Whole-genome sequencing, strain composition, and predicted antimicrobial resistance of *Streptococcus pneumoniae* causing invasive disease in England in 2017–20: a prospective national surveillance study

Joshua C D'Aeth, Marta Bertran, Fariyo Abdullahi, Seyi Eletu, Erjola Hani, Norman K Fry, Shamez N Ladhani, David J Litt

Summary

Background Surveillance of the invasive disease burden caused by *Streptococcus pneumoniae* in England is performed by the UK Health Security Agency (UKHSA). In 2017, UKHSA switched from phenotypic methods to whole-genome sequencing (WGS) approaches for pneumococcal surveillance. Here, we present the first results of national WGS surveillance, up to the start of the COVID-19 pandemic, with the aim of describing the population genomics of this important pathogen.

Methods We examined prospective national surveillance data from England, using bacterial isolates from cases of invasive pneumococcal disease (IPD) submitted to the national reference laboratory at UKHSA. A bioinformatic pipeline was developed to quality control WGS data and routinely report species and serotype. We assembled isolate data, assigned global pneumococcal sequencing clusters (GPSCs), and predicted antimicrobial resistance (AMR) profiles for isolates that passed further quality control. We collected additional data on patient outcomes and characteristics using enhanced surveillance questionnaires completed by patients' general practitioners. We used logistic regression analysis to assess the effects of various genomic and patient characteristics on the outcomes of IPD.

Findings In England, between July 1, 2017, and Feb 29, 2020, there were 15 400 cases of IPD. From these cases, 13 749 (89·3%) isolates were sequenced, passed quality control, and were included in analyses. Serotype diversity was high during the study period, with 2751 (20%) isolates serotyped as 13-valent pneumococcal conjugate vaccine (PCV13) types, whereas serotype 8 was the most prevalent serotype (n=3074 [22·4%]) overall. There were 157 GPSCs within the collection, with GSPC3 the most common, encompassing 98·7% (3033 of 3074) of serotype 8 isolates. Most isolates (n=10 198 [74·2%]) did not contain AMR-associated genes. Resistance to co-trimoxazole was the most frequently predicted resistance (n=2331 [17%]), followed by resistance to tetracycline (n=1199 [8·7%]) and β-lactams (n=1149 [8·4%]). Logistic regression analysis found the presence of AMR-associated genes significantly increased the odds of patient death (odds ratio 1·18, 95% CI 1·01–1·38). Some GPSCs were also associated with a significant increase in the odds of patient death, such as GPSC12 (1·88, 1·48–2·38). Isolates from 2018 were associated with a significant increase in the odds of patient death (1·12, 1·00–1·25), whereas younger patient age was significantly associated with a reduction in the odds of patient death compared with being aged 85 years or older.

Interpretation WGS-based surveillance has allowed us to interrogate country-wide population dynamics driving changes in pneumococcal serotype frequency. Here, we observe a stable but diverse population before the COVID-19 pandemic restrictions were enforced in England, with low rates of AMR. These findings will provide the baseline for pandemic and post-pandemic data, to collectively inform implementation and development of the vaccination programme within the country.

Funding None.

Introduction

Streptococcus pneumoniae (the pneumococcus) is an opportunistic pathogen that colonises the human nasopharynx and can be distinguished into more than 100 different serotypes on the basis of its unique capsular polysaccharide.[1] The pneumococcus is responsible for a high burden of disease both in terms of common infections, such as sinusitis and otitis media, as well as more severe infections, including bacteraemic pneumonia and meningitis.[2] The deployment of effective pneumococcal conjugate vaccines (PCVs), which targeted the most prevalent serotypes responsible for invasive pneumococcal disease (IPD), has helped to lower this burden of disease.[3–5]

In the UK, the 7-valent PCV (PCV7) was introduced into the routine childhood vaccine programme in 2006, and replaced with PCV13 in 2010.[6] Both vaccine programmes

Lancet Microbe 2025; 6: 101102

Published Online May 24, 2025 https://doi.org/10.1016/j.lanmic.2025.101102

Respiratory and Vaccine Preventable Bacterial Reference Unit, UK Health Security Agency, London, UK (J C D'Aeth PhD, S Eletu PhD, N K Fry PhD, D J Litt PhD); **Immunisation and Vaccine Preventable Diseases Division, UK Health Security Agency, London, UK** (M Bertran MSc, F Abdullahi MSc, E Hani MSc, N K Fry, S N Ladhani PhD, D J Litt); **Paediatric Infectious Diseases Research Group, St George's University of London, London, UK** (S N Ladhani)

Correspondence to: Dr Joshua C D'Aeth, Respiratory and Vaccine Preventable Bacterial Reference Unit, UK Health Security Agency, London NW9 5EQ, UK
joshua.daeth@ukhsa.gov.uk

Research in context

Evidence before this study

We searched PubMed using the terms ((((Streptococcus pneumoniae surveillance[Title/Abstract]) OR (Streptococcus pneumoniae whole genome sequencing[Title/Abstract])) OR ("pneumococcal surveillance"[Title/Abstract])) OR (pneumococcal whole genome sequencing[Title/Abstract])) OR (pneumococcal national surveillance[Title/Abstract]) on Jan 13, 2024, with no date restrictions. Only publications written in English were included. We identified 374 relevant articles from a large number of countries distributed globally, with some studies focused on specific regions within countries. Reported population-based surveillance studies were largely set up to monitor the effectiveness of pneumococcal conjugate vaccines after implementation in national or regional vaccination programmes. Whole-genome sequencing (WGS) has generally only been performed on serotypes or strains of interest, with overall surveillance performed mainly through serotyping isolates via Quellung, latex agglutination, or PCR-based methods. Some high-income countries, such as Australia and the USA, have switched to WGS-based surveillance, but so far studies published from these settings have been limited to select regions within the country. International collaborations have also sought to sequence pneumococci collected as part of national surveillance in different countries, although notably these datasets are smaller and lack longitudinal data compared with our national dataset for England.

Added value of this study

We present the first results from country-wide surveillance of invasive pneumococcal disease in England using WGS, which has produced an extensive and comprehensive dataset within a large defined geographical region over multiple years of surveillance. We used the latest typing techniques to place these isolates within an international context, allowing for global comparisons. The dataset also provides a benchmark of the pneumococcal population diversity before the perturbation of the COVID-19 pandemic and changes to childhood and adult pneumococcal vaccination programmes in England.

Implications of all the available evidence

We observed a diverse pneumococcal population causing invasive pneumococcal disease (IPD) in England, with 68 different serotypes and 157 different global pneumococcal sequencing clusters (GPSCs). However, one clonal lineage—the serotype 8, GPSC3, clonal complex 53 grouping—caused one in five IPD cases during the study period. This lineage has been emerging in other high-income countries, and its expansion should be monitored closely. Predicted rates of antimicrobial resistance (AMR) within England were low, in contrast to international datasets of pneumococcal WGS data. We also found predicted AMR to be linked to a significant increase in the odds of patient death, underpinning the importance of AMR surveillance in the pneumococcus. The evidence highlights the added benefits of WGS-based surveillance, allowing for further interrogation of the biology of this pathogen.

were associated with large and sustained declines in IPD due to the reduction of vaccine serotypes across all age groups. At the same time, however, there was an increase in IPD due to non-PCV serotypes.[5,6] In the UK, adults aged 65 years and older and individuals older than 2 years at high risk of infection are also offered a single dose of the 23-valent pneumococcal polysaccharide vaccine (PPV23). Surveillance of pneumococci causing IPD is critical for monitoring the effectiveness of these vaccines, informing immunisation policy, and developing next-generation PCVs.[7] In England, the UK Health Security Agency (UKHSA; formerly Public Health England) is responsible for national IPD surveillance. Hospitals routinely send all invasive *S pneumoniae* isolates to the UKHSA national reference laboratory for confirmation and serotyping. Previously, this was done using phenotypic identification and serotyping by slide agglutination. With the development of robust methods for in silico species typing[8] and pneumococcal serotyping[9] directly from whole-genome sequencing (WGS) data, the UKHSA changed to routine WGS of all IPD isolates in 2017.

