Ten-year longitudinal molecular epidemiology study of *Escherichia coli* and *Klebsiella* species bloodstream infections in Oxfordshire, UK

Samuel Lipworth¹,²,³*, Karina-Doris Vihta¹, Kevin Chau¹, Leanne Barker¹, Sophie George¹, James Kavanagh¹, Timothy Davies¹,², Alison Vaughan¹, Monique Andersson², Katie Jeffery², Sarah Oakley², Marcus Morgan², Susan Hopkins⁴, Timothy E. A. Peto¹,²,³,⁵, Derrick W. Crook¹,²,⁵,⁶, Ann Sarah Walker¹,⁵,⁶† and Nicole Stoesser¹,²†

Abstract

**Background:** The incidence of Gram-negative bloodstream infections (BSIs), predominantly caused by *Escherichia coli* and *Klebsiella* species, continues to increase; however, the causes of this are unclear and effective interventions are therefore hard to design.

**Methods:** In this study, we sequenced 3468 unselected isolates over a decade in Oxfordshire (UK) and linked this data to routinely collected electronic healthcare records and mandatory surveillance reports. We annotated genomes for clinically relevant genes, contrasting the distribution of these within and between species, and compared incidence trends over time using stacked negative binomial regression.

**Results:** We demonstrate that the observed increases in *E. coli* incidence were not driven by the success of one or more sequence types (STs); instead, four STs continue to dominate a stable population structure, with no evidence of adaptation to hospital/community settings. Conversely in *Klebsiella* spp., most infections are caused by sporadic STs with the exception of a local drug-resistant outbreak strain (ST490). Virulence elements are highly structured by ST in *E. coli* but not *Klebsiella* spp. where they occur in a diverse spectrum of STs and equally across healthcare and community settings. Most clinically hypervirulent (i.e. community-onset) *Klebsiella* BSIs have no known acquired virulence loci. Finally, we demonstrate a diverse but largely genus-restricted mobilome with close associations between antimicrobial resistance (AMR) genes and insertion sequences but not typically specific plasmid replicon types, consistent with the dissemination of AMR genes being highly contingent on smaller mobile genetic elements (MGEs).

* Correspondence: Samuel.lipworth@ndm.ox.ac.uk
† Ann Sarah Walker and Nicole Stoesser are joint senior authors.
¹ Nuffield Department of Medicine, University of Oxford, Oxford, UK
² Oxford University Hospitals NHS Foundation Trust, Oxford, UK
³ Oxford University Hospitals NHS Foundation Trust, Oxford, UK
⁴ Full list of author information is available at the end of the article

© The Author(s). 2021 Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.
Conclusions: Our large genomic study highlights distinct differences in the molecular epidemiology of *E. coli* and *Klebsiella* BSIs and suggests that no single specific pathogen genetic factors (e.g. AMR/virulence genes/sequence type) are likely contributing to the increasing incidence of BSI overall, that association with AMR genes in *E. coli* is a contributor to the increasing number of *E. coli* BSIs, and that more attention should be given to AMR gene associations with non-plasmid MGEs to try and understand horizontal gene transfer networks.

Keywords: Gram-negative bloodstream infections, Bacteraemia, Whole genome sequencing, *Klebsiella pneumoniae*, Virulence, Antimicrobial resistance

Background
Gram-negative bloodstream infections (GNBSIs), predominantly caused by *Escherichia coli* and *Klebsiella* spp., are a significant and increasing threat to public health. They are now the leading cause of bloodstream infection (BSI) in the UK with a substantial associated burden of morbidity and mortality [1, 2]. Despite becoming a significant public health concern and policy focus, their incidence continues to increase and it is unclear how ambitious targets to reduce this can be achieved [3].

There is a significant association between GNBSIs and antimicrobial resistance (AMR), particularly in certain globally successful multi-locus sequence types (STs), such as *E. coli* ST131 [4, 5]. Enterobacteriaceae have relatively open pan-genomes and are able to rapidly adapt to changing selection pressures (including antibiotic usage) [6, 7]. Multidrug resistance (MDR)-associated STs have been linked with prolonged hospital stay and adverse outcomes [8]. Whilst infections caused by relatively susceptible isolates still represent the majority of cases, the potential for rapid proliferation of AMR-associated clones and the dissemination of AMR genes on mobile genetic elements between lineages and species is a major concern [9].

Recent molecular epidemiology studies in the UK have replicated global findings that most *E. coli* BSIs are caused by STs 131, 95, 73 (all phylogroup B2) and ST 69 (phylogroup D) [7, 10]. One study has shown that after the emergence and expansion of ST69 and ST131 in the early 2000s, the population structure reached an equilibrium, with STs 131, 95, 73, and 69 predominating [11]. However, this study only sequenced isolates cultured prior to 2012, which may represent a critical inflection point in the incidence rate [1], which continues to increase year on year. It remains unclear whether the population structure remains in equilibrium or whether one or more STs are responsible for this increase.

For *Klebsiella* spp., it has been hypothesised that isolates broadly cause two categories of BSI [12, 13], namely, (i) multidrug resistant healthcare-associated (HA) infections caused by low virulence strains with AMR gene-associated plasmids and (ii) clinically hypervirulent community-associated (CA) infections caused by strains carrying virulence gene plasmids. The convergence of these two phenotypes might pose a significant risk to human health [14]. Whilst there are multiple studies describing in detail the epidemiology of globally distributed clones associated with AMR and hyper-virulence [5, 15], such isolates represent a minority of BSI in the UK and most of Europe [16]. A recent study in Cambridgeshire (UK) of 162 *Klebsiella pneumoniae* isolates, enriched to over-represent MDR isolates, demonstrated a predominance of globally important STs with apparent cycling of relative incidence over a 2-year period [17]. However, MDR infections represent approximately 18% of *K. pneumoniae* BSIs in England and the molecular epidemiology of most invasive disease caused by *Klebsiella* spp. has not been systematically studied [2].

