nature communications

Article

https://doi.org/10.1038/s41467-024-52459-3

Impact of pneumococcal conjugate vaccines on invasive pneumococcal disease-causing lineages among South African children

Received: 6 February 2024

Accepted: 3 September 2024

Published online: 27 September 2024

Check for updates

Cebile Lekhuleni[1,2] ✉, Kedibone Ndlangisa[1], Rebecca A. Gladstone[3], Sopio Chochua[4], Benjamin J. Metcalf[4], Yuan Li[4], Jackie Kleynhans[1,5], Linda de Gouveia[1], Scott Hazelhurst[6,7], Ana D. S. Ferreira[8], Happy Skosana[1], Sibongile Walaza[1,5], Vanessa Quan[9], Susan Meiring[9], Paulina A. Hawkins[4], Lesley McGee[4], Stephen D. Bentley[8], Cheryl Cohen[1,5], Stephanie W. Lo[8,10], Anne von Gottberg[1,2,11] & Mignon du Plessis[1,2]

Invasive pneumococcal disease (IPD) due to non-vaccine serotypes after the introduction of pneumococcal conjugate vaccines (PCV) remains a global concern. This study used pathogen genomics to evaluate changes in invasive pneumococcal lineages before, during and after vaccine introduction in South Africa. We included genomes (N = 3104) of IPD isolates from individuals aged <18 years (2005–20), spanning four periods: pre-PCV, PCV7, early-PCV13, and late-PCV13. Significant incidence reductions occurred among vaccine-type lineages in the late-PCV13 period compared to the pre-PCV period. However, some vaccine-type lineages continued to cause invasive disease and showed increasing effective population size trends in the post-PCV era. A significant increase in lineage diversity was observed from the PCV7 period to the early-PCV13 period (Simpson's diversity index: 0.954, 95% confidence interval 0.948-0.961 vs 0.965, 0.962-0.969) supporting intervention-driven population structure perturbation. Increases in the prevalence of penicillin, erythromycin, and multidrug resistance were observed among non-vaccine serotypes in the late-PCV13 period compared to the pre-PCV period. In this work we highlight the importance of continued genomic surveillance to monitor disease-causing lineages post vaccination to support policy-making and future vaccine designs and considerations.

Globally, *Streptococcus pneumoniae* (pneumococcus) continues to cause invasive diseases such as bacteraemia and meningitis in children despite the routine use of pneumococcal conjugate vaccines (PCV) in many countries[1]. Seven-valent PCV (PCV7) and 13-valent PCV (PCV13) were introduced into the routine immunisation programme in South Africa in 2009 and 2011, respectively[2,3]. Subsequently, significant reductions in vaccine serotype invasive pneumococcal disease (IPD), particularly in children <2 years of age, were observed[3]. Sustained reductions in the same age group of 32.4 to 1.9 and 14.0 to 1.0 cases per 100,000 population for PCV7 and PCV13-unique serotypes, respectively, were reported by 2019[4]. Following vaccine introduction, non-vaccine serotypes 8, 12F, 15B/C, and 16F became the most common cause of invasive disease in children aged <15 years in the PCV13 era (2012–2018)[5].

Serotype replacement has been shown to occur through two mechanisms: (1) expansion of non-vaccine-type lineages following

A full list of affiliations appears at the end of the paper. ✉e-mail: cebilel@nicd.ac.za

vaccine-type lineage declines, and (2) expansion of non-vaccine serotype components within lineages that expressed both vaccine and non-vaccine serotypes; these lineages usually predominantly expressed vaccine serotypes before PCV introduction[6,7]. The Global Pneumococcal Sequencing Project (GPS) (https://www.pneumogen.net/gps/) was established in 2011 and aimed to understand the impact of vaccination on pneumococcal populations. The GPS project currently has a collection of over 20,000 genomes from 59 countries through which the standardised, genome-wide global pneumococcal sequence cluster (GPSC) typing scheme was established to assign pneumococcal lineages. Not only is the GPSC typing scheme concordant with traditional seven-loci multi-locus sequence typing (MLST), but this approach can also reveal shared descent among MLST sequence types (ST) and clonal complexes (CC)[8]. Previous studies from South Africa showed that some pneumococcal lineages and genotypes that cause IPD in our setting have rarely been reported elsewhere[9–11]. Such observations necessitate the need for ongoing spatiotemporal genomic surveillance and the inference of past population dynamics to reveal the demographic histories and growth patterns of important drivers of disease[12].

Settings with sustained PCV immunisation programmes have benefited from reduced antibiotic-nonsusceptible IPD[13,14]. However, the expansion of lineages expressing both vaccine and non-vaccine serotypes not only contributes to IPD persistence but may also threaten the maintenance of reduced antibiotic-nonsusceptible disease given the shared genetic background of resistance traits. Before PCV7 introduction in South Africa, globally disseminated pneumococcal molecular epidemiology network (PMEN) clones such as the multidrug-resistant (MDR) Denmark^{14}-32 (CC230), penicillin-nonsusceptible South Africa^{6B}-8 (ST185), and Sweden^{1}-27 (CC217) were widely circulating[10]; with the latter persisting through the early-PCV13 period[15]. Our study aimed to perform an in-depth analysis of the genomic population structure of IPD isolates from children younger than 18 years before and after the introduction of PCVs, identify predominant lineages causing disease after vaccine introduction and describe antimicrobial resistance (AMR) patterns in the post-PCV13 era.

In this work we show that although PCVs have substantially reduced IPD, lineages that expressed both vaccine and non-vaccine serotypes in the pre-PCV period have the potential to persist post vaccination due to the expansion of non-vaccine serotype constituents, along with AMR features. Monitoring persistent vaccine-type lineages together with vaccine-escape serotypes is useful for understanding the impact of vaccination and guiding prospective vaccine formulations.

Results

National invasive pneumococcal disease surveillance

Among the 54,199 IPD episodes reported during the 16-year study period (2005–2020), age was captured for 96% (51,917/54,199) of which 33% (17,121/51,917) were among children <18 years. Of these, 12,283 (72%) cultures were submitted to the National Institute for Communicable Diseases (NICD); 96% (11,833/12,283) had viable isolates available for sequencing. A total of 3141 (27%) viable isolates were sequenced.

Among the 3141 genomes, we determined concordance between the routine phenotypic serotyping results and in silico serotypes. Isolates reacting in the pool G antisera serogroup ($n = 32$) that could not be resolved to the serotype level phenotypically were distinguished in silico to serotype 29 ($n = 1$), 34 ($n = 4$), 35A ($n = 1$), 35B ($n = 19$), 35D ($n = 3$), and 35F ($n = 4$). Serogroup 16, 22, and 25 ($n = 5$) were resolved to serotype 16F ($n = 3$), 22F ($n = 1$), and 25F ($n = 1$), respectively. Among 64 of 3141 (2%) isolates that had discordant serotyping results, 27 isolates were viable and upon repeat testing of both phenotypic and in silico serotyping, all were concordant. The remaining 37 that could not be repeated due to loss of viability were excluded. Concordance of 99% (3071/3104) was achieved between phenotypic and in silico serotypes; 1% (33/3104) was concordant to the serogroup level. Finally, a total of 3104 isolates were included in the current study. Isolate characteristics are shown in Supplementary Table 1.

Serotype distribution

Among the 3104 genomes, we identified 57 serotypes and one unencapsulated isolate (Supplementary Data 1). Overall, the ten most predominant serotypes in our selection, in order of prevalence, were serotype 19F (244/3104, 8%), 6A (241/3104, 8%), 8 (231/3104, 7%), 14 (231/3104, 7%), 1 (228/3104, 7%), 23F (224/3104, 7%), 19A (217/3104, 7%), 6B (181/3104, 6%), 15B/C (127/3104, 4%), and 12F (117/3104, 4%; Supplementary Data 1). The serotype distribution among sequenced isolates was consistent with the serotype distribution among all 11,789 IPD cases with known phenotypic serotype data from our IPD surveillance during the same period (Supplementary Figs. 1 and 2).

Before the introduction of PCVs in South Africa, IPD in children was largely caused by PCV7 serotypes 14 (122/934, 13%), 23F (112/934, 12%), 6B (95/934, 10%), and PCV13 unique serotypes 6A (108/934, 12%) and 1 (105/934, 11%). Similar serotypes predominated after the introduction of PCV7. Changes in serotype predominance were observed during the early-PCV13 period where PCV13 serotypes 19A (81/891, 9%), 1 (67/891, 8%), 23F (58/891, 7%), 6A (57/891, 6%), and non-vaccine serotype 8 (58/891, 7%) became the five most common serotypes causing disease. In the late-PCV13 period, disease-causing serotypes were largely non-vaccine serotypes, including serotype 8 (154/855, 18%), 12F (50/855, 6%), 15B/C (46/855, 5%), 16F (40/855, 5%), and PCV7 serotype 19F (66/855, 8%). Similar serotypes circulated among the <5 and 5–17-year-olds before and after PCV introduction (Supplementary Fig. 3). Serotype 8 was the leading cause of IPD in children <5 years in the late-PCV13 period accounting for 20% (146 of 739) of invasive disease whilst serotype 19F was the most common cause of IPD in the 5–17-year-olds with a frequency of 14% (14 of 102) in the same period. Overall, we observed a strong association (Cramer's $V = 0.5126$) between PCV periods (pre-PCV, PCV7, early-PCV13, and late-PCV13) and the frequency of observing vaccine and non-vaccine serotypes (Supplementary Table 2). Vaccine and non-vaccine serotype frequencies in this collection reflected similar trends in our surveillance database across all four PCV periods (Supplementary Table 3).

