Journal of Infection 91 (2025) 106612

Contents lists available at ScienceDirect

Journal of Infection

journal homepage: www.elsevier.com/locate/jinf

Bacteria and Bacterial Diseases

An electronic health record-wide association study to identify populations at increased risk of *E. coli bacteraemia*

Emma Pritchard [a,b,c,*], Karina-Doris Vihta [b,c,d], Samuel Lipworth [b,c,e], Koen B. Pouwels [c,f], Nicole Stoesser [b,c,g], Russell Hope [h], Berit Muller-Pebody [h], T. Phuong Quan [b,c,g], Jack Cregan [b,c,h], Colin Brown [h], Susan Hopkins [c,h], David W. Eyre [c,g,i,1], A. Sarah Walker [b,c,g,1]

[a] Division of Informatics, Imaging & Data Sciences, School of Health Sciences, University of Manchester, UK
[b] Nuffield Department of Medicine, University of Oxford, Oxford, UK
[c] The National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford, Oxford, UK
[d] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
[e] Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[f] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[g] The National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford, Oxford, UK
[h] United Kingdom Health Security Agency (UKHSA), London, UK
[i] Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:
Accepted 31 August 2025
Available online 3 September 2025

Keywords:
Escherichia coli
Bloodstream infections
Electronic health records
Risk factors
Population health
Infectious disease epidemiology

SUMMARY

Objectives: Escherichia coli bacteraemias have been under mandatory surveillance in the UK for fifteen years, but cases continue to rise. Systematic searches of all features present within electronic healthcare records (EHRs), described here as an EHR-wide association study (EHR-WAS), could potentially identify under-appreciated factors that could be targeted to reduce infections.
Methods: We used data from Oxfordshire, UK, and an EHR-WAS method developed for use with large-scale COVID-19 data to estimate associations between E. coli bacteraemia cases, hospital-exposed controls, and 377 potential risk factors using Poisson regression models adjusted for potential confounders for three two-year financial year (FY) periods.
Results: FY2022/23–2023/24 analysis included 757 (0.3%) cases and 276,758 (99.7%) controls. We identified six broad disease areas associated with increased or decreased E. coli bacteraemia risk. Renal/urological/urinary tract infection-related variables had the largest impact, with 47% of cases theoretically removed if these factors could be minimised. Cancer-related variables were associated with higher E. coli bacteraemia risk (1.20 times higher (95%CI 1.08–1.34) per three months closer to chemotherapy in the last year), as were gastrointestinal- and infectious disease-related variables. Cardiac/respiratory-related variables were associated with lower E. coli bacteraemia risk, whereas greater healthcare exposure showed no consistent effect. Associated factors varied across periods, but broad groups remained similar.
Conclusions: Applying an EHR-WAS approach, we show E. coli bacteraemias are largely driven by known risk factors and frailty, highlighting the importance of monitoring these factors and targeting modifiable risks where possible.

Introduction

Escherichia coli (*E. coli*) bacteraemias have been under mandatory surveillance for fifteen years in the UK, but cases continue to rise.[1] In contrast, enhanced surveillance of methicillin-resistant *Staphylococcus aureus* (MRSA) and *Clostridioides difficile* infection led to cases declining[2] by providing epidemiological insights informing targeted infection prevention strategies.[3]

* Correspondence to: Nuffield Department of Medicine, Level 7 Microbiology Research, John Radcliffe Hospital, Oxford OX3 9DU, UK.
E-mail address: emma.pritchard@ndm.ox.ac.uk (E. Pritchard).
[1] Contribution considered equal.

https://doi.org/10.1016/j.jinf.2025.106612

The UK *E. coli* bacteraemia surveillance programme collects data on age, sex, and antibiotic resistance.[4–6] Electronic health records (EHRs) offer a unique opportunity to identify populations at risk of infection more comprehensively, with regularly updated data streams facilitating this cheaply. Systematic searches of all features present within EHRs, described here as an EHR-wide association study (EHR-WAS, akin to genome-wide association studies), could identify novel, underappreciated factors to target to reduce infections.

Studies of infection using EHR data benefit from increased regional and national linkage between microbiology data and patient admissions driven by COVID-19,[7] allowing more accurate laboratory-confirmed infection ascertainment and assessment of more diverse and higher numbers of associated factors, whether causal or proxies. As laboratory results, vital sign measurements, and blood test results are often automatically uploaded to electronic systems, and reasons for inpatient admissions are coded immediately after hospital discharge, EHR data should be relatively up-to-date, enabling continuous monitoring of contemporaneous factors.

Ad hoc studies have investigated *E. coli* bacteraemia risk factors in hospital populations but often considered a limited number of factors selected a priori and were not designed for continuous monitoring.[8–10] These studies found populations at highest risk included individuals on dialysis, renal disease/failure patients, cancer patients,[8,9] and individuals with urinary catheterisation/incontinence,[9] urinary tract infections (UTI),[10] and higher comorbidity scores.[10]

Identifying more associated factors could help target interventions by directly removing or reducing a causal mechanism or protecting high-risk groups. For example, although a Phase 3 trial of a prophylactic extraintestinal pathogenic *E. coli* vaccine has recently been halted,[11] there is substantial ongoing research in this area. Importantly, analogous to machine learning, to target interventions, associations may not need to be causal, providing the underlying mechanisms they represent are generalisable.

Using EHRs, we aimed to identify populations at increased risk of *E. coli* bacteraemias using a novel EHR-WAS method which could be applied repeatedly over time, using a hospital-exposed control population to minimise missing data.

Methods

We used the Infections in Oxfordshire Research Database (IORD): a data warehouse including inpatient admissions, outpatient appointments, emergency department (ED) visits to four large teaching hospitals serving a population of 755,000; vital signs taken during hospital attendance; microbiology (positive and negative results) and biochemistry/haematology results. The hospital group provides all acute services to the region and all community and hospital laboratory and microbiology testing. IORD has approvals from the National Research Ethics Service South Central-Oxford C Research Ethics Committee (19/SC/0403), Health Research Authority and Confidentiality Advisory Group (19/CAG/0144) as a deidentified database without individual consent.

We included all admissions from 01-April-2018 to 31-March-2024, divided into three two-year periods (April to March, keeping winter months together): FY2018/19-2019/20 (i.e., 01-April-2018 to 31-March-2020), FY2020/21-2021/22, and FY2022/23-2023/24. We looked back up to five years for potential factors, hence including data from 01-April-2013.

Analysis cohort definition

Potential cases were all patients with *E. coli* cultured from blood (Fig. S1). Potential controls were all individuals who were not a case and had contact in the current period (inpatient episode, outpatient or ED visits, blood test, or microbiology sample) (Fig. S2). We included one observation per person per calendar period, selecting the first positive culture for cases and the last observation otherwise.

