adults using the specific phenotypes, we found the same dominant viruses as with mild cases - influenza dominated until COVID-19 emerged and became the dominant virus (Figure 2C). Except for 2020-21, ~5,000 severe cases were classified as influenza each season - whereas there were between ~20,000 (2023-24) and ~40,000 (2020-21) severe cases classified as COVID-19 in OpenSAFELY. As observed for mild RSV, severe RSV was detected at a lower rate - however, this number increased over time (~350 in 2017-18 vs ~1400 in 2023-24). Overlap in the classification of severe cases was rare when using the specific phenotypes (~0.5%), this increased in 2023-24 to ~1% overlap in case classification (supplementary Figure S2).

Severe sensitive phenotypes classified similar numbers of cases as COVID-19 as the specific (~500 additional cases) and a moderate number of additional severe cases as influenza (~3,000 outside of 2020-21) (Figure 2D). However, the sensitive phenotype classified a much larger number of severe cases as RSV (between ~20-100 fold increase). Notably, the key change between specific and sensitive phenotypes for severe RSV was the addition of an ICD-10 diagnosis code for unspecified acute lower respiratory infection in the first or second position in HES, suggesting that this code accounted for the additional cases classified as RSV when using the sensitive phenotype. Overlapping cases made up ~1-2.5% of all cases when using the severe sensitive phenotypes (Figure 2D, supplementary Figure S2). In most seasons we found ~500 cases of RSV being additionally classified as either influenza or COVID-19.

We found consistent results in all seasons (2016-24) for both mild (supplementary Figure S3) and severe (supplementary Figure S4) cases.

Adult and Children and Young People Cohorts

For adults, the introduction of COVID-19 increased the level of overlap in case classification (supplementary Figures S5). Overlap was increased by broader definitions for mild cases (supplementary Figure S6) much more than for severe (supplementary Figure S7). The impact of phenotype on overlap in case classification in children and young people (supplementary Figure S8) was very similar. Broader definitions led to higher levels of outcome overlap in mild cases (supplementary Figure S9), when compared to severe cases (supplementary Figure S10).

Infant Cohorts

When using the mild specific phenotype in infants, we found most cases were RSV (Figure 3A); between ~15,000-45,000 mild cases each season. Each year, we found 500-2,000 influenza cases, except for 2020-21, and ~2,000-10,000 COVID-19 cases. RSV remained dominant even during 2020-21 when COVID-19 cases were also very common. Across all seasons, some overlap was present for mild RSV, with cases additionally classified either as influenza (pre-COVID-19: ~400-800; post-COVID-19: ~1200) or COVID-19 (~1,500-6,500). This overlap was likely due to the inclusion of bronchiolitis, a condition rare outside of infant populations often used as a sentinel diagnosis for RSV.