Here, we present the results of the first 2·5 years of WGS-based surveillance, from July 1, 2017 to Feb 29, 2020, just before COVID-19 restrictions were implemented in England. We aimed to describe the population structure of the pneumococcus in England, using the latest tools to type pneumococci into lineages, allowing us to compare IPD-causing strains in England within a wider global context. We also used WGS data to identify antimicrobial resistance (AMR) genes, and to predict the AMR profile of pneumococci causing IPD.

Methods

Study design and bacterial isolates

Pneumococcal isolates were collected as part of surveillance of IPD by the national reference laboratory at UKHSA, taken from cases in England between July 1, 2017, and Feb 29, 2020. From July 1 to Sept 31, 2017, we analysed a random subset of 376 invasive pneumococcal isolates through a WGS pipeline to troubleshoot the implementation of WGS-based surveillance, using previously validated bioinformatic tools.[9] From Oct 1, 2017, we performed species confirmation and serotyping of all submitted isolates through routine use of WGS. For all confirmed cases, we contacted general practitioners to complete a short surveillance questionnaire seeking information on demographics, vaccination history, clinical presentation, and outcomes of IPD.

IPD isolates were defined as isolation of *S pneumoniae* from a normally sterile site. Clinical presentation was classified as meningitis (*S pneumoniae* detected in the cerebrospinal fluid

or pneumococcal isolation from the blood with clinical or radiological features of meningitis or a CNS focus) or non-meningitis. Serotypes were grouped as present in PCV7 (serotypes 4, 6B, 9V, 14, 18C, 19F, and 23F), PCV13 (PCV7 plus serotypes 1, 3, 5, 6A, and 7F), or PPV23 (PCV13 plus serotypes 2, 8, 9N, 10A, 11A, 12F, 15B, 17F, 20, 22F, and 33F). Non-vaccine serotypes are defined as those not present in either the PCV13 or PPV23 vaccines. Higher-valent PCV serotypes were grouped as PCV15 (PCV13 plus serotypes 22F and 33F) and PCV20 (PCV15 plus serotypes 8, 10A, 11A, 12F, and 15B). We grouped serotype 15B and 15C isolates within PCV20.

UKHSA has legal permission to process confidential information for national surveillance of communicable diseases without individual patient consent (Regulation 3 of Health Service Regulations 2002) and, as such, ethics committee approval was not required.

Procedures

We extracted and sequenced genomic DNA, as described by Kapatai and colleagues.[9] We ran the deplexed reads through a bioinformatic pipeline to identify species, serotype, and multilocus sequence type (MLST) of an isolate (appendix 1 pp 2, 8). For this study, we subjected isolates that were successfully sequenced to further quality-control steps and assembled them with shovill, version 0.9. We performed further quality control of assemblies with QUAST (version 5.2.0),[10] CheckM (version 1.2.2),[11] and PopPUNK (version 2.6.0;[12] appendix 1 pp 2–3, 9). We then assigned global pneumococcal sequencing clusters (GPSCs) to this collection using PopPUNK and the global pneumococcal sequencing project's (GPS) version 6 database. GPSCs were labelled as dominant if they contained more than 100 isolates. We predicted resistance to β-lactams using a random forest model;[13] sulfamethoxazole and trimethoprim resistance were predicted by searching for mutations in the *folP* and *folA* genes, respectively,[14] whereas other AMR genes were detected using NCBI-AMRFinderPlus, version 3.1.40[15] (appendix 1 pp 3–4). We searched for virulence factors in the GPSC3 and non-GPSC3 serotype 8 isolates, using *S pneumoniae* genes from the Virulence Factor Database[16] (appendix 1 p 4). We assigned clonal complexes (CCs) by grouping together single-locus variants of MLSTs (appendix 1 p 3).

We performed additional phylogenetic analysis on the two largest lineages, GPSC3 CC-53 isolates and GPSC12 isolates using Gubbins, version 3.3.5[17] (appendix 1 p 4). For the GPSC3 CC-53 analysis, isolates were combined with GPSC3 ST53 serotype 8 isolates from the GPS collection.[18] We gathered the clade definitions for GPSC12 from previous work.[19]

Statistical analysis

We analysed data using R, version 4.4.1, with proportions compared among age groups using Fisher's exact test. We calculated the Simpson's diversity index 1–D (SDI), where $D=\sum pi2$ and $pi2$ is the proportional abundance of species i, for serotype and GPSC compositions among different dataset subgroups. SDI ranges from 0 to 1, with a higher number indicating higher diversity. We performed a two-sample Kolmogorov-Smirnov test to compare the distribution of the within-CC core-genome distances of CC53 to other large CCs (those with ≥100 isolates). Proportions of virulence genes among three unequal groups, GPSC3 CC-53 isolates, GPSC3 non-CC-53 isolates, and non-GPSC3 serotype 8 isolates, were compared using Fisher's exact test, with a Bonferroni corrected p value of 0·00051 for multiple testing.

We constructed a logistic regression model to calculate the effect of bacterial characteristics (GPSC and AMR status [susceptible or non-susceptible]) and patient characteristics (patient age, year of sampling, and clinical presentation [meningitis or non-meningitis]) on case fatality rate. The outcome was patient death, defined as death within 30 days of first IPD sample date. We excluded five isolates with incomplete AMR profiles from the analysis. We calculated odds ratios (ORs) with 95% CIs from the logistic regression coefficients to quantify the effect of the above predictor variables on the likelihood of the outcome. We performed sensitivity analyses to assess the effect of patient vaccination status, with either PCV (either PCV7 or PCV13) or PPV23 vaccination, on case fatality rate (appendix 1 pp 5–6). We also extended sensitivity analyses to incorporate 26 GPSCs over the 24 dominant GPSCs, and replaced GPSCs with the 24 serotypes with more than 100 isolates (appendix 1 pp 5–6).

We performed a genome-wide association study (GWAS) analysis with the R package treeWAS, version 1.0,[20] to test for any significant association between an isolate's genome and case fatality rate. To mitigate the effects of population structure and patient characteristics, we limited this analysis to the GPSC12 isolates taken from cases in the group 85 years and older (appendix 1 p 6).

See Online for appendix 1

Role of the funding source

There was no funding source for this study.

Results

Between July 1, 2017, and Feb 29, 2020, there were 15 400 confirmed IPD cases in England. From these cases, 13 944 (90·5%) isolates were sequenced and 13 749 isolates (98·6% of those sequenced) passed the WGS quality control steps for inclusion in this study (appendix 1 p 9; appendix 2 p 1). The cases from this collection of 13 749 isolates followed a seasonal trend during the study period (figure 1A), peaking in the winter months, with most cases occurring among those aged 65 years and older (n=7509, 54·6%; figure 1B). In the collection, 5·6% of cases (n=776) presented with meningitis and there was a case fatality rate of 17·4% (n=2392).