As many characteristics (e.g. those based on diagnosis or procedure codes) were recorded in inpatient episodes, cases and controls were included if they had an inpatient episode in the last 5 years (y) that was not attributable to the *E. coli* bacteraemia (episode ending > 72 hours (h) before blood culture collection) to minimise reverse causality as many bacteraemias lead to admissions and factors identified from these episodes may be consequences, not causes, of infection (20% controls had inpatient episodes ≤72 h before their most recent record and were retained for analyses). Consequently, we are estimating risk of *E. coli* bacteraemias versus healthcare contact for other reasons.

Identifying EHR-wide associations

We defined six variables adjusted for in all models regardless of the magnitude of association ("core" variables): age, sex, ethnicity (white vs non-white as small numbers in the latter), deprivation (Index of Multiple Deprivation (IMD) percentile),[12] rural/urban classification, and catchment percentage (percentage of individuals in the local area visiting an Oxfordshire hospital; 0 = none, 100 = all).[13]

377 "screening" characteristics from various EHR data sources were defined from previously published research, clinical advice, and data availability (Supplementary Methods, definitions at https://github.com/EmmaPritchard/EHR-Risk-Factor-Definitions). Information from the 72 h before blood culture collection for cases was excluded to avoid reverse causality. Analogous to a GWAS approach, our analysis considered a wide range of potential risk factors without requiring prior knowledge on their impact on *E. coli* bacteraemia risk.

Variables were included as categorical and continuous parameterisations, capturing ever/never having a characteristic (within the last 5 y) and proximity of the most recent record to the current contact. Categorical effects included three levels: (i) factor recorded in the last 365 days (365d); (ii) factor recorded > 365d-5y ago; (iii) factor not recorded in IORD in the previous 5 ys. Continuous effects denoted days since the most recent record ≤365d ago. For inpatient admissions, outpatient appointments, ED visits, blood cultures, and urine tests, the number of occurrences (and length of stay for admissions) within the last 365d were included as variables, totalling 704 variables from 377 factors.

Statistical analyses

Variable selection was based on previously published methodology applied to COVID-19.[14] In brief (details in Supplementary Methods), for each period, starting with FY2022/23–2023/24 as the most relevant to the current situation post-COVID, associations between *E. coli* bacteraemias (binary yes/no, cases/controls) and "core" variables (major a priori confounders: age, sex, ethnicity, and deprivation, rural/urban classification and catchment percentage of the primary residence, the latter capturing ascertainment) were estimated using Poisson regression (log link) with cluster robust standard errors (requiring complete data for "core" variables), considering non-linearity in and pairwise interactions between "core" variables. Poisson regression was used as it estimates absolute risk directly (rather than estimating odds from logistic regression) and provides effect estimates similar to logistic regression when event rates are low, as here.[15]

Due to missing data, the large number of variables making imputation impractical, and the risk of collinearity when including many related variables simultaneously (e.g. as in backwards elimination),[14] we used a forward selection approach, adding each of

Table 1
Percentage of risk removed if all individuals (cases and controls) were simultaneously assumed to have minimal prevalence for all variables within each of the seven disease groups.

Disease area group	% risk removed assuming minimal prevalence[a]	Variables in disease area group	Prevalence in cases, % [n]	Prevalence in controls, % [n]
Renal/urological/UTI	47%	Prosthesis insertion into ureter, urine culture taken, paralysis, urine positive for *E. coli*, kidney/ureter/biliary disease, acute renal failure, urine catheter, potassium, fluid/electrolyte disease	89% [570]	58% [97,897]
Gastrointestinal/biliary	36%	Albumin, bile duct/liver surgery, alcohol dependency, digestive/anal/rectal conditions, biliary tract disease, MR pancreatic region, inpatient admission under gastroenterology	40% [258]	22% [36,006]
Infectious diseases	30%	Skin infection, CMV/EBV screen, lymphadenitis, sepsis, non-HIV infection, blood culture taken, lymphocytes	73% [466]	38% [63,661]
Other	15%	Cataract procedure, central venous catheter, weight, haemoglobin, skin disorders	31% [197]	13% [20,951]
Healthcare visits	14%	Emergency inpatient admission, inpatient admission under geriatric consultant, inpatient admission under general surgery, complex inpatient admission, emergency department visit, number of complex inpatient admissions, palliative care	84% [535]	55% [92,622]
Cancer	8%	Pancreatic cancer, bone marrow transplant, extraction of bone marrow, chemotherapy, CT chest abdomen and pelvis	19% [121]	3% [5240]
Cardiac/respiratory	8%	Myocardial perfusion scan, pneumonia, upper respiratory disease, respiratory failure, COVID-19 test	80% [512]	59% [94,391]

[a] Calculated by taking the difference between the predicted probability of being a case given recorded exposure and the predicted probability with minimal exposure (absence of factor) in cases and controls, then dividing by the predicted probability of being a case given recorded exposure (Supplementary Methods).

> 400 "screening" variables individually to the "core" model to retain as many individuals as possible. Non-linear effects of all continuous variables were considered, and levels of categorical variables grouped based on Wald tests ($p < 0.05$). Correlation between all variables with a univariable $p < 0.25$ was calculated, excluding one variable from each pair with Spearman correlation coefficient > 0.75 to reduce collinearity (selected on a case-by-case basis based on which variable was judged clinically more meaningful for potential interventions). We then used backwards elimination on variables with univariable global $p < 0.25$ (to avoid missing important variables[16]) to identify a final model (exit p-value=0.05 for main effects and 0.01 for non-linear terms, keeping the linear component in the latter), grouping categorical variables during the process as above. Gross collinearity in the final model was assessed by identifying variables where the direction of effect switched between univariable and multivariable models when both were $p < 0.05$. Collinear variables were kept unless the sign change was clearly influenced by another variable that conflicted with clinical or epidemiological reasoning (assessed on a case-by-case basis). The final model was refitted on complete cases for all selected variables.

The full model fitting process was repeated for FY2018/19-2019/20 and FY2020/21-2021/22, comparing selected variables across the three periods; analyses were conducted separately for each period to account for the potential influence of COVID-19. Variables selected within each period were fit to all other periods to assess consistency of explained variation, summarising using pseudo-R-squared values.[17]

Measuring importance

To quantify the importance of each variable, and groups of variables, in multivariable models (jointly considering prevalence and predicted risk), we calculated the percentage of risk that could be removed from the population if all individuals (cases and controls) were assumed to have the lowest-risk level of exposure ("minimal prevalence") for that variable or group of variables (Supplementary Methods).

Targeting future interventions

We assessed how our models could be used to target future, hypothetically uniformly effective interventions, e.g. vaccination. We predicted the probability of having an *E. coli* bacteraemia from the final model for each individual, identifying the optimal threshold using the Youden index (for illustrative purposes) on a receiver operating characteristic curve. We considered how using this model-derived threshold, or 10 arbitrary criteria based on age and associated factors, as criteria for vaccination would affect sensitivity, specificity, and number vaccinated (Supplementary Methods).