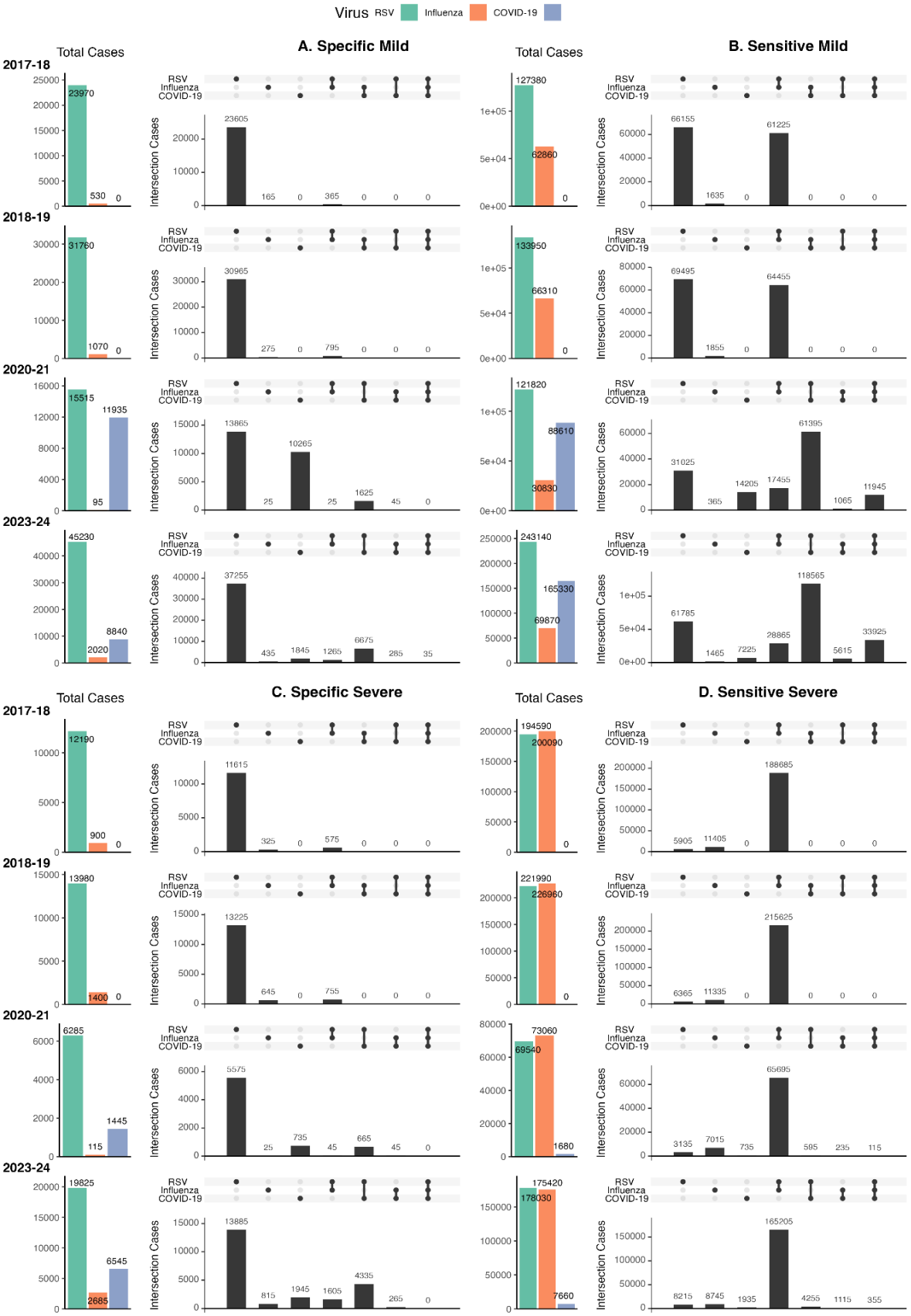

Figure 3: Classification of respiratory viruses in infants. We compared specific phenotypes (A, C) and sensitive phenotypes (B, D), where cases included RSV (green), influenza (orange) or COVID-19 (purple). Mild cases (A, B) and severe cases (C, D) were classified independently.

The mild sensitive phenotype in infants resulted in much higher numbers of cases (Figure 3B): RSV was ~5-10 fold higher, influenza was ~60-120 fold, and COVID-19 was ~8-20 fold higher. The latter two corresponded to a large number which were classified simultaneously as another virus: outcome overlap for influenza (pre-COVID-19: ~10-20 fold; post-COVID-19: ~20-600 fold) and COVID-19 (~20-60 fold). Additionally, ~30% of cases were classified as more than one virus (Figure 3, supplementary Figure S11), with a large number of mild cases being classified as all three viruses (between ~10,000-30,000).

RSV was the dominant virus when using the severe specific phenotypes, resulting in ~5,000-20,000 cases per season (Figure 3C). Influenza cases increased from ~1000 in 2017-18 to ~2,500 in 2023-24, with an exception in 2020-21. Severe COVID-19 cases were more common than influenza (~1,400 and 4,000 more cases of COVID-19 in 2020-21 and 2023-24 respectively) but still much less common than RSV. For most seasons, the degree to which cases overlapped was minimal (<1,000 cases - with cases often classified as RSV). However, the most recent season (2023-24) shows a larger number of cases, with many being classified as multiple viruses simultaneously - ~6,000 cases of influenza and/or COVID-19 overlapped with RSV cases.

Using severe sensitive phenotypes meant COVID-19 and influenza were equally dominant (Figure 3D). The number of COVID-19 cases changed very little between specific and sensitive severe phenotypes. The change in dominance was also reflected in a change in case overlap (~100-fold increase), with most cases in each season being classified as both RSV and influenza. This change, which was not seen in older adults (Figure 2), indicates a difference in how clinicians coded episodes, with pathogen-associated codes more likely in infant records. The change in overlap in severe cases was also notable, <1% vs almost 50% cases for specific vs sensitive phenotypes (supplementary Figure S11).

The differences between specific and sensitive phenotypes for mild (supplementary Figure S12) and severe (supplementary Figure S13) cases was consistent across seasons. Similarly, the overlap in case classification was consistent in infants linked to their birthing parent (supplementary Figure S14). Mild cases (supplementary Figure S15) and severe cases (supplementary Figure S16) were also impacted in the same manner as with the full infant cohort - suggesting that the clinical coding patterns in the records of this subgroup were not different.

Reinfections

Reinfections occurred at different levels depending on the virus and case severity, but rarely occurred in more than 15% of individuals (supplementary Figure S17). Phenotype sensitivity impacted the proportion of individuals being reinfected, where more sensitive phenotypes had greater reinfection counts.

Due to low counts, it was difficult to explore the proportion of reinfections that occurred within 28 days for RSV and influenza. For COVID-19, specific phenotypes consistently highlighted higher proportions of reinfections within 28 days when compared to sensitive phenotypes (supplementary Figure S18).

Case Classification Types

The percentage of overall respiratory virus cases that were classified as RSV, influenza, or COVID-19 increased over time. Prior to 2020-21, for mild cases in older adults (supplementary Figure S19), adults (supplementary Figure S20), and children and young people (supplementary Figure S21), roughly 30% of overall respiratory virus cases were classified as either RSV, influenza, or COVID-19; this rose to ~50% of cases from 2020-21 onwards. For severe cases, in these age cohorts, the percentage of overall respiratory virus cases classified as RSV, influenza, or COVID-19 remained stable at ~40%. For infants (supplementary Figure S22) and maternally linked infants (supplementary Figure S23) the percentage of mild overall cases being classified as RSV, influenza, or COVID-19 was around 10-20%, which steadily rose to ~30% from 2021-22 onwards. For severe cases in infants (irrespective of maternal linkage status), over 50% of overall respiratory cases were consistently classified as RSV, influenza, or COVID-19.

Comparison to Surveillance Data

Surveillance sources varied slightly in the number of cases reported, and years with data available (supplementary Figure S24).