See Online for appendix 2

Most isolates possessed a serotype present in PCV7 (n=391 [2·8%]), PCV13 (n=2751 [20%]), or PPV23 (10 339 [75·2%]; figure 1A). Serotype 8, included in PPV23, was the most prevalent (n=3074 [22·4%]), followed by serotypes 3 (n=1532 [11·1%]), 12F (n=1070 [7·8%]),

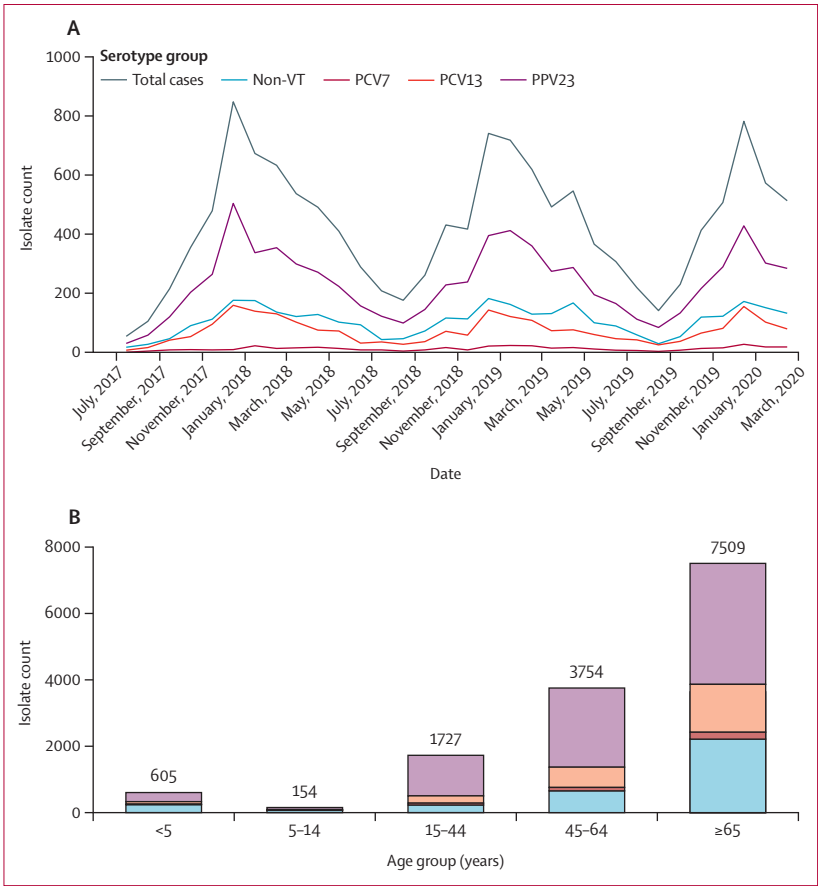

***Figure 1: Streptococcus pneumoniae* from invasive disease in England, isolate count, and isolate age grouping**
(A) The counts of isolates per month between July 1, 2017, and Feb 29, 2020. For plotting purposes, PCV13 represents the six additional serotypes compared with the PCV7 formulation, PPV23 represents the 11 additional serotypes compared with the PCV13 formulation. Non-VT refers to serotypes not included in PCV7, PCV13, or PPV23. (B) The overall numbers of isolates in each of the five defined patient age groups over the whole course of the study. Lines and bars are coloured by grouping of isolates based on serotype inclusion in pneumococcal vaccines. PCV=pneumococcal conjugate vaccine. PPV=pneumococcal polysaccharide vaccine. VT=vaccine-type.

22F (n=938 [6·8%]), and 9N (n=891 [6·5%]; appendix 1 p 10). Of the non-vaccine serotypes, serotype 15A was the most prevalent (n=540 [3·9%]). For the two higher valent vaccines licensed after our study period, PCV15 in 2021, and PCV20 in 2022, 4141 isolates (30·1%) were covered by PCV15, whereas 9288 (67·6%) isolates were covered by PCV20. The overall serotype composition remained stable throughout the surveillance period, with an SDI of 0·914 (95% CI 0·872–0·95) in 2017–18 and 0·908 (95% CI 0·870–0·951) in 2019–20. SDI stability over time was also observed across individual age groups, although the groups aged 15–44 years and 45–64 years were less diverse, with SDI scores below 0·9, compared with other age groups (appendix 1 p 11). This lack of diversity reflects the greater dominance of serotype 8 in these adult age groups compared with the groups aged younger than 5 years, 5–14 years, and 65 years and older (appendix 1 p 12).

In total, there were 157 GPSCs detected in the population, with 13 728 (99·8%) of 13 749 isolates assigned a GPSC. There were 50 GPSCs with ten or more isolates, including 24 dominant GPSCs with 100 or more isolates, which accounted for 90·3% (n=12 418) of all isolates (figure 2).

GPSC3 (n=3803 [27·7%]) was the most frequently observed, accounting for 98·7% (n=3033) of all serotype 8 isolates (figure 2), followed by GPSC12 (n=1399 [10·2%]), GPSC16 (n=911 [6·6%]), GPSC55 (n=877 [6·4%]), and GPSC19 (n=773 [5·6%]). There were no notable alterations in the dominant GPSC frequencies during the surveillance period (appendix 1 p 13). The overall SDI for GPSCs was high, at 0·894 (95% CI 0·834–0·959), remaining at this level throughout the surveillance period (appendix 1 p 14). The median number of serotypes expressed by each of the dominant GPSCs was 4·5 (IQR 2·0–8·25), with 20 of 24 dominant GPSCs expressing more than one serotype. GPSC44 expressed the most, with 13 different serotypes. Of the four dominant GPSCs that expressed only one serotype, GPSC12 was the largest, expressing only the vaccine serotype 3. GPSC17 (n=120 [0·9%]) also only expressed the vaccine serotype 19A, whereas GPSC46 (n=211 [1·53%]) only expressed serotype 16F and GPSC57 (n=183 [1·33%]) only expressed serotype 31. In total, 15 of 24 dominant GPSCs contained isolates that expressed a PCV13 serotype; the same 15 dominant GPSCs expressed a PCV15 serotype, whereas 20 of 24 expressed a PCV20 serotype. Within the GPSC12 lineage, as reported previously by Bertran and colleagues,[19] most isolates belonged to the clade IV lineage (721 [51·5%] of 1399), with clade I, which was initially more prevalent in 2017–18, now the second most common clade in the lineage (618 [44·2%] of 1399).

The dominant GPSC lineages were unequally distributed among the broad age groups within the population (p<0·001). In line with serotype 8 results above, GPSC3 isolates were over-represented among the groups aged 15–44 years, 45–64 years, and 65 years and older (appendix 1 p 15). By contrast, the GPSC5 lineage with 309 isolates, including 280 [90·6%] isolates belonging to the non-vaccine serotype 23B, was over-represented in children (<5 years and 5–14 years; appendix 1 p 15).

There were 682 different sequence types (STs) assigned by MLST, with 13 706 isolates (99·7%) assigned to a known ST, 41 isolates with novel profiles, and two isolates with incomplete locus coverage. The most prevalent ST was ST53, with 2768 isolates (20·1%), all of which were assigned to GPSC3 and were identified as serotype 8. There were 27 STs with 100 or more isolates. Almost all STs were found within a single GPSC (681 of 682), with only ST6011 having two isolates each, four in total, in the distantly related GPSC43 and GPSC69 lineages. By contrast, each dominant GPSC encompassed a median of 13·5 STs (IQR 8–19), with GPSC3 containing the highest number of STs with 53. Only one dominant GPSC, GPSC17, expressed a single ST, ST2062. This 120-isolate GPSC was the smallest of the dominant GPSCs.