All statistical models were fitted using Stata 18, with some figures generated using R version 4.4.2.

Results

Results from FY2022/23-2023/24

There were 953 potential cases and 276,758 controls in FY2022/23–2023/24 (Figs. S1, S2). 196 (21%) cases were excluded due to no prior inpatient episode < 5y ago (higher proportion outside hospital catchment and missing core variable data [Table S1]), leaving 757 cases for analysis (445 [59%] community-onset community-associated, 187 [25%] community-onset healthcare-acquired, 125 [17%] hospital-onset healthcare-associated as defined in[18]).

After fitting the core model (Fig. S3) plus each of the 377 factors, backwards elimination on those with $p < 0.25$ and investigating

Fig. 1. Adjusted associations (incidence rate ratios with 95% CIs) between selected characteristics and *E. coli* bacteraemia risk in FY2022/23–2023/24, comparing cases with controls. *Note: factor calculated from: * diagnosis codes, †procedure codes, ‡microbiology data. SHMI=Summary Hospital-level Mortality Indicator. CT Chest Abdomen Pelvis is grouped with "cancer" for illustrative purposes however it may not always be cancer-related. Factors were defined without using data from the 72 h before blood culture collection for cases to avoid reverse causality.*

collinear variables (Supplementary Results), 51 variables were selected for the final multivariable model.

Selected infectious disease-related variables were associated with higher *E. coli* bacteraemia risk (Table 1, Fig. 1). Diagnosis codes for previous lymphadenitis or microbiology tests for CMV or EBV screening ≤1 y ago were associated with higher risk: incidence rate ratio (IRR) versus > 1 y ago/never in EHR 1.7 (95% CI: 1.2–2.6) and 2.1 (1.3–3.2), respectively. Previous non-HIV infection diagnosis codes or blood cultures taken ≤5 y ago were associated with higher risk. Skin infections or septicaemia diagnosis codes 1 y ago were associated with higher risk, however, risk reduced closer to the last record (Fig. S4). Higher and lower lymphocyte levels were associated with higher risk versus the median (Fig. S5).

Several cancer-related variables were associated with higher *E. coli* bacteraemia risk (Fig. 1, Table 1), specifically pancreatic cancer diagnosis codes (IRR=2.8 versus > 1 y ago/never in EHR [95% CI

1.4–5.4]), more recent bone marrow transplant, extraction of bone marrow, chemotherapy, or computed tomography (CT) of the chest, abdomen, and pelvis.

Most selected renal/urological/UTI-related variables were associated with higher *E. coli* bacteraemia risk (Fig. 1, Table 1), e.g. prosthesis insertion into the ureter and acute renal failure diagnosis codes ≤1 y ago versus > 1 y ago/never in the EHR. Urine cultures and urine positive for *E. coli* in the last 5 y were associated with higher risk, with risk increasing further over the last year for the latter (from 1.4 (95% CI: 1.1–1.7) 1 y ago to 3.0 (2.3–3.9) 3d ago, Fig. S4). In contrast, fluid/electrolyte disorder diagnosis codes ≤5 y ago were associated with lower risk (IRR=0.75 (0.62–0.91); defined from Summary Hospital-level Mortality Indicators, including various diagnoses, e.g. volume depletion, hyperkalaemia, and hyperosmolality). Due to high amounts of missing data for HbA1c test results (typically measured in individuals with pre-diabetes/diabetes), HbA1c was excluded from backwards elimination to reduce model instability (Table S3). However, as it was highly significant univariably, we added on top of the final multivariable model, with higher HbA1c associated with higher *E. coli* bacteraemia risk (IRR 1.14 (95% CI: 1.08, 1.20) per 5 mmol/mol higher), even within normal HbA1c range (Fig. S6).

Most selected variables related to gastrointestinal disease were associated with higher *E. coli* bacteraemia risk (Table 1, Fig. 1), e.g. alcohol dependency or biliary tract disease diagnosis codes ≤5 y ago versus > 5 y ago/never in EHR. Digestive/anal/rectal condition diagnosis codes ≤1 y ago were associated with higher risk, but risk reduced closer to the last record. More recent pancreatic region magnetic resonance scans were associated with higher risk (IRR=1.3 (95% CI 1.1–1.5) per 3 months closer). Lower albumin was associated with a higher risk, plateauing from 36 g/L onwards (Fig. S5).

Most selected cardiac/respiratory-related variables were associated with lower *E. coli* bacteraemia risk (Table 1**,** Fig. 1), including more recent pneumonia or respiratory failure diagnosis codes in the last year. However, myocardial perfusion scans ≤1 y ago were associated with a higher risk (IRR=3.9 > 1 y ago/never in EHR (95% CI 2.3–6.6)).

Selected healthcare exposure variables were associated with higher and lower risk of *E. coli* bacteraemias (Fig. 1, Table 1). Emergency inpatient admissions, inpatient admissions under geriatric consultants, and general surgery inpatient admissions were associated with higher risk, while ED visits were associated with lower risk (after adjusting for other factors). More recent complex inpatient admissions were associated with higher risk; however, risk reduced by 47% (95% CI 17%–66%) per two additional complex inpatient admissions. Palliative care diagnosis codes were associated with lower risk (IRR=0.14 (0.09–0.24)).

Other variables were also associated with higher and lower *E. coli* bacteraemia risk, e.g. more recent central venous catheter procedure codes in the last year and higher weight were associated with higher risk, skin disorder diagnosis codes (within the last year) and higher haemoglobin were associated with lower risk (Table 1, Fig. 1).

Risk attribution

Impact on estimated attributable risk in cases and controls combined generally increased with prevalence in cases (Fig. 2). Renal/urological/UTI-related factors had the largest effect; 47% of the risk in cases would theoretically be removed if the whole population was like the individuals with minimal exposure (i.e. > 5 y ago/never) across all variables in this group (Table 1). These factors were highly prevalent: 89% of cases had at least one renal/urological/UTI-related factor versus 58% of controls. 87% of cases and 54% of controls had a urine culture taken in the last 5ys, with overall risk reducing by 37% if the whole population was like the individuals with no urine sample taken for culture in the last 5ys (Fig. 2). Gastrointestinal/biliary disease-related factors had the second-largest impact, with 40% of cases having at least one associated factor in this category, followed by infectious disease-related factors. Although cancer-related variables had high IRRs, their impact on risk removal was smaller, with 8% reduction assuming the whole population was like the individuals without cancer-related variables, likely due to lower prevalence (19% cases, 3% controls).