Genomic population structure and lineage-specific incidence changes

Of the 3104 isolates in our study population, 3098 (99.8) were classified into 152 existing GPSC lineages and 3032 (98%) into 629 STs. We identified six new GPSCs all of which occurred in the late-PCV13 period. Among the 152 GPSCs, 22% ($n = 33$), 22% ($n = 33$), 23% ($n = 35$), and 34% ($n = 51$) expressed only vaccine serotypes, only non-vaccine serotypes, both vaccine and non-vaccine serotypes, and rare lineages, respectively. Lineages expressing both vaccine and non-vaccine serotypes accounted for 55% (1719/3104) of the isolates in our collection. The ten most common lineages in this study accounted for 50% (1561/3104) of the isolates (Fig. 1). An interactive visualisation of lineage data is available online (https://microreact.org/ZA_snp_analysis). The most commonly expressed serotypes among the ten predominant lineages were: GPSC1 (serotype 19F, $n = 100/104$ [96%]), GPSC2 (serotype 1, 226/226 [100%]), GPSC3 (serotype 8, 220/255 [86%]), GPSC5 (serotype 35B, 60/142 [42%]), GPSC10 (serotype 14, 120/155 [77%]), GPSC13 (serotype 6A, 118/126 [94%]), GPSC14 (serotype 23F, 153/167 [92%]), GPSC17 (serotype 19A, 199/202 [99%]), GPSC21 (serotype 19F, 90/91 [99%]), and GPSC41 (serotype 6A, 88/93 [95%]) and were those largely included in the PCV13 formulation. With reference to the GPS database, GPSC17, 21, and 41 were more prevalent among African isolates at frequencies of 92% (349/380), 100% (275/275), and 100% (129/129), respectively (Supplementary Table 4).

Among children <5 years, the average incidence of vaccine-type lineages decreased significantly between the pre-PCV and PCV7

https://doi.org/10.1038/s41467-024-52459-3

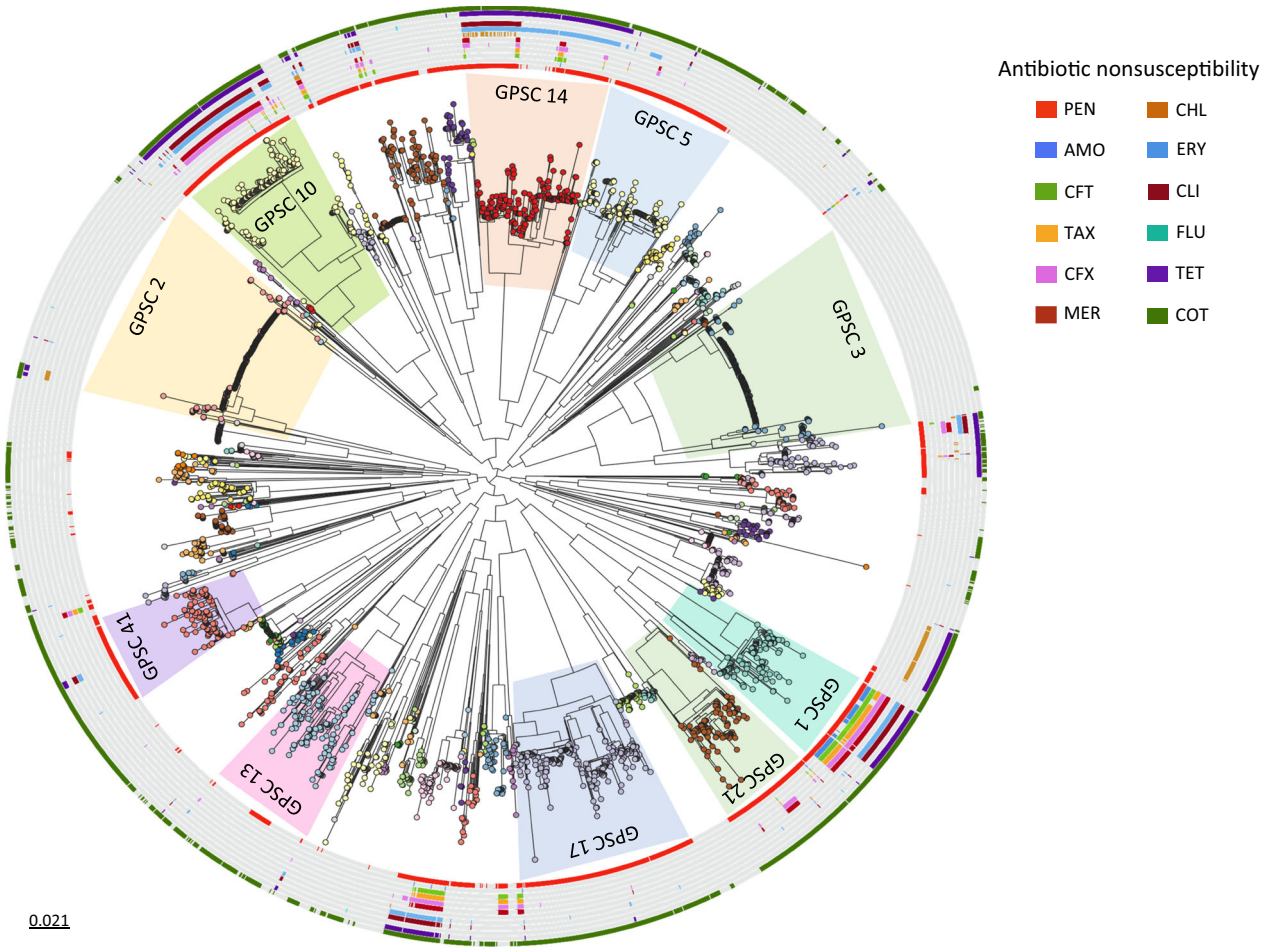


Fig. 1 | Single nucleotide polymorphism (SNP) maximum-likelihood phylogeny (*N* = 3104) of invasive pneumococcal disease isolates from South African children <18 years. The SNP tree is midpoint rooted. The 10 most predominant GPSC taxa are highlighted and listed in descending order: GPSC3 (*n* = 255/3104 [8%], GPSC2 (*n* = 226/3104 [7%], GPSC17 (*n* = 202/3104 [7%], GPSC14 (*n* = 167/3104 [5%], GPSC10 (*n* = 155/3104 [5%], GPSC5 (*n* = 142/3104 [5%], GPSC13 (*n* = 126/3104 [4%], GPSC1 (*n* = 104/3104 [3%], GPSC41 (*n* = 93/3104 [3%], and GPSC21 (*n* = 91/ 3104 [3%]. Antibiotic nonsceptibility by GPSC lineage is represented by coloured circular strips: PEN penicillin, AMO amoxicillin, CFT ceftriaxone, TAX cefotaxime, CFX cefuroxime, MER meropenem, CHL chloramphenicol, ERY erythromycin, CLI clindamycin, FLU fluoroquinolones, TET tetracycline, COT cotrimoxazole, GPSC global pneumococcal sequence clusters.

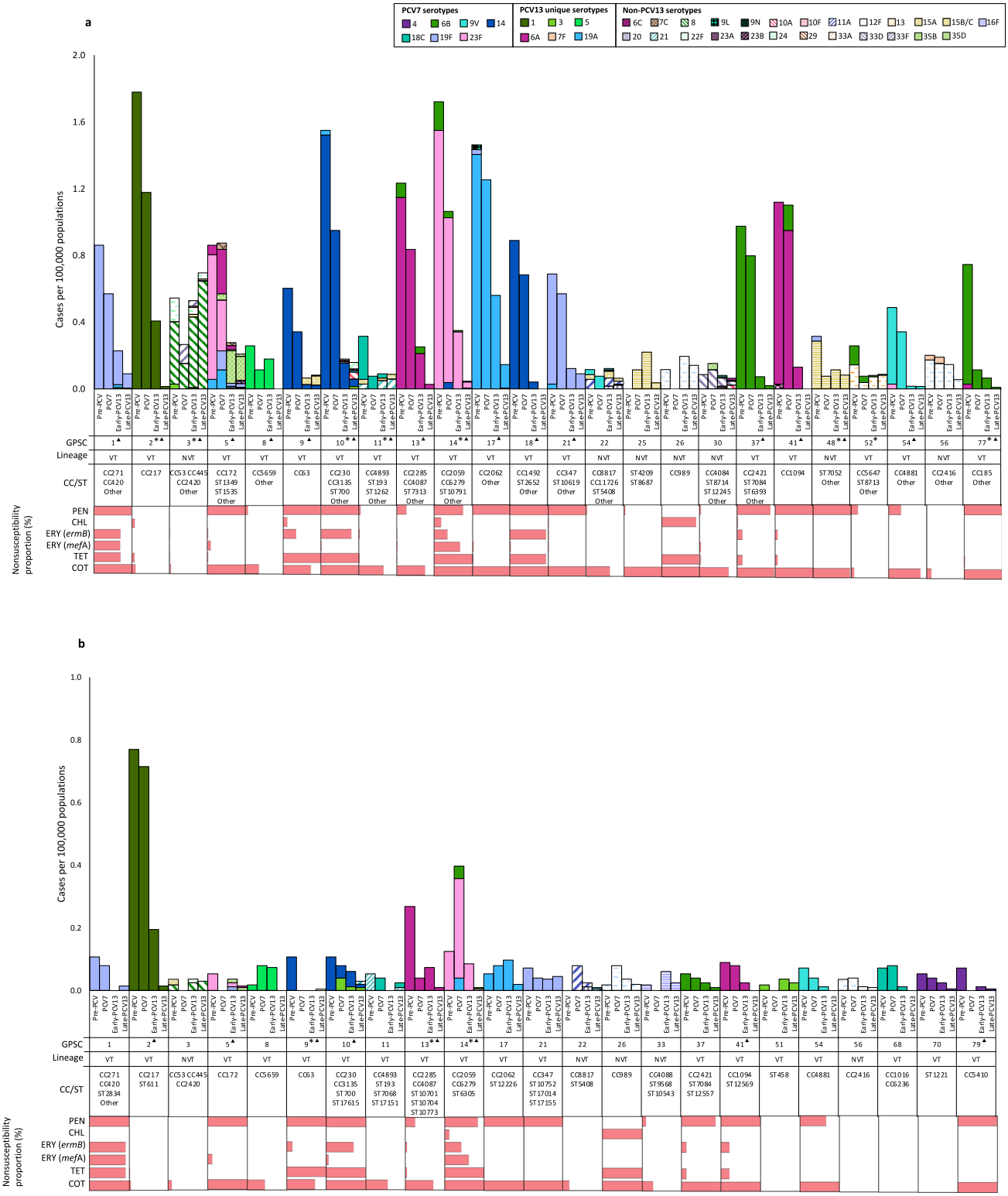
periods for GPSC2 (incidence rate ratio [IRR] 0.7, 95% confidence interval [CI] 0.9–0.5, $p = 0.04$), GPSC10 (0.6, 0.9–0.5, $p = 0.03$), GPSC11 (0.3, 0.7–0.1, $p = 0.03$), GPSC14 (0.7, 0.9–0.5, $p = 0.03$), GPSC52 (0.3, 0.8–0.1, $p = 0.04$), and GPSC77 (0.2, 0.4–0.09, $p < 0.001$, Fig. 2a and Supplementary Data 2). Similar to the <5-year-olds, the incidence of GPSC9 (0.05, 0.4–0.007, $p < 0.001$) and GPSC13 (0.2, 0.4–0.08, $p < 0.001$) declined significantly in the 5–17 year age group (Fig. 2b and Supplementary Data 3). Comparing the pre-PCV and late-PCV13 periods, both age groups showed similar incidence reduction patterns for vaccine-type lineages (Fig. 2a, b; Supplementary Data 4 and 5). For non-vaccine-type lineages, reductions in invasive disease caused by GPSC3 (0.4, 0.8–0.3, $p = 0.03$) and GPSC48 (0.3, 0.6–0.1, $p = 0.03$) were observed when comparing the pre-PCV and PCV7 periods among the <5 year age group. In contrast, significant increases in IPD caused by GPSC3 (1.8, 1.2–2.7, $p = 0.008$) were observed in the same age group when comparing the pre-PCV and late-PCV13 periods (Fig. 2a).

