Using electronic health records to improve management of *E. coli* bloodstream infections



Karina-Doris Vihta

Nuffield Department of Medicine

Corpus Christi College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2019

Using electronic health records to improve management of *E. coli* bloodstream infections

Karina-Doris Vihta, Corpus Christi College, Nuffield Department of Medicine

A thesis submitted for the degree of Doctor of Philosophy, Trinity 2019

The aim of this thesis is to use linked electronic health records of hospital admissions, and microbiology and haematology results in order to inform management of infections in Oxfordshire, and more generally worldwide. *Escherichia coli* is one of the leading bacterial pathogens causing bloodstream infections; not only is overall incidence rising, but rising resistance to commonly used empiric antibiotics is also a major concern. Despite playing such a significant role in the burden of infections, the epidemiology of *E. coli* bloodstream infections is still unclear, particularly when considering unselected populations.

I showed that the rise in *E. coli* bloodstream infections in Oxfordshire is driven by truly community-associated cases, that is, cases identified in the first 48 hours of a hospital admission, or outside of a hospital admission, who had not been admitted to the hospital in the past year. Interestingly, the rate of increase in incidence was faster the further away the previous hospital admission had been. However, rising incidence did not seem to be driven by increasing numbers of patients with evidence of a previous urinary tract infection. The number of co-amoxiclav resistant infections was rising significantly faster than the number of co-amoxiclav susceptible infections and the highest number of resistant infections in 2016 was seen in community-associated cases. However, considering 30-day mortality and various biomarkers of infection, there was no evidence for changes in the severity of infections over time. Higher co-amoxiclav use in primary care was associated with higher rates of co-amoxiclav resistant *E. coli* urinary tract infections in the subsequent year, providing the evidence needed in order to support the aim of lowering inappropriate use of broad-spectrum antibiotics in the community.

Co-amoxiclav susceptibility is particularly challenging to define with traditional laboratory testing methods, with different methodologies leading to different results both in terms of minimum inhibitory concentrations, but particularly when considering a dichotomised susceptible/resistant phenotype. Recent advances in whole genome sequencing technology and analysis tools, as well as decreases in costs, increase the potential utility of predicting phenotype from sequencing-derived genotype, particularly for challenging bacteria-drug combinations. I found that machine learning algorithms – statistical methods which learn from patterns within the data without being programmed explicitly – could predict co-amoxiclav resistance where information is extracted from assembled sequences, either through extracting information about genetic features from mapping onto resistance databases or by considering presence and absence of DNA ‘words’ (k-mers). Crucially, feature selection and expert knowledge are not required when constructing these matrices of genetic features, making these algorithms particularly appealing when considering the constant accumulation of new genetic resistance mechanisms.

Finally, by comparing the proportions of resistant bloodstream infections and infections in other less invasive sites, I found that the latter could be used as a surveillance tool for antimicrobial resistance in low- and middle-income countries, since these sites are easier to sample from and cheaper to carry out antimicrobial susceptibility testing on.

Acknowledgements

I am particularly grateful to my main supervisor, Sarah, without whom this thesis would have not been possible. She has been the dream supervisor and more, the reason for which I chose this DPhil and I can wholeheartedly say would do it all over again.

I am also grateful to the whole Modernising Medical Microbiology group for the support they have offered me throughout my DPhil. It was an incredible opportunity to be able to discuss with such passionate talented individuals, particularly Tim Peto, David Eyre, Nicole Stoesser, Daniel Wilson, Phuong Quan, Bernadette Young.

To my partner, Oliviu, who said yes to staying in the UK in order for me to pursue this lifelong ambition, and reminded me when I needed it most that I was capable of doing it. My brother, Hanzi, for inspiring me to be present and make the most out of every second, as life is a gift, and the opportunity to learn is priceless. To my mum for always wanting to hear and read about all of my work, and to my dad who gave me the opportunity to turn my dream of the University of Oxford into reality.

This work uses data provided by patients and collected by the NHS as part of their care and support. I thank all the people of Oxfordshire who contribute to the Infections in Oxfordshire Research Database (IORD).

Research Database Team: R Alstead, C Bunch, DCW Crook, J Davies, J Finney, J Gearing (community), H Jones, L O'Connor, TEA Peto (PI), TP Quan, J Robinson (community), B Shine, AS Walker, D Waller, D Wyllie. Patient and Public Panel: G Blower, C Mancey, P McLoughlin, B Nichols.

NIHR Health Protection Research Unit Steering Committee: J Coia, N French, C Marwick, M Sharland.

Funding for all Chapters

The research was funded by my NDM Prize Studentship and the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with Public Health England (PHE) [HPRU-2012-10041], and the NIHR Oxford Biomedical Research Centre.

Ethics Statement for all Chapters

IORD has Research Ethics Committee and Confidentiality Advisory Group approvals (14/SC/1069, ECC5-017(A)/2009) as a generic electronic research database, without individual patient consent.

Declaration and attributions

I, Karina-Doris Vihta, designed and conducted all the analyses presented in this thesis with the support of my supervisors Ann Sarah Walker, Tim Peto, David Eyre and Derrick Crook without whom this thesis certainly would not have been possible.

Chapter 2 attributions and acknowledgements

- Data extracted from IORD by Phuong Quan.
- Grouping of diagnostic codes done by Nicola Fawcett and Nicole Stoesser.

Chapter 3 attributions and acknowledgements

- Isolates grown and sequenced by others, and re-phenotyping done by others.
- Mapping to resistance databases done by Tim Davies.
- Phylogenetic trees were built by Sam Lipworth.
- A special thank you to Samaneh Kouchaki for her guidance on some of the coding of the machine learning algorithms.
- A special thank you to Sarah Earle and Daniel Wilson for their guidance on carrying out the bacterial genome-wide association study.

Chapter 4 attributions and acknowledgements

- A very special thank you to Nick White who saw the opportunity to answer such an important question with the data which was available from IORD, taking the goal of this thesis of informing management of infections in Oxfordshire to informing management of infections in low- and middle-income countries.
- Also, a very special thank you to Nicola C. Gordon for the very fruitful discussions on the reality of the great challenges of setting up a surveillance system for antimicrobial resistance in blood cultures in a low-income setting and therefore the motivation for the usefulness of the analyses presented in this Chapter.

Papers arising directly from this thesis

Vihta, K.-D., Stoesser, N., Llewelyn, M. J., Quan, T. P., Davies, T., Fawcett, N. J., … Walker, A.S. (2018). 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. The Lancet Infectious Diseases, 18(10), 1138–1149. https://doi.org/http://dx.doi.org/10.1016/S1473-3099%2818%2930353-0

Other papers during the course of the thesis

Zhong, L.-L., Phan, H. T. T., Shen, C., Vihta, K.-D., Sheppard, A. E., Huang, X., … Tian, G.-B. (2018). High rates of human fecal carriage of mcr-1-positive multidrug-resistant enterobacteriaceae emerge in China in association with successful plasmid families. Clinical Infectious Diseases, 66(5). https://doi.org/10.1093/cid/cix885

Table of Contents

1 INTRODUCTION 17

1.1 LINKING HEALTH DATA 17

1.2 ANTIBIOTICS AND THE IMPORTANCE OF INFECTIOUS DISEASES 21

1.2.1 Bloodstream infections in the UK 28

1.2.2 Antimicrobial susceptibility testing in the UK 32

1.2.3 Low and middle-income countries 36

1.3 THESIS OUTLINE 39

2 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 40

2.1 INTRODUCTION 40

2.2 METHODS 42

2.2.1 Statistical analysis 50

2.3 RESULTS 55

2.3.1 Impact of population and sampling on E. coli bloodstream infection incidence 70

2.3.2 Disease severity of *E. coli* bloodstream infections 75

2.3.3 Impact of previous illness on *E. coli* bloodstream infection 78

2.3.4 Antimicrobial susceptibility 83

2.3.4.1 Agreement between susceptibility testing methodologies 83

2.3.4.2 Trends in antibiotic resistant infections 84

2.3.4.3 Severity of co-amoxiclav resistant vs susceptible *E. coli* bloodstream infections 90

2.3.4.4 Associations with hospital and community antibiotic use 93

2.4 DISCUSSION 97

3 PREDICTING ANTIMICROBIAL RESISTANCE IN *E. COLI* BLOODSTREAM INFECTIONS USING MACHINE LEARNING METHODS AND BACTERIAL GENOME-WIDE ASSOCIATION STUDIES 106

3.1 INTRODUCTION 106

3.2 METHODS 107

3.2.1 Study design, population and laboratory phenotype 107
3.2.2 DNA extraction and sequencing 108
3.2.3 Predicting phenotype using WGS and machine learning algorithms 109
3.2.3.1 Input features 109
3.2.3.1.1 Presence and absence of known co-amoxiclav resistance determinants 109
3.2.3.1.2 Adding relative coverage 113
3.2.3.1.3 Adding other resistance variant detection methods for all transmissible elements in the Resfinder database, together with quality metrics 114
3.2.3.1.4 31-mers 116
3.2.3.2 Output features 119
3.2.3.3 Machine learning algorithms 120
3.2.3.3.1 Binary phenotype 121
• Base classifiers 121
• Ensemble methods 126
3.2.3.3.2 MIC 129
• Ignoring order in MICs (treating the output as multiclass) 129
• Ordinal models or threshold-based models 130
• Regression-based models (treating the output as continuous) 132
3.2.3.3.3 Feature importance 134
3.2.3.4 Pipeline for testing classifiers 134
3.2.3.4.1 Setting the threshold for calling resistance 136
3.2.3.4.2 Hyperparameter tuning 137
3.2.3.5 Performance metrics 138
3.2.4 Bacterial genome-wide association study 139
3.2.5 *Bugwas* prediction in relation to machine learning algorithms 144
3.3 RESULTS 145
3.3.1 Routine laboratory phenotypes and genotypes 145
3.3.2 Binary phenotype: comparing machine learning algorithms 146
3.3.3 Bacterial genome-wide association study 150
3.3.3.1 Binary phenotype prediction using GWAS 153
3.3.3.2 MIC prediction using GWAS 156

3.3.4 Bacterial genome-wide association study as a feature selection tool for binary prediction 161
3.3.5 Prediction of minimum inhibitory concentration 162
3.3.6 Feature importance 166
3.4 DISCUSSION 169
3.4.1 Choice of input features 169
3.4.2 Choice of output features 174
3.4.3 Choice of methods 175
3.4.4 Comparison with other studies 179
3.4.5 General issues 185
4 CAN RESISTANCE IN BLOODSTREAM INFECTIONS BE ESTIMATED FROM RESISTANCE IN OTHER TYPES OF SPECIMEN? – A NEW METHOD FOR LOW-INCOME COUNTRIES TO MONITOR ANTIMICROBIAL RESISTANCE 187
4.1 INTRODUCTION 187
4.2 METHODS 192
4.3 RESULTS 198
4.3.1 Population 198
4.3.1.1 *E. coli* infections 198
4.3.1.2 *S. aureus* infections 209
4.3.1.3 Pooling all antibiotics 219
4.3.2 Individual-level analyses 222
4.3.3 *E. coli* co-amoxiclav breakpoint 227
4.4 DISCUSSION 228
4.5 SUPPLEMENTARY MATERIAL 235
5 CONCLUSIONS AND FUTURE WORK 244
6 BIBLIOGRAPHY 258

Table of Figures

Figure 1 Timeline of antibiotic deployment and observed antibiotic resistance 22

Figure 2 Changes in the top 10 causes of death in the United States between 1900 and 1997 23

Figure 3 Global distribution of CPE 27

Figure 4 Trends in the rate of E. coli bloodstream infections in England financial years 2012/2013 to 2018/2019 29

Figure 5 Illustration of classification of infections based on previous hospital-exposure 45

Figure 6 Annual quasi-nosocomial E. coli bloodstream infections categorized according to the diagnostic code in the preceding admission 57

Figure 7 Monthly (A) E. coli bloodstream infection and (B) E. coli UTI according to recent hospital-exposure (first and recurrent infections combined) 61

Figure 8 Monthly E. coli bloodstream infection (first and recurrent infections combined) using different definitions of recency of hospital exposure 62

Figure 9 Monthly E. coli bloodstream infection according to recent hospital-exposure (A) first per patient only (B) recurrences within a patient only 63

Figure 10 Monthly E. coli UTIs according to recent hospital-exposure (A) first per patient only (B) recurrences within a patient only 66

Figure 11 Summary of incidence trends in 2016 for (A) E. coli bloodstream infections, (B) E. coli UTIs 67

Figure 12 Monthly blood samples submitted for culture regardless of result according to recent hospital-exposure (first and repeat samples per patient) 71

Figure 13 Trends in haematology/biochemistry test results and 30-day mortality following a blood culture being taken regardless of its result according to recent hospital-exposure (first and recurrent infections) 72

Figure 14 Monthly urine samples submitted for culture regardless of result according to recent hospital-exposure (first and repeat samples per patient) 73

Figure 15 Trends in haematology/biochemistry test results and 30-day mortality following E. coli bloodstream infection according to recent hospital-exposure (first and recurrent infections) 76

Figure 16 E. coli bloodstream infections according to whether (A) patient had had an E. coli UTI ≥3 days previously (B) patient had not had an E. coli UTI ≥3 days previously (C) E. coli bloodstream infections according to time from previous E. coli UTI 79

Figure 17 Annual E. coli bloodstream infections according to recent hospital-exposure and urine sample submission/results 80
Figure 18 E. coli bloodstream infections according to whether (A) patient had had a catheter urine specimen (CSU) previously (B) patient had not had a CSU previously (C) E. coli bloodstream infections according to time from previous CSU 82
Figure 19 Co-amoxiclav resistance in E. coli bloodstream infections according to different testing methods 83
Figure 20 Annual E. coli bloodstream infections that are susceptible and resistant to co-amoxiclav, with and without resistance to gentamicin and ciprofloxacin, according to recent hospital-exposure 85
Figure 21 Annual E. coli bloodstream infections according to recent hospital-exposure and antibiotic susceptibility (A) by antibiotic, (B) summary 88
Figure 22 Annual E. coli UTIs according to recent hospital-exposure and antibiotic susceptibility 89
Figure 23 Severity of co-amoxiclav-resistant vs susceptible E. coli bloodstream infections, by (A) neutrophil counts and (B) 30-day mortality across hospital exposure groups 91
Figure 24 Summary of incidence trends in 2016 for severity of co-amoxiclav resistant and susceptible E. coli bloodstream infections. 92
Figure 25 Community co-amoxiclav-resistant E. coli UTIs (A), community E. coli UTIs (B) and community urine samples submitted regardless of result (C) per 1000 patients per general practice 2012-2016 compared with co-amoxiclav usage 96
Figure 26 All co-amoxiclav-resistant E. coli UTIs regardless of hospital-exposure (A), E. coli UTIs (B) and urine samples submitted regardless of result (C), per 1000 patients per general practice 2012-2016 compared with co-amoxiclav usage 96
Figure 27 Overview of ARIBA mapping and targeted assembly pipeline 110
Figure 28 Gradient descent for optimization of the cost function 123
Figure 29 Toy example illustrating the power of kernel methods 125
Figure 30 Bias-variance trade-off 126
Figure 31 Loss functions from Top: Ordinal Immediate-Threshold (IT), Bottom: Ordinal All-Threshold (AT) 131
Figure 32 SVR prediction varying ε. 133
Figure 33 Splitting of the data during cross-validation 135

Figure 34 Additional splitting of the data for internal validation of the resistance calling threshold during cross-validation 136
Figure 35 MIC distributions of E. coli isolated from urology services in the Netherlands (1998–2005) 146
Figure 36 Amoxicillin resistance prediction performance of various machine learning classifiers applied to the first 3 input feature sets and a binary phenotype 148
Figure 37 Co-amoxiclav resistance prediction performance of various machine learning classifiers applied to the first 3 input feature sets and a binary phenotype 150
Figure 38 Leading principal components correspond to major lineages in bacterial phylogenies 151
Figure 39 Wald test of significance of lineage associations of E. coli principal components with co-amoxiclav resistance 152
Figure 40 The E. coli phylogeny annotated by co-amoxiclav resistance and principal component projections of PC 3 153
Figure 41 Phylogenetic phenotype prediction distribution 154
Figure 42 Significance of association of the presence/absence of 31-mers with co-amoxiclav resistance after controlling for population structure using LMM, mapped to E. coli CFT073 complete genome, and two ST131 candidates 155
Figure 43 Co-amoxiclav resistance prediction performance of various machine learning classifiers applied to all 5 input feature sets and a binary phenotype 162
Figure 44 Co-amoxiclav MIC prediction performance of various machine learning classifiers applied to all 5 input feature sets: essential agreement and accuracy (categorical agreement) 165
Figure 45 Co-amoxiclav MIC prediction performance of various machine learning classifiers applied to all 5 input feature sets: sensitivity and specificity 166
Figure 46 Number of (A) E. coli bloodstream infections and (B) E. coli urinary tract infections split by susceptibility to amoxicillin, co-amoxiclav, ciprofloxacin, trimethoprim, 1998-2018 200
Figure 47 Analyses of agreement between proportion of co-amoxiclav-resistant E. coli bloodstream infections and proportion of co-amoxiclav-resistant E. coli urinary tract infections 204
Figure 48 Analyses of agreement between proportion of resistant E. coli bloodstream infections and proportion of resistant E. coli urinary tract infections, for amoxicillin, co-amoxiclav, ciprofloxacin and trimethoprim .. 207