In order to investigate underlying potential pathogen genetic factors driving the increasing incidence of these infections, we analysed all *E. coli* and *Klebsiella* spp. BSIs over a decade (2008–2018) in Oxfordshire, UK, linking electronic health records and mandatory reporting data with sequencing data for all isolates, to give a detailed picture of the regional genomic epidemiology of *E. coli* and *Klebsiella* spp. BSIs. We used other publicly available sequencing datasets to contextualise our findings.

Methods
Study setting, laboratory procedures, and DNA extraction
All isolates causing BSI between September 15, 2008, and December 01, 2018 (de-duplicated to one BSI isolate per 90-day period per patient), were processed by the clinical microbiology laboratory at the John Radcliffe Hospital, Oxford, UK, using standard operating procedures, with sub-cultured stocks stored at −80°C in 10% glycerol nutrient broth. Blood cultures processed by the laboratory were taken as part of routine clinical workup. We attempted to sequence all isolates with no pre-selection according to, e.g. in vitro antibiotic susceptibility, patient factors, and date of acquisition; we hereby refer to this as an “unselected” sampling frame. 3461 *E. coli* isolates and 886 *Klebsiella* spp. isolates were successfully sequenced. All isolates were included in all parts of the relevant analysis except where otherwise indicated below. The microbiology laboratory serves all four hospitals and all community healthcare facilities within Oxfordshire, with a catchment population of
~805,000. For sequencing, isolate stocks were grown overnight on Columbia blood agar at 37°C and DNA was extracted using the QuickGene DNA extraction kit (Autogen, MA, USA) as per the manufacturer’s instructions with the addition of a mechanical lysis step (FastPrep, MP Biomedicals, CA, USA; 6m/s for 40 s). Short-read sequencing was performed using the Illumina HiSeq 2500/3000/4000/MiSeq instruments as previously described [18]. Additional, publicly available short-read sequencing data was obtained to facilitate phylogenetic comparison with previous studies as follows: 436 E. coli BSI isolates from Sydney, Australia (January 2013 to March 2016, PRJNA480723 [19]); 415 E. coli BSI isolates from Cambridge, UK (2006–2012, PRJEB4681 [11]); 481 E. coli BSI isolates from the Netherlands (2014–2016, enriched for ESBL isolates, PRJEB35000 [20]); and 162 E. coli BSI isolates from Scotland (2013–2015, PRJEB12513 [7]). These isolates were mapped to ST-specific references and phylogenetic trees generated, followed by permutation tests for geographical clustering as described in the statistical analysis section below.

**Epidemiological linkage**

Isolate data was linked to laboratory and electronic health record data via the Infections in Oxfordshire Research Database (IORD). Data on suspected infectious focus and patient provenance was acquired via linked local infection control records which had been submitted to Public Health England as part of the mandatory surveillance programme; such data was available for 400 Klebsiella spp. and 2773 E. coli isolates. Healthcare-associated (HA) BSI were defined as occurring >48 h post-hospital admission or ≤30 days since hospital discharge; other cases were defined as community-associated (CA) [21].

**Genomic analysis**

All programmes were run using default settings unless indicated. Raw reads were assembled using Shovill (v1.0.4) [22] with assemblies <4Mb and >7Mb excluded from further analysis, because these were thought to represent possible assembly errors given the typical genome size for these species (4-6.5Mb) [22]. De novo assemblies were annotated with Prokka (v1.14.6) [23]; AMR genes, virulence factors, and plasmid replicons were identified using the Resfinder [24], VFDB [25], ISFinder [26], and PlasmidFinder [27] databases with Abricate (v0.9.8) (--min-id 95 --min-cov 95) [28], and Kleborate for Klebsiella spp. [12, 29–31]. Integrons were located using the IntegronFinder [32] tool and annotated by Mash [33] comparison to reference sequences (minimum containment ≥ 0.95, accessions in Additional File 1). Multi-locus sequence types (MLST) were determined in silico using the MLST tool (v2.17.6) for E. coli and Kleborate for Klebsiella spp. [34, 35], and isolates were assigned to sequence types (ST) based on allelic profiles catalogued in the Achtman and Pasteur schemes respectively [36, 37]. Phylogroups were determined using the ClermonTyping tool [38]. To fully utilise the resolution provided by whole genome sequencing, we additionally used fastbaps [39] to partition the population using the core gene alignment produced by Panaroo [40].

Mapping to MLST-specific reference genomes was performed using Snippy (v4.6.0) [41]; genomes were acquired from NCBI (appendix). Gubbins (v2.3.4) was used to build recombination-corrected phylogenies [42]. Time-scaled phylogenies were created using the BactDating [43] library in R (version 3.6.0) under a relaxed gamma model with a minimum of 100,000 Markov chain Monte Carlo iterations [43]. Significance of temporal signal was assessed with root-to-tip plots and 10,000 random permutations of tip dates. Effective sample size was assessed using the CoDa package [44]. Evolutionary distinctiveness (ED) was calculated using the Picante package in R [45]; low ED scores indicate closely related genomes. All bioinformatics was performed using the Oxford University Biomedical Research Computing Facility.

Contigs were classified as being of likely chromosomal or plasmid origin using MLplasmids with a 0.7 probability cut-off [46]. All plasmid/chromosomal contigs for a given isolate were binned into separate multi-fasta files and the distances between these calculated using Dashing [47]. The pairwise distance matrix was filtered on a distance of 0.71 (the median plasmidome similarity of blacTX-M-15 carrying K. pneumoniae ST490 isolates) and then clustered into “plasmidome groups” using the LinkComm package in R [48]. The ecology of categorical groups was compared with a permutational multivariate analysis of variance (permanova) test in the Vegan package in R [49].