The STs were further grouped together into CCs (figure 3). There were 29 CCs that contained 100 or more isolates, with the GPSC3 CC-53 lineage, encompassing the

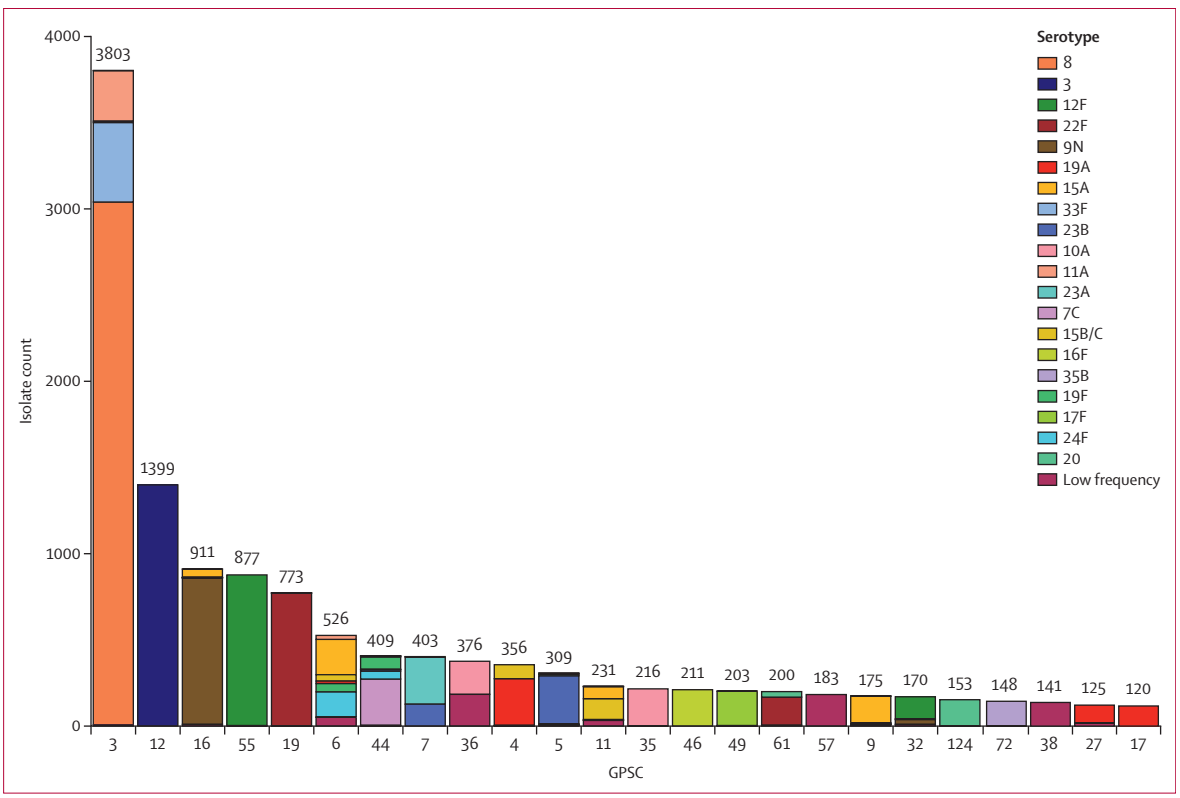

Figure 2: **Overall global pneumococcal sequencing cluster counts, split by serotype, within the collection**
The total number of isolates is shown above each bar. The bars are coloured by serotypes within a GPSC. Only the 20 most frequent serotypes are highlighted, the rest are grouped into the low-frequency category. Serotypes are ordered by frequency in the key, with serotype 8 the most common, followed by serotype 3, while serotype 20 is the least. GPSC=global pneumococcal sequencing cluster.

ST53 and single locus variant strains, being the largest in the collection. CC-53 formed a monophyletic clade of 3027 isolates in the core-genome distance tree (figure 3). This CC-53 clade had a lower median within core-genome distance of 0·00012 single nucleotide polymorphisms (SNPs) per site (IQR $7{\cdot}3\times10^{-5}$) compared with the median within-CC core-genome distance of 0·00031 SNPs per site (IQR 0·00167) for the other 28 CCs with 100 or more isolates; these two distance distributions were significantly different (p<0·0001).

The GPSC3 CC-53 lineage had a higher proportion of isolates containing the *nanA* virulence factor, involved in endothelial invasion than did non-CC-53 GPSC3 isolates and non-GPSC3 serotype 8 isolates (CC-53 *vs* non-CC-53 GPSC p<0·0001; CC-53 *vs* non-GPSC3 serotype 8 p<0·0001; appendix 1 p 16). The GPSC3 CC-53 isolates from this study also did not form a monophyletic clade when global serotype 8 GPSC3 CC-53 isolates were compared with our collection (appendix 1 pp 6, 17). These external GPS isolates were collected between 1999 and 2015.

The frequency of predicted AMR was low across the population, with 10 198 (74·2%) of 13 749 isolates containing no AMR-associated genes (figure 3; table). Within the GPSC3 CC-53 serotype 8 lineage, 3006 (99·3%) of 3027 isolates were predicted to be wholly susceptible. Of the dominant GPSCs, seven were predicted to be majority resistant, and all 24 had at least one isolate predicted to be resistant (appendix 1 p 18). The most commonly predicted resistance was to co-trimoxazole, with 2331 (17·0%) of 13 749 isolates containing mutations in either *folA* for the constituent sulfamethoxazole resistance (n=1209 [8·8%]), in *folP* for trimethoprim resistance (n=16 [0·1%]), or in both *folA* and *folP* (n=1106 [8%]). The second most common predicted resistance was to tetracycline, with 1199 isolates (8·7%) predicted to be resistant, primarily due to isolates containing the *tet(M)* resistance gene (807 [67·3%] of 1199) or the *tet(32)* gene (388 [32·4%] of 1199). The third most common predicted resistance was to β-lactams, with 1149 (8·4%) of 13 749 isolates predicted to be resistant. Among these 1149 isolates predicted resistant to β-lactams, there were 205 unique PBP types, including novel variants, as defined previously by Li and colleagues.[21] The most common resistant profile was PBP1a-7, PBP2b-67, PBP2x-1 with 147 (12·8%) of 1149 β-lactam-resistant isolates containing this profile, all of which were within GPSC5 (appendix 1 p 6). Of the dominant-GPSCs, GPSC5, GPSC9, and GPSC17 exhibited very high proportions (≥98%) of their isolates predicted to be β-lactam resistant, while 17 of the 24 dominant-GPSCs contained at least one resistant isolate.