Descriptive Comparisons

The timing of peaks in surveillance data was reflected in cases from both specific and sensitive phenotypes (Figure 4 and supplementary Figure S25). The relative scale of peaks varied across viruses and phenotypes (supplementary Figure S26); which is expected as OpenSAFELY covers 42% of the population. For RSV, phenotyped cases exceeded those recorded in surveillance data (often over 30 fold). Initially, similar numbers of cases were recorded in specific influenza phenotypes and surveillance data. However, in more recent years EHR phenotypes captured fewer influenza cases; roughly half of cases reported in surveillance data - in line with the population proportion in OpenSAFELY. The sensitive phenotype led to ~5-10-fold higher influenza cases than in surveillance data. Between 2020-22, surveillance data reported very few cases of RSV and influenza while the EHR phenotypes recorded cases - irrespective of severity and phenotype (Figure 4). This corresponds to a period in which respiratory virus presentations were likely to be tested for COVID-19 but may not have been coded as such by clinicians. Mild cases of COVID-19 reported via OpenSAFELY aligned closely, in the timing of peaks, with surveillance data; although 4 times more cases were recorded in surveillance data (supplementary Figure S21). Some peaks in phenotyped severe COVID-19 cases were not shown in surveillance data, likely due to an ICD-10 code used for COVID-19 (U07.2) which was informed by clinical diagnosis without test confirmation.

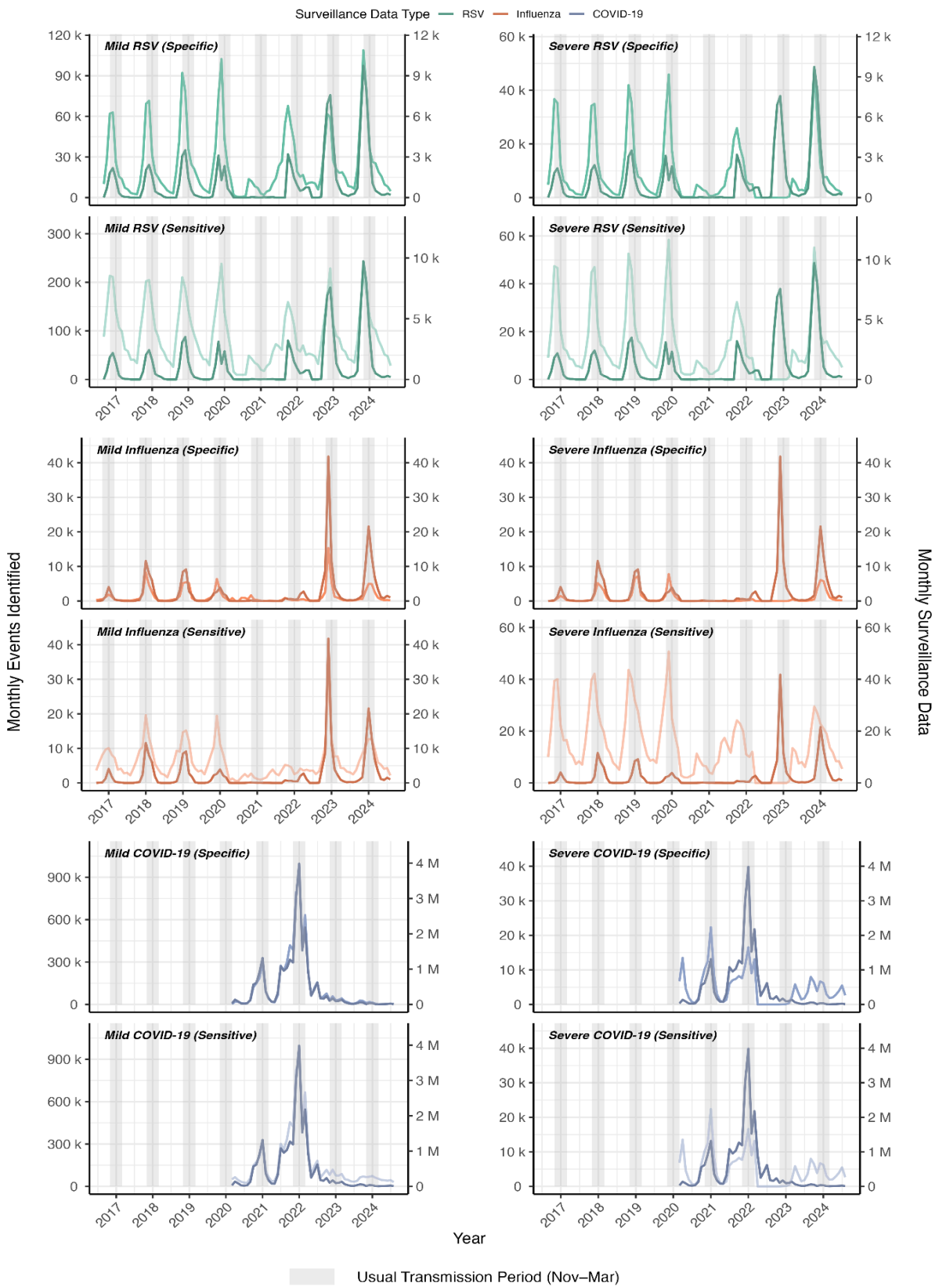

Figure 4: Comparison of phenotyped cases in OpenSAFELY (left axis) and surveillance data (right axis). WHO RespiMart surveillance data for RSV and influenza were for England and for COVID-19 were for the UK. Results are shown separately for mild (left panels) and severe cases (right panels), with specific cases appearing in upper panels and sensitive cases in lower panels for each virus - where RSV is green, influenza is orange, and COVID-19 is purple. Surveillance data are indicated by a darker shade, with specific cases being lighter and sensitive cases being lightest.