Before PCV introduction GPSC5, 9, 10 and 11 predominantly expressed PCV13 serotypes. After PCV introduction, especially in the early- and late-PCV13 periods, these lineages largely expressed non-PCV13 serotypes (Fig. 2). GPSC5 (CC172) largely expressed vaccine serotype 23F (*n* = 29/33, 88%) with no observation of non-vaccine serotype 35B in the pre-PCV period in this collection. In the late-PCV13 period, replacement serotype 35B (CC172) was the commonly expressed serotype within this lineage (34/49, 69%) with no observation of serotype 23F (Supplementary Fig. 4). Similarly, both GPSC9 (CC63) and GPSC10 (CC230) expressed vaccine serotype 14 (27/27, 100% and 59/60, 98%) in the pre-PCV and PCV7 periods (9/9, 100%; 26/27, 96%), respectively (Supplementary Figs. 5 and 6). However, in the late-PCV13 period, GPSC9 largely expressed non-vaccine serotype 15A (13/19, 68%); GPSC10 expressed a combined proportion of 49% (20/41) of non-vaccine serotype 10A (11/41, 27%) and serogroup 24 (9/41, 22%), in addition to serotype 14 (12/41, 29%, Supplementary Figs. 5 and 6). Similar patterns were observed for GPSC11 (Supplementary Fig. 7).

Phylodynamic analysis

We observed significant temporal signals ($p < 0.001$) for the persistent vaccine-type lineages of interest (Supplementary Data 6) that continued to circulate in the late-PCV13 period, allowing for coalescent analysis. Effective sample sizes (ESS) of greater than 200 were also obtained for these lineages. Past population dynamics for GPSC5 (CC172, serotype 35B) and GPSC17 (CC2062, serotype 19A) had comparable temporal trends with steady increases in the effective population size in the years after 1990 to the end of the study period in 2020 (Fig. 3a, d). This observation was supported by clonal expansion findings where the time-calibrated phylogeny for GPSC5 showed the highest probability for clonal expansion in 1992 (credible interval: 1981–1998). Similarly, GPSC17 had several clonal expansion events with

https://doi.org/10.1038/s41467-024-52459-3



the earliest expansion occurring in 1953 (1936-1965) followed by several others that coincided with the increasing effective population size in 1989 (1982–1994) and 1991 (1984–1996, Supplementary Fig. 8a, d). Although the effective population sizes for GPSC9 (serotype 14 and 15A) and GPSC10 (serotype 14, serogroup 24) showed similar increases around the 1990s, a decreasing trend in the effective population size for GPSC9 was observed before the introduction of PCV7 followed by a slight increase after PCV13 introduction (Fig. 3b). For GPSC10, however, this decrease coincided with the introduction of PCV7 in 2009 (Fig. 3c). Clonal expansion analysis for GPSC9 and GPSC10 revealed that the vaccine and non-vaccine-serotype variants in these lineages diverged before PCV introduction with non-vaccine serotype 15A prominently expanding in 2005 (2001–2009), shortly before PCV7 introduction (Fig. 4a, b).

Simpson's diversity index trends

We further explored serotype and lineage diversity indices which were generally high across the study period, ranging

https://doi.org/10.1038/s41467-024-52459-3

Fig. 2 | Global Pneumococcal Sequence Cluster (GPSC) dynamics among invasive disease isolates from children in South Africa (*N* = 3104). a children <5 years (*n* = 2680) and **b** children 5–17 years (*n* = 424). Invasive pneumococcal disease incidence per 100,000 population is plotted by GPSC lineages stratified into four vaccine periods (pre-PCV [2005–2008], PCV7 [2009–2010], early-PCV13 [2011–2014], and late-PCV13 [2015–2020]). Serotypes included in the 13-valent pneumococcal conjugate vaccine are shown by solid colour fills and non-vaccine serotypes by hatched pattern colour fills. Lineages with <40 isolates are not shown. A vaccine-type (VT) lineage was defined as having ≥50% vaccine serotype(s) in the pre-PCV period; non-vaccine-type (NVT) lineage as having >50% non-vaccine serotype(s) in the pre-PCV period. Lineage-specific clonal complexes (CC) and sequence types (ST) are shown underneath the graph. Either Poisson regression or negative binomial regression was used to calculate the incidence rate ratios (IRR) of GPSCs by vaccine period. If neither model fit, the IRR was calculated using the average annual incidence (Supplementary Data 2–5). Significant changes in the average incidence between the pre-PCV and PCV7 or late-PCV13 periods are represented by an asterisk or triangle, respectively. Significance was determined at <0.05 using a two-sided *p*-value where applicable. Multiple testing (>10 tests) was corrected using the Benjamini–Hochberg false discovery rate of 5%. Antibiotic nonsusceptibility proportions by lineage are shown in red horizontal bars. Penicillin (PEN) nonsusceptibility was predicted based on the penicillin-binding protein types: *pbp*1a, *pbp*2x, *pbp*2b; chloramphenicol (CHL), based on the presence of chloramphenicol acetyltransferase gene, *cat*; erythromycin (ERY), based on the presence of erythromycin resistance methylase gene *ermB* or macrolide efflux pump gene *mefA*; tetracycline (TET), by the presence of *tet*M or *tet*S/M gene with no promoter region interruptions; cotrimoxazole (COT), by the presence of mutation I100L in *folA* and/or indel within amino acid residue 56–67 in *folP*.