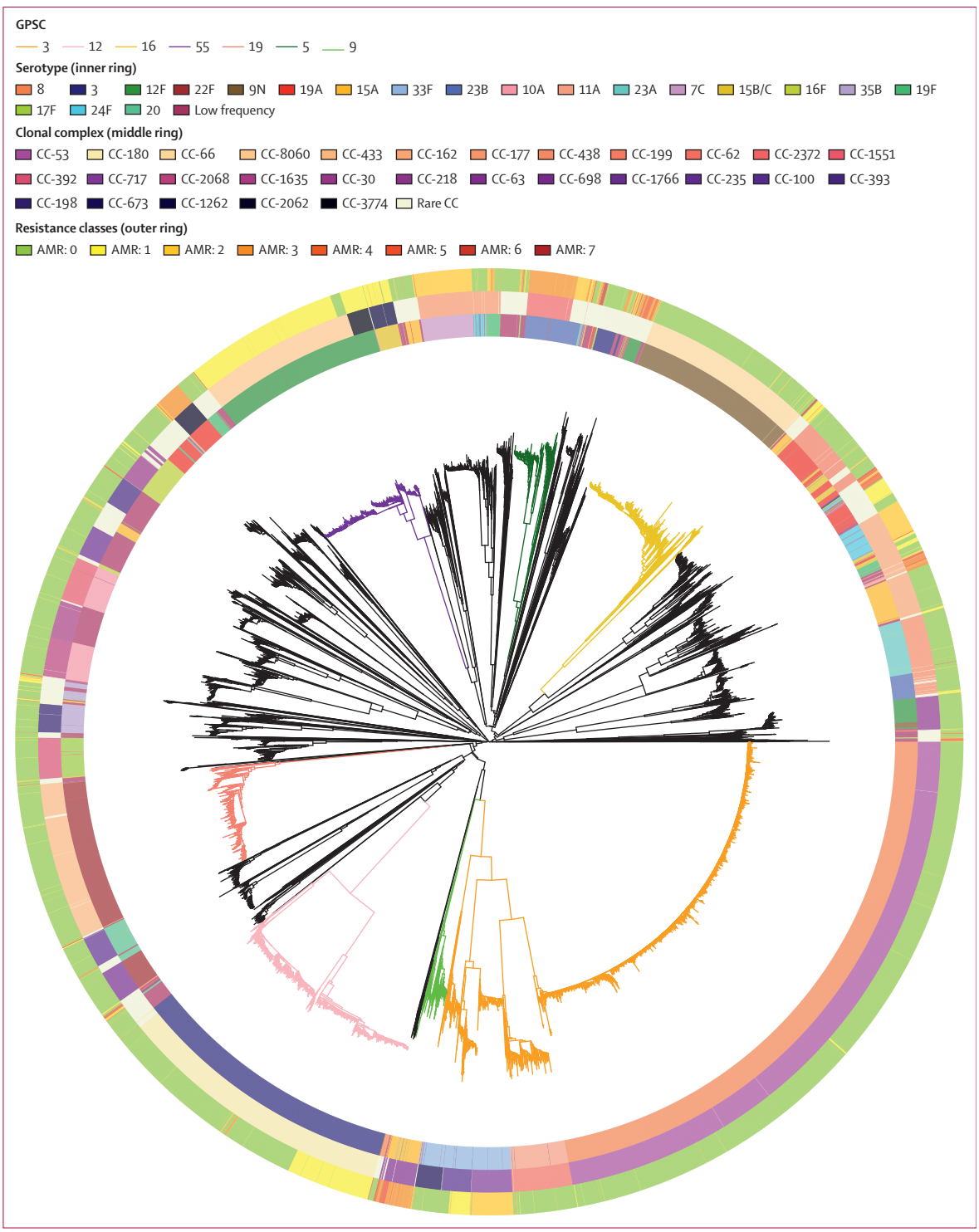

Figure 3: **Core-genome distance phylogeny of the collection**
A core-genome distance phylogeny created using RapidNJ of the 13 749 isolates in the collection. Branches of the phylogeny are coloured by dominant GPSC, with only seven select lineages labelled. The inner annotation ring represents the serotype of the tips of the tree, with only the 20 most frequent serotypes annotated. The middle ring represents the CC of the tip isolates, with only CCs represented by greater than 100 isolates annotated. The outer ring represents the number of classes of antibiotic, as defined by the NCBI AMRFinderPlus tool, to which an isolate is predicted to be resistant. GPSC=global pneumococcal sequencing cluster. CC=clonal complex. AMR=antimicrobial resistance. NCBI=National Centre for Biotechnology Information.

	Total	β-lactam	Tetracycline	Erythromycin	Macrolide	Chloramphenicol	Quinolone	Aminoglycoside	Quaternary ammonium	Streptothricin	Trimethoprim	Sulfamethoxazole
3	3803	11 (0·3%)	211 (5·5%)	14 (0·4%)	207 (5·4%)	0	4 (0·1%)	0	1 (0%)	0	3 (0·1%)	108 (2·8%)
12	1399	0	406 (29·0%)	1 (0·1%)	15 (1·1%)	12 (0·9%)	0	0	1 (0·1%)	0	1 (0·1%)	1 (0·1%)
16	911	17 (1·9%)	14 (1·5%)	11 (1·2%)	4 (0·4%)	9 (1·0%)	2 (0·2%)	0	0	0	13 (1·4%)	60 (6·6%)
55	877	107 (12·2%)	0	0	0	0	3 (0·3%)	0	1 (0·1%)	0	0	723 (82·4%)
19	773	0	2 (0·3%)	1 (0·1%)	0	0	6 (0·8%)	0	0	0	1 (0·1%)	1 (0·1%)
6	526	67 (12·7%)	31 (5·9%)	48 (9·1%)	16 (3·0%)	0	1 (0·2%)	1 (0·2%)	0	0	212 (40·3%)	226 (43·0%)
44	409	16 (3·9%)	33 (8·1%)	0	34 (8·3%)	0	1 (0·2%)	2 (0·5%)	0	1 (0·2%)	270 (66·0%)	270 (66·0%)
7	403	8 (2·0%)	2 (0·5%)	5 (1·2%)	2 (0·5%)	1 (0·2%)	1 (0·2%)	0	0	0	5 (1·2%)	5 (1·2%)
36	376	1 (0·3%)	0	2 (0·5%)	0	0	1 (0·3%)	0	0	0	0	1 (0·3%)
4	356	2 (0·6%)	7 (2·0%)	3 (0·8%)	6 (1·7%)	0	0	0	0	0	2 (0·6%)	1 (0·3%)
5	309	306 (99·0%)	11 (3·6%)	1 (0·3%)	10 (3·2%)	0	1 (0·3%)	1 (0·3%)	0	0	227 (73·5%)	292 (94·5%)
11	231	2 (0·9%)	3 (1·3%)	1 (0·4%)	3 (1·3%)	0	0	0	0	0	8 (3·5%)	117 (50·6%)
35	216	0	0	0	0	0	1 (0·5%)	0	0	0	0	1 (0·5%)
46	211	0	1 (0·5%)	0	1 (0·5%)	0	1 (0·5%)	0	0	0	3 (1·4%)	17 (8·1%)
49	203	0	0	2 (1·0%)	0	0	0	0	0	0	0	1 (0·5%)
61	200	2 (1·0%)	7 (3·5%)	3 (1·5%)	1 (0·5%)	2 (1·0%)	0	0	0	0	1 (0·5%)	5 (2·5%)
57	183	1 (0·5%)	0	0	0	0	0	0	0	0	0	0
9	175	173 (98·9%)	175 (100%)	4 (2·3%)	172 (98·3%)	1 (0·6%)	5 (2·9%)	3 (1·7%)	0	3 (1·7%)	32 (18·3%)	40 (22·9%)
32	170	1 (0·6%)	1 (0·6%)	0	0	0	0	0	0	0	0	1 (0·6%)
124	153	0	0	0	0	0	1 (0·7%)	0	0	0	0	3 (2·0%)
72	148	0	1 (0·7%)	0	0	0	0	0	0	0	1 (0·7%)	0
38	141	6 (4·3%)	7 (5·0%)	6 (4·3%)	0	0	1 (0·7%)	0	0	0	6 (4·3%)	6 (4·3%)
27	125	80 (64·0%)	0	1 (0·8%)	0	0	0	0	0	0	0	1 (0·8%)
17	120	120 (100%)	0	0	0	0	0	0	0	0	120 (100%)	120 (100%)
Other	1331	229 (17·2%)	287 (21·6%)	76 (5·7%)	133 (10·0%)	68 (5·1%)	10 (0·8%)	7 (0·5%)	0	5 (0·4%)	217 (16·3%)	315 (23·7%)
Total	13 749	1149 (8·4%)	1199 (8·7%)	179 (1·3%)	604 (4·4%)	93 (0·7%)	39 (0·3%)	14 (0·1%)	3 (<0·1%)	9 (0·1%)	1122 (8·2%)	2315 (16·8%)

Data are N or n (%). The number of isolates predicted to be resistant to 11 different antimicrobial classes among the 24 dominant GPSCs, arranged in order of size, followed by an other category containing isolates not present in the 24 dominant GPSCs. GPSCs=global pneumococcal sequencing clusters.

Table: **Distribution of resistance classes among GPSCs**

Aside from tetracycline and β-lactam resistance, macrolide resistance was the next most frequently predicted (604 [4·4%] of 13 749 isolates). Resistance to both tetracycline and macrolides was frequently found in the same isolate (n=598 [99%] of 604 macrolide-resistant isolates). In GPSC9, all 175 isolates contained the *tet(M)* gene encoding tetracycline resistance and 172 also contained a macrolide resistance gene, either *erm(B)* or *mef(A)*. These genes were often present within a *Tn916*-like element (appendix 1 p 7).