Quantitative Comparisons

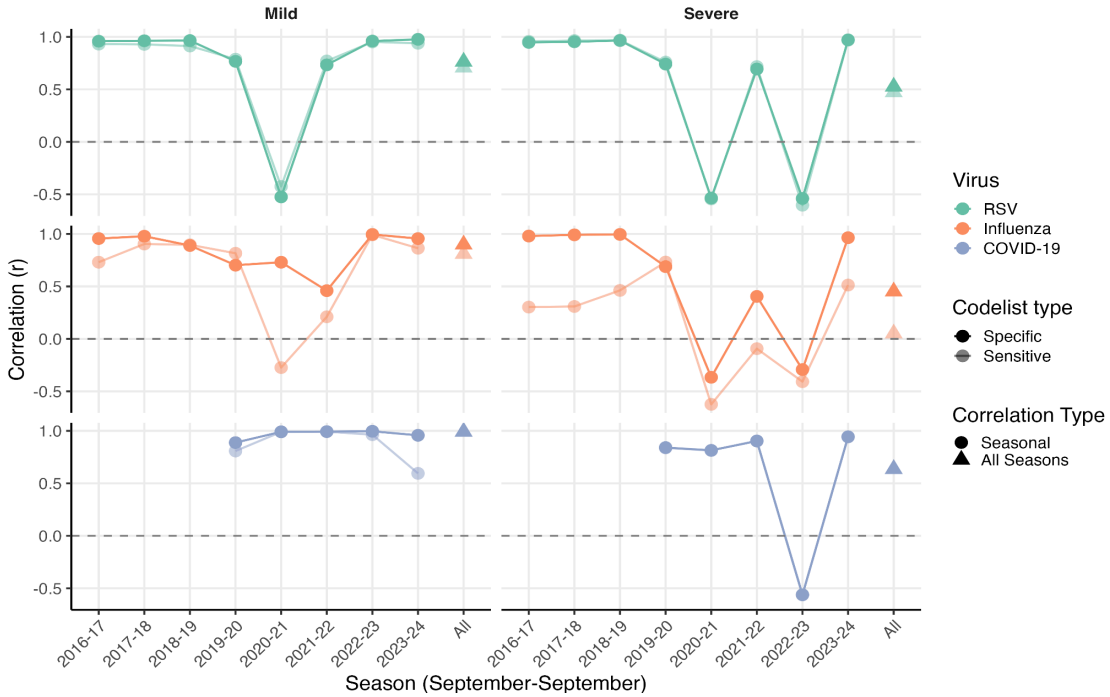

Figure 5: Pearson's correlation coefficient for each virus, severity, and phenotype combination. Tests were performed individually for each season. The colour of the line indicates the virus (RSV in green, influenza in orange, and COVID-19 in purple); the darkness of the shade indicates phenotype, where lighter shades are sensitive phenotypes; the shape of the point indicates whether the correlation is on a seasonal level or for all seasons (seasonal indicated by a filled circle and all seasons indicated by a filled triangle).

We found high temporal correlation between surveillance data and cases recorded in OpenSAFELY (Figure 5). The degree of correlation varied both overall (supplementary Table S7) and annual cohort (supplementary Figure S27). The years affected by lockdown and unusual viral circulation patterns had much lower correlation for RSV and influenza. During this period surveillance data was largely unavailable while low numbers of cases were still being recorded in EHRs - as RSV and influenza were circulating at much lower levels. Low counts meant that correlations were hard to accurately calculate. In the same period, testing for COVID-19 was prevalent and reflected milder cases to a higher degree than severe. As a result, correlation between surveillance data for severe COVID-19 was consistently lower than for mild COVID-19. When freely available testing ceased, correlations for mild and severe COVID-19 became more similar. For RSV and COVID-19, correlations did not vary much across phenotypes; however, for influenza, sensitive phenotypes were less correlated with surveillance data than specific phenotypes; particularly for severe cases (0.75 vs 0.9).

Discussion

We developed phenotypes to classify and differentiate cases of RSV, influenza, and COVID-19 from routinely-collected health data in England, and compared these with national surveillance data. Prior work has validated phenotypes for acute respiratory infections for surveillance and early warning systems (26), and here we focus on differentiating the key viruses which contribute towards acute respiratory infections. Our study has detailed methods for classifying

cases for three key respiratory viruses in healthcare data across primary and secondary care, without the use of testing information. These methods could be valuable for quantifying the burden of these viruses experienced in primary care, and will deepen understanding of how these common respiratory viruses impact patients and GP practices. This is crucial for understanding the healthcare burden and patterns of primary and secondary care utilisation of these major respiratory viruses, for which vaccines are in use.

In older adults, the specific phenotypes classified similar numbers of mild and severe cases for both RSV and influenza, and sometimes severe cases exceeded mild cases. This may be due to extremely limited testing in primary care. Conversely, mild COVID-19 cases far exceeded severe cases because testing was widely available outside of hospital settings during 2020-22. Pathogen-specific ICD-10 coding may underestimate case burden (39), whereas pathogen-associated ICD-10 codes may lead to case misclassification. While overlap in severe case classification was rare, even when using the sensitive phenotypes, this does not mean that misclassification was not present. In mild cases, sensitive phenotypes led to increased overlap in classification of COVID-19 and RSV cases. However, surveillance data suggests that these were COVID-19 cases being misclassified as RSV by our sensitive phenotypes. This highlights how difficult it is to distinguish viruses leading to respiratory infection in EHR data and how pandemic shocks can disrupt these surveillance efforts.

Comparisons in older adults of the specific phenotype and the broader case definitions of the sensitive phenotype changed which virus appeared dominant (influenza vs RSV). While there is existing evidence that RSV burden may be underestimated (40), the large discrepancy between RSV and influenza burden shown by sensitive phenotypes suggests poor influenza ascertainment (4). In mild cases this was likely driven by symptom-based codes in the sensitive RSV phenotype. In severe cases this was likely from ICD-10 code J22 (unspecified acute lower respiratory infection) which was used only for RSV as informed by previous research (41).