Figure 49 Number of (A) S. aureus bloodstream infections (B) S. aureus non-blood cultures split by susceptibility to ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin, tetracycline, 1998-2018 210
Figure 50 Number of (A) S. aureus sterile site cultures (excluding blood cultures) (B) S. aureus non-sterile site cultures split by susceptibility to ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin, tetracycline, 1998-2018 211
Figure 51 Proportion ciprofloxacin-resistant blood cultures vs non-blood cultures, sterile site cultures and non-sterile site cultures 213
Figure 52 Proportion erythromycin-resistant blood cultures vs non-blood cultures, sterile site cultures and non-sterile site cultures 214
Figure 53 Agreement between proportion of resistant S. aureus bloodstream infections and proportion of resistant S. aureus non-sterile site cultures, for ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin, tetracycline 215
Figure 54 Proportion resistant non-blood cultures versus proportion resistant blood cultures per year for all antibiotics 220
Figure 55 Counts (left) and proportions (right) of drug-years with varying proportions of resistant blood cultures classified by the proportion with resistant urine/non-sterile site cultures 221
Figure 56 Venn diagram of E. coli urinary tract infections, E. coli bloodstream infections, S. aureus non-sterile site cultures, S. aureus non-blood cultures and S. aureus bloodstream infections 223
Figure 57 Co-amoxiclav MIC distribution per month post EUCAST breakpoint change 227
Figure 58 Agreement between proportion of amoxicillin-resistant E. coli bloodstream infections and proportion of amoxicillin-resistant E. coli urinary tract infections 235
Figure 59 Agreement between proportion of ciprofloxacin-resistant E. coli bloodstream infections and proportion of ciprofloxacin-resistant E. coli urinary tract infections 236
Figure 60 Agreement between proportion of trimethoprim-resistant E. coli bloodstream infections and proportion of trimethoprim-resistant E. coli urinary tract infection 237
Figure 61 Agreement between proportion of ciprofloxacin-resistant S. aureus bloodstream infections and proportion of ciprofloxacin-resistant S. aureus non-sterile site cultures 238
Figure 62 Agreement between proportion of erythromycin-resistant S. aureus bloodstream infections and proportion of erythromycin-resistant S. aureus non-sterile site cultures 239

Figure 63 Agreement between proportion of flucloxacillin-resistant S. aureus bloodstream infections and proportion of flucloxacillin-resistant S. aureus non-sterile site cultures 240
Figure 64 Agreement between proportion of gentamicin-resistant S. aureus bloodstream infections and proportion of gentamicin-resistant S. aureus non-sterile site cultures 241
Figure 65 Agreement between proportion of penicillin-resistant S. aureus bloodstream infections and proportion of penicillin-resistant S. aureus non-sterile site cultures 242
Figure 66 Agreement between proportion of tetracycline-resistant S. aureus bloodstream infections and proportion of tetracycline-resistant S. aureus non-sterile site cultures 243

Table of Tables

Table 1 Primary diagnostic code for the previous admission for quasi-nosocomial E. coli bloodstream infections . 58

Table 2 Relative contribution of recurrent E. coli bloodstream infections and E. coli UTIs to total numbers in 2016 by recent hospital-exposure 65

Table 3 Summary of current (2016) annual incidence rate ratios 68

Table 4 Overall E. coli bloodstream infection incidence trends in 2016, unadjusted and standardized to the sex and age population of Oxfordshire in 1998 74

Table 5 Incidence trends for first E. coli bloodstream infection per patient, unadjusted and standardized to the sex and age population of Oxfordshire in 1998 74

Table 6 Yearly numerators/denominators (percentages) for 30-day mortality following E. coli bloodstream infection, co-amoxiclav resistant E. coli bloodstream infections and 30-day mortality following E. coli bloodstream infection in co-amoxiclav resistant versus susceptible cases 77

Table 7 Total number and percentage of E. coli bloodstream infections and E. coli UTIs tested for each antibiotic and resistant to each antibiotic over the whole period and in 2016 86

Table 8 Cross-correlation between hospital antimicrobial prescribing and co-amoxiclav-resistant nosocomial E. coli bloodstream infection 93

Table 9 Cross-correlation between hospital antimicrobial prescribing and co-amoxiclav-resistant E. coli bloodstream infections regardless of hospital-exposure 94

Table 10 Distribution of co-amoxiclav MICs within 976 E. coli bloodstream infections 145

Table 11 Co-amoxiclav phenotype and predicted genotype from GWAS using $\log_2$(MIC) as the outcome and first strategy for interval censoring 157

Table 12 Co-amoxiclav phenotype versus predicted genotype from GWAS using $\log_2$(MIC) as the outcome and second strategy for interval censoring 158

Table 13 Co-amoxiclav phenotype versus predicted genotype from GWAS using $\log_2$(MIC) as the outcome and third strategy for interval censoring 159

Table 14 Co-amoxiclav phenotype versus predicted genotype from GWAS using $\log_2$(MIC) as the outcome (third interval censoring strategy): excluding ST131 E. coli bloodstream infections 160

Table 15 Top 20 features for co-amoxiclav MIC prediction from 'known + relative coverage' feature sets using the Ordinal SE classifier and the squared weights of the features for ranking 167

TABLE 16 TOP 20 FEATURES FOR CO-AMOXICLAV MIC PREDICTION FROM 'KNOWN + RELATIVE COVERAGE' USING THE ORDINAL SE CLASSIFIER AND RECURSIVE FEATURE ELIMINATION FOR RANKING 168

TABLE 17 SUMMARY OF KEY RESULTS FOR AGREEMENT BETWEEN PROPORTION RESISTANT IN BLOOD CULTURES VS NON-STERILE SITE CULTURES 218

TABLE 18 PERFORMANCE OF USING SUSCEPTIBILITY RESULTS WITHIN THE SAME PATIENT IN OTHER SITES OF INFECTIONS AS A MARKER FOR BLOODSTREAM INFECTION SUSCEPTIBILITY 226

1 Introduction

1.1 Linking health data

> *"How the people of England live is one of the most important questions that can be considered; and how – of what causes, and at what ages – they die is scarcely of less account; for it is the complement of the primary question teaching men how to live a longer, healthier and happier life."*
>
> (Farr 1875)

The idea of linking separate data entries related to the health of individuals within a whole population in order to be able to conduct both cross-sectional and longitudinal studies was first popularized by William Farr (Farr 1875): "by abstraction they are reduced to mere units undergoing changes as purely physical […] the analysis of the elementary facts observed in their various relations to time and place will shed new light on the more complicated phenomena of national life". His role in promoting the linkage of healthcare data is recognized through the naming of the original Farr Institute of Health Informatics Research (2013-2018) ("Farr Institute" 2019). The institute brought together 21 academic institutions and health partners across the UK in order to encourage collaboration and interdisciplinary research required in order to share, link and analyse the large and complex National Health Service (NHS) datasets with the goal of guiding improvements in healthcare at a patient and population level. This has now metamorphosed into Health Data Research UK ("Health Data Research UK" 2019) which continues with the same goal of conducting high-quality research in order to improve the quality and length of life. This is done by collecting, curating and linking vast amounts of healthcare data, including genomics, and providing the research expertise required in order to hone the potential of such data.

One of the first studies aiming to link health data on a large scale was the Oxford record linkage study (ORLS), a pilot study set up to develop a health register for the residents of the County Borough of Oxford, Abingdon Borough and Abingdon Rural Area, comprising a population of 325,000 individuals (Acheson 1964). It brought together information on births, deaths, domiciliary and hospital confinement as well as discharges from hospital from a number of different health sources recorded after the 1st of January 1962. Even though this was initially done by hand, the possibility of linkage by an electronic computer had already been mentioned in the literature (Newcombe, James, and Axford 1957). By 1991, the ORLS provided the opportunity for studies of trends over time in hospital admissions covering six districts in Southern England; each admission was matched to a patient, allowing for example for identification of multiple admissions; it also included diagnosis at admission. It enabled the study of admissions to ophthalmology between 1975 and 1985, with numbers increasing from a total of 3,527 admissions for 3,180 patients to 4,102 admissions for 3,564 patients, that is an increase of 16% in the number of admissions and of 12% in the number of patients (Ferguson et al. 1991). They showed that the increase was restricted to patients aged 65 years or older with a +4.8% average annual change in admission rates with a main diagnosis of cataract. This seemed to be associated with changes in referral practice which followed advances in surgical procedures and anaesthesia. The ORLS gave the opportunity to study many other conditions such as pneumococcal pneumonia (Fedson and Baldwin 1982), cancers of the large bowel (Prager 1978), and schizophrenia (Herrman, Baldwin, and Christie 1983). It has led to over 100 publications in total with useful insights still being gained by studies based on these historical records to this day (Goldacre, Pakpoor, and Goldacre 2017). The original ORLS was based on data collected from 1963 to 1998. From 1998 the Hospital Episode Statistics (HES) started collecting data for admitted patient care

for all of England and was later extended to also include outpatient and accident and emergency data (NHS Digital 2019).

The Oxford Record Linkage Study was a condition-agnostic database – it linked health data regardless of the underlying condition, and was not, for example restricted to patients with cancer or pneumonia. However, in the early 2000s hospitals faced the challenge of managing a large outbreak of methicillin resistant *Staphylococcus aureus* (MRSA). This underlined the importance of linking hospital and microbiological data in order to inform the management of infections through continuous surveillance for early detection of outbreaks and guidance of treatment and interventions supported by data. In the Oxford hospitals, this was first done by Dr David Wyllie in 2005 based on "static" data extracts from separate hospital admission and microbiology databases, leading to a highly cited publication in the BMJ (David H. Wyllie, Peto, and Crook 2005). Between the 1 April 1997 and 31 March 2003, 1467 *S. aureus* bloodstream infections were identified in the Oxford teaching hospitals. 493/1467 (34%) of these were methicillin resistant. They were able to show that only a quarter of MRSA bloodstream infections were identified at admission (compared to almost a half of the methicillin susceptible *S. aureus*), and that the majority of these (91%) had had previous hospital-exposure. In a later study, they were able to show that MRSA cases were driving the increase in *S. aureus* bloodstream infections as well as increasing 30-day mortality following infection compared to methicillin susceptible *S. aureus* (crude odds ratio 1.49) (David H. Wyllie, Crook, and Peto 2006). In further analyses of these same linked data from Oxford hospitals, Walker et al showed that the incidence of nosocomial cases (identified more than 2 days after admission) of MRSA isolated from non-sterile sites could be used for the surveillance of nosocomial MRSA bloodstream infections

and as numbers are approximately 8 times higher, changes in rate could be identified much faster and potentially considering smaller areas (S. Walker et al. 2008). These papers showed the power that such linked data could have in guiding the management of infectious diseases.

The importance of this kind of data linkage led to the creation in 2009 of the Infections in Oxfordshire Research Database (IORD), a completely anonymised database enhancing such linked records by also linking them to haematology results in a continuously updating database (Finney et al. 2011). IORD contains anonymised records of all admissions to the Oxford University Hospitals National Health Service (NHS) Foundation Trust, Oxfordshire which provides all acute care, microbiology, and pathology services in the region of approximately 680 000 individuals. All datasets containing patient identifiable data are linked together within the NHS systems and only anonymised information is exported and stored within IORD. As more data begins to be stored electronically, data sources such as electronic prescribing in hospitals (from October 2015) are also being added to IORD, further expanding the number of hypotheses that can be tested in order to improve management of infections. In the future, information on interactions with primary care will also be available. Since 2009, 14 publications have arisen from 58 projects supported by IORD, with an updated list of approved projects being made available online https://oxfordbrc.nihr.ac.uk/research-themes- overview/antimicrobial-resistance-and-modernising-microbiology/iord-projects/.With records going back to 1998, by April 2019 IORD contained approximately 500 million medical records from approximately 4 million unique individuals. The increasingly rich data contained within IORD allows monitoring of the rates of infection with various pathogens and of the evolution of these pathogens,

particularly under selective pressures such as antibiotic use. This is crucial in order to make informed decisions for the management of infectious diseases.

1.2 Antibiotics and the importance of infectious diseases

In September 1928, Alexander Fleming noticed on a plate of bacteria and accidentally grown mould that there was one area around the Penicillium mould in which bacteria had been killed. Over 10 years later, in February 1941 enough purified penicillin was produced for the first trial in humans to begin and the leading researcher, Howard Florey, together with Charles Fletcher, treated the first patient with penicillin at the Radcliffe Infirmary in Oxford (Wood 2019). Up until 1975, quite a few more new antibiotics were developed and deployed, however, antibiotic resistance to each of these new drugs also developed over time (**Figure 1**) (Clatworthy, Pierson, and Hung 2007). Bacteria develop resistance through mutation of existing genes, i.e. vertical evolution, and/or by acquiring new genes from other species, i.e. horizontal gene transfer, and the use of antibiotics applies selective pressure for these resistant bacteria to increase. After 1975, there was a gap in the discovery of new antibiotics with the exception of a few narrow-spectrum ones, that is antibiotics which are targeted towards a specific bacterial infection (**Figure 1**) (Clatworthy, Pierson, and Hung 2007). Therefore, developing tools which could guide clinicians in making an informed decision on antibiotic treatment both within the individual and at a population level quickly became of crucial importance. As bacteria continue to develop resistance, existing tools for antibiotic resistance prediction need to be re-evaluated periodically and if required, new updated tools need to be developed.

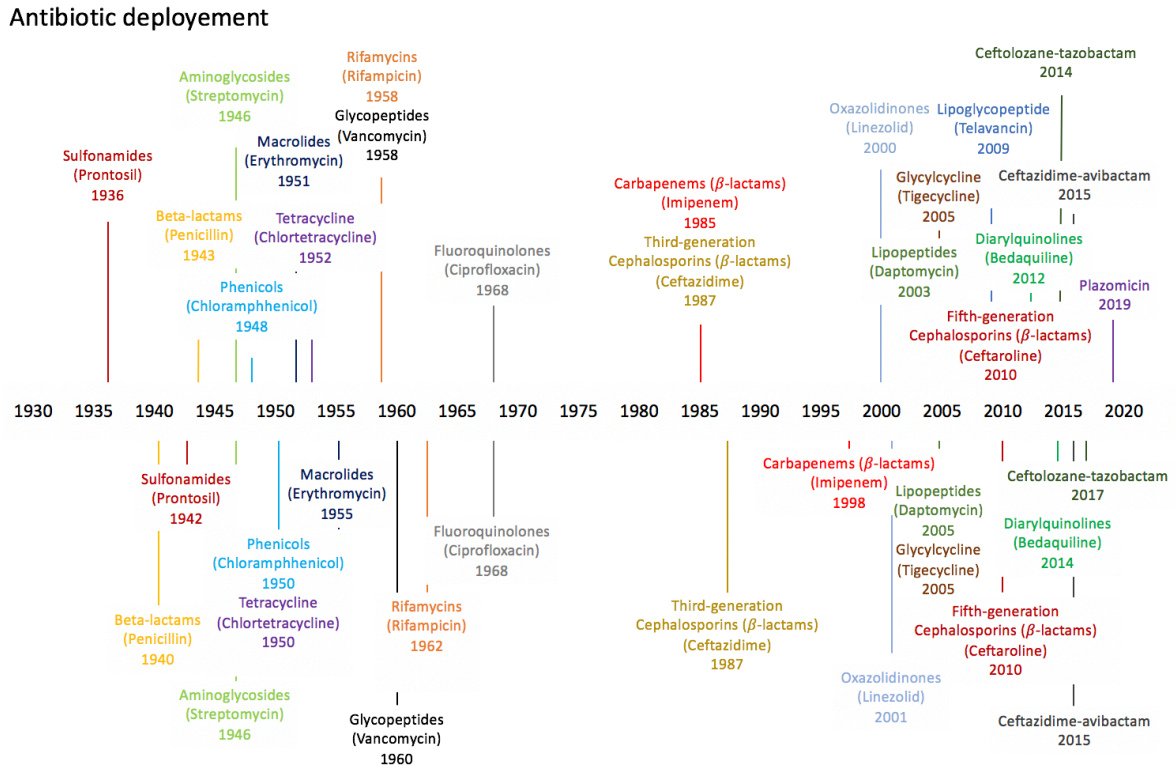


Figure 1 Timeline of antibiotic deployment and observed antibiotic resistance

Note: Adapted after (Clatworthy, Pierson, and Hung 2007).