A pan genome wide association study (PGWAS) was conducted using PySeer [50] with a linear mixed model utilising relatedness distances inferred from a core genome phylogeny to correct for population structure. A discriminant analysis of principal components (DAPC) was performed using the Adegenet library in R [51]. Gene co-occurrence was examined using graphs constructed using the iGraph library in R [52]. For each species, edge lists with all genes/plasmids/insertion sequences with a Pearson correlation coefficient spp. >0.5 were created and communities detected using single linkage.

**Statistical analysis**

To describe molecular epidemiological trends, STs were arbitrarily categorised as “rare” (≤10 isolates over the study period, or untypeable STs), “intermediate” (11-100 isolates), and for E. coli, “sub-major” (101-300 isolates) and “major” (≥300 isolates). For Klebsiella spp, we considered the most prevalent ST (i.e., ST490) in isolation. Stacked negative binomial regression models with
clustered standard errors were used to compare rates of change over time (STATA v16 [53]) [54]. Incomplete years (2008 and 2018) are excluded from this part of the analysis. All other statistical tests were performed in R v3.6 using the Stats package unless indicated [55]. Charlon score was calculated in the Comorbidity [56] R package using all ICD10 codes associated with episodes in the year prior to specimen collection dates. To test for geographical/healthcare setting structure in MLST groups we conducted a permutation test similar to that previously described [57]. In brief, the ratio of median distances between isolates from the same centre and different centres was calculated. This observed ratio was compared to a permuted distribution created by 1000 tip label randomisations. The number of permuted values at least as extreme as the observed value divided by the number of permutations carried out was calculated to give a one-sided test of statistical significance.

**Results**

*Escherichia coli* bloodstream infections are stably dominated by major lineages across community and healthcare-associated settings, but with diverse sporadic lineages accounting for almost half of cases, and evidence of sub-lineage replacement in major STs

From September 2008 to December 2018, 3461 *E. coli* isolates from 3196 patients were sequenced. Major STs, namely ST73 (n=574), ST131 (457), ST95 (320), and ST69 (314), comprised 48% of all isolates; sub-major STs, ST12 (144), and ST127 (113), 7% of all isolates; intermediate STs (32 STs) 24% of all isolates; and rare STs (304 STs) 21% of all isolates (Fig. 1, Additional File 1: Fig. S1). The incidence rate ratio per year (IRRy, i.e. the relative increase in incidence per year) of all these categories increased over time (major IRRy=1.13 (95% CI 1.10–1.17), sub-major IRRy=1.16 (1.10–1.22), intermediate IRRy=1.15 (1.12–1.18), rare IRRy=1.10 (1.06–1.14)) (Fig. 1A). There was no evidence that the proportion of BSIs caused by major or sub-major STs changed over the study period (Fig. 1B). There was evidence that the incidence of BSIs caused by intermediate STs increased slightly faster versus all other isolates and by rare isolates slightly slower (p heterogeneity <0.05).

After sub-stratifying the major STs 69, 73, 95, and 131 using fastbaps, there was some evidence of sub-lineage (i.e. genomic groups below the level of STs) replacement. Whilst the biological significance of this is not clear, it is interesting that the two sub-lineages of ST69 displaying a relative increase/decrease in incidence had different O-antigen serotypes (fastbaps cluster 16, 38/38 O17, IRRy 0.95 (95% CI 0.87–1.05), and fastbaps cluster 19, 50/67 O15, IRRy 1.31 (95% CI 1.19–1.43)). However, this was not the case for the other two sub-lineages with some evidence of replacement in ST95 (fastbaps clusters 48 and 49; Additional File 1: Fig.S2/ Additional File 1: TableS1).

![Fig. 1](image.png)

**Fig. 1** Population dynamics of *E. coli* (A, B) and *Klebsiella* spp. (C, D) STs over time. A Absolute number and B proportions of *E. coli* BSIs caused by the four major STs (STs 131/95/73/69), sub-major STs (STs 127/12), intermediate STs, and rare STs. C Absolute number and D proportions of *Klebsiella* spp. BSIs caused by ST490, intermediate, and rare STs. NB years 2008 and 2018 are incomplete (see the “Methods” section)
The majority of E. coli BSIs were CA-BSIs (2104/3461, 61%), with no evidence of variation in the proportion caused by major STs between CA-BSI and HA-BSI (990/2104, 47%, vs 661/1357, 49%, p=0.4). Relatively few isolates came from patients resident in a nursing/care home (169 (7%) of 2301 with data available). The four major E. coli STs accounted for a significantly greater proportion of these cases (97/169, 57% vs 1009/2132, 47% p=0.01); however, there was no evidence of large-scale nursing/care home-associated BSI outbreaks (Additional File 1: Fig. S3). E. coli CA-BSIs occurred in slightly older median age 76 (IQR 63–85) vs 70 (55–80), p<0.001) and less comorbid individuals (median Charlson score 1 (IQR 0–2) vs 1 (IQR 1–2), p=0.003) than HA-BSIs.

**Klebsiella spp. bloodstream infections are caused by a diverse representation of sub-species with significant intra-species diversity, with the exception of K. pneumoniae ST490, causing a transient, clonal, local outbreak**

Amongst the 886 successfully sequenced Klebsiella spp. isolates, a large number of sub-species were observed, including K. pneumoniae (n=528 [60%]), Klebsiella varicola (n=112 [13%]), Klebsiella michiganensis (n=90 [10%]), Klebsiella oxytoca (n=59 [7%]), Klebsiella aerogenes (n=38 [4%]), Klebsiella grimontii (n=30 [3%]), Klebsiella quasipneumoniae (n=28 [3%]), and Klebsiella africana (n=1 [0.1%]). In stark contrast to E. coli, 738/886 (83%) of Klebsiella spp. isolates belonged to rare STs (Fig. 1, Additional File 1: Fig. S4). The multidrug-resistant K. pneumoniae ST490 was the most prevalent Klebsiella ST and the only one with >40 isolates; its incidence decreased over the study period (IRRy 0.78, 95%CI 0.68–0.89). This ST is rarely seen in other studies, consistent with a transient but relatively large, local outbreak. Notably 16/45 (36%) of isolates from this ST were community-onset.