In infants, we found low numbers of influenza cases using specific phenotypes, suggesting these cases were difficult to ascertain in primary care; however, the sensitive phenotype had higher overlap, suggesting broader case definitions made influenza hard to distinguish from other viruses. Sensitive phenotypes also resulted in more cases for RSV and COVID-19; however, as with influenza, this often corresponded to greater overlap in classification compared with the specific phenotypes. This is likely due to bronchiolitis codes, which, while a very common presentation for RSV (17,18), could also be listed in cases with influenza or COVID-19 codes leading to simultaneous classification with RSV.

Clinical coding approaches appeared to differ by age cohort. In infants, overlap in case classification suggested that pathogen-associated codes were more prevalent in both primary and secondary care records. In secondary care, younger individuals may be more likely to have pathogen-associated codes such as bronchiolitis or upper respiratory tract infection (42). By contrast, for hospitalised older adults, pathogen-associated codes may be less prevalent - potentially due to higher levels of hospital testing in this population. Usage of pathogen-associated codes in secondary care may reduce as age increases, with older adults often having 'pneumonia and influenza' codes (42).

When evaluating phenotypes we found that reinfections were relatively rare across the population. A high proportion of COVID-19 reinfections occurred within 28 days, suggesting that the episode length definition (14 days) might have been too short. We also found that a high proportion of overall respiratory virus cases were not classified as RSV, influenza, or COVID-19. This was indicative of how common undifferentiated respiratory codes were, which is in line with both historic and current findings on RSV, influenza, and COVID-19 prevalence. A previous study found that ~30% of NHS Direct respiratory calls could be attributed to RSV or influenza, with variation by age (43). Additionally, recent virological data shows that RSV, influenza, and COVID-19 accounted for less than half of respiratory virus samples, with proportions differing by age (44).

We found the timing and size of epidemics changed over time in both surveillance data and in EHRs. In addition to differences in epidemic sizes during this time, this change may also be in part due to changes in clinical coding behaviours, surveillance data reporting, or testing practices. Testing, for example, was limited to hospital settings for RSV and influenza until recently. However, for COVID-19, NHS testing outside of hospital was common during 2020-2022. Therefore, specific phenotypes yielded fewer mild cases of RSV and influenza but for COVID-19 may have ascertained cases which did not present to health services - e.g. very mild or asymptomatic cases.

The phenotypes outlined here could be expanded to include other key respiratory viruses such as adenovirus or rhinovirus, and future development and strengthening of the findings could include securing linkage of testing data, including vaccine history of patients, and the potential for analysis of free text in the EHR, which is not available at this time. Downstream analyses using these phenotypes also have the potential to investigate patient-level characteristics associated with clinical coding specificity, including understanding of the major health inequalities in burden of these viruses in England.

Strengths and Limitations

We leveraged very large scale EHR data through OpenSAFELY with detailed coding on clinical events, and medications prescribed within primary care, plus linked secondary care including emergency attendance, which gave high granularity for designing phenotypes. We did not have access to free text fields, testing data, dispensation of prescriptions, or medications administered within secondary care. Free text constitutes up to 80% of data within health records (45), such that case ascertainment may underestimate true burden.

While designing EHR phenotypes we identified difficulties associated with the absence of linked testing data alongside uncertainty surrounding information flow between primary and secondary care. Notably, we observed a high proportion of undifferentiated respiratory viruses across age groups in both primary and secondary care settings, highlighting the potential for more systematic pathogen-specific coding in settings where testing data are available. By enforcing temporal progression of case severity from mild to severe we also ensured that those who presented only to secondary care were not classified in primary care due to communication across systems. Another difficulty was distinguishing symptom types within a limited number of codelists (e.g. one codelist for possible RSV symptoms); in order to enforce

the presence of two distinct symptoms in broader phenotypes, symptom-based codelists would need to be used. However, this adds further layers of computational complexity.

Conclusion

EHR data phenotypes can be used to classify mild and severe cases of three key respiratory viruses. This can provide a detailed picture of respiratory virus burden within primary and secondary care, particularly for vaccine-preventable respiratory diseases. For younger patient populations, broader case definitions had higher chances of misclassification than specific phenotypes. The relative scale and timing of peaks in EHR-recorded cases were highly correlated with surveillance data for all three viruses. The phenotypes outlined here could be a powerful adjunct to existing surveillance efforts, supporting further epidemiological evaluation of respiratory viruses.

References

1. Taylor S, Taylor RJ, Lustig RL, Schuck-Paim C, Haguinet F, Webb DJ, et al. Modelling estimates of the burden of respiratory syncytial virus infection in children in the UK. BMJ Open. 2016 June 1;6(6):e009337.
2. Reeves RM, Hardelid P, Panagiotopoulos N, Minaji M, Warburton F, Pebody R. Burden of hospital admissions caused by respiratory syncytial virus (RSV) in infants in England: A data linkage modelling study. J Infect. 2019 June 1;78(6):468–75.
3. Fleming DM, Taylor RJ, Lustig RL, Schuck-Paim C, Haguinet F, Webb DJ, et al. Modelling estimates of the burden of Respiratory Syncytial virus infection in adults and the elderly in the United Kingdom. BMC Infect Dis. 2015 Oct 23;15:443.
4. FLEMING DM, TAYLOR RJ, HAGUINET F, SCHUCK-PAIM C, LOGIE J, WEBB DJ, et al. Influenza-attributable burden in United Kingdom primary care. Epidemiol Infect. 2016 Feb;144(3):537-47.
5. Cromer D, van Hoek AJ, Jit M, Edmunds WJ, Fleming D, Miller E. The burden of influenza in England by age and clinical risk group: A statistical analysis to inform vaccine policy. J Infect. 2014 Apr 1;68(4):363-71.
6. Covid inquiry: The UK pandemic in numbers. BBC News [Internet]. 2020 Mar 6 [cited 2025 Jan 6]; Available from: https://www.bbc.com/news/uk-51768274
7. Coronavirus (COVID-19) latest insights - Office for National Statistics [Internet]. [cited 2025 Jan 6]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19latestinsights/hospitals
8. GOV.UK [Internet]. [cited 2025 Jan 6]. Surveillance of influenza and other seasonal respiratory viruses in the UK, winter 2022 to 2023. Available from: https://www.gov.uk/government/statistics/annual-flu-reports/surveillance-of-influenza-and-other-seasonal-respiratory-viruses-in-the-uk-winter-2022-to-2023
9. What Are Codelists and How Are They Constructed? | Bennett Institute for Applied Data Science [Internet]. 2023 [cited 2025 Jan 6]. Available from: https://www.bennett.ox.ac.uk/blog/2023/09/what-are-codelists-and-how-are-they-constructed/
10. CDC. Classification of Diseases, Functioning, and Disability. 2024 [cited 2025 Jan 6]. ICD-10. Available from: https://www.cdc.gov/nchs/icd/icd-10/index.html