Antimicrobial resistance, i.e. the ability of bacteria to survive antibiotics which were previously effective at killing them and hence treating the infection, is now among the top 10 world health challenges (World Health Organisation 2019). Developing tools in order to help prevent overuse and inappropriate use of antibiotics is crucial in order to try and slow down the development of antimicrobial resistance (S. Davies, Watson, and Shallcross 2015).

Considering changes in pattern of illness in high-income countries, in 1900 the leading causes of death were infectious diseases, while by 1997 chronic illnesses dominated among the top 10 causes of death (Cohen 2000). By 2016, 88% of deaths in high-income countries were caused by non-communicable diseases (World Health Organisation 2018). Globally, the only communicable diseases still in the top 10 causes of death in high-income countries

by 2016 were lower respiratory infections, with a crude death rate of 40 per 100,000 population (World Health Organisation 2018). This massive shift took place partly due to the development and deployment of antibiotics, together with other factors such as improved hygiene and sanitation, immunizations, but also safer food and water as well as living environments (Cohen 2000).

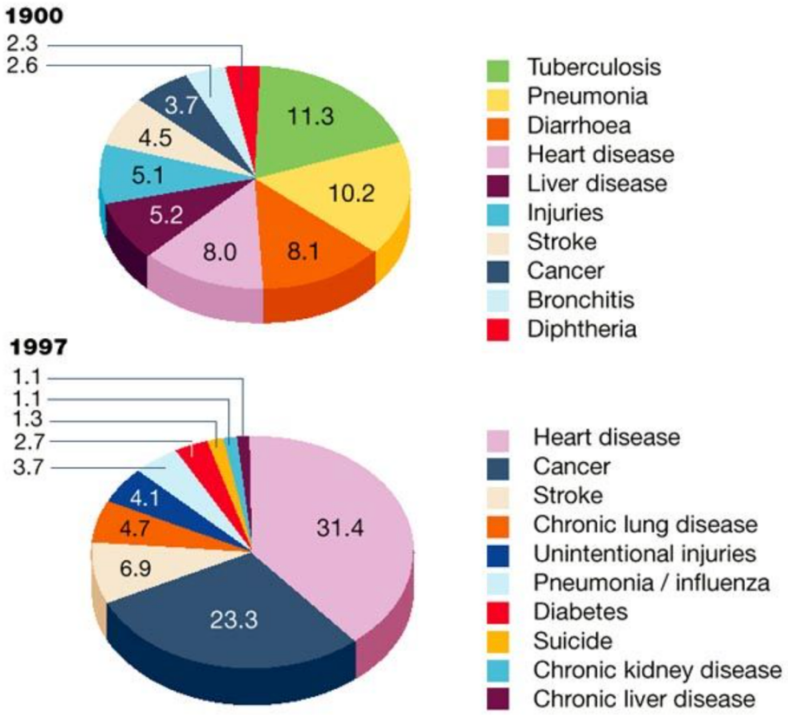


Figure 2 Changes in the top 10 causes of death in the United States between 1900 and 1997

Note: Source: (Cohen 2000).

The Global Burden of Disease Study 2017 (GBD 2017 Causes of Death Collaborators 2018) is a high quality study of trends in age-standardised numbers and rates of death across 282 different causes, covering 195 countries and territories between 1980 and 2017. They found that the overall decrease in the rate of death due to communicable diseases was driven by reductions in some of the main causes such as HIV/AIDS, tuberculosis, diarrhoeal diseases, and malaria. Focusing on lower respiratory infections and diarrhoeal disease, the decrease

in deaths observed between 2007 and 2017 was, however, only observed in age groups under 50 years, surpassing however the increases observed in the over 50 age groups.

Therefore, infectious diseases are currently no longer one of the major concerns in terms of mortality or morbidity when focusing on high-income countries. Despite this, a study in Oxfordshire showed how an increasing number of hospital admissions are being caused by infectious diseases, with the fastest increases in the number of admissions due to respiratory and gastrointestinal infections (Quan et al. 2016). Focusing on community-acquired pneumonia, they showed that the increase could not be attributed to an ageing population alone, neither could it be explained by changes in admission practice or in diagnostic codes. Unfortunately, drivers of the increase in the number of admissions due to respiratory and gastrointestinal infections were not analysed in detail in this paper, which also classified reason for admission based on diagnostic codes. Given higher risks of infection with waning immunity (Meyer 2005), an ageing population is likely to explain some, but most likely not all, of the increase. Increasing co-morbidity, particularly chronic obstructive pulmonary disease or other structural lung disease, also increases the risk of respiratory infection in older individuals, so may also contribute. Furthermore, increasing antimicrobial resistance rates are raising concerns about the availability of safe modern healthcare procedures in a "post-antibiotic era" (O'Neil 2014). For example, many cancer treatments supress the immune system of the patient making them more susceptible to infections. Surgery is also known to increase the risk of infection and therefore prophylactic antibiotics are currently being administered to patients in order to reduce the risk of infection. During the 20th century, there has been a 50-fold decrease in maternal deaths,

and much of this is attributable to the effectiveness of antimicrobials (O'Neil 2014). Antimicrobial resistance thus threatens all these healthcare areas.

Epidemiologically, there have been rises in resistance prevalence worldwide. This is as a result of the complex interplay between the human, animal and environmental microbiome and the associated effects of antibiotic usage. The relative importance of each to the rise in antimicrobial resistance is difficult to discern. The molecular characterization of genes has shown that resistance to certain antibiotics can be conferred by a variety of mechanisms, many times independently of each other, across these different compartments. Coupling epidemiological data with molecular characterization has enabled researchers to study transmission faster and with increased confidence.

One of the major pathogens raising health concerns are carbapenemase-producing *Enterobacteriaceae* (CPE) (**Figure 3**). These organisms express enzymes such as IMP, VIM, KPC, NDM, and/or OXA-48-like carbapenemases, that break down most beta-lactam antibiotics, making them ineffective for treating infections with such pathogens. CPE can be found in the environment which therefore represents a potential transmission reservoir, but in high-income countries infections are generally nosocomially-acquired. A molecular epidemiological investigation of an outbreak occurring in a 619-bed tertiary care hospital in Virginia, USA showed how 16 CPE isolates collected from 14 patients over a period of 8 months all carried a bla_{KPC} plasmid and could all be genetically linked to the index case (Mathers et al. 2011). However, further work in the same institution showed that the bla_{KPC} gene was moving rapidly between different plasmids and host strains (A. E. Sheppard et al. 2016). Similar situations occurred in other hospital settings (Snitkin et al. 2012; Decraene et

al. 2018). Sinks have been shown to be the potential source of CPE transmission in hospitals (De Geyter et al. 2017; Leitner et al. 2015). Patient relocation and plumbing replacement were associated with the control of a KPC-producing *E. coli* outbreak occurring in the Central Manchester University Hospital NHS Foundation Trust, United Kingdom over 8 years (Decraene et al. 2018). However, environmental contamination with CPE as well as CPE infections quickly occurred after the intervention, albeit at lower rates. Similarly, adding lids to hoppers (toilet-link disposal units) in Virginia was associated with reductions in CPE infections (Mathers et al. 2018). Besides sinks, wastewater and sewage have also been shown to be CPE reservoirs, and unfortunately treatment of wastewater has been shown to not prevent CPE from entering the environment (Ludden et al. 2017).

However, in low and middle-income countries lack of water infrastructure creates different challenges, particularly when CPE are present in the environment. In India, an estimated 21% of communicable diseases are water-related (The World Bank 2017). In a cohort of children in rural India, extended-spectrum beta-lactamase producing *E. coli* were found in all but one sample (Purohit et al. 2019), demonstrating the ubiquity of carriage of drug-resistant bacteria and suggesting in future, levels of CPE carriage could rise to similar levels.

Worryingly, it has already been shown that exposure of carbapenemase producing *Klebsiella pneumoniae* to ceftazidime-avibactam (a combination antibiotic only deployed in 2015) can lead to resistance-conferring mutations in the plasmid-borne bla_{KPC-3} gene (Shields et al. 2017). Interestingly, however, in 2/3 patients susceptibility to meropenem was restored during the course of treatment with ceftazidime-avibactam. This demonstrates the complex interplay between resistance to different types of antimicrobials. Antimicrobial resistance

surveillance can help prescribers make data-driven choices on treatment, prolonging the utility of existing drugs.

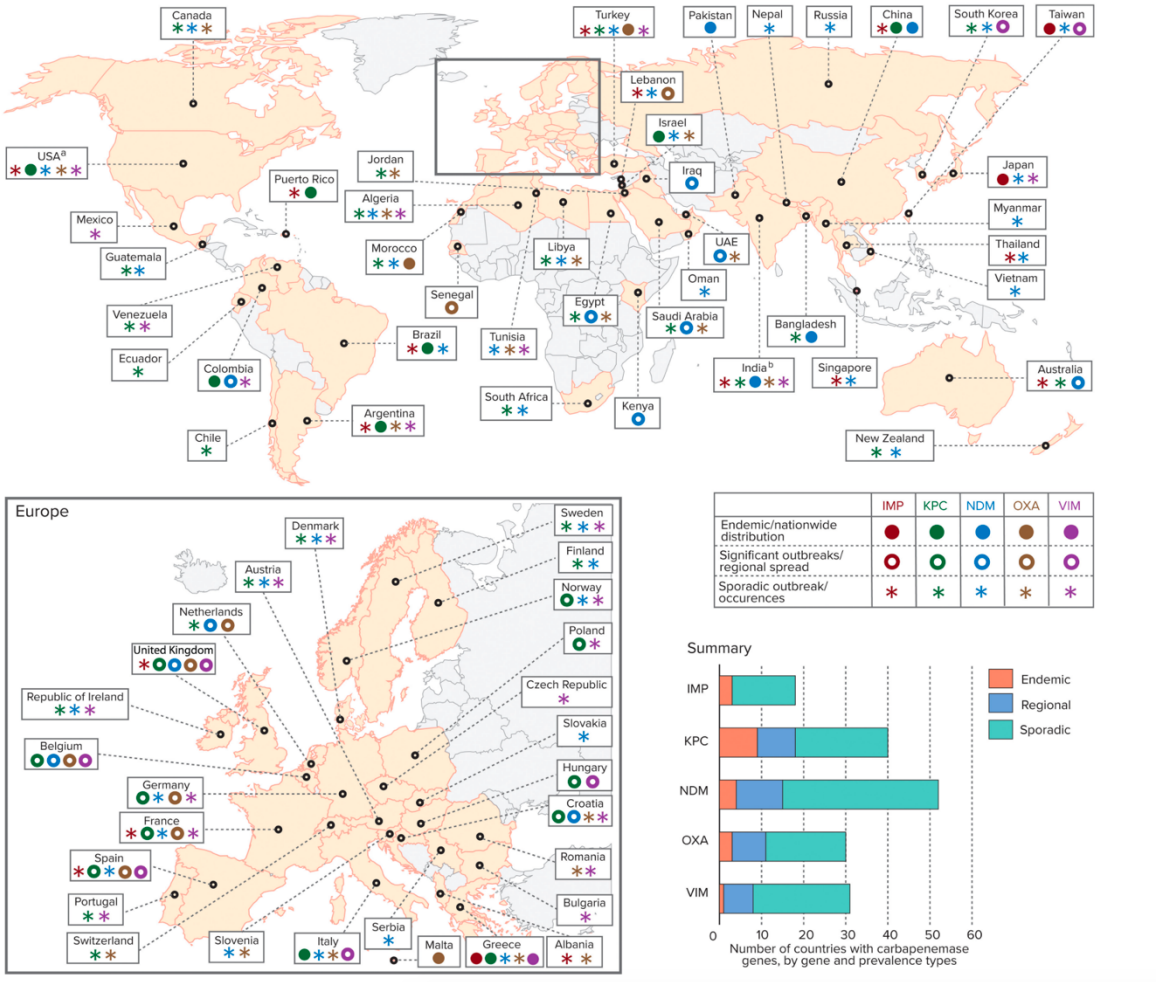


Figure 3 Global distribution of CPE

Note: Source: (Logan and Weinstein 2017).

One of the challenges in understanding patterns of resistance is that organisms causing infections are rarely identified in these common infections (Quan et al. 2016). Furthermore, there are also concerns about ascertainment bias as relevant diagnostic samples are not always taken. For these reasons, rates of resistance are most often considered in bloodstream infections, as due to the severity of these infections ascertainment bias is

expected to be minimal. A further advantage of considering bloodstream infections is that they also reflect the most important infections in terms of morbidity/mortality.

1.2.1 Bloodstream infections in the UK

During the 1990s and 2000s the majority of bloodstream infections were being caused by Gram-positive bacteria; however, by 2013 there was a shift towards Gram-negative bacteria. Laboratories in England, Wales and Northern Ireland have been voluntarily reporting bloodstream infections to the Health Protection Agency since 1975 (Health Protection Agency 2008b). Since 2003, the pathogen causing the majority of bloodstream infections has been *Escherichia coli* (Gram-negative bacteria), outnumbering those caused by *S. aureus* (Health Protection Agency 2008c, 2008a). Based on voluntary reports, there was a 44% increase in *E. coli* bloodstream infections between 2003 and 2011. However, as these are voluntary reports, much of the increase could simply be driven by an increase in the number of samples being submitted or in the number of laboratories reporting data. Reporting therefore became mandatory in June 2011 and the rate of *E. coli* bloodstream infection per 100,000 populations has been continuously rising ever since, from 60.4 in 2012/13 to 77.7 in 2018/19 (**Figure 4**). Between 1 April 2018 and 31 March 2019, only 7,632/43,242 (17.6%) *E. coli* bloodstream infections were taken more than 48 hours after patients were admitted to hospital, i.e. were hospital-onset cases. The rate of hospital-onset cases was fairly stable throughout the whole 7 years, indicating that the increasing rate is being driven by patients presenting to hospital with an *E. coli* bloodstream infection. This is in stark contrast to MRSA bloodstream infections and *Clostridium difficile* infections where the majority of cases are hospital-onset, and also where a substantial number of non-hospital-onset (or “community”) cases are in patients with recent hospital exposure

(Hassoun, Linden, and Friedman 2017; DePestel and Aronoff 2013; Public Health England 2019). However, from the mandatory data collected at a national level it is not possible to ascertain the number of community cases that have had recent hospital exposure and the number of true community cases which have not had recent hospital exposure.

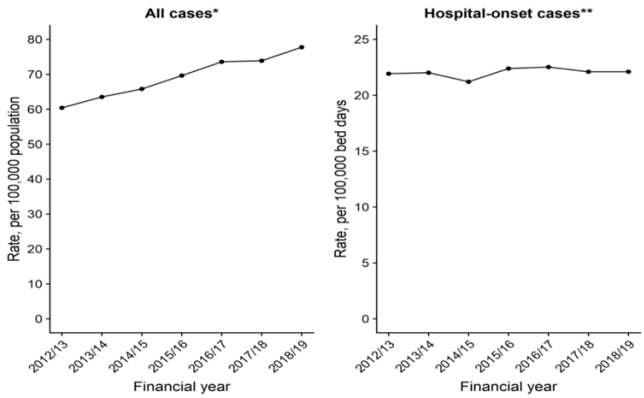


Figure 4 Trends in the rate of E. coli bloodstream infections in England financial years 2012/2013 to 2018/2019

Note: Source: (Public Health England 2019).

In a previous study, and the only longitudinal population-representative study in the UK, IORD allowed for information on all *E. coli* bloodstream infections occurring in Oxfordshire between 1 January 1999 and 15 June 2011 to be extracted and therefore trends in total rates, rates of resistant/susceptible isolates and outcome of infections over time to be estimated (Schlackow, Stoesser, et al. 2012). Cases were split into hospital-associated – defined as having their *E. coli* positive blood culture taken more than 48 hours after being admitted or having been in the hospital in the previous year – and hospital-naïve. They found an increase in the incidence of *E. coli* bloodstream infections throughout the decade, with an accelerated increase in rate from around 2006. Their analysis showed that this increase was being driven by an increase in the number of resistant isolates in both hospital-associated and hospital-naïve infections, adding to a constant burden of infection due to

susceptible bacteria; however, they defined “resistance” as resistant to any of co-amoxiclav, gentamicin, ciprofloxacin and cefotaxime. Similar trends were identified focusing particularly on co-amoxiclav resistance (70% of the resistant isolates, with resistance defined as above), and when considering co-amoxiclav resistant/susceptible *E. coli* urinary tract infections.