Targeting vaccination

Using the model-derived threshold to assign a hypothetically uniformly effective, targeted intervention, such as vaccination, yielded high combined sensitivity and specificity, with 88% of controls not selected for vaccination (specificity) and 77% of cases selected for vaccination (sensitivity) (Fig. 3, Fig. 4). In contrast, simply using age-based thresholds lowered both sensitivity and specificity. 66,834 people would be selected for vaccination using an age≥65 y criteria, compared with 41,676 and 20,909 using age≥75 y and model-derived thresholds, respectively. Vaccinating those with specific associated factors, plus those aged≥75 y, increased sensitivity, but lowered specificity by varying amounts, e.g. additionally vaccinating those with urine cultures taken ≤5 y ago increased sensitivity to 94%, but dropped specificity to 37%. Adding in those with blood cultures taken ≤5 y ago increased the sensitivity (85%) and reduced specificity by a smaller amount (60%). Vaccinating those aged≥75 y or with any of the seven associated factors returned low specificity (30%) and high sensitivity (97%), vaccinating more people (n=117,564). The combined sensitivity and specificity of a hypothetical vaccine was highest when applying the model-derived threshold from FY2022/23–2023/24 to the FY2018/19–2019/20 and FY2020/21–2021/22 data, compared with thresholds based on age and/or specific risk factors (Fig. S7).

Results from other calendar periods

While different individual variables were identified across the three two-year periods, similar categories were consistently selected (Fig. 5). In total, 113 variables were selected after backwards elimination in one or more periods, 83 (73%) in only one two-year period, 19 (17%) in two periods, and 11 (10%) in all three periods (Table S5). Variables selected in all three two-year periods had strong and high associations with *E. coli* bacteraemias, including urine positive for *E. coli*, blood culture taken, and chemotherapy. Many variables only selected in one period had similar characteristics selected in other periods; e.g., renal failure diagnosis codes in FY2018/19–2019/20 versus acute renal failure diagnosis codes in FY2020/21–2021/22 and FY2022/23–2023/24, and any cancer or rectal cancer diagnosis codes in FY2020/21–2021/22, liver cancer or rectal cancer diagnosis codes in FY2018/19–2019/20, and pancreatic cancer diagnosis codes in FY2022/23–2023/24. The largest difference was the increase in the cardiac/respiratory groups from FY2020 due to COVID-19, resulting in large reductions in risk removed in FY2020/21–2021/22, likely due to changes in hospital population composition (Fig. 5). Further, factors related to neurological/psychosis disorders were identified in FY2018/19–2019/20 and FY2020/21–2021/22 but not in FY2022/23–2023/24; however, their impact was small (risk removed=−9% and 8%, respectively) (Supplementary Results).

Prevalence of variables remained mostly stable across all periods (Figs. S14-S16). Most variables which varied between periods related to COVID-19, e.g., increases in previous hospital SARS-CoV-2 tests, non-HIV infection diagnosis codes, and respiratory failure diagnosis codes from FY2018/19–2019/20 to FY2020/21–2021/22.

Variation explained reduced modestly fitting variables selected in FY2018/19–2019/20 (pseudo-R-squared=22.1%) to data from FY2020/21–2021/22 (pseudo-R-squared=18.2%) and FY2022/

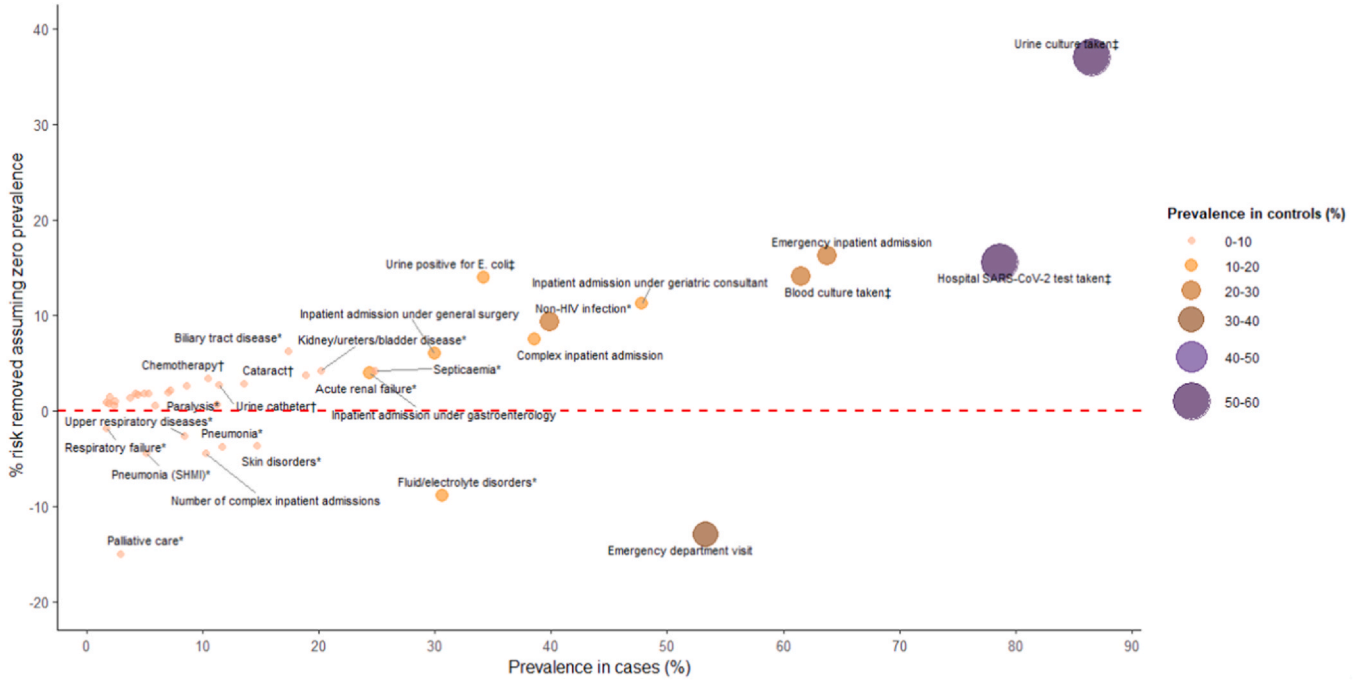


Fig. 2. Percentage of risk removed if all individuals (cases and controls) were assumed to have minimal prevalence for each characteristic individually, plotted against the prevalence of that characteristic in cases in FY2022/23–2023/24. *Note: Circle size and colour represent prevalence in controls. Factor calculated from: * diagnosis codes, †procedure codes, ‡microbiology data. The red dashed line indicates no change in risk by assuming minimal prevalence in cases and controls. Blood tests and traits are not shown on the above graph as there is no corresponding prevalence. They have the following risk removal: albumin 25%, haemoglobin 13%, lymphocytes 4%, weight −0.1%, potassium −3%. Factors were defined without using data from the 72 h before blood culture collection for cases to avoid reverse causality.*

23–2023/24 (pseudo-R-squared=19.5%) (Table S6). Results were similar in other periods.