11. Vuokko R, Vakkuri A, Palojoki S. Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review. JMIR Med Inform. 2023 Feb 6;11(1):e43750.
12. COVID-19 | UKHSA data dashboard [Internet]. [cited 2026 Feb 3]. Available from: https://ukhsa-dashboard.data.gov.uk/respiratory-viruses/covid-19
13. Definitions [Internet]. Rethinking Clinical Trials. [cited 2025 Jan 6]. Available from: https://rethinkingclinicaltrials.org/chapters/conduct/electronic-health-records-based-phenotyping/definitions/
14. Wellcome Open Research | Open Access Publishing Platform [Internet]. 2022 [cited 2026 Feb 19]. OpenSAFELY: Representativeness of electronic health record platform ... Available from: https://wellcomeopenresearch.org/articles/7-191
15. Emergency attendances - OpenSAFELY documentation [Internet]. [cited 2026 Feb 2]. Available from: https://docs.opensafely.org/data-sources/ecds/#metadata
16. Respiratory syncytial virus | Treatment summaries | BNF content published by NICE [Internet]. [cited 2023 Nov 20]. Available from: https://bnf.nice.org.uk/treatment-summaries/respiratory-syncytial-virus/
17. nhs.uk [Internet]. 2017 [cited 2024 Feb 5]. Bronchiolitis. Available from: https://www.nhs.uk/conditions/bronchiolitis/
18. Rybak A, Cohen R, Kramer R, Béchet S, Delobbe JF, Dagrenat V, et al. Respiratory Syncytial Virus in Outpatient Children with Bronchiolitis: Continuous Virus Circulation During the Nonepidemic Period. Pediatr Infect Dis J. 2023 Dec;42(12):e488–90.
19. Wishaupt JO, van der Ploeg T, de Groot R, Versteegh FGA, Hartwig NG. Single- and multiple viral respiratory infections in children: disease and management cannot be related to a specific pathogen. BMC Infect Dis. 2017 Jan 11;17(1):62.
20. Price O, Birrell FA, Mifsud EJ, Sullivan SG. Epidemiology of repeat influenza infection in Queensland, Australia, 2005–2017. Epidemiol Infect. 2022 Jan;150:e144.
21. Molecular Analysis of Respiratory Syncytial Virus Reinfections in Infants from Coastal Kenya | The Journal of Infectious Diseases | Oxford Academic [Internet]. [cited 2024 Jan 4]. Available from: https://academic.oup.com/jid/article/193/1/59/870060
22. Tang CY, Wang Y, McElroy JA, Li T, Hammer R, Ritter D, et al. Reinfection with two genetically distinct SARS-CoV-2 viruses within 19 days. J Med Virol. 2021;93(10):5700–3.
23. Rohde G, Wiethege A, Borg I, Kauth M, Bauer TT, Gillissen A, et al. Respiratory viruses in exacerbations of chronic obstructive pulmonary disease requiring hospitalisation: a case-control study. Thorax. 2003 Jan 1;58(1):37–42.
24. Hanania NA, King MJ, Braman SS, Saltoun C, Wise RA, Enright P, et al. Asthma in the elderly: Current understanding and future research needs—a report of a National Institute on Aging (NIA) workshop. J Allergy Clin Immunol. 2011 Sept 1;128(3):S4–24.
25. Fitzner J, Qasmieh S, Mounts AW, Alexander B, Besselaar T, Briand S, et al. Revision of clinical case definitions: influenza-like illness and severe acute respiratory infection. Bull World Health Organ. 2018 Feb 1;96(2):122–8.
26. Elson WH, Jamie G, Wimalaratna R, Forbes A, Leston M, Okusi C, et al. Validation of an acute respiratory infection phenotyping algorithm to support robust computerised medical record-based respiratory sentinel surveillance, England, 2023. Eurosurveillance. 2024 Aug 29;29(35):2300682.