Interestingly, rises in co-amoxiclav resistant bloodstream and urinary tract infections took place the quarter before a shift in empiric antibiotic hospital-prescribing practices from second and third-generation cephalosporins towards co-amoxiclav and gentamicin as the empiric antibiotic treatment for sepsis. This was done as *C. difficile* incidence rates were rising and cephalosporins were known to be associated with development of *C. difficile*. There were no changes in antimicrobial susceptibility testing in 2006 that could have explained this increase. There were, however, no changes in the severity of the *E. coli* bloodstream infections infection when looking at albumin, creatinine, neutrophils and urea levels at when the blood culture was taken, as well as 7-day and 28-day mortality following the bloodstream infection (Schlackow, Stoesser, et al. 2012). Considering climatic data, they showed a relationship with increasing temperature, with an increase in mean weekly temperature of 5^0C being associated with a 5% increase in the incidence of *E. coli* bloodstream infections, but found no evidence of association with rainfall. The authors underlined the importance of continuously monitoring trends, particularly in resistant organisms, as well as the analysis of these trends with antimicrobial prescribing practices, both within an individual and at a population level.

Other important studies on *E. coli* bloodstream infection in the UK have had to focus on much shorter periods of time, and have therefore not been able to estimate trends. For example, in the winter of 2012/2013 a 3-month enhanced sentinel surveillance study was carried out in 35 National Health Service (NHS) hospitals in order to augment the relatively basic information which was available on the epidemiology of *E. coli* bloodstream infections from the English national mandatory surveillance programme (age, sex, time to onset among inpatients and voluntarily reported primary focus of infection) (Abernethy et al. 2017). Out of 1688 *E. coli* bloodstream infections included in this study, 865 (51%) had the underlying focus reported as "urogenital tract", with the next most common foci being "hepatobiliary" (although accounting for only 264 (16%) cases) and "unknown" (only 252 (15%) cases). 690/865 (80%) of those with a urogenital tract focus were recorded as having a urinary tract infection. 1263/1688 (75%) *E. coli* bloodstream infections were identified from samples taken in the first 2 days of the admission or outside of an admission. 930 (55%) had a healthcare-exposure in the month before the bloodstream infection, 458 (27%) had no healthcare-exposure in the month before and this information was not available for the remaining 300 (18%) cases. Considering particular types of healthcare-exposure in the 4 weeks prior to the bloodstream infection, 546 (32%) had received antibiotics, 681 (40%) had not and this information was not available for the remaining 461 (27%) cases. 354 (21%) had a urinary catheter *in situ*, inserted, removed or manipulated in the 7 days prior to the bloodstream infection, 1206 (71%) did not and this information was not available for the remaining 128 (8%) cases. Overall this study suggested that interventions should be considered in both community and hospitals settings, with a particular focus on improving urinary catheter use and appropriate empiric treatment of urinary tract infections. However, the relatively large amount of missing data (8-27% across different data items)

shows the challenges in trying to conduct large-scale surveillance based on manual data collection.

Finally, a prospective observational study of 679 Gram-negative bloodstream infections occurring between October 2013 and March 2014 in ten acute English hospitals showed that the primary predictors of outcome were patient and disease factors. Inappropriate empiric antimicrobial treatment occurred in 34% of the cases, however, there was no evidence for any association with increased mortality (Fitzpatrick et al. 2016).

1.2.2 Antimicrobial susceptibility testing in the UK

One important challenge, not particularly discussed by the two multi-centre studies above, is that traditional laboratory antimicrobial susceptibility testing methods often differ between laboratories. For longitudinal studies, they can also differ within a laboratory over time. The former makes comparisons between resistance rates in different regions challenging, while the later can make the analysis of trends over time challenging. Different methods can give significantly different results. Looking at 160 clinical *E. coli* isolates and determining co-amoxiclav minimum inhibitory concentrations (MIC) by broth microdilution using a fixed 2 mg/L clavulanate concentration (European Committee on Antimicrobial Susceptibility Testing – EUCAST) and a 2:1 ratio (Clinical and Laboratory Standards Institute – CLSI), essential agreement (i.e. MICs within a doubling dilution between the two methods) was only 25.6% (Díez-Aguilar et al. 2015). Focusing particularly on extended-spectrum beta-lactamase producing isolates, considering EUCAST breakpoints (≤8 mg/L susceptible, >8 mg/L resistant) 45% of isolates were susceptible and 55% were resistant, while when considering CLSI breakpoints (≤8 mg/L susceptible, 16 mg/L intermediate, ≥32 mg/L

resistant) 90% of isolates were susceptible and 10% were intermediate (Díez-Aguilar et al. 2015). Focusing on the differences between EUCAST and CLSI testing methodologies, and the impact of a laboratory switching from CLSI to EUCAST, Leverstein-van Hall et al compared changes in resistance rates in 3 laboratories which switched, to those seen in 17 laboratories which continued to use CLSI (Leverstein-van Hall et al. 2013). In the laboratories which switched methodology, resistance rates increased from 19% in 2010 to 31% in 2011 ($p<0.0001$), while in the others laboratories resistance decreased from 20% to 17% ($p<0.01$). Looking at clinical outcome in 171 patients they found that EUCAST MIC results correlated better with clinical outcome than Etests (bioMerieux, France) which use a fixed concentration of clavulanate, as does CLSI.

Recent years have shown an increasing interest in whole genome sequencing (WGS) of pathogens and their role in streamlining and improving diagnostic microbiology (Didelot et al. 2016). In particular, given the differences between susceptibility testing methods outlined above, one obvious question is whether resistance can be more accurately predicted from sequencing genotype rather than phenotypically. In the first proof of principle study investigating this, Stoesser et al looked at all extended-spectrum cephalosporin-resistant *E. coli* and *Klebsiella pneumoniae* bloodstream infections occurring in patients at the Oxford University Hospitals NHS Trust, Oxfordshire, UK, between January 2008 and November 2010 (74 *E. coli* isolates) or June 2011 (69 *K. pneumoniae* isolates) which could be retrieved (Stoesser et al. 2013). Even though this was a fairly small study and it was conducted on isolates which were expected to have multidrug-resistant phenotypes, they were able to use WGS and heuristic rules to predict antimicrobial susceptibility for amoxicillin, co-amoxiclav, gentamicin, ciprofloxacin, ceftriaxone, ceftazidime, meropenem

with sensitivity (proportion of truly resistant isolates which were predicted resistant) across all antibiotics for both species of 0.96 [95% confidence interval 0.94-0.98], and specificity (proportion of truly susceptible isolates which were predicted susceptible) 0.97 [0.95-0.98]. Extending this work, Davies et al looked at an unselected population of 976 *E. coli* bloodstream infections occurring between January 2013 and August 2015 in Oxfordshire, UK (T. J. Davies et al. 2019). By carrying out repeat EUCAST and CLSI tests on a subset of samples using one of the "gold standard" reference methods (agar dilution), he showed both poor agreement between the two methods when considering co-amoxiclav resistance, but also poor reproducibility between repeated tests using the same method. 27/261 (10%) of isolates were deemed resistant by both EUCAST and CLSI, 116/261 (44%) were deemed susceptible by both EUCAST and CLSI, 39/261 (15%) were deemed resistant by EUCAST, but susceptible by CLSI, with the remaining 79/261 (30%) being deemed resistant by EUCAST, but intermediate by CLSI. Therefore, overall agreement between susceptible/non-susceptible was only 85%. 158/261 (61%) of isolates had at least one doubling dilution between the minimum and maximum MIC when re-tested by EUCAST, and only 73/261 (28%) when re-tested by CLSI. However, when the MICs were dichotomised into susceptible/resistant results, 40/261 (15%) were classified as both susceptible and resistant by EUCAST repeat testing, while 31/216 (21%) changed between the susceptible/intermediate/resistant CLSI categories. They constructed a linear mixed model and carried out careful variable selection in order to predict co-amoxiclav MIC from WGS and compared this with that obtained from routine clinical laboratory automated antimicrobial susceptibility testing using BD Phoenix™, Beckton, Dickinson and Company (equivalent to EUCAST). They tested the model on 704 out of the remaining 715 isolates on which agar dilution testing was not conducted, but which had no new genetic features. The

predicted MIC was within one doubling dilution for 681/704 (97%) of isolates and agreed exactly with the routine clinical laboratory MIC for 548/704 (78%) of isolates. This study showed that dichotomising the phenotype for co-amoxiclav resistance raises critical challenges, producing poorly reproducible results regardless of susceptibility testing method. However, a detailed understanding of the genetic mechanisms together with very intensive modelling could predict co-amoxiclav MICs accurately from WGS.

In contrast, machine learning algorithms are statistical models which learn from patterns in data in order to make accurate predictions without being explicitly programmed for the task. The term was first coined in 1959 by Arthur L. Samuel (Samuel 1959) and much of the theory has existed for a very long time. However, it is the exponential increase in storage capacity and computational power that gave machine learning the opportunity to achieve tasks such as defeating the Go world champion in 2016 (Silver et al. 2016).

The routine use of WGS is potentially coming within reach for diagnostic microbiology as advances in sequencing technologies and the required analysis tools have led to an increase in speed and a significant lowering of costs. WGS is already being routinely used by the NHS in order to diagnose, treat appropriately (through accurate resistance prediction) and type (to assess potential transmission) infections with *Mycobacterium tuberculosis* and it is the first healthcare system in the world to do this at such a scale (Genomics England 2017). This has reduced diagnosis times, and time to availability on information which guides choice of antibiotic treatment, from what previously could have been a month to just over a week, as well as providing the possibility to analyse spread of infections (Genomics England 2017; Quan et al. 2018).

With an increasing number of drugs for which resistance rates are increasing, and with large amounts of data collected from WGS, the question whether machine learning can predict antimicrobial resistance from WGS naturally arises, because this could be a tool which does not require careful variable selection or expert knowledge. One of the largest studies to date considering machine learning algorithms in order to predict antimicrobial susceptibility from WGS looked at 13402 *Mycobacterium tuberculosis* isolates with susceptibility results for 11 antibiotics (Kouchaki et al. 2018). They extracted information from WGS by aligning each isolate to a reference strain and considering single nucleotide polymorphisms (SNPs) in genes known to be associated with resistance to one or more drugs. Prediction performance was compared to what they termed a "direct association" method which calls an isolate resistant if one or more known resistance-conferring mutations are present within the isolate, and susceptible otherwise (T. M. Walker et al. 2015). Prediction performance ranged from 76% sensitivity and 87% specificity for moxifloxacin to 92% sensitivity and 98% specificity for isoniazid. By comparison to the direct association method, machine learning algorithms were able to increase sensitivity, at the cost of a slight decrease in specificity, particularly for drugs for which there is less information available on genetic mechanisms of resistance. For example, looking at pyrazinamide resistance prediction, the best performing machine learning algorithm increased sensitivity from 43% to 88% compared to the direct association method, while specificity decreased from 98% to 89%.

1.2.3 Low and middle-income countries

In low-income countries, in 2016 over 50% of deaths were attributed to communicable, maternal, neonatal and nutritional conditions, with lower respiratory infections and

diarrhoeal diseases as the top two causes (World Health Organisation 2018). In lower-middle income countries, the top two causes of death were ischaemic heart disease and stroke, as in high-income countries; however, infections (lower respiratory infections, tuberculosis, diarrhoeal diseases) were still in the top 10 causes of death and causing a disproportionate burden (World Health Organisation 2018).

While in high-income countries WGS is already being used routinely for *Mycobacterium tuberculosis* infections and studies are showing that this is potentially coming within reach for other species, vast regions in low-income countries still lack basic laboratory facilities required in order to safely carry out antimicrobial susceptibility testing of urine cultures and of samples from other non-invasive sites. Massive efforts on multiple fronts are required in order to build laboratories which are able to process blood cultures and further challenges are presented when considering building an infrastructure which would allow surveillance of resistance rates.

The Global Antimicrobial Resistance Surveillance System (GLASS) was launched in October 2015 by the World Health Organization (WHO) (World Health Organization 2015) in order to support the Global Action Plan on Antimicrobial Resistance adopted in May 2015 at the Sixty-eighth World Health Assembly (World Health Organisation 2015). The aim is to collect, analyse and report standardized data related to antimicrobial resistance such as resistance rates and antimicrobial consumption rates in order to inform actions at all levels: local, national, regional and global. 13/83 (16%) GLASS enrolled countries are classified as low-income countries by the World Bank 2019 (World Health Organization 2019b). Developing methods which can help low-income countries collect antimicrobial resistance data within a

short timeframe and realistic investment size is crucial in order to be able to provide such data for the global surveillance of antimicrobial resistance, but also in order to guide local antimicrobial prescribing.

1.3 Thesis outline

The goal of this thesis is to inform management of infections both in Oxfordshire, but also worldwide, using routinely collected electronic health records. I decided to focus on *E.coli* bloodstream infections given this organism's importance worldwide. I first explore a number of hypotheses regarding what is driving the increase in their number in the Oxfordshire region by triangulation of different types of electronic health record data, with the goal of identifying whether there are groups that would be small enough for a targeted intervention, yet big enough to have an impact on the overall trend (Chapter 2). This first results Chapter (Chapter 2) identifies potential issues with the existing antimicrobial susceptibility testing methods, particularly when focusing on co-amoxiclav. Therefore, in my second results Chapter (Chapter 3) I aim to explore whether machine learning algorithms could predict co-amoxiclav resistance using WGS. I focus particularly on machine learning algorithms as these are methods which do not rely on careful feature selection and model building, do not require expert knowledge and can be easily applied to updated datasets. While in high-income countries using WGS routinely is potentially within reach and is already being used for *M. tuberculosis* diagnostics in the NHS, vast regions in low-income countries still lack laboratory capacity to carry out routine testing of blood cultures. Therefore, in my third results Chapter (Chapter 4), I aim to build a tool for the surveillance of trends in antimicrobial resistance rates in bloodstream infections in low-income countries based on other sites of infection which are easier to sample from and also carry out antimicrobial susceptibility testing on.

The thesis concludes with implications for policy makers and directions for future work.

2 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

The work presented in this Chapter was published in 2018 in Lancet Infectious Diseases (Vihta et al. 2018). Most of the tables and figures are as published in the paper and its supplementary material.

2.1 Introduction

Escherichia coli (*E. coli*) is one of the most important causes of bloodstream infection (Public Health England 2015b) and is also one of the pathogens in which antimicrobial resistance (AMR) is a major concern (O'Neill 2014) with resistance rates rising across Europe (Buetti et al. 2017; European Centre for Disease Prevention and Control 2016; van der Mee-Marquet et al. 2015). As noted in Chapter 1, *E. coli* bloodstream infections reported to Public Health England using their voluntary reporting scheme rose by 44% between 2003-2011 (Gerver et al. 2014). After mandatory reporting was introduced in July 2011, this increased by a further 28% by July-September 2016, to 78.8 cases/100,000 population (Public Health England 2016a).

Similarly to other regions of the world, a large proportion (>70%) of *E. coli* bloodstream infections in England are identified within two days of admission (Public Health England 2016a), which is one of the standard definitions used for community-associated infections.

However, the impact of the timing of previous hospital-exposure on trends in *E. coli* bloodstream infection has not been comprehensively investigated, with only two relevant previous studies, one in the Calgary Health Region 2000-2006 (Laupland et al. 2008), and another in Oxfordshire in 1999-2011 (Schlackow, Stoesser, et al. 2012) (described in detail in Chapter 1) subclassifying infections within two days of admission based on previous healthcare-exposure. The relevance of such older studies to current trends is unclear. The source of *E. coli* bloodstream infection may also differ by recency of hospital-exposure. In a recent study, ~50% of *E. coli* bloodstream infections in England were considered most likely due to urinary tract infections (UTIs) (Abernethy et al. 2017). Another common foci of infection in this study was gastrointestinal, and this was found to be more common in inpatients than in outpatients (Public Health England 2016a).

All-cause mortality in the 30 days following *E. coli* bloodstream infection is ~16% (Public Health England 2015c). Given the impact of increasing AMR on outcomes as described in Chapter 1 (O'Neill 2014), there are important concerns that this could rise further. In the earlier study in Oxfordshire, rises in incidence of *E. coli* bloodstream infection through 2011 were essentially confined to infections that were resistant to ciprofloxacin, co-amoxiclav, cefotaxime and/or aminoglycosides (Schlackow, Stoesser, et al. 2012). The drivers of rising resistant *E. coli* bloodstream infection, and *E. coli* bloodstream infection more generally, are unclear, with increased antibiotic usage or resistance implicated in some, but not all, studies (van der Mee-Marquet et al. 2015; Livermore et al. 2000; Hay et al. 2005; Kahlmeter, Menday, and Cars 2003; Livermore et al. 2013; Lishman et al. 2016; Ironmonger et al. 2018; Ho et al. 2012). An alternative explanation is that the increases in incidence are a natural consequence of increased risk of infection with age, combined with an aging population

(van der Mee-Marquet et al. 2015). Co-amoxiclav in particular is one of the most frequently used antibiotics in England (Public Health England 2017), where it is commonly used as empiric antibiotic treatment, particularly for community-acquired pneumonia and undifferentiated sepsis in hospitals, as well as for prophylaxis, although individual hospital and primary-care guidelines vary (Llewelyn et al. 2015). Hence, trends in co-amoxiclav resistance are particularly important.