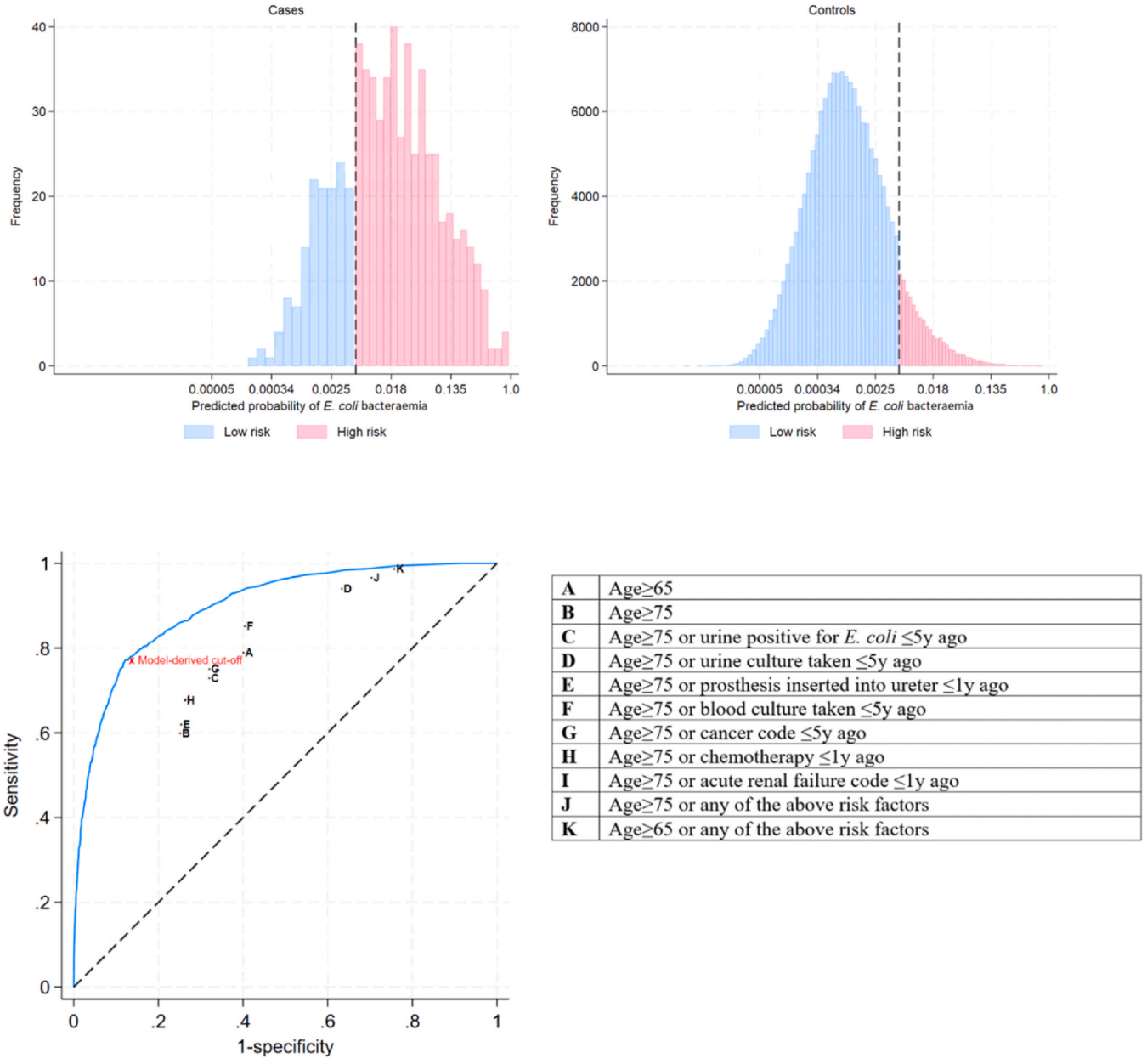
Discussion

Using our EHR-WAS approach, we identified EHR-derived factors associated with *E. coli* bacteraemias over six years. Many associations identified reflect known risk factors, e.g., in FY2022/23–2023/24, previous infectious diseases, cancer, renal/urological/UTI, and gastrointestinal/biliary-related factors were generally associated with higher risk. Respiratory illnesses were associated with lower risk of *E. coli* bacteraemia, likely reflecting common reasons for attending hospital unrelated to *E. coli* bacteraemias, given controls were hospital-exposed to reduce missing data.[19] Previous healthcare attendances were associated with higher bacteraemia risk, e.g. previous emergency inpatient admissions, and lower risk, e.g. higher numbers of complex inpatient admissions, potentially reflecting survivor bias. *E. coli* bacteraemia risk was associated with blood test results, including higher risk for individuals with lower albumin. Considering prevalence and predicted risk, targeting individuals with common associated factors, such as previous urine cultures taken, may be useful, although these were also common in controls. Associations differed across successive two-year periods, possibly due to the influence of COVID-19 on hospitalisation; however, similar groups of factors were identified, suggesting underlying risk likely remained similar.

Many factors identified in this study associated with higher *E. coli* bacteraemia risk were common markers or contributors to frailty, e.g., low albumin, low haemoglobin, frequent previous healthcare attendances, and chronic illnesses, including renal failure, cancer, and gastrointestinal conditions. This characterisation differs from MRSA and *C. difficile*, where surveillance reduced incidence, likely as these infections are caused by healthcare-associated acquisition and antimicrobial use that could be targeted by interventions. In contrast, *E. coli* bacteraemias often have an intrinsic origin from gut flora, and events leading to bacteraemia are more common in those experiencing frailty. This highlights the complexity of reducing *E. coli* bacteraemias, with interventions having to incorporate the multifaceted nature of frailty rather than there being a "silver bullet". Some specific factors identified could be targeted further, e.g. urinary catheters, identified here and in other studies.[9,20,21] Avoiding unnecessary catheter use, removing catheters when no longer needed,[22] and prioritising bladder outflow obstruction surgery could reduce risks. Patients with cancer were at increased risk,[23] reflecting risks from surgery and chemotherapy, which form a necessary part of treatment, but where there are still opportunities to mitigate risks, e.g., implementing better hand hygiene and using full-barrier precautions during central line catheter insertion, which has previously been shown to reduce infection risk.[24] We found higher *E. coli* bacteraemia risk in individuals within the normal and pre-diabetic HbA1c range, suggesting that checking HbA1c levels in people with urinary infections could help identify those at increased *E. coli* bacteraemia risk.

Using factors identified in the final multivariable models could potentially improve targeting of future prophylactic vaccines if these are effective. Compared to age-based thresholds, model-based targeting improved specificity, meaning fewer low-risk individuals would be unnecessarily vaccinated, reducing costs and any vaccine-associated risks. We assumed the vaccine would be uniformly effective; however, it may be less effective in highest-risk groups with impaired immune responses, reducing the number of bacteraemias prevented (Fig. S17). We used the Youden index to identify a model-based cut-off, balancing sensitivity and specificity, which may not be ideal. While Decision Curve Analysis[25] could be more appropriate, determining how to balance costs and benefits of this hypothetical vaccination is unclear.