Logistic regression analysis found that, when controlling for the GPSC of an isolate, age group, clinical presentation, and year of diagnosis, the presence of genes associated with resistance to one or more classes of antibiotic was associated with a 1·18-fold (95% CI 1·01–1·38; p=0·035) higher odds of death (figure 4). There were four dominant GPSCs that also had a significantly increased case fatality rate compared with the reference GPSC19 lineage, which tracked closest to the dominant GPSC overall case fatality rate of 17·2% (appendix 1 p 19). Most notably, GPSC12 isolates were associated with a 1·88-fold (95% CI 1·48–2·38; p<0·0001) higher odds of death. Isolates taken during 2018 were also associated with a 1·12-fold (95% CI 1·0–1·25; p=0·047) higher odds of death. Younger age was significantly associated with lower case fatality rate, with all age groups having lower odds of death compared with the reference group aged 85 years and older (figure 4).

In the sensitivity analyses, factoring in a patient's PCV vaccine status confounded the effect of age on case fatality rate for the youngest age groups (0–4 years, 5–9 years, and 10–14 years; appendix 2 pp 2, 4), with these age groups no longer associated with a significant decrease in the odds of death because of the very high vaccine uptake among cases (544 [98·2%] of 554, 95 [100%] of 95, and 35 [85·4%] of 41, respectively). PPV23 vaccine status, however, did not confound the association between age and a decreased odds of death (appendix 2 pp 3–4).

Increasing the number of dominant GPSCs to include the 96-isolate GPSC83, all of which expressed a serotype 3 capsule, revealed that, in contrast to the serotype 3 GPSC12 lineage, GPSC83 had no significant change in the odds of death compared with the reference GPSC19 lineage (appendix 2 p 5). Finally, when modelling the most frequent serotypes instead of GPSCs, certain serotypes were associated with significant changes in the odds of death, with serotype 3 increasing the odds (OR 1·85, 95% CI 1·33–2·62; p=0·0026; appendix 2 p 6), whereas the presence of AMR genes was no longer found to be associated with a significantly higher odds of death (OR 1·09, 95% CI 0·95–1·26; p=0·21; appendix 2 p 6).

A GWAS analysis on the 302 IPD cases caused by GPSC12 isolates in the group 85 years and older found no significant associations between core-genome SNPs and death, or accessory gene presence or absence and death (appendix 1 p 20).

Discussion

This study reports results from the initial period of using WGS for enhanced national surveillance of IPD in England. We have sequenced more than 5000 invasive pneumococcal isolates annually during the 2·5-year pre-pandemic period. With the added detail of WGS, we found that although the pneumococcal population causing IPD is diverse in England, the most common lineage (GPSC3 CC-53) represents a successful international clonal expansion. Additionally, the strain composition in the population remained stable during the surveillance period, providing a baseline before the COVID-19 pandemic and the change in the childhood PCV13 immunisation programme from a 2 + 1 to a 1 + 1 schedule.[19] Finally, we used the WGS data to predict AMR, finding low rates overall, with resistance genes identified mainly in select lineages, such as GPSCs 5, 9, and 17. However, the presence of AMR genes was linked to a significant increase in the case fatality rate.

The paraphyletic nature of the GPSC3 CC-53 serotype 8 lineage in England, with respect to country of origin, is evidence of a remarkably successful international clonal expansion. Countries across Europe and Africa have been reporting an increase in serotype 8 IPD following the introduction of PCV13·[22–26] In Spain and Denmark, an over-representation of serotype 8 isolates among adults aged 15–65 years was also observed, similar to the trend in England reported herein, which might indicate greater transmission of this serotype within adult age groups.[24,27] Furthermore, of these serotype 8 isolates, the GPSC3 CC-53 lineage is commonly reported to be driving the spread of IPD cases.[22,25] Our investigation into the virulence factors present in the lineage revealed an over-representation of the *nanA* gene, a neuraminidase important for endothelial invasion, compared with both other serotype 8 isolates and non-serotype 8 GPSC3 isolates.[28] The over-representation of *nanA* could be a reason for the greater invasiveness of this lineage, with recent modelling work by Løchen and colleagues also showing the serotype 8 GPSC3 combination to be highly invasive.[29] This higher invasiveness might help to explain why the GPSC3 CC-53 lineage appears to be a very successful international clonal expansion, despite moderate rates of nasopharyngeal carriage in adults and very low carriage rates in children.[30]

In our dataset, the prevalence of predicted resistance to at least one antibiotic class from WGS analysis was 25·8%. This prevalence contrasts with that of the diverse GPS collection, where 61% of isolates were predicted to be resistant to at least one antibiotic class.[18] Our logistic regression modelling of AMR gene presence, along with other patient and isolate data, showed a significant association between AMR-associated gene presence and higher case fatality rate. This finding is broadly in line with other recent large-scale European and global modelling studies, which predict resistance to increase the relative risk of death.[31,32] However, in our sensitivity analyses, we found that when controlling for serotype instead of GPSC, AMR-associated gene presence no longer had a significant effect on a patient's odds of death. Here, although serotype is thought to be the most important pneumococcal virulence factor, an isolate's GPSC might more accurately control for

Variable	Samples (n [%])	Died (n [%])	aOR (95% CI)	p value
Total	12 413 (100%)	2129 (17·2%)		
GPSC				
GPSC19	772 (6·2%)	129 (16·7%)	Ref (1)	
GPSC11	231 (1·9%)	43 (18·6%)	1·25 (0·82–1·86)	0·29
GPSC12	1399 (11·3%)	385 (27·5%)	1·88 (1·48–2·38)	<0·0001
GPSC124	153 (1·2%)	24 (15·7%)	1·11 (0·67–1·79)	0·67
GPSC16	911 (7·3%)	171 (18·8%)	1·28 (0·98–1·66)	0·065
GPSC17	120 (1·0%)	15 (12·5%)	0·85 (0·45–1·52)	0·60
GPSC27	125 (1·0%)	18 (14·4%)	0·90 (0·50–1·56)	0·73
GPSC3	3801 (30·6%)	454 (11·9%)	0·90 (0·72–1·12)	0·35
GPSC32	170 (1·4%)	14 (8·2%)	0·63 (0·33–1·10)	0·13
GPSC35	216 (1·7%)	33 (15·3%)	1·20 (0·77–1·84)	0·41
GPSC36	376 (3·0%)	73 (19·4%)	1·46 (1·04–2·03)	0·026
GPSC38	141 (1·1%)	31 (22·0%)	1·30 (0·81–2·04)	0·26
GPSC4	356 (2·9%)	64 (18·0%)	1·20 (0·85–1·69)	0·29
GPSC44	409 (3·3%)	87 (21·3%)	1·26 (0·90–1·77)	0·17
GPSC46	211 (1·7%)	51 (24·2%)	1·51 (1·02–2·20)	0·036
GPSC49	203 (1·6%)	39 (19·2%)	1·14 (0·75–1·71)	0·54
GPSC5	308 (2·5%)	49 (15·9%)	1·01 (0·67–1·50)	0·97
GPSC55	877 (7·1%)	110 (12·5%)	0·89 (0·65–1·23)	0·49
GPSC57	183 (1·5%)	57 (31·1%)	1·89 (1·28–2·76)	0·0011
GPSC6	526 (4·2%)	98 (18·6%)	1·03 (0·76–1·41)	0·83
GPSC61	200 (1·6%)	28 (14·0%)	0·78 (0·49–1·22)	0·30
GPSC7	403 (3·2%)	76 (18·9%)	1·12 (0·81–1·55)	0·49
GPSC72	148 (1·2%)	33 (22·3%)	1·38 (0·87–2·14)	0·16
GPSC9	174 (1·4%)	47 (27·0%)	1·43 (0·93–2·19)	0·10
Age (years)				
≥85 years	1997 (16·1%)	758 (38·0%)	Ref (1)	
0–4 years	554 (4·5%)	26 (4·7%)	0·08 (0·05–0·12)	<0·0001
5–9 years	95 (0·8%)	5 (5·3%)	0·09 (0·03–0·21)	<0·0001
10–14 years	41 (0·3%)	2 (4·9%)	0·09 (0·01–0·28)	0·0007
15–19 years	70 (0·6%)	2 (2·9%)	0·05 (0·01–0·17)	<0·0001
20–29 years	310 (2·5%)	11 (3·5%)	0·07 (0·03–0·12)	<0·0001
30–39 years	692 (5·6%)	45 (6·5%)	0·13 (0·09–0·17)	<0·0001
40–49 years	1089 (8·8%)	86 (7·9%)	0·16 (0·12–0·20)	<0·0001
50–59 years	1714 (13·8%)	170 (9·9%)	0·20 (0·16–0·23)	<0·0001
60–64 years	1061 (8·5%)	134 (12·6%)	0·25 (0·20–0·30)	<0·0001
65–69 years	1193 (9·6%)	155 (13·0%)	0·25 (0·21–0·30)	<0·0001
70–74 years	1263 (10·2%)	216 (17·1%)	0·35 (0·29–0·41)	<0·0001
75–79 years	1185 (9·5%)	222 (18·7%)	0·38 (0·32–0·45)	<0·0001
80–84 years	1149 (9·3%)	297 (25·8%)	0·58 (0·49–0·68)	<0·0001
Predicted AMR				
Susceptible	9308 (75·0%)	1546 (16·6%)	Ref (1)	
Resistant	3105 (25·0%)	583 (18·8%)	1·18 (1·01–1·38)	0·035
Meningitis				
No	11705 (94·3%)	2038 (17·4%)	Ref (1)	
Yes	708 (5·7%)	91 (12·9%)	1·22 (0·96–1·54)	0·11
Year				
2019	4817 (38·8%)	796 (16·5%)	Ref (1)	
2017	1862 (15·0%)	312 (16·8%)	1·02 (0·88–1·19)	0·77
2018	4752 (38·3%)	846 (17·8%)	1·12 (1·00–1·25)	0·047
2020	982 (7·9%)	175 (17·8%)	1·04 (0·86–1·26)	0·65