I therefore aimed to investigate possible drivers of changes in *E. coli* bloodstream infection incidence and antibiotic susceptibilities in Oxfordshire over the last two decades, while stratifying for hospital-exposure, particularly focussing on co-amoxiclav resistance. I had three hypotheses:

- first, that increases may be due to features of the at-risk population – I therefore explored the impact on trends in incidence of demographics, recurrent vs first infections, and increased ascertainment (submission of samples for culture),
- second, that increases may be due to features of individuals' healthcare history – I therefore explored the impact on trends in incidence of previous urine cultures, catheter specimens, admission diagnoses, antibiotic usage
- and third, that increases may be due to features of the bacteria – I therefore explored mortality following infection, severity of the initial presenting infection, and resistance to multiple antibiotics.

2.2 Methods

The Infections in Oxfordshire Research Database (IORD) (Finney et al. 2011) records all admissions to the Oxford University Hospitals National Health Service Foundation Trust

(OUH), Oxfordshire, UK, from April 1997 to date, linked by patient with microbiology and biochemistry/haematology results. The four OUH hospitals provide all acute care, microbiology and pathology services in the region (approximately 680,000 individuals). Out-of-hospital mortality was determined by updates from a national information system (the NHS Spine).

The primary study outcome was *E. coli* bloodstream infection, defined as *E. coli* isolated from blood cultures taken from 1 January 1998 to 31 December 2016 inclusive, including polymicrobial cultures, without age restriction and de-duplicated within 14 days of each index positive following mandatory reporting guidelines (Public Health England 2016b). I included *E. coli* isolated from blood from pure and mixed/polymicrobial cultures in the primary outcome as mixed/polymicrobial cultures comprised a substantial minority, 763/5706 (13%) *E. coli* bloodstream infections over the study, and previous hospital-exposure, resistance and other factors are also relevant to them. Of these 763 mixed/polymicrobial cultures, 187 (25%) were infections with *E. coli* and only plausible contaminants, including Coagulase negative staphylococcus; *Streptococcus viridans*, *oralis*, *salivarius*, *mitis*, and unspecified streptococcus; diphtheroids; Propionibacterium species; and Bacillus species. Of the 576 polymicrobial *E. coli* bloodstream infections with at least one other plausible pathogen, 412 (72%) other pathogens were likely gastrointestinal including *Klebsiella pneumoniae* and *oxytoca*, Enterococcus species, *Enterococcus faecalis* group D, *Proteus mirabilis*, Bacteroides fragilis, Enterobacter species, gastrointestinal anaerobes, and yeast. The percentage of polymicrobial infections (using total *E. coli* bloodstream infections per month as offset) did not vary over calendar time (annual

incidence rate ratio (IRR)=1.00 [95% CI 0.98, 1.01], p=0.73). Seasonal variation in polymicrobial infections was not analysed.

As *E. coli* bloodstream infections are hypothesised to commonly arise from UTIs, for context I also analysed *E. coli* UTIs, defined as pure culture from urine of $>10^4$ colony-forming-units/ml, de-duplicated within 90 days to avoid over-counting ongoing infections (Butler et al. 2007).

I classified *E. coli* bloodstream infections/*E. coli* UTIs as 'nosocomial' if samples were taken more than 48h after admission up until discharge (Garner et al. 1988). All other *E. coli* bloodstream infections/*E. coli* UTIs were classified as 'community', 'quasi-community' or 'quasi-nosocomial' if the last hospital discharge was >1 year, 31-365 days, or 0-30 days prior to the samples being taken respectively (Laupland et al. 2008; D H Wyllie et al. 2007; Public Health England and NHS Improvement 2017; Cardoso et al. 2014). That is, *E. coli* bloodstream infections/*E. coli* UTIs were classified as:

(i) 'community', if the sample was taken in the first 48 hours of an admission or outside of an admission and the last hospital discharge was >1 year prior to the sample being taken

(ii) 'quasi-community', if the sample was taken in the first 48 hours of an admission or outside of an admission and the last hospital discharge was between 31 and 365 days prior to the sample being taken

(iii) 'quasi-nosocomial', if the sample was taken in the first 48 hours of an admission or outside of an admission and the last hospital discharge was between 0 and 30 days prior to the sample being taken

(iv) 'nosocomial', if the sample was taken more than 48 hours after admission up until discharge.

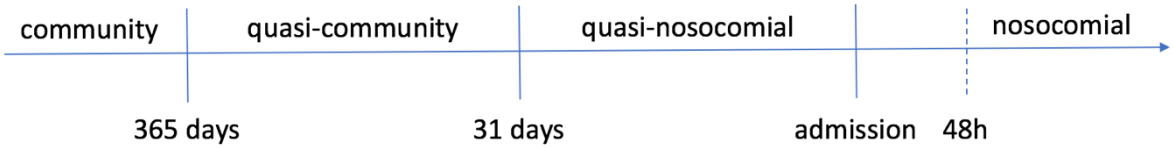


Figure 5 Illustration of classification of infections based on previous hospital-exposure

I also calculated incidences of first ever (since 1998) and recurrent *E. coli* bloodstream infections in order to investigate the hypothesis that change in trends was being driven mainly by recurrences, i.e. the same patient acquiring a/multiple new infection/s (at least 14 days after the previous *E. coli* bloodstream infection). I used a strict definition of nosocomial *E. coli* bloodstream infection ending at discharge in order to investigate the group whose *E. coli* bloodstream infection had not actually been identified during hospitalisation. Some of this could however reflect inaccuracies in timing of discharge and sample collection – however, a relatively small number, 44/1132 (4%), of patients with a quasi-nosocomial *E. coli* bloodstream infections were recorded as having been discharged in the 24 hours preceding the blood culture being taken (147/1132 (13%) were discharged in the 48 hours preceding the blood culture being taken).

To assess whether changes in trend were being driven by demographic changes such as ageing and population growth, I standardised incidence against the 1998 Oxfordshire age-sex population (estimates from the UK Office for National Statistics) (ONS 2018). Estimates from the UK Office for National Statistics were not available for 2016 so I used a linear extrapolation of the previous two years. To assess whether changes in how bloodstream

and urine infections were being ascertained could be driving changes in trends, I considered the incidence of blood/urine cultures, regardless of result, and also additionally considered the impact of standardising outcome for culture rates. Both standardisations were done using inverse probability weighting.

As a proxy for the effect of changes in bacterial virulence, that is to assess whether changes in bacterial virulence could be driving changes in trends, I considered separately mortality in the 30 days after sample collection, and levels of monocytes, neutrophils, lymphocytes, C-reactive protein (CRP), creatinine and urea at sample collection (using the closest value within [-2,+2] days). Plots illustrate a representative selection of these biomarkers. The hypothesis is that over time less severe bloodstream infections are being identified and as a result driving the increase. I do not have genotypic information and cannot exclude smaller subpopulations/strains causing more or less severe infections. I did however look for overall changes in *E. coli* bloodstream infections occurring in patients who died within 30 days after the sample collection, and separately in those who did not die, and there was no evidence for clinically meaningful changes in overall median severity over time in these two subgroups either. To assess the impact of AMR, which might also affect treatment outcomes, I investigated *E. coli* bloodstream infections reported by the microbiology service laboratory at OUH as resistant to amoxicillin, co-amoxiclav, trimethoprim, gentamicin, ciprofloxacin, ceftriaxone, ceftazidime, piperacillin-tazobactam and meropenem, and *E. coli* UTI reported as resistant to amoxicillin, co-amoxiclav, trimethoprim, ciprofloxacin, nitrofurantoin and cephalexin. The choice of these drugs was made because these were the only ones consistently tested throughout the study period. Before February 2013, antimicrobial susceptibility was tested in the OUH microbiology service laboratory using disk

diffusion in an uncontrolled inoculum using a control; in February 2013 this was replaced by automated susceptibility testing with the BD Phoenix™ Automated Microbiology System, Beckton Dickinson, Franklin Lakes, NJ, USA using European Committee on Antimicrobial Susceptibility testing (EUCAST) breakpoints, using disk diffusion direct from blood in an uncontrolled inoculum as an early flag. In December 2013, disk diffusion in a controlled inoculum using the British Society for Antimicrobial Chemotherapy (BSAC) diameter zones was introduced for selected samples in addition to BD Phoenix™. The majority of the samples collected after February 2013 would therefore have both a disk diffusion and a BD Phoenix™ result. Where multiple results were available for one sample, the Phoenix™ result was used in preference to the disk diffusion result as most disk diffusion results were uncontrolled; otherwise any resistant result was used in preference to susceptible results. I also analysed the agreement in yearly percentage of co-amoxiclav resistant *E. coli* bloodstream infections by splitting the isolates into: tested by disk diffusion using BSAC diameter zones (DD) only, tested by BD Phoenix™ using EUCAST (MIC) only, tested by both DD and MIC and resistant by DD and either susceptible or resistant by MIC, tested by both DD and MIC and resistant by MIC and either susceptible or resistant by DD, and tested by both DD and MIC and resistant by either. This was done both due to a considerable rise in the percentage of co-amoxiclav resistant isolates in 2013 which coincided with the changing of testing methodologies (see Results), and concerns about changes in trends in resistance being driven by these changes in testing methodology having been previously raised in the literature (Leverstein-van Hall et al. 2013; Díez-Aguilar et al. 2015).

Guidelines recommend empirical treatment for uncomplicated UTIs and for urine samples to be sent for microbiological testing only from individuals with clinical treatment failure,

frequent or recurrent UTI or with possibly resistant infections (Public Health England 2017). To investigate the impact of UTIs on rates of bloodstream infections, I first classified *E. coli* bloodstream infections by whether the patient had ever had an *E. coli* UTI ≥3 days previously. However, symptomatic UTIs where *E. coli* was not isolated may also be due to *E. coli*, particularly if cultures show mixed growth (since this may reflect contamination from faecal/vaginal flora) and even if cultures are negative, particularly if antibiotics have been taken before sample collection, data not available to me through the hospital databases. Even though urinary specimens should only be sent for microbiological testing if there is clinical suspicion of a UTI (Public Health England 2015a), 43% of mixed growth or culture negative urine samples taken around an *E. coli* bloodstream infection (within [-30,+2] days of the blood collection date) did not have a completed request code making it difficult to assess whether there really was clinical suspicion of urinary infection before or around the bloodstream infection. To investigate the combined potential contribution of antecedent UTIs to rising *E. coli* bloodstream infection incidence, I therefore hierarchically classified *E. coli* bloodstream infections as:

(i) 'likely urine-associated', if they either had an *E. coli*-positive urine culture, or if they had mixed growth or negative urine culture with a relevant request code (mentioning UTI or other urinary symptoms, dysuria, urosepsis, pyelonephritis, positive dipstick), within [-30,-3] days of the *E. coli* bloodstream infection sample

(ii) 'urosepsis', if they either had an *E. coli*-positive urine culture, or if they had mixed growth or negative urine culture with a relevant request code, within (-3,+2] days of the *E. coli* bloodstream infection sample (but not (i), i.e. no pre-existing evidence of a

urine infection which could have potentially been prevented from becoming urosepsis)

(iii) 'unlikely urine-associated', if they had a urine culture positive for other pathogens within [-30,+2] days of the *E. coli* bloodstream infection sample, or if no urine was taken for culture within [-30,+2] days of the *E. coli* bloodstream infection sample (but not (i) or (ii))

(iv) 'unknown', if they had a mixed growth or negative urine culture and either an irrelevant or no request code within [-30,+2] days of the *E. coli* bloodstream infection sample (but not (i), (ii) or (iii))).

Sensitivity analyses included definitions based on urine cultures up to 100 days before the *E. coli* bloodstream infection sample rather than 30 days, with similar results (data not shown).

To investigate the contribution of catheters, I classified *E. coli* bloodstream infections by whether a catheter urine specimen had been submitted up to and including the day of blood collection (regardless of result).

To investigate the contribution of previous admission characteristics, I classified quasi-nosocomial *E. coli* bloodstream infections by whether the primary diagnostic code of the previous admission was infection-related, or by whether any diagnostic code (primary/secondary) included UTI. The primary diagnostic codes from the previous admission were grouped by Nicola Fawcett and Nicole Stoesser into 'cardiovascular disorder', 'neurological disorder', 'dermatological/rheumatological disorders', 'endocrine disorder', 'obstetrics and gynaecology disorder', 'haematological disorder', 'malignancy',

'gastrointestinal disorder', 'orthopaedic disorders including trauma', 'poisoning', 'renal and urological disorders', 'respiratory disorder', 'other' (Quan et al. 2016).

2.2.1 Statistical analysis

Counts of *E. coli* bloodstream infection/*E. coli* UTI per month were modelled using negative-binomial regression (incorporating overdispersion), assuming the same underlying population (i.e. no offset), namely the population of Oxfordshire. Mortality in the 30 days following *E. coli* bloodstream infection and CRP≥156mg/L (the minimum upper threshold above which levels were not quantified by the laboratory for the majority of the period under study, see below) at *E. coli* bloodstream infection (binary outcomes) were modelled using poisson regression (to estimate rate ratios analogous to those for the incidence trends), and absolute values of other continuous test results were modelled using median quantile regression. Poisson and quantile regression used sample date as the underlying time scale and adjusted for age and sex.

Changes in trends in all outcomes were estimated using iterative sequential regression (ISR) (Schlackow, Walker, et al. 2012), and compared between outcomes using stacked regression (Lunn and McNeil 1995). The ISR algorithm first modelled the outcome using samples taken between 1 January 1998 and 1 January 1999, and compared a model with one trajectory over calendar time in the outcome to a model allowing this trajectory to change 6 months after the start of observation. If the model with two trajectories was not a better fit (determined by a Bayesian Information Criterion being lower by at least 3.84 (the critical value to detect a significance level of 0.05 with a test and one degree of freedom)), an additional six months' observations (to June 1999) were included. Then the model with one trajectory was compared to models with 2 trajectories with either June 1998 or January

1999 as the breakpoint, again considering whether any model with a change in trajectory substantially improved model fit. Any breakpoint that improved model fit was fixed, and then an additional six month's data included. This process was iterated up to December 2016 (end of study period). I chose to use this method as it enables an unknown number of multiple changes in trend to be estimated efficiently, in contrast to traditional grid search algorithms which require the number of changes in trend to be fixed, and, in addition, also require every possible combination of changepoints to be estimated, making this very computationally intensive even for a small number of changepoints.

For antibiotic resistance trends, due to the smaller number of observations, counts per year were modelled rather than per month, first considering samples taken between 1 January 1998 and 1 January 2002, and then successively every year through 31 December 2016. Fewer than 1% of susceptibility results were missing for each antibiotic tested, with the exception of trimethoprim for which blood cultures were not tested October-December 2014. Analyses therefore used a probability weight of 4/3 for the incidence of trimethoprim-resistant *E. coli* bloodstream infections in 2014; all other analyses of incidence of resistant bloodstream infections/UTIs were based on observed data only (i.e. complete cases).

For analysis of levels of monocytes, neutrophils, lymphocytes, CRP, creatinine and urea at sample collection (continuous outcomes, closest value within [-2,+2] days of sample collection), continuous test results were first truncated at the 1st and 99th percentile; median values were modelled using quantile regression to avoid influence from outliers. All analyses of test results were restricted to complete cases; for *E. coli* bloodstream infection completeness was 93% for neutrophils, 93% for CRP (available post-2000 only), 95% for

creatinine and 93% for urea. CRP was reported with different upper thresholds over the study period, and approximately half the values were consistently above the upper threshold. CRP was therefore considered as a binary rather than continuous outcome, namely as CRP ≥156 mg/L (as 156 mg/L was the minimum upper threshold used over 1998-2016). In January 2009, the creatinine analysis method changed in the laboratory (Schlackow, Walker, et al. 2012); models adjusted for this change using a step-function.

Out-of-hospital mortality was determined by routine updates from the national information system used by the NHS (NHS Spine) in the underlying data sources. 2% (87/5706) of patients with *E. coli* bloodstream infections could not be identified in NHS Spine (for example, due to incomplete identifiers in the underlying NHS data sources, particularly in older historical data where NHS numbers were not used consistently); again mortality analyses were restricted to complete cases where out-of-hospital mortality was available. All analyses of laboratory parameters and 30-day mortality were adjusted for age and sex (which led to 5 community cases being dropped as sex was unknown). To investigate whether there was any evidence of differential severity in susceptible versus resistant cases, I extended these models for 30-day mortality following *E. coli* bloodstream infection and neutrophils at presentation with *E. coli* bloodstream infection to additionally include a binary factor for whether the *E. coli* bloodstream infection was co-amoxiclav-resistant versus co-amoxiclav-susceptible, and the interaction between this factor and calendar time (as represented by exact date of blood sample collection). The interaction term tests whether there is any evidence that these severity markers are changing differently over time in co-amoxiclav-resistant versus co-amoxiclav-susceptible *E. coli* bloodstream infections. As there was no evidence of such heterogeneity, the interaction term was

removed from the model and a main effect of co-amoxiclav-resistant versus co-amoxiclav-susceptible *E. coli* bloodstream infection on 30-day mortality following *E. coli* bloodstream infection and neutrophils at presentation with *E. coli* bloodstream infection estimated.