The majority of Klebsiella spp. BSIs were HA-BSI (510/882 [missing data for 4 isolates], 59%). As with E. coli few K. pneumoniae cases were in patients resident in a nursing/care home (10/151, 7%) though these data were not available for most isolates. There was no difference in the proportion of intermediate STs amongst HA-BSI vs CA-BSI (90/510 18% vs 58/372 16%, p=0.5). Klebsiella spp. CA-BSI cases were older (median age 76 years (IQR 65–85) vs 66 years (49–75) for HA-BSI; p<0.001) but also less comorbid than HA-BSI cases (median Charlson score 1 (IQR 0–2) vs 2 (1–3); p<0.001). A large number (265/341, missing data for 31) of CA-BSIs occurred in relatively healthy individuals (Charlson score ≤2, i.e. predicted 10-year survival ~90%). In contrast to previous reports suggesting K. quasipneumoniae/K. varicola may be less virulent than K. pneumoniae [13], in our study there was no evidence of differences in crude 30-day mortality between the Klebsiella subspecies with ≥20 isolates (111/524 (21%) K. pneumoniae, 28/112 (25%) K. varicola, 23/89 (26%) K. michiganensis, 9/57 (16%) K. oxytoca, 10/36 (28%) K. aerogenes, 7/28 (25%) K. quasipneumoniae, and 7/29 (24%) K. grimontii; exact p=0.5, missing data for 9 isolates), nor the proportion of CA-BSI (230/527 (44%), 48/112 (43%), 34/89 (38%), 27/58 (47%), 11/28 (39%), 9/27 (25%), and 12/30 (40%) respectively; exact p=0.4, missing data for 4 isolates).

**No evidence of intra- or inter-regional clustering of major Escherichia coli lineages causing BSI in either community- or healthcare-associated settings**

For the four largest E. coli STs, we compared the phylogeny of BSI isolates in this study to those in previous studies over a similar timeframe (see Methods). There was no obvious phylogenetic clustering of HA or CA isolates (Fig. 2), supported by permutation testing, confirming that for the major STs these were indeed randomly distributed across the phylogeny by geography and healthcare/community setting (Additional File 1: Table S2). Additionally, for all four of the major E. coli STs, there was no difference in the ED scores between HA and CA cases and therefore no evidence of healthcare-associated adaptation and subsequent transmission (Additional File 1: Table S3). Our PGWAS did not identify any significant hits separating CA and HA isolates (Additional File 1: Fig. S5), including those that had been previously identified in other studies [7].

**Urinary and hepatobiliary sources of E. coli and Klebsiella spp. BSI predominate, with major E. coli STs over-represented in urinary-associated E. coli BSIs**

The urinary tract was the most common physician-identified source of infection in both species. In E. coli, these were strongly associated with the presence of the *pap* group of genes (Additional File 1: Table S4) which were over-represented in major STs (916/1651 (55%) vs 394/1810 (22%) isolates, p < 0.001). The hepatobiliary tract was the second most common source for both *E. coli* and *Klebsiella* spp., followed by other gastrointestinal infections for *E. coli* and the respiratory tract for *Klebsiella* spp. (Additional File 1: Fig. S6a). For *E. coli* BSIs, there was some evidence that the incidence of BSIs not attributed to the urinary tract increased faster than those thought to be of urinary origin (IRRy 1.45, 95%CI 0.98–2.17 vs IRRy 1.36, 95%CI 0.88–0.88 respectively; *p* heterogeneous=0.005, Additional File 1: Fig. S6b). The proportion of BSIs caused by the five major *E. coli* STs was greater for BSIs with urinary compared to other sources (586/1071 [55%], vs 520/1229 [42%], p<0.001); this was not the case for *Klebsiella* spp. where the STs causing BSI attributable to all sources were diverse.
Multi-drug resistant E. coli BSIs are increasing, likely driven by the association of AMR genes within lineages, and ceftriaxone resistance is driven by the expansion of resistant ST131 and ST73 sub-lineages

For E. coli BSIs, considering total effects of resistance to individual antibiotics, regardless of mechanism, there was evidence of increasing incidence of BSI carrying AMR genes conferring resistance to amoxicillin IRR\textsuperscript{y}=1.17 (95%CI 1.14–1.20), co-amoxiclav IRR\textsuperscript{y}=1.15 (1.08–1.24), ceftriaxone IRR\textsuperscript{y}=1.26 (1.17–1.36), gentamicin IRR\textsuperscript{y}=1.24 (1.15–1.33), trimethoprim IRR\textsuperscript{y}=1.17 (1.14–1.20), and ciprofloxacin (IRR\textsuperscript{y}=1.23 (1.17–1.29); Fig. 3); there was only a single isolate carrying a gene encoding for carbapenem resistance (ST10, bla\textsubscript{OXA-48}) in 2014. For ceftriaxone, co-amoxiclav, and gentamicin, the incidence trends of isolates not carrying genes conferring resistance to these antibiotics increased faster than those carrying resistance (\textit{p}_\text{heterogeneity} \leq 0.01).