While our models estimated associations between factors and *E. coli* bacteraemias rather than causal relationships, these associations could still inform interventions, provided they are reliable and generalisable. Targeting interventions at broader groups with higher



A	Age≥65
B	Age≥75
C	Age≥75 or urine positive for *E. coli* ≤5y ago
D	Age≥75 or urine culture taken ≤5y ago
E	Age≥75 or prosthesis inserted into ureter ≤1y ago
F	Age≥75 or blood culture taken ≤5y ago
G	Age≥75 or cancer code ≤5y ago
H	Age≥75 or chemotherapy ≤1y ago
I	Age≥75 or acute renal failure code ≤1y ago
J	Age≥75 or any of the above risk factors
K	Age≥65 or any of the above risk factors

Fig. 3. The predicted probability distribution in the low- and high-risk groups, as defined using the Youden index for cases (top left panel) and controls (top right panel), and Receiver Operating Characteristic (ROC) curve for the predicted probability from the logistic regression model with vaccination criteria marked using letters (bottom panel). *Note: Area Under the Curve (AUC) = 89.8% (95% CI: 88.6%–91.0%).*

overall risk rather than focusing on specific procedures may be effective. For example, we found that recent pancreatic region magnetic resonance scans were associated with higher bacteraemia risk. While this does not imply causation, these individuals may have some underlying characteristic putting them at higher risk, therefore, interventions could be targeted at these patients. Conversely, low estimated risk in those receiving palliative care may reflect differential ascertainment due to fewer samples taken in this group, rather than a protective effect. This highlights the importance of interpreting associations cautiously, while recognising their potential to guide public health strategies, even without direct causal evidence.

A challenge with our approach is that multiple features identified in EHR may represent a unified clinical pathway or condition, e.g. bone marrow aspirate, chemotherapy, bone marrow transplant, and testing for EBV/CMV may all be experienced by patients undergoing bone marrow transplants. Risks may also be captured in different ways, e.g. procedure codes for pre-operative imaging, procedure codes for surgery, and cancer diagnostic codes. This can make models more difficult to interpret and mean that different variables representing similar underlying factors were identified across different periods, as we observed. Clustering of related variables before or during model fitting could improve this.

A key study limitation was its conduct within a single hospital group, albeit large and covering around 1% of the UK population. Additionally, to reduce missing data, we restricted our analyses to individuals with inpatient episodes, excluding many individuals from the control group who were likely less comorbid, potentially causing underestimating associations. However, most cases were retained, and we observed little difference in demographic model estimates when considering a broader, healthcare-based control group.[19] We did not have access to community prescriptions and did not include hospital prescriptions as potential factors. Previous antibiotic use has been associated with higher *E. coli* bacteraemia

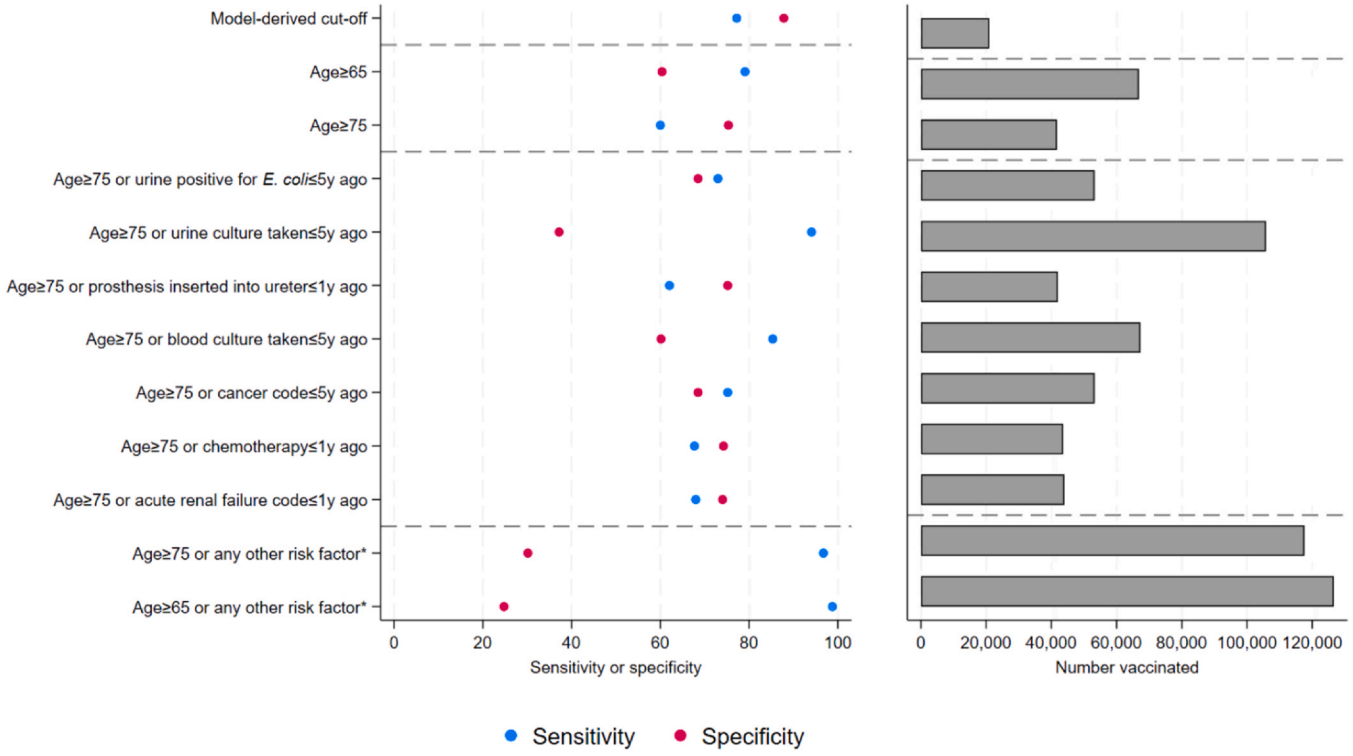


Fig. 4. Sensitivity and specificity (left) and the number of people vaccinated with a hypothetical uniformly effective vaccine (right) using different vaccination criteria. **Any of the seven individual risk factors presented in the main panel, specifically urine positive for E. coli (≤5 y ago), urine culture taken (≤5 y ago), prosthesis inserted into the ureter (≤1 y ago), blood culture taken (≤5 y ago), diagnosis code for cancer (≤5 y ago), procedure code for chemotherapy (≤1 y ago), or any diagnosis code for acute renal failure (≤1 y ago). Note: For clarity, horizontal dashed lines separate model-derived cut-offs, age groups, specific risk factors, and all risk factors combined.*

incidence.[26] Other drugs, including anti-depressants, can also affect the gut microbiome,[27] though their impact on bacteraemia risk is unclear. Including prescriptions may be useful if community prescribing data is available. While we considered 377 potential risk factors, this was not an exhaustive list, and other characteristics could be added in future research. Further, we did not consider mortality as an outcome in our study and therefore cannot comment on trends in case fatality rates despite observed rises in cases.