Incidence trends in different subgroups, or for different outcomes, were compared using stacked regression (Lunn and McNeil 1995). In brief, these methods "stack" the data for different regression models; for example, for *E. coli* bloodstream infections, incidence in each of the four different healthcare exposure subgroups over calendar time, on top of each other. Individual model-specific trends are then calculated but across the entire stacked dataset, using robust variance adjustment to account for the same independent variable (here month) occurring repeatedly in the stacked dataset. Because these model-specific trends are calculated within the same stacked dataset, they can then be compared using standard Wald tests.

In order to estimate a simple univariable association between hospital antimicrobial prescribing (standardized using WHO defined daily doses (DDD) (World Health Organization 2019a)) and nosocomial co-amoxiclav resistant bloodstream infection incidence, analogous to a Spearman rho for two continuous factors, I calculated a bivariate cross-correlation, i.e. the correlation between one series at time t and another series at time t - k as a function of the time t and lag k. Because of differences in the time periods in which (quarterly) antibiotic prescribing data were available, I was only able to include financial years 2003-2014. For each class of antibiotics, and all antibiotics combined, I considered a time lag of 0 (ie same quarter), and all quarters up to -3 and +3 (where -1/4 means antibiotic use in previous quarter against bloodstream infections in current quarter). In July 2010, the

hospital co-amoxiclav formulation changed from 250mg amoxicillin and 125mg clavulanate to 500mg amoxicillin and 125mg clavulanate affecting DDD because of the different strengths. Prior to July 2010, hospital practice was to prescribe an additional 250mg amoxicillin tablet with the original co-amoxiclav formulation to obtain the same total dosage as in the single co-amoxiclav formulation after July 2010. This was evident in the data which demonstrated a decrease in raw amoxicillin DDDs in July 2010 concurrent with the change in formulation (because it was no longer being prescribed with the original co-amoxiclav formulation) and a corresponding increase in co-amoxiclav DDDs in July 2010 (as an additional 250mg amoxicillin was being counted as a co-amoxiclav DDD rather than an amoxicillin DDD). I therefore adjusted raw co-amoxiclav and amoxicillin DDDs before July 2010 to count the additional amoxicillin prescribed with the old co-amoxiclav formulation as a co-amoxiclav DDD, making assignment consistent over the whole study period.

To estimate associations between annual community urine sample submission, community *E. coli* UTI and community co-amoxiclav-resistant *E. coli* UTI incidence, and co-amoxiclav use in primary care, I used backwards elimination to identify the most parsimonious model including co-amoxiclav DDD per 1000 registered patients in the cases' general practice in the current and previous year together with their interaction with the calendar year trend, adjusting for general practice and including the number of patients per general practice per year as an offset in negative-binomial regression models. Data on antibiotic prescribing and numbers of registered patients for each general practice were obtained from the Health and Social Care Information Centre, but were only available from January 2011-December 2016 only. I was therefore able to consider annual outcomes from 2012-2016 only. I was not able to consider co-amoxiclav resistant *E. coli* bloodstream infection incidence as numbers were

too small in individual general practices over the period where antibiotic data were available. There was a sharp increase in the incidence of co-amoxiclav resistant *E. coli* UTIs in 2013 (incidences of 753, 686, 2045, 2533, 2404, 2368 over years from 2011 to 2016). Therefore, in these analyses restricted to 2012-2016 I allowed for this different incidence in 2012 using a step function, and estimated time trends in addition to this.

All models excluded 13 general practices, 8 which had missing data for at least one of the years and 5 which submitted less than 151 urine samples over 2011-2016 (all others submitted over 308 urine samples). I also obtained 2017 demographics for each practice from the Health and Social Care Information Centre and considered the effect of including the proportion aged over 65 years and the proportion of males in each general practice as explanatory variables, without general practice, to see whether these characteristics affected associations between co-amoxiclav resistant UTIs and antibiotic usage and explained the varying effects of general practice.

I conducted analyses using R 3.2.2 (R Core Team 2017), and STATA 14.1 for stacked regression and probability weighted analyses.

2.3 Results

After 14-day de-duplication, from 1998-2016, 5706 *E. coli* bloodstream infections occurred in 5215 patients, i.e. 491 (9%) were recurrences (relapse and/or reinfection). Median (IQR) age at infection was 73 (58-83) years and 2915 (51%) occurred in women (all infections). Recurrences occurred a median (IQR) 144 (39-577) days apart: of 391 patients with recurrences, 324 (83%) had one and 52 (13%) had two (range 1-8).

Overall, incidence of *E. coli* bloodstream infection increased year-on-year (annual incidence rate ratio (IRR)=1.06 [95% CI 1.05, 1.06]). Most (5393 (95%)) *E. coli* bloodstream infection cases were admitted to OUH before the blood culture was taken or within 24 hours following the blood culture being taken; the remainder were mostly sampled in emergency departments/community hospitals. However, only 1365 (24%) *E. coli* bloodstream infections were 'nosocomial' (sample taken ≥48h post-admission). A further 1132 (20%) were 'quasi-nosocomial' (discharged up to 30 days previously), 1346 (24%) were 'quasi-community' (discharged 31-365 days previously) and 1863 (33%) were 'community' cases (discharged >1 year previously or never previously admitted to OUH).

The 1132 quasi-nosocomial *E. coli* bloodstream infection were most commonly previously admitted for malignancy (395, 35%), gastrointestinal disorders (177, 16%), and renal/urological disorders (164, 14%) (**Table 1**). There was no major variability in the reasons for these previous admissions over time (**Figure 6A**). Renal/urological malignancies were classified as malignancies. There were only 6 in total, 4 malignant neoplasms of prostate and 2 of kidney, except renal pelvis. Patients with malignancies are immunocompromised and therefore at high risk of bloodstream infections. Consequently, long-term antibiotic prophylaxis is commonly used in these patients, which however presents risks as a result of microbiota alteration and selection of resistant pathogens (Liss and Cornely 2016). However, the proportion of quasi-nosocomial *E. coli* bloodstream infections admitted for malignancy if anything decreased over time (2% yearly decrease [95% CI -5%, 0%], p=0.04). One potential explanation could be the greater awareness of the risk associated with ineffective antibiotic prophylaxis which could lead to antibiotic resistant infections; I do not have information on what antibiotic prophylaxis was being used over the study period so cannot explore this

further. There was no evidence that the previous admission was shorter in the 1132 quasi-nosocomial *E. coli* bloodstream infection cases than the quasi-community *E. coli* bloodstream infection cases (median 2.0 [IQR 0.3, 7.9] days vs 2.3 [0.3, 8.2] days respectively, ranksum p=0.15). There was, however, strong evidence that quasi-nosocomial *E. coli* bloodstream infections with a UTI diagnostic code or an infectious primary diagnostic code for the previous admission were rising faster than those without (heterogeneity p=0.005, p<0.001 respectively, **Figure 6B&C**). Despite the fact that these types of quasi-nosocomial *E. coli* bloodstream infection were rising faster than other cases, they still comprised <25% of quasi-nosocomial *E. coli* bloodstream infections in total.

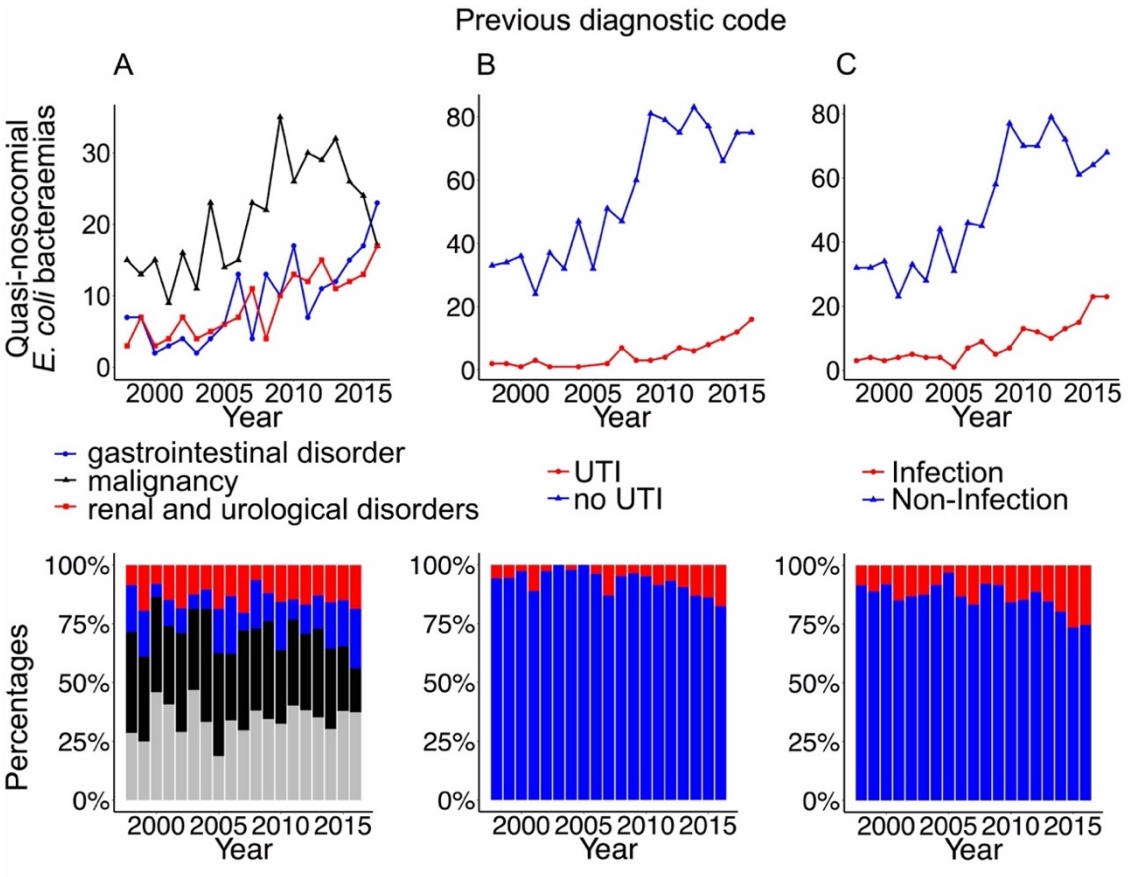


Figure 6 Annual quasi-nosocomial E. coli bloodstream infections categorized according to the diagnostic code in the preceding admission

Note: (A) by the three main categories of primary diagnostic codes from the previous admission, (B) by having a UTI in any of the diagnostic codes from the previous admission, and (C) having the primary diagnostic code of the previous admission as an infection versus non-infection. See **Table 1** for all diagnostic code categories.

Table 1 Primary diagnostic code for the previous admission for quasi-nosocomial E. coli bloodstream infections

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
Cardiovascular disorder	1	2	3	2	2	4	3	1	4	3	8	7	5	3	13	4	2	4	4	75
Neurological disorder	1	0	2	2	1	1	2	0	0	1	2	4	2	5	3	0	2	3	3	34
Dermatological or rheumatological disorders	0	1	1	0	1	1	2	0	1	2	3	1	0	3	1	2	1	2	1	23
Endocrine disorder	1	1	1	0	0	0	0	0	1	0	1	0	0	0	1	2	0	1	0	9
Gastrointestinal disorder	7	7	2	3	4	2	4	6	13	4	13	10	17	7	11	12	15	17	23	177
Gynaecological or obstetric disorder	1	1	1	3	1	1	4	1	2	1	1	0	1	3	0	2	4	2	1	30
Haematological disorder	1	0	0	1	2	0	1	0	0	0	2	2	3	2	4	2	0	1	0	21
Malignancy	15	13	15	9	16	11	23	14	15	23	22	35	26	30	29	32	26	24	17	395
Orthopaedic disorders including trauma	0	0	3	0	0	2	1	2	1	2	2	2	5	4	1	4	5	7	3	44
Poisoning	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3
Renal and urological disorders	3	7	3	4	7	4	5	6	7	11	4	10	13	12	15	11	12	13	17	164
Respiratory disorder	2	1	0	3	2	4	0	1	3	4	3	7	4	5	3	4	5	7	12	70
Dermatological disorder	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Other	3	2	5	0	2	2	3	1	5	3	1	6	7	8	8	10	4	6	9	85

Incidence trends for *E. coli* bloodstream infections varied substantially with hospital-exposure (**Figure 7A**), with overall increases clearly driven by community and quasi-community hospital-exposure groups, and no evidence of different incidence trends between these two groups in 2016 (heterogeneity p=0.27). By contrast, quasi-nosocomial and nosocomial *E. coli* bloodstream infections increased more slowly. Results were similar using alternative categorisations of nosocomial as taken >48 hours after admission (rather than up to and including two calendar days) (**Figure 8**), and dividing the remaining *E. coli* bloodstream infections that were either taken in the first 48 hours of an admission or outside an admission according to their last discharge date being more than two years before, between two years and one year, between one year and 90 days, between 90 days and 30 days and less than 30 days; as well as thresholds of 90 and 365 days. These sensitivity analyses support current incidence increases becoming larger and larger the longer the time since last hospital exposure.

Regarding the changes in incidence observed around 2010, the only change in practice that I am aware of around this time is the change in co-amoxiclav formulation which would have not affected the hospital usage as clinicians were prescribing an additional 250mg of amoxicillin before the change in formulation; however it would have represented quite a significant change in the total co-amoxiclav DDD GP usage.

Considering only the *first E. coli* bloodstream infection per patient or subsequent *E. coli* bloodstream infections (**Figure 9**) gave broadly similar results. Year-on-year increases in the incidence of *first E. coli* bloodstream infection became smaller (but still significant) the more recent the hospital exposure. Quasi-community *recurrent E. coli* bloodstream infection

increased faster than *first* quasi-community *E. coli* bloodstream infections (heterogeneity $p<0.0001$). In contrast, the stable current trend in the incidence of all quasi-nosocomial *E. coli* bloodstream infections was driven by declines in *recurrent* quasi-nosocomial *E. coli* bloodstream infections coupled with increases in *first* quasi-nosocomial *E. coli* bloodstream infections.

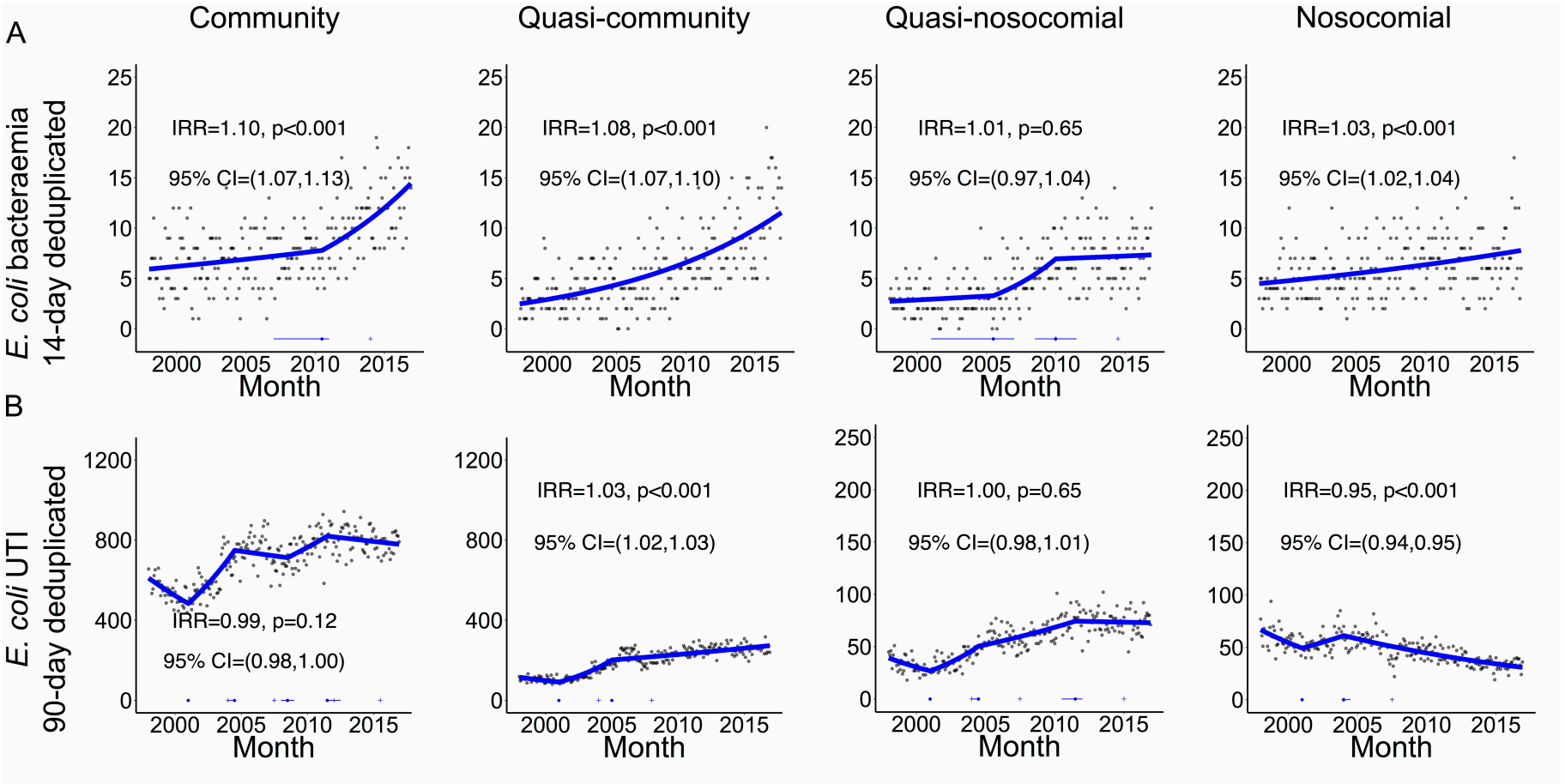


Figure 7 Monthly (A) E. coli bloodstream infection and (B) E. coli UTI according to recent hospital-exposure (first and recurrent infections combined)

Note: only counting *E. coli* bloodstream infection recurrences occurring >14 days after an index positive, and *E. coli* UTI recurrences occurring >90 days after an index positive. Thick blue line represents the estimated incidence by iterative sequential regression (ISR). Blue lines at the base of the graph represent 95% CI around the breakpoints estimated by the ISR model, while the crosses represent the month when the breakpoint was identified. IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016.