A major contributing factor to this finding was the overall decline of the MDR ST490 lineage and the emergence of a more susceptible ST490 sub-lineage between 2005 and 2010 which had lost the \textit{aac(3)-IIa}, \textit{aac(6\textsuperscript{'})-Ib-cr}, \textit{bla}\textsubscript{OXA-1} and \textit{tet(a)} genes (Additional File 1: Fig.S7). Genes encoding for resistance to ceftriaxone were significantly more common in intermediate STs (STs with >10 isolates in the dataset) (54/148, 36% vs 73/738, 10% \textit{p}<0.001) and in HA-BSIs (85/510, 17% vs 42/372, 11% \textit{p}=0.03). Similarly MDR isolates (resistant to \geq 3 antibiotic classes) were more common in intermediate vs rare STs (89/366, 24% vs 80/520, 15%, \textit{p}=0.001) and HA-BSI vs CA-BSI (109/510, 21% vs 59/372, 16%, \textit{p}=0.049).

Drug-susceptible Klebsiella spp. BSIs are increasing faster than drug-resistant strains, with ceftriaxone-resistance largely healthcare-associated

For Klebsiella spp., there was only evidence of increasing incidence of isolates carrying genes/mutations conferring resistance to co-amoxiclav IRR\textsuperscript{y}=1.07 (95%CI 1.02–1.12) and ciprofloxacin IRR\textsuperscript{y}=1.11 (1.06–1.16) (Fig. 3). In contrast, the incidence of genes conferring resistance to ceftriaxone IRR\textsuperscript{y} 0.99 (0.92–1.09) and trimethoprim IRR\textsuperscript{y} = 1.03 (0.93–1.15) were stable whilst gentamicin IRR\textsuperscript{y} = 0.86 (0.76–0.97) decreased. For ceftriaxone, co-amoxiclav, and gentamicin, the incidence trends of isolates not carrying genes conferring resistance to these antibiotics increased faster than those carrying resistance (\textit{p}_\text{heterogeneity} \leq 0.01).

A major contributing factor to this finding was the overall decline of the MDR ST490 lineage and the emergence of a more susceptible ST490 sub-lineage between 2005 and 2010 which had lost the \textit{aac(3)-IIa}, \textit{aac(6\textsuperscript{'})-Ib-cr}, \textit{bla}\textsubscript{OXA-1} and \textit{tet(a)} genes (Additional File 1: Fig.S7). Genes encoding for resistance to ceftriaxone were significantly more common in intermediate STs (STs with >10 isolates in the dataset) (54/148, 36% vs 73/738, 10% \textit{p}<0.001) and in HA-BSIs (85/510, 17% vs 42/372, 11% \textit{p}=0.03). Similarly MDR isolates (resistant to \geq 3 antibiotic classes) were more common in intermediate vs rare STs (89/366, 24% vs 80/520, 15%, \textit{p}=0.001) and HA-BSI vs CA-BSI (109/510, 21% vs 59/372, 16%, \textit{p}=0.049).
Virulence factors are strongly structured by ST amongst *E. coli* but not *Klebsiella* spp. BSI isolates with no evidence in *Klebsiella* spp. of a difference in virulence gene carriage between community vs hospital acquired BSIs

The distribution of virulence factors strongly reflected the underlying clonal population structure in *E. coli* isolates, with segregation by ST, confirmed with a discriminant analysis of principal components (DAPC) (Fig. 4). Most notably phylogroup B2 isolates were separated from the rest of the population by the presence of *chuA*, *chuX* (involved in haem utilisation), and *espL1/espX4* (elements of the type 3 secretion system). Given this and the broad equilibrium in the *E. coli* BSI population structure demonstrated above, it is unlikely that increased carriage of certain virulence factors explains the increasing incidence of these infections. In keeping with this, there was a near-perfect correlation between frequency of carriage of any given virulence gene across all isolates in 2009 vs 2018 (Pearson correlation=0.99; *p* < 0.001).

No such structuring by virulence factors was observed for *Klebsiella* spp.; instead, virulence factors were widely distributed amongst 79 known STs and 14 isolates with no assigned ST (Additional File 1: Fig. S8). Notably, of the 372 CA-BSIs, only 65 (17%) carried known virulence factors. There was no difference in the proportion of CA-BSI and HA-BSI isolates carrying ≥1 virulence factor (65/372 [17%] vs 92/510 [18%]; *p*=0.9) nor carrying colibactin and/or aerobactin (15/372 [4%] vs 25/510 [5%]; *p*=0.7) (Fig. 5).

**Plasmid replicon and plasmidome analyses reflect diversity consistent with a highly mobile accessory genome within species**

Given the known importance of plasmids in the carriage and transmission of AMR genes, we sought to understand the role they might play in the relative success of major STs. Analysis of plasmid replicon profiles showed these were largely genus-restricted, with some overlap (Additional File 1: Fig. S9). However, plasmid replicons were not structured by host lineage in *E. coli* or *Klebsiella* spp., with the exception of *K. pneumoniae* ST490 which was discriminated by the presence of the...
Fig. 4 Virulence elements are structured by ST/Phylogroup in E. coli. A Discriminant analysis of principal components plot showing clear separation of phylogroups by their virulence factor content. B Loading plot showing genes contributing most to the discrimination of phylogroup B2 from other phylogroups. C Heatmap of presence absence of virulence genes (x-axis) by major ST and phylogroups (y-axis).

Fig. 5 Distribution of virulence factors and antimicrobial resistance classes across community acquired (left) and healthcare associated (right) Klebsiella spp. isolates. Virulence factor scores were assigned by Kleborate as follows: virulence score 0 = no acquired loci, 1 = yersiniabactin, 2 = yersiniabactin and colibactin (or colibactin only), 3 = aerobactin (without yersiniabactin or colibactin), 4 = aerobactin with yersiniabactin (without colibactin), and 5 = yersiniabactin, colibactin and aerobactin.
IncFIA(H1) plasmid type (also identified in 23 E. coli isolates; Additional File 1: Fig. S10).