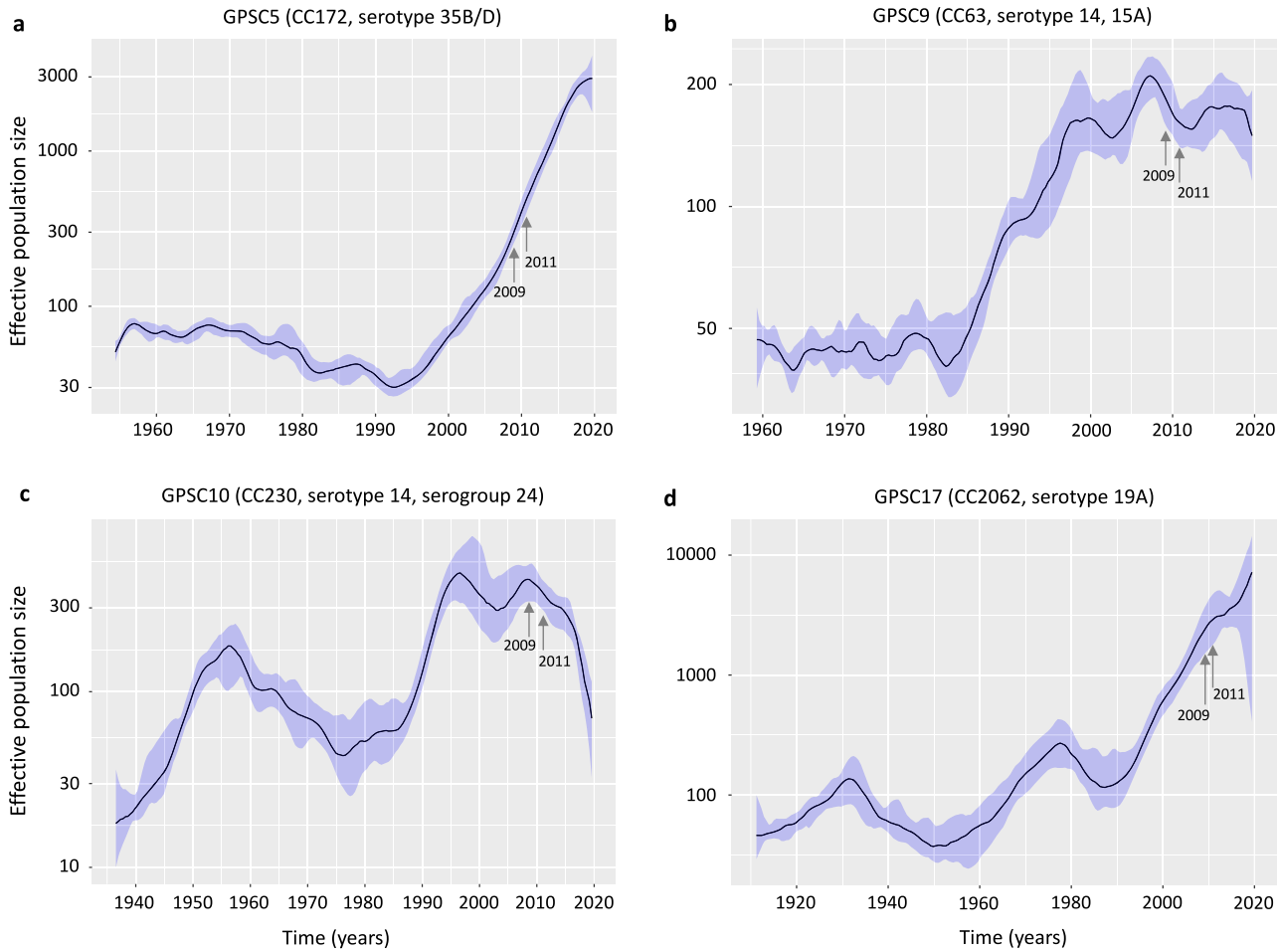


Fig. 3 | Temporal trend analysis of important vaccine-type lineages causing persistent invasive disease in the late-PCV13 period inferred using skygrowth. a–d Past population size dynamics as a function of time inclusive of the pneumococcal conjugate vaccine (PCV) introduction years shown by the grey arrows (PCV7 in 2009 and PCV13 in 2011) in South Africa. Time-calibrated phylogenies were used for inference with root dates as start time. Root probabilities and the number of analysed genomes per lineage are shown in Supplementary Data 6. The clonal complexes (CC) and serotypes expressed by each lineage are shown in each panel title. The solid line represents the posterior median population size. The shaded purple area represents the 95% credible intervals. The effective population size on the y-axis is expressed in log scale. GPSC global pneumococcal sequence cluster. *Only serotypes that had 5 or more genomes are included in the (**a**–**d**) panel titles.

between 0.903–0.993 and 0.917–0.971, respectively (Fig. 5a, c). Compared to the pre-PCV and PCV7 years, significant increases in serotype diversity were observed in the early-PCV13 years, with the highest increase occurring in the PCV13 introduction year, 2011 (Simpson's diversity index [SDI] 0.993, 95% CI 0.991–0.995). Declining serotype diversity trends were evident when comparing 2011 through to the late-PCV13 years, with diversity approaching pre-PCV estimates (Fig. 5a). However, significant increases in lineage diversity were observed between 2011 and 2013 (0.948, 0.937–0.960 vs 0.971, 0.963–0.979, Fig. 5c). Pooled lineage diversity trends reflected those of pooled serotype diversity trends both with significant differences when comparing the PCV7 and early-PCV13 periods (pooled lineage SDI 0.954, 95% CI 0.948–0.961 vs 0.965, 0.962–0.969 and pooled serotype SDI 0.915, 95% CI 0.907–0.924 vs 0.953, 0.950–0.957). Similarly, significant differences were also observed when comparing the

https://doi.org/10.1038/s41467-024-52459-3

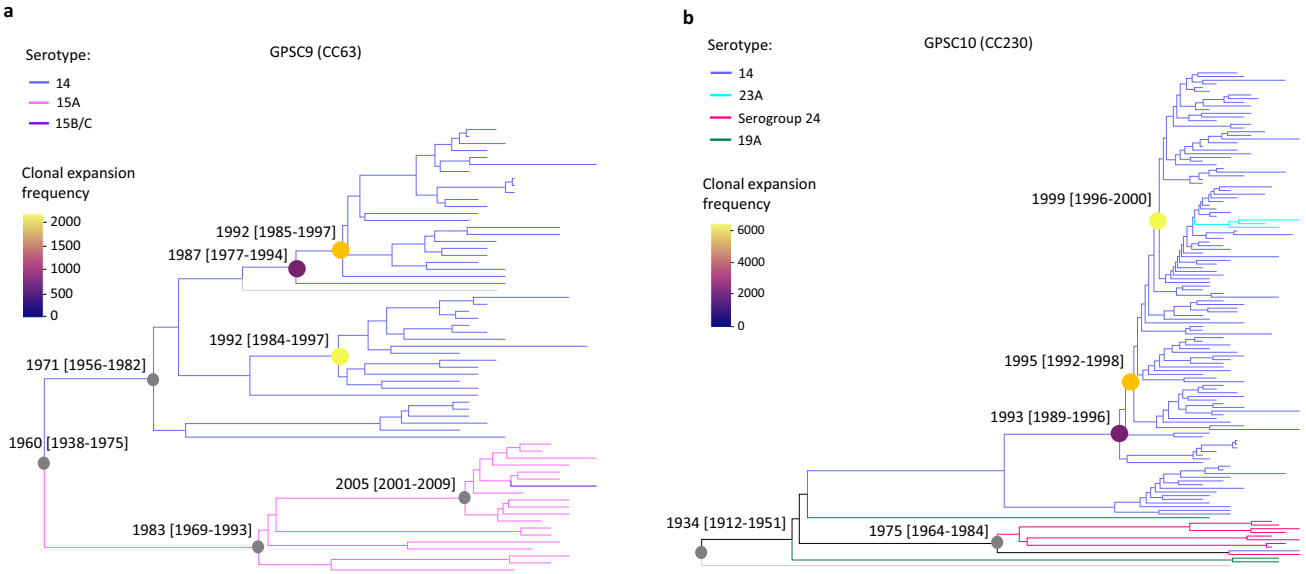


Fig. 4 | Clonal expansion dynamics of Global Pneumococcal Sequence Cluster (GPSC) lineages inferred using CaveDive. Time-resolved phylogenies for GPSC9 (**a**) and GPSC10 (**b**) showing within lineage clonal expansions with clades coloured by serotype. GPSC references are coloured in grey. The parental nodes with high probabilities for clonal expansions are highlighted by the clonal expansion frequency colour scale and the estimated expansion year with 95% credible intervals is included. Other parental nodes are highlighted in grey including the root date. GPSC global pneumococcal sequence cluster, CC clonal complex.