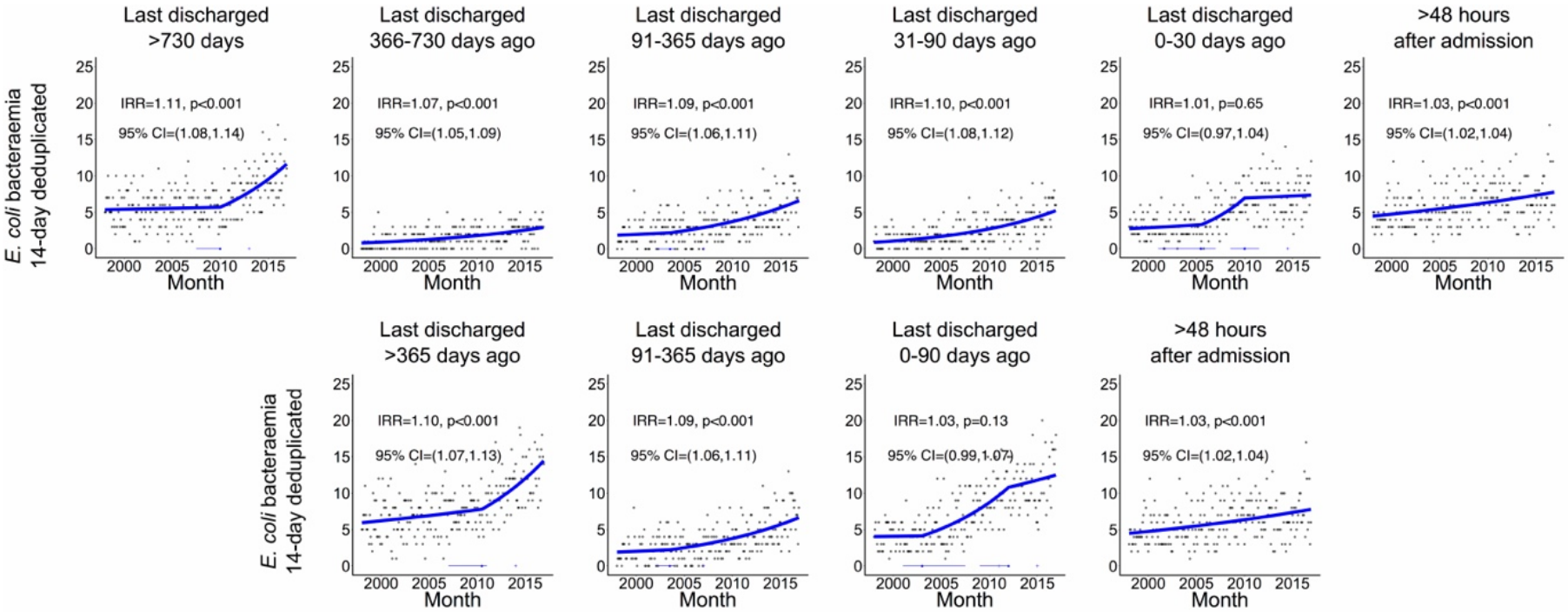


Figure 8 Monthly E. coli bloodstream infection (first and recurrent infections combined) using different definitions of recency of hospital exposure

Note: only counting *E. coli* bloodstream infection recurrences occurring >14 days after an index positiveIRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016.

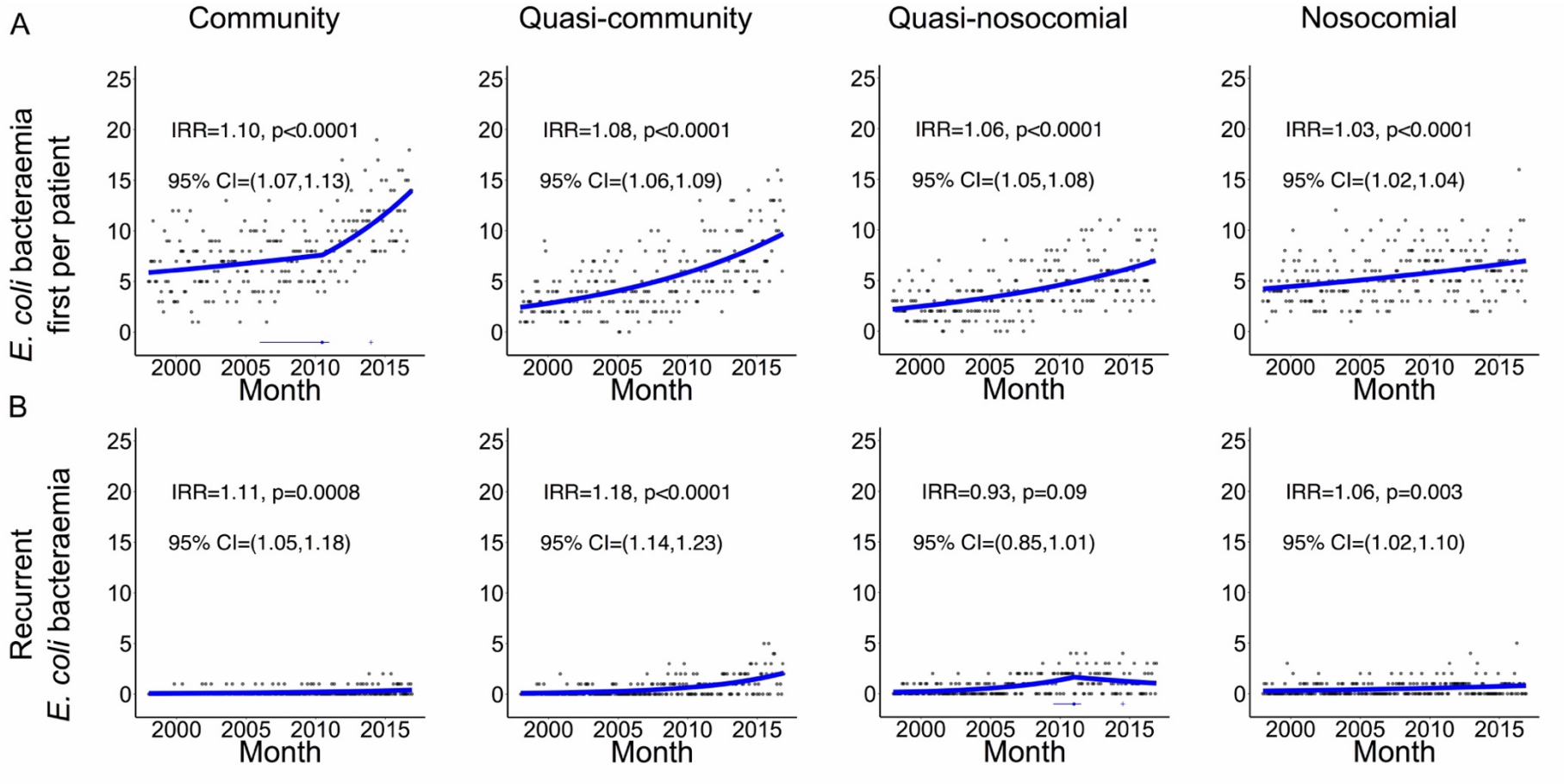


Figure 9 Monthly E. coli bloodstream infection according to recent hospital-exposure (A) first per patient only (B) recurrences within a patient only

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. See **Table 3** for heterogeneity tests between first vs subsequent E. coli bloodstream infections and **Figure 11A** for a summary of all current trends.

After 90-day de-duplication, 228376 *E. coli* UTIs occurred in 137075 patients, that is, 91301 (40%) UTIs were recurrences (relapse/re-infection). Median (IQR) age at UTI was 58 (32-76) years and 198373 (87%) occurred in women. Recurrences occurred a median (IQR) 457 (200-1119) days apart: of the 41371 (30%) patients with recurrences, 22011 (53%) had one and 8742 (21%) had two (range 1-33). 12898 (9%) patients had two *E. coli* UTI within six months. *E. coli* UTIs were predominantly community (160359, 70%), and less commonly quasi-community (44283, 19%), quasi-nosocomial (12764, 6%) or nosocomial (10970, 5%).

Incidence of *E. coli* UTI increased between 1998 and 2016 in the community, quasi-community and quasi-nosocomial groups, although current trends were fairly stable; however, incidence of nosocomial *E. coli* UTIs declined significantly (**Figure 7B**). Furthermore, increases in the other hospital-exposure groups were accounted for entirely by substantial increases in recurrent *E. coli* UTI episodes, with decreasing overall trends in first *E. coli* UTI per patient (**Figure 10**).

In 2016, therefore, recurrences accounted for at least half of community, quasi-community and quasi-nosocomial *E. coli* UTIs, and around a fifth of quasi-community and quasi-nosocomial *E. coli* bloodstream infections (**Table 2**).

Table 2 Relative contribution of recurrent E. coli bloodstream infections and E. coli UTIs to total numbers in 2016 by recent hospital-exposure

	Community	Quasi-community	Quasi-nosocomial	Nosocomial
	Recurrent/total (%)	Recurrent/total (%)	Recurrent/total (%)	Recurrent/total (%)
Bloodstream infections	4/163 (2%)	24/164 (15%)	17/91 (19%)	11/98 (11%)
UTIs	4682/9464 (49%)	2003/3097 (65%)	472/885 (53%)	148/416 (36%)

Table 3 and **Figure 11A&B** summarise all current incidence trends (in 2016) and heterogeneity tests between different types of *E. coli* infections (bloodstream and urine).

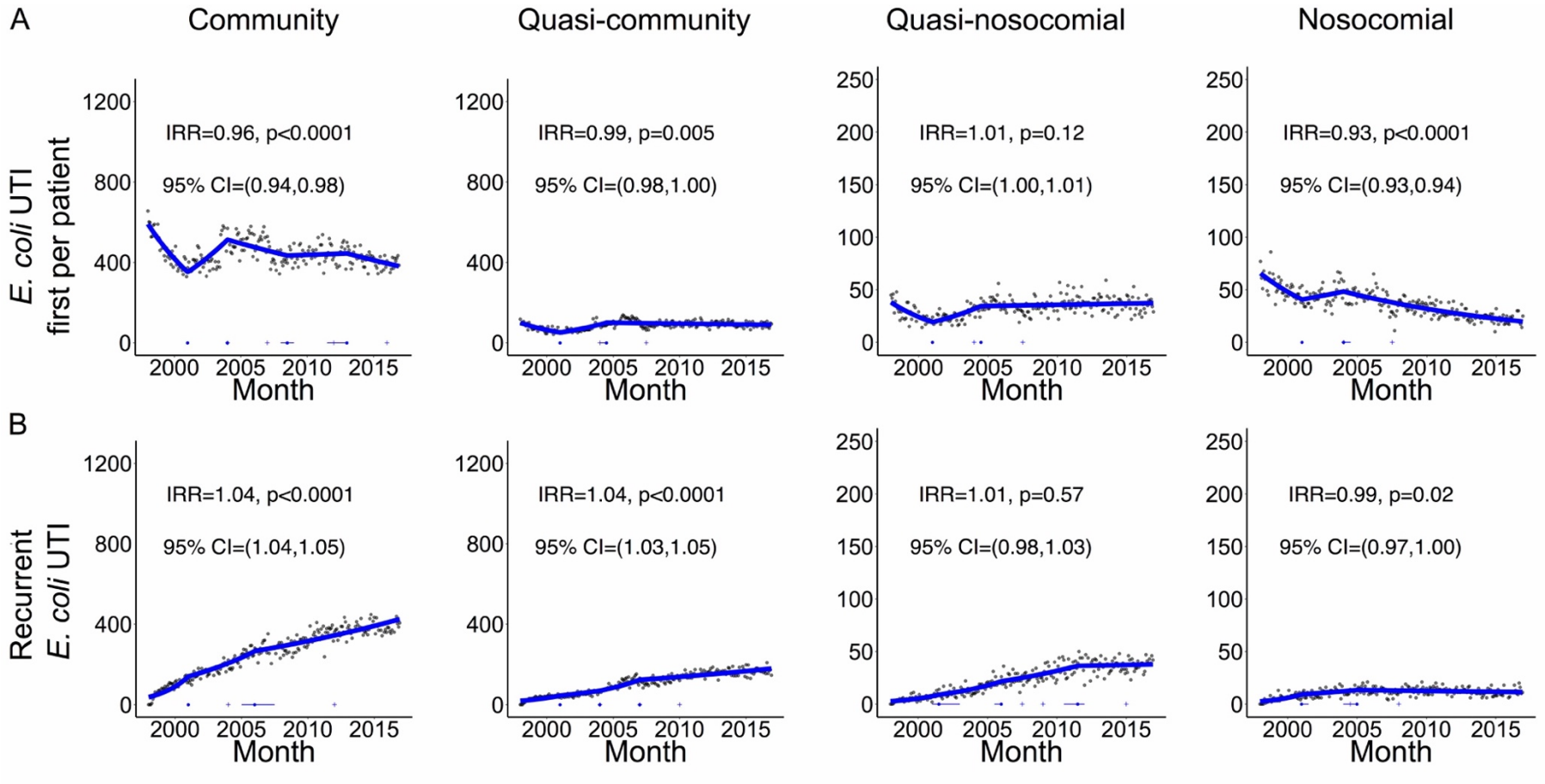


Figure 10 Monthly E. coli UTIs according to recent hospital-exposure (A) first per patient only (B) recurrences within a patient only

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. See **Table 3** for heterogeneity tests between first vs subsequent *E. coli* UTIs and **Figure 11B** for a summary of all current trends.

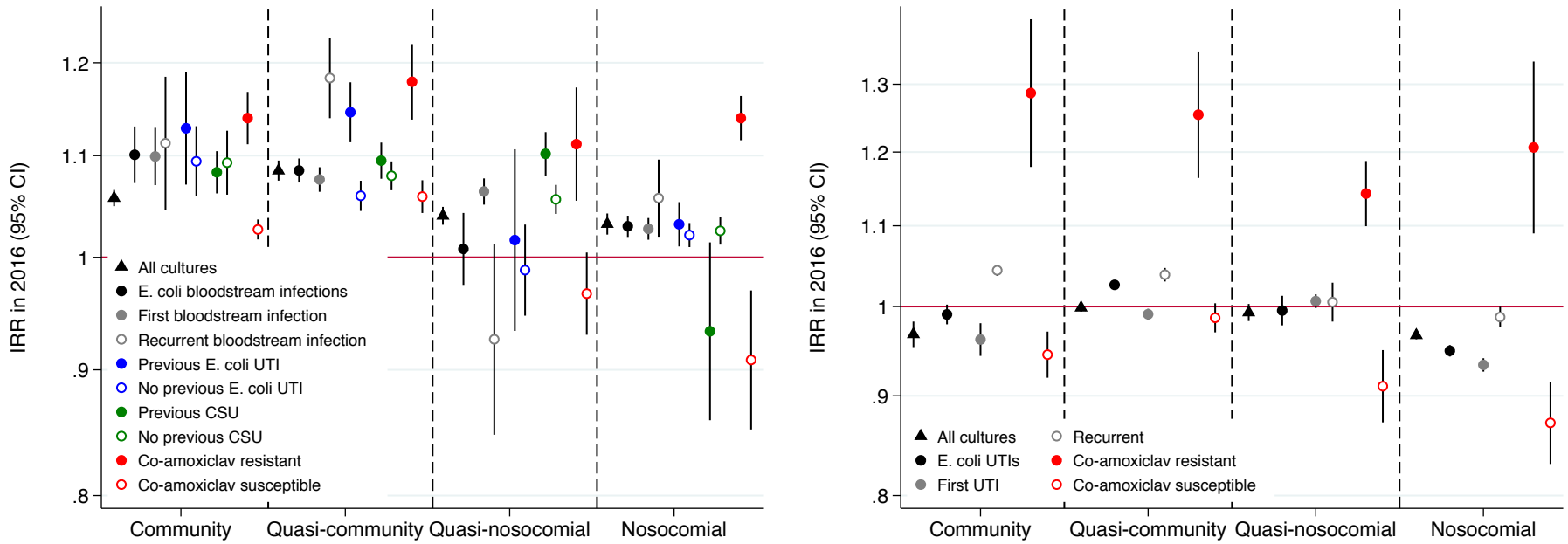


Figure 11 Summary of incidence trends in 2016 for (A) E. coli bloodstream infections, (B) E. coli UTIs

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. CSU=catheter urine specimen. See **Table 3** for numbers and heterogeneity tests.