Using our approach to predict plasmidome population structure based on k-mer similarity of contigs binned as plasmid, we observed a striking degree of plasmidome diversity within highly genomically related isolates/STs for all species (Additional File 1: Fig. S9); for 656 isolates with a core genome similarity >0.99 to another isolate in the dataset, the median plasmidome similarity was only 0.51 (IQR 0.28–0.86). Isolates with a near-identical plasmidome (>0.99 similarity, n=115 isolates) did however have highly similar chromosomes (median plasmidome similarity: 0.97 [IQR 0.94–0.99]). Compared with other STs, major E. coli STs had larger plasmidomes (median size 106,766 vs 97,432, p<0.001) which belonged to more “plasmidome groups” (see methods) (median 3 vs 2, p<0.001). For both E. coli and Klebsiella spp., there was no evidence of different plasmid populations between those associated with HA- vs CA-BSI (p=0.4).

Finally we analysed the co-occurrence of AMR genes, plasmid replicon types, insertion sequences (ISs), and integron-associated markers within E. coli and Klebsiella spp. The networks formed were notable for the widespread co-occurrence of AMR genes, insertion sequences and (for some classes of AMR genes) integrons, but not usually specific plasmid types (Additional File 1: Figs. S11/12). This suggests that most common AMR genes are found in a diverse range of genetic contexts (e.g. multiple plasmid types, chromosomally integrated), and that horizontal gene transfer of important AMR genes in these isolates is largely facilitated by smaller, non-plasmid MGEs such as transposons/ISs/integrons, which are difficult to reliably evaluate with short-read data [58].

Discussion
In this unbiased longitudinal sequencing study of all (90-day-deduplicated) E. coli and Klebsiella spp. BSIs in Oxfordshire over a decade (2008–2018), we highlight the similarities and differences in the molecular epidemiology of these species. Overall, the increasing incidence of E. coli and Klebsiella spp. BSIs in Oxfordshire is not explained by the expansion of a single ST, and much of the burden (40%) of Klebsiella spp. disease is caused by non-K. pneumoniae species. Although six lineages stably account for nearly half of all E. coli BSIs and a clonal outbreak was observed in K. pneumoniae, many BSIs in our setting are caused by diverse strains with diverse accessory genomes.

We found no genomic evidence supporting the stratification of E. coli isolates into HA-BSI vs CA-BSI. Strains causing healthcare-associated BSIs may not be specifically healthcare-acquired but may still be healthcare-provoked (for example by the presence of indwelling devices or relative immunosuppression). Interventions targeting only HA infections (such as that of the UK government to reduce these by 50% by 2021 [3]) might have limited efficacy without considering the wider ecology of these species. Whilst the incidence of E. coli BSIs attributed to the urinary tract increased over the study period, there was evidence that the rate of this increase was slower compared to other sources. This possibly reflects some success of infection prevention interventions such as catheter care, but also highlights the multi-faceted approach required to reduce the overall incidence rate. More work is required to understand CA-Klebsiella spp. BSIs because known markers of virulence found in CA disease (e.g. aerobactin, yersiniabactin, colibactin) in other studies [14] were not seen in most of our cases. Importantly therefore, existing markers may therefore be insensitive to detect the emergence of clinically hypervirulent, CA, multiderug-resistant strains.

For E. coli, our analysis is consistent with previous studies demonstrating a broadly stable population structure at the MLST level with the predominance of STs 69, 73, 95, and 131 [7, 11]. Within this however, we demonstrated some replacement in sub-clades of major STs over the study period. This may represent an adaptive strategy within major STs that allows them to continue to maintain their niche and evolve to survive changing environmental and immunological selection pressures. As has been previously noted, virulence factors in E. coli were structured by the underlying phylology [59, 60]. Differences in carriage of haem utilisation genes were a notable factor discriminating isolates in major STs from the rest of the population, suggesting that iron metabolism might be an important selective pressure in determining invasive potential, a finding supported by a recent GWAS of a mouse model [61]. In contrast, for Klebsiella spp., we demonstrated that most infections are caused by sporadic STs which individually only accounted for only a small proportion of the overall population causing disease.

In E. coli, the incidence of isolates carrying genes conferring resistance against commonly used antimicrobial classes increased faster than those without; in general, the opposite was true for Klebsiella spp. The most commonly isolated MDR-ST in Klebsiella spp. (ST490) significantly decreased in incidence over the study period, lending credence to the idea that the relentless expansion of MDR clades is not inevitable, although it is unclear what interventions may have helped in causing its decline, particularly given that approximately a third of cases were community onset. The increasing incidence of gentamicin/ceftriaxone resistance in predominantly CA-E. coli BSIs seems paradoxical given that these antibiotics are rarely used in this setting. One explanation is that they are co-located on MGEs with other AMR
genes encoding resistance to antibiotics more commonly used in the community (e.g. amoxicillin, trimethoprim). An alternative hypothesis might be increased exposure to third-generation cephalosporins due to the rise of ambulatory/“hospital-at-home” medical pathways where the once-a-day dosing of ceftriaxone makes it a relatively widely prescribed antibiotic in these settings. Regardless, overall, Klebsiella strains causing BSI appear to be exposed to declining antibiotic selection pressures compared to E. coli, suggesting that they are maintained and selected for in distinct ecological niches.