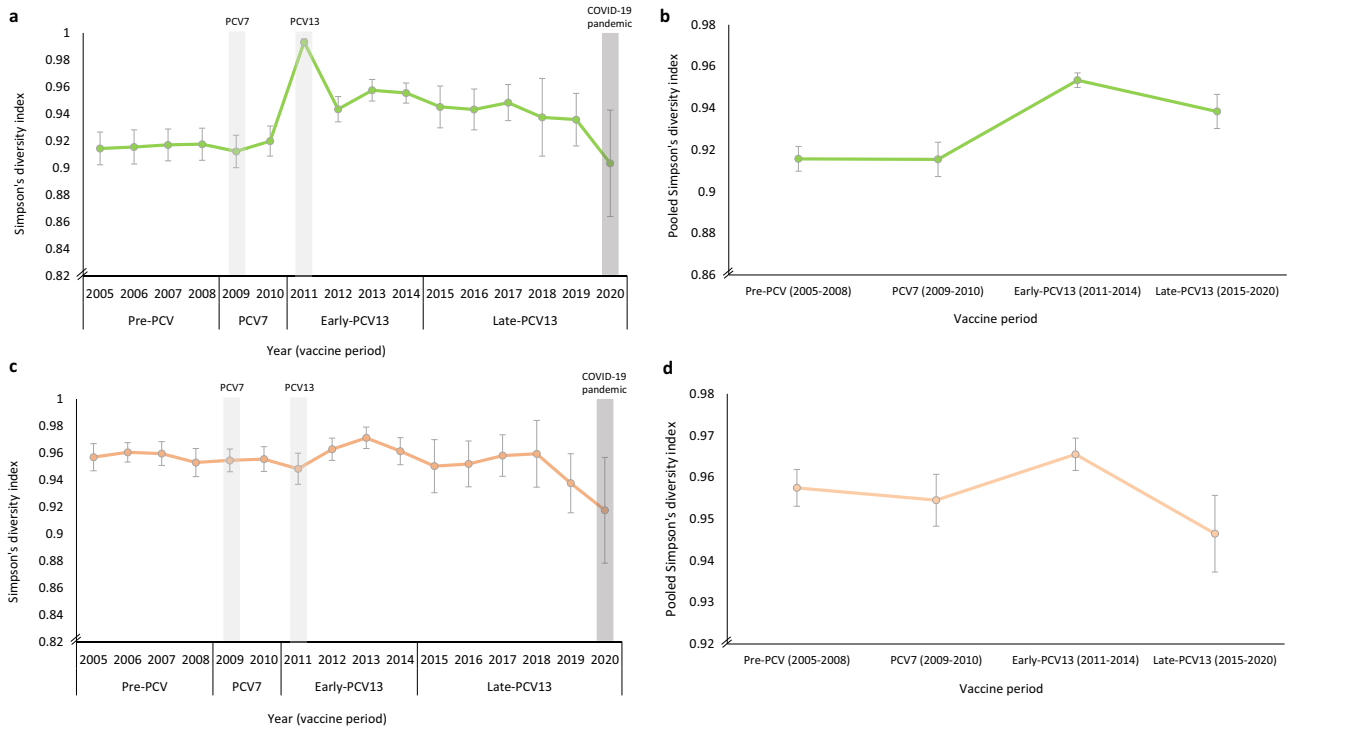


Fig. 5 | Simpson's diversity index (SDI) estimates for invasive disease serotypes and lineages. a Serotype diversity trends by year showing fluctuations before and after the introduction of 7- and 13-valent pneumococcal conjugate vaccines (PCV7 and PCV13) in 2009 and 2011, respectively (highlighted in light grey); the COVID-19 pandemic year (2020) is highlighted in dark grey. **b** Pooled serotype SDI estimates by vaccine period showing statistically significant differences between the PCV7 and early-PCV13 periods and early-PCV13 and late-PCV13 periods. **c** Lineage diversity trends by year before and after the introduction of PCV7 and PCV13 in 2009 and 2011, respectively (highlighted in light grey); the COVID-19 pandemic year (2020) is highlighted in dark grey. **d** Pooled lineage SDI estimates by vaccine period showing statistically significant differences between the PCV7 and early-PCV13 period and early-PCV13 and late-PCV13 periods. Estimates were derived at a 95% confidence interval among invasive disease isolates ($N = 3104$) from children <18 years in South Africa during the pre-PCV ($n = 924$), PCV7 ($n = 457$), early-PCV13 ($n = 882$), and late-PCV13 ($n = 841$) periods. SDI estimates represent the measure of diversity with standard error included at each data point.

early- and late-PCV13 periods (0.965, 0.962–0.969 vs 0.946, 0.937–0.956 and 0.953, 0.950–0.957 vs 0.938, 0.930–0.947, Fig. 5b, d).

Antimicrobial resistance

AMR concordances comparing phenotypic antimicrobial susceptibility testing (AST) and in silico AMR predictions are shown in Supplementary Table 5. Among the 3104 IPD isolates, 2058 (66%) had at least one resistance determinant for the antibiotics investigated (Fig. 1). Overall, reductions in the prevalence of predicted antibiotic resistance among all classes of antibiotics, except chloramphenicol, were observed when comparing the pre-PCV and late-PCV13 periods (Table 1). In contrast, predicted antibiotic resistance to penicillin, erythromycin and MDR increased significantly in invasive disease caused by non-vaccine serotypes (Table 1). Among the 1249 non-vaccine serotypes, 232 (19%) had PBP gene combinations that conferred penicillin resistance. Of these, serotype 35B (GPSC5, $n = 60$), 15B/C (GPSC48, $n = 46$), 15A (GPSC9/GPSC168, $n = 40$), 10A (GPSC10, $n = 12$), 23B (GPSC102, $n = 12$) and serogroup 24 (GPSC10, $n = 9$) accounted for 77% (179/232) of the isolates and these serotypes were among the predominant non-vaccine serotypes circulating in the late-PCV13 period (Supplementary Figs. 1 and 2). Erythromycin resistance was identified in 50 (4%) of 1249 non-vaccine serotype isolates with *ermB*-mediated resistance occurring in 35 (70%) of 50 cases compared to *mefA*-mediated resistance. Of the 120 MDR non-vaccine serotype isolates, 73 (61%) were from the late-PCV13 period (Table 1). The three most common lineages driving MDR increase were GPSC9 (serotype 15A; 13/73, 18%), GPSC10

Table 1 | Changes in the prevalence of antibiotic nonsusceptibility among South African children, comparing the pre-PCV and late-PCV13 periods

Antibiotic Resistance mechanism	Pre-PCV[a] (2005–2008)	PCV7 (2009–2010)	Early-PCV13 (2011–2014)	Late-PCV13[a] (2015–2020)	Adjusted *p*-value Linear regression[b]
Overall (*n*)	924	457	882	841	
Penicillin	507 (55%)	266 (58%)	345 (39%)	293 (35%)	<0.0001
pbp:1a–2b–2x[c]	507 (55%)	266 (58%)	345 (39%)	293 (35%)	
Erythromycin	203 (22%)	112 (25%)	118 (13%)	86 (10%)	<0.0001
ermB or *ermB/mefA*	151 (16%)	81 (18%)	75 (9%)	66 (8%)	
mefA	52 (6%)	31 (7%)	43 (5%)	20 (2%)	
Clindamycin	151 (16%)	81 (18%)	75 (9%)	66 (8%)	<0.0001
ermB or *ermB/mefA*	151 (16%)	81 (18%)	75 (9%)	66 (8%)	
Chloramphenicol	29 (3%)	14 (3%)	46 (5%)	35 (4%)	0.27
cat	29 (3%)	14 (3%)	46 (5%)	35 (4%)	
Tetracycline	246 (27%)	119 (26%)	159 (18%)	137 (16%)	<0.0001
*tet*M	200 (22%)	102 (22%)	139 (16%)	131 (16%)	
*tet*S/M	46 (5%)	17 (4%)	20 (2%)	6 (1%)	
Cotrimoxazole	637 (69%)	338 (74%)	537 (61%)	479 (57%)	<0.0001
*folA*_I100L	17 (2%)	6 (1%)	7 (1%)	4 (0%)	
*folP*_aa_insert_57–70	170 (18%)	73 (16%)	162 (18%)	152 (18%)	
*folA*_I100L, *folP*_aa_insert_57-70	450 (49%)	259 (57%)	368 (42%)	323 (38%)	
Multidrug resistance[d]	239 (26%)	126 (28%)	157 (18%)	133 (16%)	<0.0001
Non-vaccine serotypes only (*n*)	143	60	415	631	
Penicillin	17 (12%)	5 (8%)	70 (17%)	140 (22%)	0.0066
pbp:1a–2b–2x[c]	17 (12%)	5 (8%)	70 (17%)	140 (22%)	
Erythromycin	1 (1%)	0 (0%)	12 (3%)	37 (6%)	0.032
ermB or *ermB/mefA*	1 (1%)	0 (0%)	7 (2%)	27 (4%)	
mefA	0 (0%)	0 (0%)	5 (1%)	10 (2%)	
Clindamycin	1 (1%)	0 (0%)	7 (2%)	27 (4%)	0.07
ermB or *ermB/mefA*	1 (1%)	0 (0%)	7 (2%)	27 (4%)	
Chloramphenicol	5 (3%)	2 (3%)	30 (7%)	31 (5%)	0.49
cat	5 (3%)	2 (3%)	30 (7%)	31 (5%)	
Tetracycline	8 (6%)	3 (5%)	41 (10%)	77 (12%)	0.026
*tet*M	8 (6%)	3 (5%)	41 (10%)	75 (12%)	
*tet*S/M	0 (0%)	0 (0%)	0 (0%)	2 (0.3%)	
Cotrimoxazole	65 (45%)	26 (43%)	212 (51%)	315 (50%)	0.35
*folA*_I100L	1 (1%)	0 (0%)	1 (0%)	4 (1%)	
*folP*_aa_insert_57–70	36 (25%)	16 (27%)	94 (23%)	123 (19%)	
*folA*_I100L, *folP*_aa_insert_57–70	28 (20%)	10 (17%)	117 (28%)	188 (30%)	
Multidrug resistance[d]	6 (4%)	2 (3%)	39 (9%)	73 (12%)	0.012

[a]Vaccine periods of interest for generalised linear regression analysis.
[b]A generalised linear regression model to detect a correlation of significance between the introduction of PCV and changes in antibiotic nonsusceptibility prevalence, comparing the pre-PCV and late-PCV13 periods.
[c]Mutations in *pbp*1a, *pbp*2b, *pbp*2x, or a combination of mutations that confer penicillin nonsusceptibility.
[d]Multidrug resistance was defined as nonsusceptibility to three or more classes of antibiotics.

(serotype 10A; 11/73, 15% and serogroup 24; 9/73, 12%), and GPSC26 (serotype 12F, 30/73, 41%, Supplementary Table 6). GPSC9 and GPSC10 were both resistant to penicillin, tetracycline, and cotrimoxazole, with GPSC10 additionally resistant to erythromycin. All of serogroup 24 MDR isolates were from the late-PCV13 period in this study.

Discussion

Our study captures the genomic population structure and temporal trends among invasive pneumococci in South African children across a 16-year period which spans pre- and post-PCV eras. We highlight significant reductions in globally disseminated vaccine-type GPSC lineages following the introduction of PCV7 and PCV13. We show that the persistence of IPD after PCV13 introduction is, in part, attributed to both between and within lineage expansion of non-vaccine serotypes. Through diversity analysis, we show supporting evidence of pneumococcal population perturbation after vaccine introduction which reached equilibrium 11 years after the first use of PCV in 2009. We also show that although the prevalence of antibiotic nonsusceptibility decreased overall, significant increases in nonsusceptibility to penicillin, erythromycin, and MDR were observed in IPD due to non-vaccine serotypes expressed by both vaccine- and non-vaccine-type lineages in the late-PCV13 period.