27. Tamborska AA, Higgins R, Boukari Y, Kingsley V, Ojedele L, Oreagba K, et al. OpenCodeCounts: An open-access, interactive online tool and R package for analysing clinical code usage in England [Internet]. medRxiv; 2025 [cited 2026 Feb 25]. p. 2025.10.14.25338005. Available from: https://www.medrxiv.org/content/10.1101/2025.10.14.25338005v1 doi:10.1101/2025.10.14.25338005
28. OpenCodelists [Internet]. [cited 2026 Feb 2]. Available from: https://www.opencodelists.org/
29. Davidson JA, Banerjee A, Smeeth L, McDonald HI, Grint D, Herrett E, et al. Risk of acute respiratory infection and acute cardiovascular events following acute respiratory infection among adults with increased cardiovascular risk in England between 2008 and 2018: a retrospective, population-based cohort study. Lancet Digit Health. 2021 Dec 1;3(12):e773–83.
30. Graham S, Walker JL, Andrews N, Nitsch D, Parker EPK, McDonald H. Identifying markers of health-seeking behaviour and healthcare access in UK electronic health records. 2024 Sept 1 [cited 2025 Oct 24]; Available from: https://bmjopen.bmj.com/content/14/9/e081781
31. RespiMart [Internet]. [cited 2025 May 19]. Available from: https://www.who.int/tools/RespiMart
32. GOV.UK [Internet]. 2021 [cited 2025 Aug 14]. Search: Respiratory Lab Reports. Available from: https://www.gov.uk/search/all?keywords=respiratory+lab+reports
33. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg. 2018 May;126(5):1763.
34. Estimates of the population for the UK, England, Wales, Scotland, and Northern Ireland - Office for National Statistics [Internet]. [cited 2025 Nov 21]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland
35. Profile of the older population living in England and Wales in 2021 and changes since 2011 - Office for National Statistics [Internet]. [cited 2026 Jan 5]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/ageing/articles/profileoftheolderpopulationlivinginenglandandwalesin2021andchangessince2011/2023-04-03
36. Death registrations and populations by sex, age and IMD quintile, England and Wales: 2001 to 2020 - Office for National Statistics [Internet]. [cited 2026 Feb 3]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/1324deathregistrationsandpopulationsbysexageandimdquintileenglandandwales2001to2020
37. Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, vanStaa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. J Public Health. 2014 Dec 1;36(4):684–92.
38. Andrews CD, Mathur R, Massey J, Park R, Curtis HJ, Hopcroft L, et al. Consistency, completeness and external validity of ethnicity recording in NHS primary care records: a cohort study in 25 million patients' records at source using OpenSAFELY. BMC Med. 2024 July 10;22(1):288.

39. Cai W, Tolksdorf K, Hirve S, Schuler E, Zhang W, Haas W, et al. Evaluation of using ICD-10 code data for respiratory syncytial virus surveillance. Influenza Other Respir Viruses. 2020;14(6):630–7.
40. Ramirez J, Carrico R, Wilde A, Junkins A, Furmanek S, Chandler T, et al. Diagnosis of Respiratory Syncytial Virus in Adults Substantially Increases When Adding Sputum, Saliva, and Serology Testing to Nasopharyngeal Swab RT–PCR. Infect Dis Ther. 2023 June 1;12(6):1593–603.
41. Agyei-Manu E, Atkins N, Nundy M, St-Jean C, Gornall-Wick A, Birley E, et al. Characteristics of influenza, SARS-CoV-2, and RSV surveillance systems that utilise ICD-coded data: a systematic review. J Glob Health. 15:04177.
42. Usher Institute [Internet]. 2024 [cited 2026 Jan 20]. Deliverables | Respiratory Syncytial Virus Research. D1.6 RSV healthcare burden in young children and the elderly in 6 European countries before and since the emergence of the COVID-19 pandemic. Available from: https://usher.ed.ac.uk/respiratory-syncytial-virus-research/promise/deliverables#uoe_featurebox_e6e4d63f274f14ce0e7d2929eaf48eda9
43. Cooper DL, Smith GE, Edmunds WJ, Joseph C, Gerard E, George RC. The contribution of respiratory pathogens to the seasonality of NHS Direct calls. J Infect. 2007 Sep;55(3):240–8.
44. GOV.UK [Internet]. [cited 2026 Feb 18]. National flu and COVID-19 surveillance report: 12 February 2026 (week 7). Available from: https://www.gov.uk/government/statistics/national-flu-and-covid-19-surveillance-reports-2025-to-2026-season/national-flu-and-covid-19-surveillance-report-12-february-2026-week-7
45. The untapped potential of clinical free-text: understanding privacy risk in electronic health records | Dataloch [Internet]. [cited 2026 Feb 25]. Available from: https://dataloch.org/our-impact/case-studies/untapped-potential-clinical-free-text-understanding-privacy-risk-electronic

Data Availability

All data were linked, stored and analysed securely using the OpenSAFELY platform, https://www.opensafely.org/, as part of the NHS England OpenSAFELY COVID-19 service. Data include pseudonymised data such as coded diagnoses, medications and physiological parameters. No free text data are included. No GP data from patients who have registered a Type-1 Opt out with their GP surgery were included in this study. Detailed pseudonymised patient data is potentially re-identifiable and therefore not shared. Primary care records managed by the GP software provider, TPP, were linked to hospital episode statistics (HES) and ONS death data through OpenSAFELY.

NHS England is the data controller of the NHS England OpenSAFELY COVID-19 Service; TPP is the data processor; all study authors using OpenSAFELY have the approval of NHS England (1). This implementation of OpenSAFELY is hosted within the TPP environment which is accredited to the ISO 27001 information security standard and is NHS IG Toolkit compliant (2).

Patient data has been pseudonymised for analysis and linkage using industry standard cryptographic hashing techniques; all pseudonymised datasets transmitted for linkage onto OpenSAFELY are encrypted; access to the NHS England OpenSAFELY COVID-19 service is via a virtual private network (VPN) connection; the researchers hold contracts with NHS England and only access the platform to initiate database queries and statistical models; all database activity is logged; only aggregate statistical outputs leave the platform environment following best practice for anonymisation of results such as statistical disclosure control for low cell counts (3).

The service adheres to the obligations of the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018. The service previously operated under notices initially issued in February 2020 by the the Secretary of State under Regulation 3(4) of the Health Service (Control of Patient Information) Regulations 2002 (COPI Regulations), which required organisations to process confidential patient information for COVID-19 purposes; this set aside the requirement for patient consent (4). As of 1 July 2023, the Secretary of State has requested that NHS England continue to operate the Service under the COVID-19 Directions 2020 (5). In some cases of data sharing, the common law duty of confidence is met using, for example, patient consent or support from the Health Research Authority Confidentiality Advisory Group (6).