In E. coli BSI isolates, there was no genetic signal of adaptation to either healthcare or community environments, suggesting that these are not relevant niches for selection, contrary to a previous study [7]. However, this study included small numbers of isolates (n=162) and may have been unable to fully account for population structure. Furthermore our analysis suggests that, contrary to true nosocomial pathogens [62], the ecology of the E. coli plasmidome is similar for HA/CA infections. Our findings support the hypothesis that the diversity observed in any given epidemiological strata (e.g. age groups/infection focus/healthcare setting) is sampled from the same common ecological pool of isolates with invasive potential rather than representing specialised adaptation to any given setting. There are of course exceptions to this, such as localised outbreaks within care facilities and hospital environments [63]. However, our data suggests that at least in Oxfordshire, these outbreaks do not contribute significantly to the epidemiology of E. coli BSIs. Importantly, iatrogenic, non-pathogen-associated factors promoting invasive infection (e.g. urinary catheterisation) should be minimised to reduce the incidence of E. coli BSIs.

For Klebsiella spp., the traditional view that HA infections are opportunistic and caused by MDR isolates and CA infections are caused by isolates carrying one or more specific virulence genes (e.g. yersiniabactin, colibactin, aerobactin, salmochelin), which appears to be an over-simplification in our setting [12, 13]. The majority of CA Klebsiella spp. infections contained none of the known genetic markers of hypervirulence, and these genes are unlikely to be a reliable method of surveillance for emerging strains with a propensity to cause invasive CA disease, at least in our setting. Whilst ESBL and MDR isolates were significantly more common in HA isolates, about a third were CA (even using our fairly conservative definition) which significantly challenges the prevailing dogma that these are largely opportunistic HA strains.

Our study was limited by the inability to reconstruct closed genomes/plasmids using short-read sequencing data and our plasmidome analysis should therefore be interpreted with caution. The widely used PlasmidFinder database allows some inferences to be made using short read data; however, there are many untypeable plasmids that are not reflected in this database [64]. Similarly the MLplasmids classification algorithm used is trained on a limited reference set which may lead to some erroneous contig classifications. Additionally, analysis of the entire plasmidome may be too crude to identify the nuances of similarities/differences, particularly for small, shared plasmids as these represent only a small proportion of the overall plasmidome. Whilst we used a relatively conservative definition of healthcare-associated (i.e. within 30 days of hospital admission), this may have failed to capture the longer-term impacts of earlier hospital admissions/other healthcare exposures.

Conclusions

The contrasting epidemiology of E. coli and Klebsiella spp. BSIs suggests different reservoirs, selection pressures, and modes of acquisition for these genera. Separating strategies to reduce the incidence of these infections are likely to be required, and consideration of the community as a reservoir is important. Given the stable population structure demonstrated here, vaccines may be a promising prospect to reduce the incidence of E. coli and Klebsiella pneumoniae BSI. For E. coli, such a vaccine (ExPEC4/10V, Janssen Pharmaceuticals [65]) is in clinical trials and we have recently demonstrated its potential to provide protection against most E. coli isolates causing BSI in Oxfordshire [66]. The lack of reproducibility of several findings from previous studies and the poor sensitivity of current molecular markers for hyper-virulent Klebsiella surveillance highlights the critical importance of unselected sampling frames when making epidemiological inferences.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00947-2.

Additional file 1: Fig S1. Gene presence/absence heatmap showing AMR gene presence/absence against the core genome phylogeny for E. coli. Fig S2: Time-scaled phylogenies for ST131, ST95, ST73 and ST69. Fig S3: Possible transmission within nursing homes. Fig S4: Gene presence/absence heatmap showing AMR gene presence/absence against the core genome phylogeny for Klebsiella spp. Fig S5: Manhattan plots of a pan-genome wide association study of the association of genes with community/healthcare associated onset. Fig S6: Proportions of presumed infectious foci for CA and HA BSI. Fig S7: Timescaled phylogenies for ST131, ST95, ST73 and ST69. Fig S8: Phylogenetic tree of Klebsiella spp annotated with species and virulence score. Fig S9: top panel - plasmid types in the PlasmidFinder database identified in the major/other E. coli/Klebsiella spp., bottom left - plot showing kmer based plasmidome similarity (y-axis) against chromosome similarity (x-axis) for isolates of the same MLST. Fig S10: DAPC plots for Klebsiella spp. Fig S11: Networks of genes/plasmids/insertion sequences commonly co-occurring in E. coli. Fig S12: Networks of genes commonly co-occurring in Klebsiella spp. Table S1: Incidence rate ratios for sub-lineage of major STs identified by fastbaps. Table S2: SNP ratios (median within/between
region) and (median HA/Ia) were calculated for each ST. Table S3: Evolutionary distinctiveness (ED) scores for community-associated (CA) and healthcare-associated (HA) isolates among major E. coli STs. Table S4: Top hits from a pan-genome-wide association study (PGWAS) performed using Pyseer [30] of the association of gene presence/absence with pysi-cian identified BS1 source.

Acknowledgements

We express our thanks to colleagues in the Oxford University Hospitals NHS Foundation Trust clinical microbiology laboratory who performed routine identification of the isolates in this study and catalogued them for storage. We are also indebted to Dai Griffiths who managed the isolate collected for many years. This work uses data provided by patients and collected by the UK’s National Health Service as part of their care and support. We thank all the people of Oxfordshire who contribute to the Infections in Oxfordshire Research Database.

Research Database Team: L Butcher, H Boseley, C Crichton, DW Crook, D Eyre, O Freeman, J Gearing (community), R Harrington, K Jeffery, M Landray, A Pal, TEA Petoe, TP Quan, J Robinson (community), J Sellors, B Shine, AS Walker, D Waller. Patient and Public Panel: G Blower, C Mancey, P McLoughlin, and B Nichols.

Authors’ contributions

ASW, NS, DC, and TEAP planned and managed isolate collection and acquisition funding for the study. SL and KDV performed data linkage, cleaning, and analysis. SL wrote the first draft of the manuscript. KC, LB, SG, JK, TD, and AV performed the laboratory work. KJ and MA facilitated access to isolates and resources. ASW, SH, NS, DC, and TEAP provided supervision. All authors read and approved the final manuscript.