The decline in lineages predominantly expressing PCV13 serotypes is consistent with earlier epidemiological trend studies evaluating the impact of PCVs in South Africa, the USA and the UK[3,4,16–19]. However, similar to South Africa, vaccine-type lineages such GPSC5, 9, 10, and 17 continue to cause invasive disease in the USA and the UK[7,11,20–22]. The GPS project database of 21,155 genomes from 59 countries revealed that three of the ten most common disease-causing lineages (GPSC17, 21 and 41) among South African children are endemic to the African continent, each with a contribution of at least 90% in the GPS collection (https://www.pneumogen.net/gps/gps-database-overview/). These lineages, particularly GPSC17 and 21 are important drivers of residual vaccine-serotype disease in South Africa expressing vaccine serotypes 19A and 19F, respectively[23]. Serotype 19A is a genetically diverse serotype that was a major cause of IPD in children in the USA after PCV7 introduction, and earlier studies identified antimicrobial-resistant emergent genotypes that expanded due to positive antibiotic selective pressure[24,25]. For example, the MDR CC320/271 clone outcompeted the antimicrobial susceptible clone CC199[25]. Similar findings were reported in Canada[23], and some Asian countries[26]. By contrast, serotype 19A isolates from South Africa were historically represented by the CC2062 (GPSC17) clone[10], and this remains the predominant clone after PCV introduction. Similarly, the persistence of serotype 3 IPD was driven by ST458 (GPSC51), a clone that is predominant in South Africa but rarely reported elsewhere[9,10]. GPSC17 (CC2062), 21 (CC347), and 51 (CC458) require continued genomic surveillance in South Africa and other African countries as these lineages express vaccine-escape serotypes[19,22]. The persistence of serotype 3, however, was not surprising as PCV13 has been shown to have reduced vaccine effectiveness against this serotype[19]. Serotype 3 continues to have a steady incidence in South Africa particularly in the elderly where serotype replacement was evident[4]. As expected, vaccine introduction in South Africa was accompanied by significant increases in serotype diversity in the early-PCV13 period followed by declines towards baseline trends in the late-PCV13 period. Pooled serotype and lineage diversity trends followed similar patterns supporting the phenomenon of expected increases in the diversity of the pneumococcal population following the perturbation of genetic composition due to vaccine introduction[27,28].

Apart from residual vaccine-serotype IPD, non-vaccine serotypes play an important role in the persistence of invasive disease in children globally in the post-PCV era. Genomic surveillance of all serotypes is useful for understanding the genetic background of lineages driving persistence of non-vaccine serotype disease. We observed non-vaccine serotype IPD due to both within and between lineage expansions. Among the ten most common lineages in our collection, non-vaccine-type GPSC3 (predominantly expressing serotype 8) substantially contributed to IPD in children after vaccine introduction. Before PCV7 introduction in South Africa, serotype 8 was the second most common non-vaccine serotype causing IPD in children aged <2 years (after serotype 15B/C)[10]. After PCV13 introduction, serotype 8 became the most common cause of IPD in both children and adults[5], moving from a frequency of 6% in the early-PCV13 period to 17% in the late-PCV13 period (data not shown). In England and Wales, serotype 8 IPD frequency increased substantially and accounted for 20% of disease after PCV13 introduction[29]. Non-vaccine serotype 12F (predominantly expressed by GPSC26 and 56), 15B/C (GPSC25 and 48), and 16F (GPSC33 and 46) were also frequently isolated from children with IPD in the current study after PCV13 introduction. Previous data from South Africa showed that serotype 15B/C was significantly associated with increased in-hospital deaths and serotype 16F was most commonly isolated among HIV-infected children[5,16]. Serotype 8, 12F and 15B/C are included in higher valency vaccines PCV20 and PCV24. The vaccine IVT-25 covers in addition serotype 16F with a trade-off of PCV13 serotype 6A.

Previous studies have reported on evidence of shared genetic backgrounds among isolates of different pneumococcal serotypes[8,30]. We observed within lineage non-vaccine serotype expansions among GPSC5, 9, 10 and 11. The expansion of non-vaccine serotype 35B within GPSC5 in the late-PCV13 period in this study coincided with earlier findings where serotype 35B increases were accompanied by a four-fold increase in penicillin nonsusceptibility in children aged <5 years[11]. According to the global antimicrobial use data, the use of broad-spectrum penicillins in South Africa increased substantially from a defined daily dose (DDD) of 2023 per 1000 population in the year 2000, to 3794 DDD in 2015 (https://resistancemap.onehealthtrust.org/). It is, therefore, reasonable to suggest that antibiotic selective pressure has contributed, in part, to the persistence of penicillin-nonsusceptible lineages. Similar patterns of penicillin-nonsusceptible lineages were observed in Nepal and France for GPSC9 and GPSC10, respectively[21,31]. The historic population dynamics of GPSC9 and GPSC10 together with GPSC5 (serotype 35B) and GPSC17 (serotype 19A) suggest that penicillin nonsusceptibility likely promoted the expansion of these lineages in the past. Declines were observed in the effective population sizes of these lineages around the time when penicillin was widely used in children to treat streptococcal infections in the late 1950s (https://resistancemap.onehealthtrust.org/Timeline), but this was followed by an exponential increase as penicillin resistance emerged. In South Africa, pneumococcal infections are mainly treated with penicillin or ampicillin. The persistence of lineages that are nonsusceptible to these antibiotics may pose challenges in IPD management in our healthcare facilities.

The expansion and extinction of GPSC lineages underlie the trends in serotype incidence and AMR through the AMR genes these lineages harbour in disease. Identifying lineages that express non-vaccine serotypes that are increasing in incidence may be an early indication of which non-vaccine serotypes may be of concern in disease, particularly those that carry resistance genes against multiple antibiotic classes and require continuous monitoring. Before vaccine introduction in South Africa (2000–2005), PCV7 serotypes plus serotype 6A accounted for 94% of erythromycin-nonsusceptible isolates, with serotype 14 accounting for 40% of invasive disease[32]. These findings correspond with the high *ermB* prevalence among serotype 14 isolates expressed by the GPSC10 (CC230) lineage, and 15A isolates expressed by GPSC9 (CC63) in the current study. Macrolide-resistant GPSC10 (CC230, serotype 14 or Denmark^{14}-32) is a MDR lineage that continues to cause disease in South Africa in the post-PCV13 era and is responsible for the global spread of MDR serotype 24F[21]. Although the

expansion of non-vaccine serotype 15A (GPSC9) and serogroup 24 (GPSC10) after PCV introduction in South Africa does not outweigh the reduction of vaccine serotype 14 expressed by both lineages, the time-calibrated phylogenies of these lineages suggest that these non-vaccine serotypes diverged long before PCV introduction and successfully propagated to current times when expansion was facilitated by the removal of vaccine serotype 14. The high prevalence of cotrimoxazole resistance in this study is not surprising since this antibiotic is widely used as prophylaxis in HIV seropositive patients in South Africa[33].

This study included approximately 20% of viable IPD isolates from individuals aged <18 years, with sampling bias towards younger children aged ≤2 years. Thus, our findings may not accurately capture the vaccine impact among older children, especially the 5–17 year olds. In addition, sampling limits our ability to confidently identify all emerging genotypes. Therefore, we may have missed detecting earlier genomic changes which may lead to an underestimation of circulating lineages. Our AMR findings may also not fully reflect resistance trends due to the usage of a sampled study population. The dichotomisation of vaccine-type and non-vaccine-type lineages may be sensitive, partly due to geotemporal differences in serotype distributions. Nonetheless, the impact of vaccination may be more accurately reflected in the ≤2 years as this is the target age group for vaccination given the high disease burden. Our study collection was a snapshot of the Group for Enteric, Respiratory and Meningeal Disease Surveillance in South Africa (GERMS-SA) database across all four vaccine periods, hence the expectation that drawn conclusions closely reflect those of the surveillance programme. Although not fully assessed, we hypothesise that there is no systematic bias for cases diagnosed by PCR alone. The use of the standardised GPSC lineage typing scheme to evaluate vaccine impact proved useful as it offers global comparability at the genome level. Due to the high concordance among phenotypic and in silico predictions, the use of whole-genome sequencing (WGS) to enrich our surveillance through the routine sequencing of IPD isolates or culture-negative specimens is a possibility in our setting going forward.

In conclusion, pneumococcal genomic surveillance provides a powerful means to evaluate the impact of interventions with respect to genomic characteristics. In South Africa, starting in 2024, PCV13 will be replaced by the more affordable SII PCV10 (Serum Institute of India, Pneumosil), which excludes serotypes 3, 4, and 18C in the formulation. Although serotypes 4 and 18C were less important in infants and children before PCV introduction, they were, however important in adult populations that will no longer benefit from the indirect protective effect against these serotypes. Serotype 3 remained largely unaffected by PCV13. Genomic surveillance in all age groups will be particularly important in our setting to monitor the impact of switching to a lower valency vaccine as well as monitoring non-vaccine serotype fluctuations, such as serotype 8. Collaborative genomic epidemiology efforts within the African continent are useful for the in-depth evaluation of endemic lineages that are contributing to both residual vaccine serotype and non-vaccine serotype disease. Not only will this effort improve geographical representation, it may also contribute towards future vaccine designs and inform policy making within the African countries.

Methods

National invasive pneumococcal disease surveillance and sample selection

Pneumococcal isolates were obtained as part of GERMS-SA, a national, laboratory-based surveillance programme for IPD that commenced in 1999 and includes all nine South African provinces[34]. Over 200 participating primary diagnostic laboratories voluntarily submit patient demographic data, clinical specimens, and/or *S. pneumoniae* isolates to the reference laboratory at the NICD. Quarterly audits are performed for the public sector to assess the completeness of reported cases, with unreported cases added as audit cases. Pneumococcal disease syndromes that do not routinely have specimens taken, such as pneumonia, are placed under ascertainment. An IPD case was defined as the detection of *S. pneumoniae* from a normally sterile-site specimen (e.g., blood, cerebrospinal fluid, joint fluid, pleural fluid)[5]. Viable isolates routinely undergo phenotypic characterisation, as previously described[11] (Supplementary Methods) and are cryopreserved at −70 °C. Ethical approval for national IPD surveillance was obtained through the Human Ethics Research Committee (medical) at the University of the Witwatersrand, Johannesburg, South Africa (protocol no. M1809107), and ethical clearance was also obtained for this study (protocol No. M210749).