Another limitation was the relatively small number of cases, although we found highly significant effects. The observed variation in associated factors over two-year periods may reflect true changes or model instability, which could be assessed using bootstrapping.[28] Further, screening many variables increased the risk of type 1 error; however, our analysis focused on identifying new risk factors, and broadly similar risk factor groups were observed across all three financial year periods, adding confidence to identified associations. To increase statistical power, future studies could use larger, national-level datasets which would also improve generalisability, allow analysis stratified by subgroups of interest (e.g. community-acquired, hospital-onset cases), and reduce the risk of missing bacteraemias; using Oxfordshire data alone may miss individuals receiving care outside the region. Although *E. coli* bacteraemias are

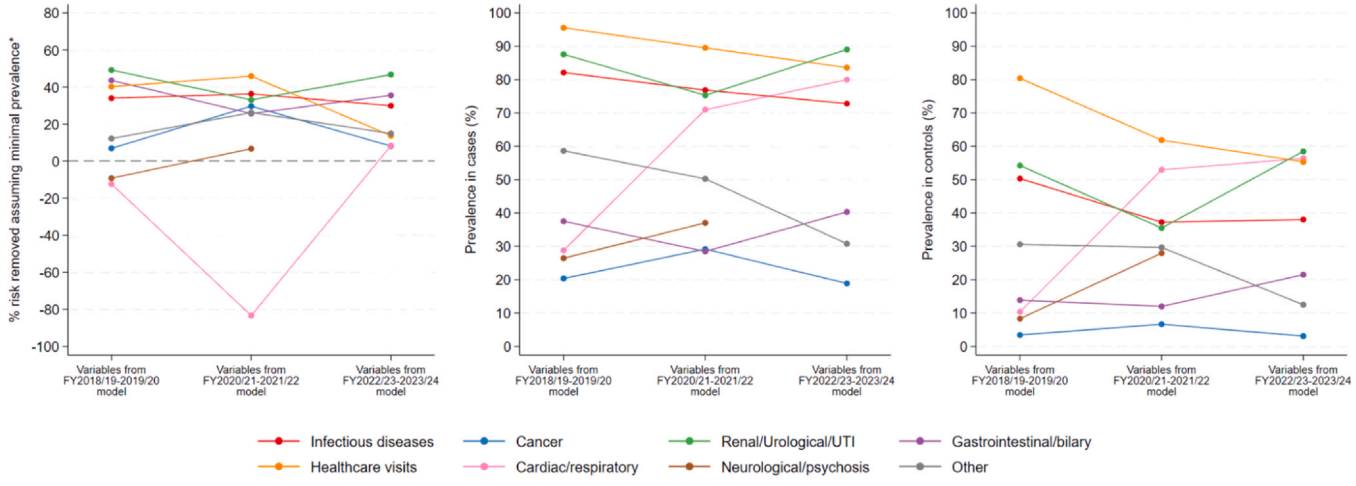


Fig. 5. Percentage of risk removed assuming minimal prevalence across all individuals (cases and controls) (left), and prevalence in cases (middle) and controls (right) across FY2018/19–2019/20, FY2020/21–2021/22, and FY2022/23–2023/24. **calculated by taking the difference between the predicted probability of being a case given recorded exposure and the predicted probability with minimal exposure (in cases and controls), then dividing by the predicted probability of being a case given recorded exposure (Supplementary Methods). Negative means risk increased at the population level.*

captured by UKHSA's Second Generation Surveillance System, which has previously been linked to Hospital Episode Statistics, the national dataset would not include key factors identified in this study, such as blood tests, vital signs, and culture-negative microbiology results. Selecting cases and controls while balancing missing data and bias is complex. We excluded cases without previous inpatient episodes; their lower hospital interactions may mean different risk factors and interventions are needed. However, without access to primary care data, many factors for these individuals could not be derived. Future studies could investigate factors in populations with varying amounts of healthcare contact.

Overall, we found that *E. coli* bacteraemias were largely associated with known risk factors and frailty, explaining why enhanced surveillance has not led to reduced incidence. Our study also demonstrates an EHR-WAS approach that can be used with EHR data to identify associated factors without constraining the search by prior knowledge. This may have applications to other infectious diseases and particularly how associated factors change over time.

Funding

This study was funded by the National Institute for Health Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at Oxford University in partnership with the UK Health Security Agency (UKHSA) (NIHR200915) and the NIHR Biomedical Research Centre, Oxford. DWE is supported by a Robertson Fellowship. ASW is an NIHR Senior Investigator. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or the UKHSA.

Author contributions

The study was designed and planned by EP, DWE, and ASW. EP conducted the statistical analysis of the data. EP, DWE, and ASW drafted the manuscript and all authors contributed to interpretation of the data and results and revised the manuscript. All authors approved the final version of the manuscript.

Data availability

The datasets analysed during the current study are not publicly available as they contain personal data but are available from the Infections in Oxfordshire Research Database (https://oxfordbrc.nihr.ac.uk/research-themes-overview/antimicrobial-resistance-and-modernising-microbiology/infections-in-oxfordshire-research-database-iord/), subject to an application and research proposal meeting the ethical and governance requirements of the Database. For further details on how to apply for access to the data and for a research proposal template, please email iord@ndm.ox.ac.uk.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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 Peto, TP Quan, J Robinson (community), J Sellors, B Shine, AS Walker, D Waller. Patient and Public Panel: G Blower, C Mancey, P McLoughlin, B Nichols.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jinf.2025.106612.