Funding

The research was supported by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance (NIHR200915) at the University of Oxford in partnership with Public Health England (PHE) and by Oxford NIHR Biomedical Research Centre. T Petoe and AS Walker are NIHR Senior Investigators. The report presents independent research funded by NIHR. The views expressed in this publication are those of the authors and not necessarily those of the NHS, NIHR, the Department of Health or Public Health England. The computational aspects of this research were funded from the NIHR Oxford BRC with additional support from the Wellcome Trust Core Award Grant Number 203411/2/16/Z. SL is supported by a Medical Research Council Clinical Research Training Fellowship. KC is Medical Research Foundation-funded.

Availability of data and materials

All sequencing data has been deposited in the NCBI under project accession number PRJNA604975 [67] at https://www.ebi.ac.uk/ena/browser/view/PRJNA604975. Additional metadata (e.g. sample provenance, collection date, suspected source) and output from bioinformatics pipelines (e.g. resistance/virulence gene calls, MLST/phylogroup predictions) used in the analysis is available at https://github.com/samlipworth/Oxford_Ecoli_Kleb [68].

Declarations

Ethics approval and consent to participate

The Infections in Oxfordshire Research Database (IORD; https://oxfordbrc.nihr.ac.uk/research-themes/overview/antimicrobial-resistance-and-modernising-microbiology/infections-in-oxfordshire-research-database-iord/) has generic Ethics Committee, Health Research Authority and Confidentiality Advisory Group approvals (19/SC/0403, 19/CAG/0144) which facilitate the pseudo-anonymised linkage of routinely collected NHS electronic healthcare record data from the Oxford University Hospitals NHS Foundation Trust Clinical Systems Data Warehouse and research data (e.g. sequencing data) from the Antimicrobial Resistance and Modernising Microbiology Theme of the Oxford NIHR Biomedical Research Centre, Oxford. IORD links records by a specific, random, number ensuring that no patient-identifiable information is shared with researchers using this resource. We sequenced bacterial isolates from bloodstream infections that are routinely stored by the John Radcliffe Hospital Microbiology laboratory. In the UK, bacterial isolates (such as those sequenced in this study) routinely cultured from human clinical samples do not require ethical approval for analysis under the provisions of the Human Tissue Act as they do not contain any material considered to be human tissue.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

1Nuffield Department of Medicine, University of Oxford, Oxford, UK. 2Oxford University Hospitals NHS Foundation Trust, Oxford, UK. 3John Radcliffe Hospital, Oxford OX3 9DU, UK. 4National Infection Service, Public Health England, Colindale, London, UK. 5NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK. 6NIHR Biomedical Research Centre, Oxford, UK.

Received: 22 January 2021 Accepted: 3 August 2021
Published online: 03 September 2021

References

1. Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, Fawcett NJ, et al. Trends over time in Escherichia coli bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998–2016: a study of electronic health records. Lancet Infect Dis. 2018;18(10):1138–49. https://doi.org/10.1016/S1473-3099(18)30353-0.

2. English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR). Public Health England; 2019. Available from: https://assets. publishing.service.gov.uk/government/uploads/system/uploads/attachment_ data/file/843129/English_Surveillance_Programme_for_Antimicrobial_Utilisa tion_and_Resistance_2019.pdf

3. Public Health England. Escherichia coli (E. coli): guidance, data and analysis. GOV.UK. 2010 [cited 2018 May 20]. Available from: https://www.gov.uk/ government/collections/escherichia-coli-e-coli-guidance-data-and-analysis

4. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, et al. Evolutionary history of the global emergence of the Escherichia coli Epidemic Clone ST131. MBio. 2016;7:e02162.

5. David S, Reuter S, Harris SR, Glaser C, Feltwell T, Argimon S, et al. Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. Nat Microbiol. 2019;4(11):1919–29. https://doi.org/10.1038/s41564-019-0462-z.

6. Wyres RL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae: PLOS Genet. 2019;15:e1008114.

7. Goswami C, Fox S, Holden M, Connor M, Leanard A, Evans TJ. Genetic analysis of invasive Escherichia coli in Scotland reveals determinants of healthcare-associated versus community-acquired infections. Microb Genom. 2018;4. mgen.microbiologyresearch.org; Available from: https://doi.org/10.1099/mgen.0.000190.

8. Blandy O, Honeyford K, Gharti M, Thomas A, Ramzan F, Ellington MJ, et al. Factors that impact on the burden of Escherichia coli bacteraemia: multivariable regression analysis of 2011-2015 data from West London. J Hosp Infect. 2019;101(2):120–8. https://doi.org/10.1016/j.jhin.2018.10.024.

9. Kazy Gordon A, Phan HTT, Lipworth SI, Cheong E, Gottlieb T, George S, et al. Genomic dynamics of species and mobile genetic elements in a prolonged bladIMP-4-associated carbapenemase outbreak in an Australian hospital. J Antimicrob Chemother. 2020;75(4):873–82. https://doi.org/10.1093/jac/dkaa256.

10. Day MJ, Dournith M, Abemethy J, Hope R, Reynolds R, Wain J, et al. Population structure of Escherichia coli causing bacteraemia in the UK and Ireland between 2001 and 2010. J Antimicrob Chemother. 2016;71(8):2139–42. https://doi.org/10.1093/jac/dkv145.

11. Kallonen T, Brodick HJ, Harris SR, Corander J, Brown NM, Martin V, et al. Systematic longitudinal survey of invasive Escherichia coli in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. Genome Res. 2017; Available from:27(8):1437–49. https://doi.org/10.1101/gr.216606.116.
Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.