From 2005 through 2020, 54,199 IPD episodes were reported, of which 93% (50,468/54,199) were culture-positive at the primary diagnostic laboratory, and 72% (36,549/50,468) had viable isolates available for further characterisation. For WGS, stratified random sampling was used to select approximately 300 isolates per year within predetermined age categories: ~150/300 (50%) isolates from children aged 0–2 years and ~75/300 (~25%) isolates each from individuals aged 3–5 years and >5 years. Only age stratification was used. The selection was purposively biased toward the <2 years age category due to the higher disease burden in this age group that is targeted for vaccination. Among 4834 sequenced isolates from all ages, 3141 genomes were from children <18 years of age, of which 3104 (99%) were included in this study.

Whole-genome sequencing and processing

WGS was performed at the Wellcome Sanger Institute (Hinxton, UK) using Illumina platforms (either HiSeq or NovaSeq), and at the NICD Sequencing Core Facility (Johannesburg, South Africa) using the NextSeq Illumina sequencer (Illumina, San Diego, CA, USA). Genomic data were analysed as previously described[8]. Briefly, paired-end reads were assembled and annotated using the Sanger pipeline (Supplementary Methods). Raw data were deposited into the European Nucleotide Archive (Supplementary Dataset 1). In silico serotype and ST were inferred using SeroBA, version 1.0.1[35], and MLSTcheck, version 2.1.1706216[36]. respectively. GoeBURST housed in PHYLOViZ (version 2.0) was used to assign CCs[37]. A CC was defined as a group of related STs sharing at least six of the seven alleles with the founder ST; a founder was defined as the ST with the highest number of single-locus variants (SLV) in the CC group.

Species-wide phylogenetic analysis

Species-wide analysis was performed by mapping sequencing reads from each isolate against the *S. pneumoniae* reference genome ATCC 700669[38], (NCBI accession code FM211187) using Smalt (version 0.7.4). The SNP-sites tool was used to reduce the pseudo-genome alignment to variant sites[39]. The resulting SNP alignment was used to construct a maximum-likelihood phylogenetic tree using the generalised time-reversible (GTR) model in FastTree2[40]. Microreact was used for visualisation[41].

Lineage-specific analyses

GPSCs were determined using PopPUNK (https://poppunk.net/) version 2.4.0[42], and the GPS reference database version 6 (https://www.pneumogen.net/gps/assigningGPSCs.html). GPSC-specific phylogenies were constructed by mapping Illumina reads of isolates forming a lineage against previously prepared lineage-specific references using Burrow-Wheeler Alignment (BWA) v0.7.17-r1188[8,43]. Gubbins was used to identify recombination blocks and to create recombination-free phylogenies using the GTR model in RAxML[44,45]; Microreact and Phandango were used for visualisation of phylogenies and recombination, respectively[41,46].

Dating and phylodynamic analysis

The BactDating R package was used to determine the presence of a temporal signal of persistent vaccine-type lineages, including those that were predominantly expressing non-vaccine serotypes in the late-PCV13 period using recombination-free phylogenies to perform root-to-tip regression[47]. BactDating was ran with three replicates and one equal tip date through 100 million MCMC chains sampled every 1000 states with a burn-in of 10 million using the arc model with a coalescent prior for the Bayesian inference of ancestral dates[47]. The three replicate MCMC chains showed convergence when analysed with the Gelman diagnostic of approximately 1 for alpha, sigma and mu using the coda R package[48]. We assessed the temporal signal by using the modelcompare function of BactDating to compare the first replicate model to a model with equal isolate dates but ran under the same parameters. We then used the effectiveSize function of the coda R package to evaluate whether the ESS on the first replicate model was more than 200. Time-calibrated phylogenies were used as input for the skygrowth R package to determine past effective population sizes of the lineages of interest using a Bayesian MCMC approach that ran for 100,000 iterations, allowing for population size estimations of approximately 50 time points prior to the start of our sampling date[49]. We then used the CaveDive R package to test for evidence of within-lineage clonal expansion, and the model was ran for 100 million MCMC iterations[50].

In silico antimicrobial resistance determinant analyses

In addition to phenotypic AST, AMR was also predicted from the genome. Sequencing data for each isolate was run through the CDC pneumococcal typing pipeline for the detection of resistance determinants and in silico minimum inhibitory concentration predictions (MICs, Supplementary Methods)[51]. Predicted MDR was defined as resistance to three or more of the following antibiotic classes: beta-lactams (penicillin), macrolides (erythromycin), lincosamides (clindamycin), chloramphenicol, tetracyclines (tetracycline), and sulphonamides (trimethoprim-sulfamethoxazole or cotrimoxazole). This study reports AMR data predicted from the genome.

Statistical analysis

PCV periods were defined as pre-PCV (2005–2008), PCV7 (2009–2010), early-PCV13 (2011–2014), and late-PCV13 (2015–2020). A vaccine-type lineage was defined as having ≥50% PCV13 serotypes, and a non-vaccine-type lineage as having >50% non-PCV13 serotypes in the pre-PCV period, as previously described[7]. Rare lineages were defined as GPSCs that were neither designated vaccine-type, nor non-vaccine-type and did not have a large enough genome count to be assigned. The chi-squared test was used to evaluate the frequency of vaccine and non-vaccine serotypes between the pre-PCV and late-PCV13 periods. Cramer's *V* was used to measure the association between PCV periods (pre-PCV, PCV7, early-PCV13, and late-PCV13) and the frequency of vaccine and non-vaccine serotypes. The study population was corrected for sampling framework by year and by age group (<5 years and 5–17 years) using selected sample proportions against the GERMS-SA database. The overall incidence rate of lineages was imputed by assuming that the proportion of lineages by year and age group as determined through WGS was similar to that of the total number of IPD cases identified through surveillance for each year and age group. Poisson regression was used to calculate IRRs at 95% CI of imputed individual GPSC lineages using the IPD incidence rate estimated per year for the three vaccine periods (PCV7, early-PCV13, and late-PCV13), compared to pre-PCV period. To correct for overdispersion, if observed, the robust standard error for a minor violation of the Poisson regression assumption measured by goodness-of-fit at *p*-value ranging between 0.01 and 0.05 was reported, or the negative binomial regression measured at Poisson goodness-of-fit *p*-value < 0.01 was used. If none of the models fit, IRR was calculated using the average annual incidence for the three vaccine periods. To avoid IRR equating to zero or infinity, a constant number of one was added to all estimated IPD cases for when a GPSC was not observed in any of the three vaccine periods. Statistical significance was at <0.05 significance level using a two-sided *p*-value. Benjamini–Hochberg false discovery rate of 5% was used to correct for multiple testing when number of tests exceeded ten. Lineage and serotype diversity over time were analysed using Simpson's diversity index 1-D (SDI) with values ranging from 0 to 1; higher values indicate higher diversity. We pooled both serotype and lineage diversity indices for comparing the overall diversity patterns by PCV period between and within serotype and lineage levels. All statistical analyses were performed in R version 4.2.1. R scripts were adapted from a public repository (https://github.com/rgladstone/GPSCs/tree/master/lo_et_al).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The authors confirm that all relevant data supporting the findings of this study are included within the paper and its supplementary information files. The sequencing reads for the isolates used in this study are available in the European Nucleotide Archive and the isolate accession numbers and metadata are provided in the Supplementary Dataset 1 and on FigShare.

Code availability

All the data analysis tools and methods used in this study are publicly available and are clearly described in the Methods section and the supplementary files. Derivative R scripts are available at https://github.com/cebile-lekh/pcv_impact_manuscript[52].

References

1. Savulescu, C. et al. Effect of high-valency pneumococcal conjugate vaccines on invasive pneumococcal disease in children in SpIDnet countries: an observational multicentre study. *Lancet Respir. Med.* **5**, 648–656 (2017).
2. Madhi, S. A., Cohen, C. & von Gottberg, A. Introduction of pneumococcal conjugate vaccine into the public immunization program in South Africa: translating research into policy. *Vaccine* **30**, C21–C27 (2012).
3. von Gottberg, A. et al. Effects of vaccination on invasive pneumococcal disease in South Africa. *Vaccine* **371**, 4200–4208 (2014).
4. von Gottberg, A. et al. Long-Term Impact of Pneumococcal Conjugate Vaccines from National, Active, Laboratory-Based Surveillance of Invasive Pneumococcal Disease, South Africa. https://doi.org/10.2139/SSRN.4519527 (2005–2019).
5. Müller, A. et al. Streptococcus pneumoniae serotypes associated with death, South Africa, 2012-2018. *Emerg. Infect. Dis.* **28**, 166–179 (2022).
6. Chochua, S. et al. Invasive serotype 35B pneumococci including an expanding serotype switch lineage, United States, 2015–2016. *Emerg. Infect. Dis.* **23**, 922–930 (2017).
7. Lo, S. W. 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.* **19**, 759–769 (2019).
8. Gladstone, R. A. et al. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* **43**, 338–346 (2019).
9. Mothibeli, K. M. et al. An unusual pneumococcal sequence type is the predominant cause of serotype 3 invasive disease in South Africa. *J. Clin. Microbiol.* **48**, 184–191 (2010).
10. Ndlangisa, K. M. et al. Population snapshot of streptococcus pneumoniae causing invasive disease in South Africa prior to