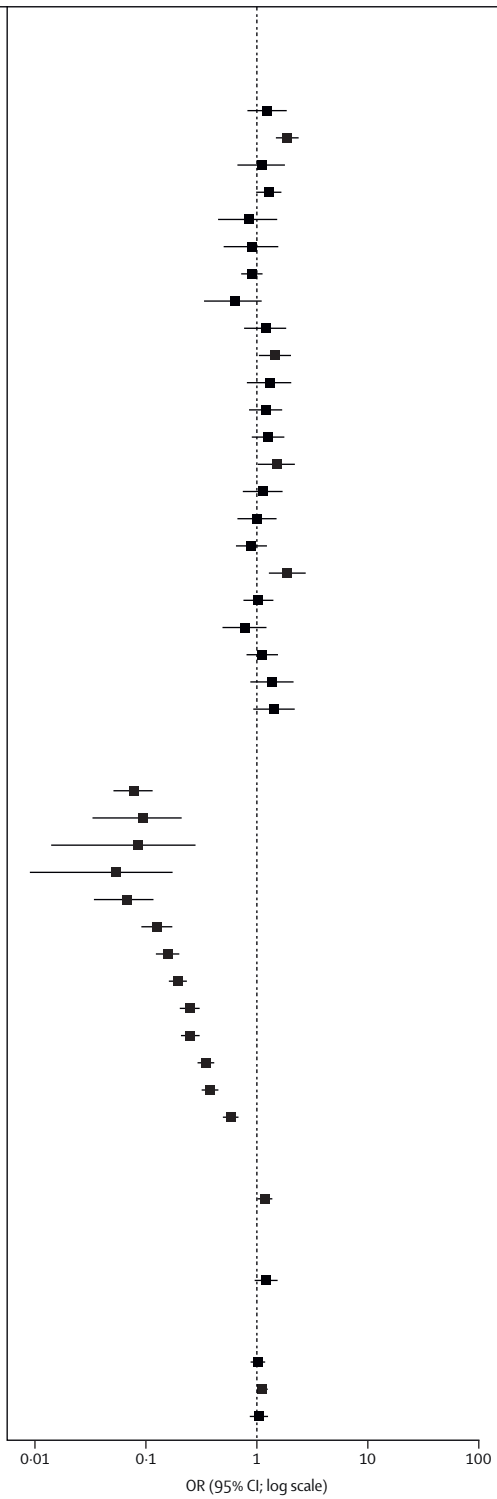

Figure 4: **Analysis of factors associated with patient death**
Forest plot showing the association among categorical variables (GPSC, patient age, AMR status, meningitis status, and year of case isolation) and patient death within 30 days of laboratory case confirmation. Only p values less than 0·05 were considered significant. Adjusted ORs with 95% CIs are displayed. AMR=antimicrobial resistance, here predicted genotypically. GPSC=global pneumococcal sequencing cluster. aOR=adjusted odds ratio.

both serotype and the diversity of other virulence factors that can influence the outcome of an infection.[33] For instance, we observed that serotype 3 was associated with a significantly higher odds of death; however, of the two large GPSCs that solely expressed serotype 3—GPSC12 and GPSC83—only GPSC12 was associated with a significantly higher odds of death. This association suggests the serotype lineage combination is key in determining an isolate's potential effects on mortality.

Our regression results were in contrast with the GWAS on GPSC12 isolates we performed, which showed no significant association between genetic variation and case fatality rate. Our GWAS results align with a previous GWAS focusing on pneumococcal isolates, which found no links between pathogen genetics and disease severity.[34] This previous GWAS also included data from the patient genome and identified a relationship between disease severity and certain markers of human genomic variation.[34] This highlights a limitation of our study, in that we use a limited number of variables to capture patient variation. Considering other factors, such as patient comorbidities, might allow future studies to gauge the effect of bacterial characteristics more accurately on mortality.

An additional limitation of this study was the sole use of in silico-predicted AMR profiles. For instance, AMR rates from phenotypic testing of a subset of UK isolates, as reported by the European Antimicrobial Resistance Surveillance Network (EARS-Net), were lower for the same period. EARS-Net reported only 5·3–5·6% of isolates were identified as penicillin non-susceptible during 2017–19,[35] compared with the predicted 8·4% rate in our collection. Our findings highlight the difficulties of in-silico resistance prediction, particularly with regard to the complex gene interactions that underpin penicillin resistance in pneumococci.[36] However, although the presence of resistance-associated alleles might not wholly predict the expression of β-lactam resistance, in-silico AMR predictions will allow us to monitor trends in the respective genes and their movement across lineages over time.

The work presented herein shows the benefit gained by using WGS in national surveillance. Switching to WGS has allowed the UKHSA to use the most recent strain-typing methods for pneumococci using genomic data, GPSCs, and allowed us to explore other genetic factors, such as AMR, which can be key in the expansion of strains causing invasive disease. We have shown how these data can be used to answer key clinical questions around IPD, for instance what genomic features of the pneumococcus are linked to increased mortality. Moving forward, these data can be used to inform decisions around the roll-out and development of vaccines, giving us a greater understanding of the lineages and serotypes associated with increased mortality or AMR.