Taken together, these provide the legal bases to link patient datasets using the service. GP practices, which provide access to the primary care data, are required to share relevant health information to support the public health response to the pandemic, and have been informed of how the service operates.

This study was approved by the HRA (IRAS ID: 340706) and LSHTM Ethics (29879).

1. The NHS England OpenSAFELY COVID-19 service - privacy notice. NHS Digital (Now NHS England). https://digital.nhs.uk/coronavirus/coronavirus-covid-19-response-information-governance-hub/the-nhs-england-opensafely-covid-19-service-privacy-notice (accessed 4 July 2023). ↩
2. Data Security and Protection Toolkit - NHS Digital. NHS Digital (Now NHS England). https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/data-security-and-protection-toolkit (accessed 4 July 2023) [archived here]. ↩
3. ISB1523: Anonymisation Standard for Publishing Health and Social Care Data. NHS Digital (Now NHS England). https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data (accessed 4 July 2023) [archived here]. ↩
4. Coronavirus (COVID-19): notice under regulation 3(4) of the Health Service (Control of Patient Information) Regulations 2002 – general. 2022. https://www.gov.uk/government/publications/coronavirus-covid-19-notification-of-data-controllers-to-share-information/coronavirus-covid-19-notice-under-regulation-34-of-the-health-service-control-of-patient-information-regulations-2002-general--2 (accessed 5 July 2023). ↩
5. Secretary of State for Health and Social Care - UK Government. COVID-19 Public Health Directions 2020: notification to NHS Digital. https://digital.nhs.uk/about-nhs-digital/corporate-information-and-documents/directions-and-data-provision-notices/secretary-of-state-directions/covid-19-public-health-directions-2020 (accessed 4 July 2023) [archived here]. ↩
6. Confidentiality Advisory Group. Health Research Authority. https://www.hra.nhs.uk/about-us/committees-and-services/confidentiality-advisory-group/ (accessed 4 July 2023) [archived here]. ↩

Code Availability

All code is shared openly for review and re-use under MIT open license in the following repository: https://github.com/opensafely/disparities-comparison.

Acknowledgements

We would like to thank the PPIE advisory group for their insightful contributions to this research. We are grateful to the named contributors, Yvonne Alder, Sarah Barley-McMullen, Janice Elliott, Michael Natt, Farheen Yameen, as well as to those members who have chosen not to be named, whose perspectives were integral to shaping this study. We would like to acknowledge the contribution of Sophie Graham and Francis Collette-White for providing support on the design of COVID-19 primary care codelists. We would also like to acknowledge the continual support from the tech-team at the Bennett Institute for Applied Data Science, as well as fellow researchers using OpenSAFELY for their insight and advice. We are very grateful for all the support received from the TPP Technical Operations team

throughout this work, and for generous assistance from the information governance and database teams at NHS England and the NHS England Transformation Directorate.

EP was funded by National Institute for Health and Care Research (NIHR); grant number: NIHR303287. RME was supported by the Medical Research Council (MR/X033260/1), and the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Health Analytics & Modelling, a partnership between the UK Health Security Agency, Imperial College London and LSHTM (grant code NIHR207404). CWG is supported by a Wellcome Career Development Award (225868/Z/22/Z). REC is funded by the Medical Research Council (MR/X033260/1), and the NIHR Health and Social Care Delivery Research programme (NIHR158218). EPKP is funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation (NIHR207408), a partnership between UK Health Security Agency and the London School of Hygiene and Tropical Medicine. EPKP has also received funding as a consultant for Tulane University as part of the Safe in Pregnancy and Childhood Study, supported by the Safety Platform for Emergency Vaccines, Task Force for Global Health, and The Coalition for Epidemic Preparedness Innovations. EPKP received reimbursements for providing input on an expert report for the UK COVID-19 Inquiry. WJH is funded by grants from the Wellcome Trust (311535/Z/24/Z) and NIHR (NIHR209347, NIHR206900). He also contributes to the NHS OpenSAFELY Data Analytics Service, funded by NHS England. Funders did not play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the author(s) and not necessarily those of the NIHR, UK Health Security Agency or the Department of Health and Social Care.

The OpenSAFELY platform is principally funded by grants from NHS England [2023-2025]; The Wellcome Trust (222097/Z/20/Z [2020-2024] and 311535/Z/24/Z [2025-2031]); The Medical Research Council (MRC) (MR/V015737/1 [2020-2021]). Additional contributions to OpenSAFELY have been funded by grants from: Medical Research Council (MRC) via the National Core Study programme Longitudinal Health and Wellbeing strand (MC_PC_20030, MC_PC_20059 [2020-2022]) and the Data and Connectivity strand (MC_PC_20058 [2021-2022]); The National Institute for Health Research (NIHR) and the Medical Research Council (MRC) via the CONVALESCENCE programme (COV-LT-0009, MC_PC_20051 [2021-2024]); NHS England via the Primary Care Medicines Analytics Unit [2021-2024].

Author Contributions

EP, RME, EPKP, WH, CWG, and JKQ contributed to the conception and design of the work; BG, SB, AM, DE, WH contributed to data acquisition; EP, RME, EPKP, WH, DE, REC, SBMM, FY, MN, YA contributed to the analysis and/or interpretation of the data; BG, SB, AM, DE, RE, WH contributed to the creation of the software used within the work; and EP, RME, EPKP, WH, REC, CWG, JKQ, SBMM, FY, MN, YA contributed to the drafting and revision of the work.

Competing Interests

REC holds personal shares in AstraZeneca.