introduction of pneumococcal conjugate vaccines. *PLoS ONE* **9**, 1–11 (2014).
11. Ndlangisa, K. M. et al. A Streptococcus pneumoniae lineage usually associated with pneumococcal conjugate vaccine (PCV) serotypes is the most common cause of serotype 35B invasive disease in South Africa, following routine use of PCV. *Microb. Genom.* **8**, 1–10 (2022).
12. Ho, S. Y. W. & Shapiro, B. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* **11**, 423–434 (2011).
13. Golden, A. R. et al. Comparison of antimicrobial resistance patterns in *Streptococcus pneumoniae* from respiratory and blood cultures in Canadian hospitals from 2007-16. *J. Antimicrob. Chemother.* **74**, IV39–IV47 (2019).
14. Kaur, R., Pham, M., Yu, K. O. A. & Pichichero, M. E. Rising pneumococcal antibiotic resistance in the post-13-valent pneumococcal conjugate vaccine era in pediatric isolates from a primary care setting. *Clin. Infect. Dis.* **72**, 797–805 (2021).
15. du Plessis, M. et al. Phylogenetic analysis of invasive serotype 1 pneumococcus in South Africa, 1989 to 2013. *J. Clin. Microbiol.* **54**, 1326–1334 (2016).
16. Cohen, C. et al. Effectiveness of the 13-valent pneumococcal conjugate vaccine against invasive pneumococcal disease in South African children: a case-control study. *Lancet Glob. Health* **5**, e359–e369 (2017).
17. Whitney, C. G. et al. Decline in invasive pneumococcal disease after the introduction of protein–polysaccharide conjugate vaccine. *N. Engl. J. Med.* **348**, 1737–1746 (2003).
18. Moore, M. R. 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.* **15**, 301–309 (2015).
19. Andrews, N. J. et al. Serotype-specific effectiveness and correlates of protection for the 13-valent pneumococcal conjugate vaccine: a postlicensure indirect cohort study. *Lancet Infect. Dis.* **14**, 839–846 (2014).
20. Hawkins, P. A. et al. A global genomic perspective on the multidrug-resistant Streptococcus pneumoniae 15A-CC63 sub-lineage following pneumococcal conjugate vaccine introduction. *Microb. Genom.* **9**, mgen000998 (2023).
21. Lo, S. W. et al. Emergence of a multidrug resistant and virulent Streptococcus pneumoniae lineage mediates serotype replacement after PCV13. *medRxiv* https://doi.org/10.1101/2021.11.24.21266813 (2021).
22. Corcoran, M. et al. Using genomics to examine the persistence of Streptococcus pneumoniae serotype 19A in Ireland and the emergence of a sub-clade associated with vaccine failures. *Vaccine* **39**, 5064–5073 (2021).
23. Pillai, D. R. et al. Genome-wide dissection of globally emergent multi-drug resistant serotype 19A *Streptococcus pneumoniae*. *BMC Genomics* **10**, 1–13 (2009).
24. Hanage, W. P. et al. Clonal replacement among 19A Streptococcus pneumoniae in Massachusetts, prior to 13 valent conjugate vaccination. *Vaccine* **29**, 8877–8881 (2011).
25. Beall, B. W. et al. Shifting genetic structure of invasive serotype 19A pneumococci in the United States. *J. Infect. Dis.* **203**, 1360–1368 (2011).
26. Shin, J., Baek, J. Y., Kim, S. H., Song, J. H. & Ko, K. S. Predominance of ST320 among *Streptococcus pneumoniae* serotype 19A isolates from 10 Asian countries. *J. Antimicrob. Chemother.* **66**, 1001–1004 (2011).
27. Hanage, W. P. et al. Evidence that pneumococcal serotype replacement in Massachusetts following conjugate vaccination is now complete. *Epidemics* **2**, 80–84 (2010).
28. Corander, J. et al. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat. Ecol. Evol.* **1**, 1950–1960 (2017).
29. Ladhani, S. N. 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.* **18**, 441–451 (2018).
30. Croucher, N. J. et al. Evidence for soft selective sweeps in the evolution of pneumococcal multidrug resistance and vaccine escape. *Genome Biol. Evol.* **6**, 1589–1602 (2014).
31. Kandasamy, R. et al. Effect of childhood vaccination and antibiotic use on pneumococcal populations and genome-wide associations with disease among children in Nepal: an observational study. *Lancet Microbe* **3**, e503–e511 (2022).
32. Wolter, N., von Gottberg, A., du Plessis, M., de Gouveia, L. & Klugman, K. P. Molecular basis and clonal nature of increasing pneumococcal macrolide resistance in South Africa, 2000-2005. *Int. J. Antimicrob. Agents* **32**, 62–67 (2008).
33. Soeters, H. M., Von Gottberg, A., Cohen, C., Quan, V. & Klugman, K. P. Trimethoprim-sulfamethoxazole prophylaxis and antibiotic non-susceptibility in invasive pneumococcal disease. *Antimicrob. Agents Chemother.* **56**, 1602–1605 (2012).
34. Huebner, R. E. et al. Laboratory surveillance for Haemophilus influenzae type B meningococcal, and pneumococcal disease. *South Afr. Med. J.* **89**, 924–925 (1999).
35. Epping, L. et al. SeroBA: Rapid high-throughput serotyping of Streptococcus pneumoniae from whole genome sequence data. *Microb. Genom.* **4**, 1–6 (2018).
36. Page, A. J., Taylor, B. & Keane, A. J. Multilocus sequence typing by blast from de novo assemblies against PubMLST. *J. Open Source Softw.* **1**, 118 (2016).
37. Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* **186**, 1518–1530 (2004).
38. Croucher, N. J. et al. Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae* Spain23F ST81. *J. Bacteriol.* **191**, 1480–1489 (2009).
39. Page, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056 (2016).
40. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
41. Argimón, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genom.* **2**, e000093 (2016).
42. Lees, J. A. et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
44. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
45. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
46. Hadfield, J. et al. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* **34**, 292–293 (2018).
47. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).
48. Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *journal.r-project.org R news* (2006).

49. Volz, E. M. & Didelot, X. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Syst. Biol.* **67**, 719–728 (2018).
50. Helekal, D., Ledda, A., Volz, E., Wyllie, D. & Didelot, X. Bayesian inference of clonal expansions in a dated phylogeny. *Syst. Biol.* **71**, 1073 (2022).
51. Metcalf, B. J. et al. Strain features and distributions in pneumococci from children with invasive disease before and after 13-valent conjugate vaccine implementation in the USA. *Clin. Microbiol. Infect.* **22**, 60.e9–60.e29 (2016).
52. Lekhuleni, C. Impact of pneumococcal conjugate vaccines on invasive pneumococcal disease-causing lineages among South African children. *GitHub* https://doi.org/10.5281/zenodo.13284763 (2024).

Acknowledgements

This work was supported in part by a Fogarty International Center Global Infectious Disease research training grant, National Institutes of Health, to the University of Pittsburgh and National Institute for Communicable Diseases (D43TW011255), through AvG. Genome sequencing was funded by the Bill & Melinda Gates Foundation (grant code OPP1189062), through SDB, and Wellcome Trust (grant reference 206194) through the Global Pneumococcal Sequencing project, SDB, and the SEQAFRICA project which is funded by the Department of Health and Social Care's Fleming Fund using UK aid to AvG. The views expressed in this publication are those of the authors and not necessarily those of the UK Department of Health and Social Care or its Management Agent, Mott MacDonald. We would like to thank the CRDM surveillance team at the NICD for assisting with the retrieval of isolates, phenotypic characterisation, and managing and cleaning up the GERMS-SA database. We would like to thank the Global Pneumococcal Sequencing Consortium for their collaborative efforts and intellectual input to the current work. Thank you to Hsueh-Chien Raymond Cheng for assisting with the phylodynamics work of the paper. We would also like to thank members of the sequencing facility and Pathogen Informatics teams at Wellcome Sanger Institute for informatics support. We also thank the Sequencing Core Facility at the NICD for the generation of additional genomic data.

Author contributions

C.L., M.d.P., A.v.G., S.W.L., and K.N. conceptualised the study. C.L. did the literature review, performed data analysis, produced figures and tables, and wrote the paper. M.d.P., A.v.G., S.W.L., R.A.G., K.N., S.C., B.J.M., Y.L., and J.K. contributed to data analysis and interpretation. L.d.G., H.S., A.v.G., C.C., S.M., S.W., V.Q., S.W.L., S.D.B., A.D.S.F., L.M., and P.A.H. contributed to data collection. S.W.L., A.D.S.F., and S.H. provided software and computational support. All authors reviewed and/or edited the paper. C.L. and S.W.L. had access to raw data. C.L., M.d.P., A.v.G., and K.N. made the final decision to submit the paper. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Competing interests

A.v.G. reports grants from NIH (training grant), US CDC, SEQAFRICA/Fleming Fund, and chairperson of the NITAG for SA (NAGI: National Advisory Group for Immunisation). C.C. reports grants from the US CDC, BMGF, Wellcome, SA MRC, Sanofi Pasteur, and PATH outside the submitted work. J.K. reports support for meetings and travel from NIH, outside the submitted work. R.A.G. reports a grant from Pfizer outside the submitted work. S.WL reports the Robert Austrian Research Award, sponsored by Pfizer, outside the submitted work. S.M. reports EMGuidance Technologies educational event/presentation, non-paid consultant for MSD, and non-paid educational lectures for Sanofi, outside the submitted work. V.Q. reports MSD support for meetings and travel, outside the submitted work. S.W. reports grants from US CDC and BMGF, outside the submitted work. The remaining authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52459-3.

Correspondence and requests for materials should be addressed to Cebile Lekhuleni.

Peer review information *Nature Communications* thanks Cheryl Andam, Palle Valentiner-Branth and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases, a division of the National Health Laboratory Service, Johannesburg, South Africa. [2]School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [3]Department of Biostatistics, University of Oslo, Oslo, Norway. [4]Division of Bacterial Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA. [5]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [6]School of Electrical & Information Engineering, University of the Witwatersrand, Johannesburg, South Africa. [7]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [8]Parasites and Microbes Programme, The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [9]Division of Public Health Surveillance and Response, National Institute for Communicable Diseases, A division of the National Health Laboratory Service, Johannesburg, South Africa. [10]Milner Centre for Evolution, Department of Life Sciences, University of Bath, Bath, UK. [11]Division of Medical Microbiology, Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa. ✉e-mail: cebilel@nicd.ac.za