Contributors

DJL, SNL, and NKF were responsible for the conceptualisation of the study. JCD, MB, FA, SE and EH were involved in data curation. MB, JCD, and FA accessed and verified the data. JCD conducted the genomic and statistical analysis. JCD wrote the first draft of the manuscript, which was edited and reviewed by all authors. All authors have seen and approved the final manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

The Immunisation and Vaccine Preventable Diseases Division has provided vaccine manufacturers with post-marketing surveillance reports on pneumococcal and meningococcal infection which the companies are required to submit to the UK licensing authority in compliance with their Risk Management Strategy. A cost recovery charge is made for these reports. The Respiratory and Vaccine Preventable Bacteria Reference unit of the UKHSA has received grants from vaccine manufacturers for investigator-led research projects on the pneumococcus. All other authors declare no competing interests.

Data sharing

Reads for the 13 749 isolates which passed further assembly quality control were deposited in the sequence read archive with BioProject accession numbers PRJNA1034002 and PRJNA1027675 (appendix 2 p 1). The code used for the statistical analyses and plotting of figures can be found at: https://gitlab.phe.gov.uk/Joshua.DAeth/daeth_lancet_microbe_pneumococcus_analysis.

Acknowledgments

We thank Nick J Andrews from the UKHSA for providing advice and assistance with the statistical analyses presented in this work.

References

1 Ganaie F, Saad JS, McGee L, et al. A new pneumococcal capsule type, 10D, is the 100th serotype and has a large *cps* fragment from an oral Streptococcus. *MBio* 2020; **11:** e00937-20.
2 Henriques-Normark B, Tuomanen EI. The pneumococcus: epidemiology, microbiology, and pathogenesis. *Cold Spring Harb Perspect Med* 2013; **3:** a010215.
3 Mackenzie GA, Hill PC, Jeffries DJ, et al. Impact of the introduction of pneumococcal conjugate vaccination on invasive pneumococcal disease and pneumonia in The Gambia: 10 years of population-based surveillance. *Lancet Infect Dis* 2021; **21:** 1293–302.
4 Moore MR, Link-Gelles R, Schaffner W, et al. Effect of use of 13-valent pneumococcal conjugate vaccine in children on invasive pneumococcal disease in children and adults in the USA: analysis of multisite, population-based surveillance. *Lancet Infect Dis* 2015; **15:** 301–09.
5 Waight PA, Andrews NJ, Ladhani SN, Sheppard CL, Slack MPE, Miller E. Effect of the 13-valent pneumococcal conjugate vaccine on invasive pneumococcal disease in England and Wales 4 years after its introduction: an observational cohort study. *Lancet Infect Dis* 2015; **15:** 535–43.
6 Ladhani SN, Collins S, Djennad A, et al. Rapid increase in non-vaccine serotypes causing invasive pneumococcal disease in England and Wales, 2000–17: a prospective national observational cohort study. *Lancet Infect Dis* 2018; **18:** 441–51.
7 Ladhani SN, Andrews N, Ramsay ME. Summary of evidence to reduce the two-dose infant priming schedule to a single dose of the 13-valent pneumococcal conjugate vaccine in the national immunisation programme in the UK. *Lancet Infect Dis* 2021; **21:** e93–102.
8 Chattaway MA, Dallman TJ, Larkin L, et al. The transformation of reference microbiology methods and surveillance for *Salmonella* with the use of whole genome sequencing in England and Wales. *Front Public Health* 2019; **7:** 317.
9 Kapatai G, Sheppard CL, Al-Shahib A, et al. Whole genome sequencing of *Streptococcus pneumoniae*: development, evaluation and verification of targets for serogroup and serotype prediction using an automated pipeline. *PeerJ* 2016; **4:** e2477.
10 Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018; **34:** i142–50.
11 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; **25:** 1043–55.
12 Lees JA, Harris SR, Tonkin-Hill G, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res* 2019; **29:** 304–16.

13 Li Y, Metcalf BJ, Chochua S, et al. Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. *BMC Genomics* 2017; **18:** 621.

14 D'Aeth JC, van der Linden MPG, McGee L, et al. The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. *eLife* 2021; **10:** e67113.

15 Feldgarden M, Brover V, Haft DH, et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob Agents Chemother* 2019; **63:** e00483-19.

16 Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res* 2022; **50:** D912–17.

17 Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015; **43:** e15.

18 Gladstone RA, Lo SW, Lees JA, et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* 2019; **43:** 338–46.

19 Bertran M, D'Aeth JC, Abdullahi F, et al. Invasive pneumococcal disease 3 years after introduction of a reduced 1+1 infant 13-valent pneumococcal conjugate vaccine immunisation schedule in England: a prospective national observational surveillance study. *Lancet Infect Dis* 2024; **24:** 546–56.

20 Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS Comput Biol* 2018; **14:** e1005958.

21 Li Y, Metcalf BJ, Chochua S, et al. Penicillin-binding protein transpeptidase signatures for tracking and predicting β-lactam resistance levels in *Streptococcus pneumoniae*. *MBio* 2016; **7:** 7.

22 Lo SW, Gladstone RA, van Tonder AJ, et al. Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. *Lancet Infect Dis* 2019; **19:** 759–69.

23 Tin Tin Htar M, Morato Martínez J, Theilacker C, Schmitt H-J, Swerdlow D. Serotype evolution in western Europe: perspectives on invasive pneumococcal diseases (IPD). *Expert Rev Vaccines* 2019; **18:** 1145–55.

24 Hansen CB, Fuursted K, Valentiner-Branth P, Dalby T, Jørgensen CS, Slotved H-C. Molecular characterization and epidemiology of *Streptococcus pneumoniae* serotype 8 in Denmark. *BMC Infect Dis* 2021; **21:** 421.

25 Sanz JC, Rodríguez-Avial I, Ríos E, García-Comas L, Ordobás M, Cercenado E. Increase of serotype 8, ST53 clone, as the prevalent strain of *Streptococcus pneumoniae* causing invasive disease in Madrid, Spain (2012-2015). *Enferm Infecc Microbiol Clin* 2020; **38:** 105–10.

26 Müller A, Kleynhans J, de Gouveia L, et al. *Streptococcus pneumoniae* serotypes associated with death, South Africa, 2012–2018. *Emerg Infect Dis* 2022; **28:** 166–79.

27 de Miguel S, Domenech M, González-Camacho F, et al. Nationwide trends of invasive pneumococcal disease in Spain from 2009 through 2019 in children and adults during the pneumococcal conjugate vaccine era. *Clin Infect Dis* 2021; **73:** e3778–87.

28 Weiser JN, Ferreira DM, Paton JC. *Streptococcus pneumoniae*: transmission, colonization and invasion. *Nat Rev Microbiol* 2018; **16:** 355–67.

29 Løchen A, Truscott JE, Croucher NJ. Analysing pneumococcal invasiveness using Bayesian models of pathogen progression rates. *PLOS Comput Biol* 2022; **18:** e1009389.

30 Goldblatt D, Andrews NJ, Sheppard CL, et al. Pneumococcal carriage following PCV13 delivered as one primary and one booster dose (1 + 1) compared to two primary doses and a booster (2 + 1) in UK infants. *Vaccine* 2023; **41:** 3019–23.

31 Murray CJL, Ikuta KS, Sharara F, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* 2022; **399:** 629–55.

32 Cassini A, Högberg LD, Plachouras D, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis* 2019; **19:** 56–66.

33 Mitchell AM, Mitchell TJ. *Streptococcus pneumoniae*: virulence factors and variation. *Clin Microbiol Infect* 2010; **16:** 411–18.

34 Lees JA, Ferwerda B, Kremer PHC, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun* 2019; **10:** 2176.

35 European Centre for Disease Prevention and Control. Antimicrobial resistance in the EU/EEA (EARS-Net) - Annual Epidemiological Report 2019. Stockholm, 2020.

36 Dewé TCM, D'Aeth JC, Croucher NJ. Genomic epidemiology of penicillin-non-susceptible *Streptococcus pneumoniae*. *Microb Genom* 2019; **5:** 5.