References

1. UK Health Security Agency. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) report; 2024, 100282. ⟨https://www.gov.uk/government/publications/english-surveillance-programme-antimicrobial-utilisation-and-resistance-espaur-report⟩ (accessed 12 June 2024).
2. UK Health Security Agency. Annual epidemiological commentary: Gram-negative, MRSA, MSSA bacteraemia and C. difficile infections, up to and including financial year 2023 to 2024; 2024. ⟨https://www.gov.uk/government/statistics/mrsa-mssa-and-e-coli-bacteraemia-and-c-difficile-infection-annual-epidemiological-commentary/annual-epidemiological-commentary-gram-negative-mrsa-mssa-bacteraemia-and-c-difficile-infections-up-to-and-including-financial-year-2023-to-2024#meticillin-resistant-staphylococcus-aureus-bacteraemia-mrsa⟩ (accessed 26 February 2025).
3. Johnson AP, Davies J, Guy R, Abernethy J, Sheridan E, Pearson A, et al. *Mandatory surveillance of methicillin-resistant Staphylococcus aureus (MRSA) bacteraemia in England: the first 10 years. J Antimicrob Chemother* 2012;**67**(4):802–9.
4. UK Health Security Agency. Mandatory healthcare associated infection surveillance: data quality statement for April 2019 to March 2020; 2022. Accessed on: 29th August 2025. https://assets.publishing.service.gov.uk/media/62a095fdd3bf7f03667c65ea/mandatory-healthcare-associated-infection-surveillance-data-quality-statement-FY2019-to-FY2020.pdf.
5. UK Health Security Agency. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR). Report 2021 to 2022; 2022. Accesed on: 29th August 2025. https://webarchive.nationalarchives.gov.uk/ukgwa/20231002172235/https://www.gov.uk/government/publications/english-surveillance-programme-antimicrobial-utilisation-and-resistance-espaur-report.
6. UK Health Security Agency. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR). Report 2022 to 2023; 2023. Accessed on: 29th August 2025. https://webarchive.nationalarchives.gov.uk/ukgwa/20240201222414/https://www.gov.uk/government/publications/english-surveillance-programme-antimicrobial-utilisation-and-resistance-espaur-report.
7. Bhattacharya A, Collin SM, Stimson J, Thelwall S, Nsonwu O, Gerver S, et al. *Healthcare-associated COVID-19 in England: a national data linkage study. J Infect* 2021;**83**(5):565–72.
8. Laupland KB, Gregson DB, Church DL, Ross T, Pitout JD. *Incidence, risk factors and outcomes of Escherichia coli bloodstream infections in a large Canadian region. Clin Microbiol Infect* 2008;**14**(11):1041–7.
9. Jackson LA, Benson P, Neuzil KM, Grandjean M, Marino JL. *Burden of community-onset Escherichia coli bacteremia in seniors. J Infect Dis* 2005;**191**(9):1523–9.
10. Song J, Walters A, Berridge D, Akbari A, Evans M, Lyons RA. *Risk factors for Escherichia coli bacteraemia: a population-based case-control study. Lancet* 2017;**390**:S85.
11. Sanofi. Press Release: Update on extraintestinal pathogenic E. coli vaccine phase 3 clinical study; 2025. ⟨https://www.sanofi.com/en/media-room/press-releases/2025/2025-02-13-06-00-00-3025576⟩ (accessed 26 February 2025).
12. Ministry of Housing, Communities & Local Government. The English Indices of Deprivation 2019 (IoD2019); 2019. Accessed on: 29th August 2025. https://assets.publishing.service.gov.uk/media/5d8e26f6ed915d5570c6cc55/IoD2019_Statistical_Release.pdf.
13. Office for Health Improvement & Disparities . NHS Acute (Hospital) Trust Catchment Populations; 2022. ⟨https://app.powerbi.com/view?r=eyJrIjoiODZmNGQ0YzItZDAwZi00MzFiLWE4NzAtMzVmNTUwMThmMTVlIiwidCI6ImVlNGUxNDk5LTRhMzUtNGIyZS1hZDQ3LTVmM2NmOWRlODY2NiIsImMiOjh9⟩ (accessed 31 Jan 2024).
14. Pritchard E, Jones J, Vihta KD, Stoesser N, Matthews PPC, Eyre DW, et al. *Monitoring populations at increased risk for SARS-CoV-2 infection in the community using population-level demographic and behavioural surveillance. Lancet Reg Health Eur* 2022;**13**:100282.
15. Zou G. *A modified poisson regression approach to prospective studies with binary data. Am J Epidemiol* 2004;**159**(7):702–6.
16. Mickey RM, Greenland S. *The impact of confounder selection criteria on effect estimation. Am J Epidemiol* 1989;**129**(1):125–37.
17. McFadden D. Conditional Logit Analysis of Qualitative Choice Behavior: Institute of Urban and Regional Development, University of California; 1973. Accessed on: 29th August 2025. https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf.
18. GOV.UK. Quarterly epidemiological commentary: Mandatory Gram-negative bacteraemia, MRSA, MSSA and C. difficile infections (data up to April to June 2023); 2025. ⟨https://www.gov.uk/government/statistics/mrsa-mssa-gram-negative-bacteraemia-and-cdi-quarterly-report/quarterly-epidemiological-commentary-mandatory-gram-negative-bacteraemia-mrsa-mssa-and-c-difficile-infections-data-up-to-april-to-june-2023⟩ (accessed 12 March 2025).

19. Pritchard E, Vihta KD, Pouwels KB, Lipworth S, Hope R, Muller-Pebody B, et al. *The effect of population selection criteria on model estimates and data missingness in electronic health record studies. medRxiv* 2025. Feb 12:2025-02 https://www.medrxiv.org/content/10.1101/2025.02.10.25321999v1.
20. Rodríguez-Baño J, Picón E, Gijón P, Hernández JR, Ruíz M, Peña C, et al. *Community-onset bacteremia due to extended-spectrum beta-lactamase-producing Escherichia coli: risk factors and prognosis. Clin Infect Dis* 2010;**50**(1):40–8.
21. Trautner BW, Darouiche RO. *Catheter-associated infections: pathogenesis affects prevention. Arch Intern Med* 2004;**164**(8):842–50.
22. Meddings J, Rogers MA, Krein SL, Fakih MG, Olmsted RN, Saint S. *Reducing unnecessary urinary catheter use and other strategies to prevent catheter-associated urinary tract infection: an integrative review. BMJ Qual Saf* 2014;**23**(4):277–89.
23. Zhang Q, Gao HY, Li D, Li Z, Qi SS, Zheng S, et al. *Clinical outcome of Escherichia coli bloodstream infection in cancer patients with/without biofilm formation: a single-center retrospective study. Infect Drug Resist* 2019;**12**:359–71.
24. Pronovost P, Needham D, Berenholtz S, Sinopoli D, Chu H, Cosgrove S, et al. *An intervention to decrease catheter-related bloodstream infections in the ICU. N Engl J Med* 2006;**355**(26):2725–32.
25. Vickers AJ, Elkin EB. *Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making* 2006;**26**(6):565–74.
26. van der Mee-Marquet NL, Blanc DS, Gbaguidi-Haore H, Dos Santos Borges S, Viboud Q, Bertrand X, et al. *Marked increase in incidence for bloodstream infections due to Escherichia coli, a side effect of previous antibiotic therapy in the elderly. Front Microbiol* 2015;**6**:646.
27. Letchumanan V, Thye AY-K, Tan LT-H, Law JWF, Johnson D, Ser HL, et al. *IDDF2021-ABS-0164 Gut feelings in depression: microbiota dysbiosis in response to antidepressants. Gut* 2021:A49–50.
28. Royston P, Sauerbrei W. *Bootstrap assessment of the stability of multivariable models. Stata J* 2009;**9**(4):547–70.