Table 3 Summary of current (2016) annual incidence rate ratios

	Insert	Community	Quasi-community	Quasi-nosocomial	Nosocomial
		IRR (95% CI)	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)
All *E. coli* bloodstream infection	Figure 7	**1.10 (1.07-1.13)**	**1.08 (1.07-1.10)**	1.01 (0.97-1.04)	**1.03 (1.02-1.04)**
According to previous *E. coli* bloodstream infection					
First *E. coli* bloodstream infection *	Figure 9	**1.10 (1.07-1.13)**	**1.08 (1.06-1.09)**	**1.06 (1.05-1.08)**	**1.03 (1.02-1.04)**
Recurrent *E. coli* bloodstream infection	Figure 9	**1.11 (1.05-1.18)**	**1.18 (1.14-1.23)**	0.93 (0.85-1.01)	1.06 (1.02-1.10)
Heterogeneity first vs recurrence *E. coli* bloodstream infection		p=0.70	p<0.0001	p=0.004	p=0.14
All blood cultures (regardless of result)	Figure 12	**1.06 (1.05-1.07)**	**1.08 (1.07-1.09)**	**1.04 (1.03-1.05)**	**1.03 (1.02-1.04)**
Heterogeneity *E. coli* bloodstream infection vs all blood cultures		P=0.0006	p=0.92	p=0.05	p=0.76
According to previous *E. coli* UTI					
All *E. coli* bloodstream infection with previous *E. coli* UTI **	Figure 16	**1.13 (1.07-1.19)**	**1.15 (1.11-1.18)**	1.02 (0.93-1.11)	1.03 (1.01-1.05)
All *E. coli* bloodstream infection with no previous *E. coli* UTI	Figure 16	**1.09 (1.06-1.13)**	**1.06 (1.04-1.07)**	0.99 (0.95-1.03)	**1.02 (1.01-1.03)**
Heterogeneity by previous *E. coli* UTI		p=0.66	p<0.0001	p=0.73	p=0.02
According to previous catheter urine specimen (CSU)					
All *E. coli* bloodstream infection with previous CSU	Figure 18	**1.08 (1.06-1.10)**	**1.09 (1.08-1.11)**	**1.10 (1.08-1.12)**	0.93 (0.86-1.01)
All *E. coli* bloodstream infection with no previous CSU	Figure 18	**1.09 (1.06-1.13)**	**1.08 (1.06-1.09)**	**1.06 (1.04-1.07)**	**1.03 (1.01-1.04)**
Heterogeneity by previous CSU		p=0.61	p=0.18	P=0.0002	p=0.03
According to co-amoxiclav susceptibility					
Co-amoxiclav resistant *E. coli* bloodstream infection	Figure 20	**1.14 (1.11-1.17)**	**1.18 (1.14-1.22)**	**1.11 (1.05-1.17)**	**1.14 (1.12-1.16)**

	Insert	Community	Quasi-community	Quasi-nosocomial	Nosocomial
		IRR (95% CI)	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)
Co-amoxiclav susceptible *E. coli* bloodstream infection	Figure 20	**1.03 (1.02-1.04)**	**1.06 (1.04-1.07)**	0.97 (0.93-1.00)	0.91 (0.85-0.97)
Heterogeneity		p<0.0001	p<0.0001	p<0.0001	p<0.0001
30-day mortality: all *E. coli* bloodstream infection	Figure 15	0.99 (0.96-1.01)	0.99 (0.96,1.01)	0.98 (0.95,1.00)	0.98 (0.96,1.00)
CRP >156 mg/L: all *E. coli* bloodstream infection	Figure 15	0.99 (0.98,1.01)	0.99 (0.97,1.01)	1.01 (0.98,1.03)	1.00 (0.98,1.02)
30-day mortality: co-amoxiclav susceptible *E. coli* bloodstream infection	Figure 23†	0.98 (0.96,1.01)	0.99 (0.97,1.02)	0.97 (0.94,1.00)	0.96 (0.94,0.99)
30-day mortality: co-amoxiclav resistant *E. coli* bloodstream infection	Figure 23†	1.00 (0.94,1.06)	0.96 (0.91,1.02)	1.00 (0.93,1.06)	0.98 (0.95,1.02)
All *E. coli* UTI	Figure 7	0.99 (0.98-1.00)	**1.03 (1.02-1.03)**	1.00 (0.98-1.01)	**0.95 (0.94-0.95)**
According to previous *E. coli* UTI					
First *E. coli* UTI *	Figure 10	**0.96 (0.94-0.98)**	0.99 (0.98-1.00)	1.01 (1.00-1.01)	**0.93 (0.93-0.94)**
Recurrent *E. coli* UTI	Figure 10	**1.04 (1.04-1.05)**	**1.04 (1.03-1.05)**	1.01 (0.98-1.03)	0.99 (0.97-1.00)
Heterogeneity		p<0.0001	p<0.0001	p=0.99	p<0.0001
All urine cultures (regardless of result)	Figure 14	**0.99 (0.98-0.99)**	**1.01 (1.01-1.01)**	**1.02 (1.01-1.02)**	1.02 (1.01-1.04)
According to co-amoxiclav susceptibility					
Co-amoxiclav resistant *E. coli* UTI	Figure 22	**1.29 (1.18-1.40)**	**1.25 (1.16-1.35)**	**1.14 (1.10-1.19)**	**1.21 (1.09-1.34)**
Co-amoxiclav susceptible *E. coli* UTI	Figure 22	**0.94 (0.92-0.97)**	0.99 (0.97-1.00)	**0.91 (0.87-0.95)**	**0.87 (0.83-0.91)**
Heterogeneity		p<0.0001	p<0.0001	p<0.0001	p<0.0001

* First ever recorded per patient between 1998-2016; all other subsequent cases counted as recurrences
** Any *E. coli* UTI 3 or more days prior to the *E. coli* bloodstream infection.
† No evidence of heterogeneity therefore **Figure 15** shows pooled mortality trends across susceptible and resistant *E. coli* bloodstream infection
Note: showing annual incidence rate ratios estimated by ISR in 2016; bold p<0.001, underline p between 0.001-0.05

2.3.1 Impact of population and sampling on E. coli bloodstream infection incidence

Blood culture submission rates increased substantially between 1998 and 2016 for community/quasi-community/quasi-nosocomial groups (**Figure 12**), raising the possibility that observed increases in *E. coli* bloodstream infections were driven by increases in the use of blood cultures as a diagnostic test. However, there was no suggestion that the indications for blood culture changed with time: changes in neutrophils and CRP when cultures (positive or negative) were taken were small and not clinically meaningful and 30-day mortality post blood culture sampling was stable (**Figure 13**). Further, increases in community blood culture submission rates were significantly smaller than increases in community *E. coli* bloodstream infections (heterogeneity p=0.0006, **Figure 11A**, **Table 3**). In contrast, urine sample submission was more stable over time (**Figure 14**).

An alternative hypothesis is that incidence increases are driven by an aging population. However, standardising for age and sex explained only between 10% and 26% of the increases in overall incidence of *E. coli* bloodstream infections (**Table 4**), and of first-per-patient *E. coli* bloodstream infections (**Table 5**) depending on the hospital-exposure group. Standardising additionally for blood cultures taken explained only between 9% and 28% of the increase in incidence of all or first-per-patient *E. coli* bloodstream infections. Moreover, the greatest percentage was explained in nosocomial *E. coli* bloodstream infections which were rising slowest and the least in community *E. coli* bloodstream infections which were rising fastest (**Table 4**, **Table 5**).

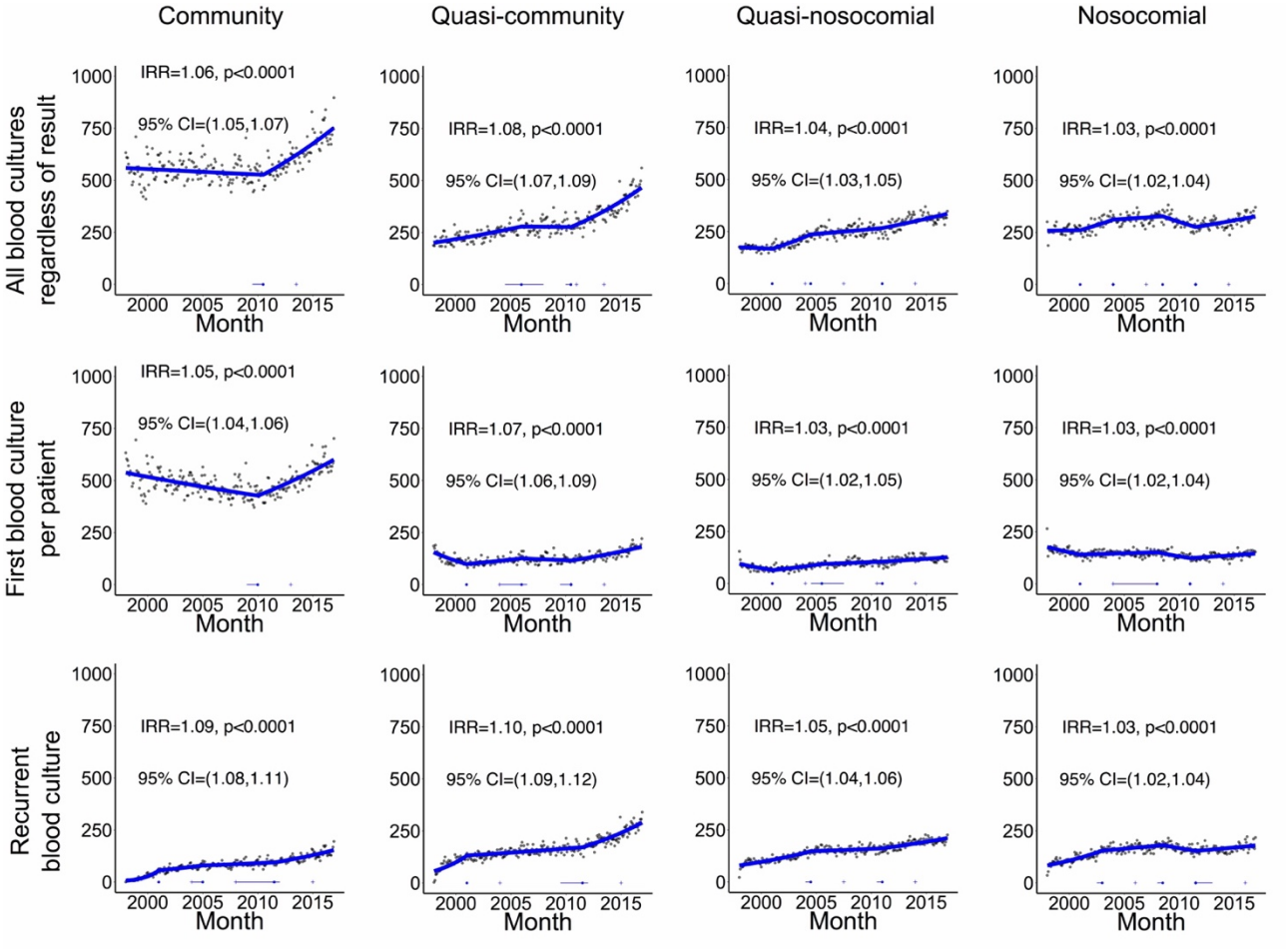


Figure 12 Monthly blood samples submitted for culture regardless of result according to recent hospital-exposure (first and repeat samples per patient)

Note: including repeat samples submitted >14 days after an index sample to match de-deduplication strategy for bloodstream infections. IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016.

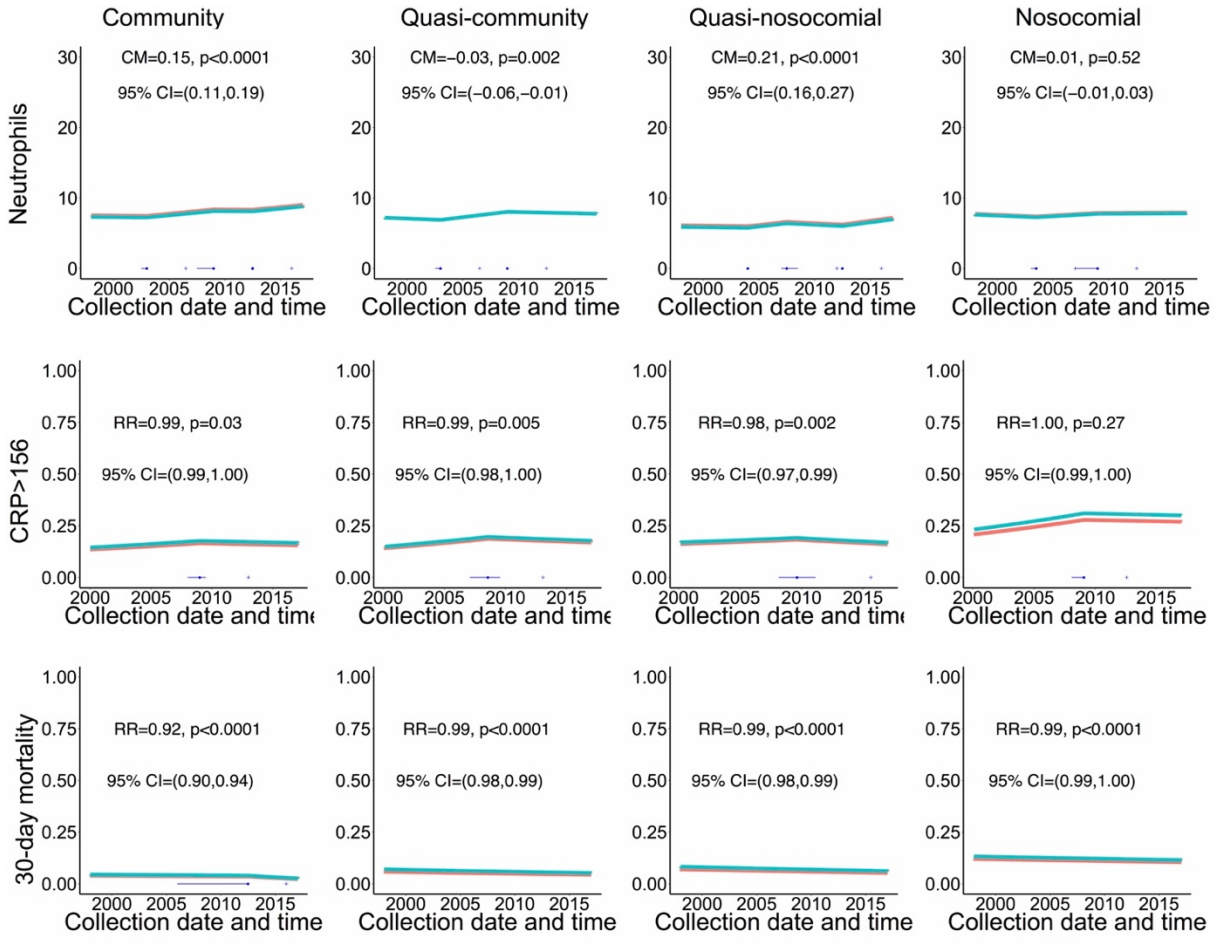


Figure 13 Trends in haematology/biochemistry test results and 30-day mortality following a blood culture being taken regardless of its result according to recent hospital-exposure (first and recurrent infections)

Note: including repeat samples submitted >14 days after an index sample. Fitted lines are for men (blue) and women (red) at the mean age 51 years, RR=annual rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. CM=annual change in median in 2016.

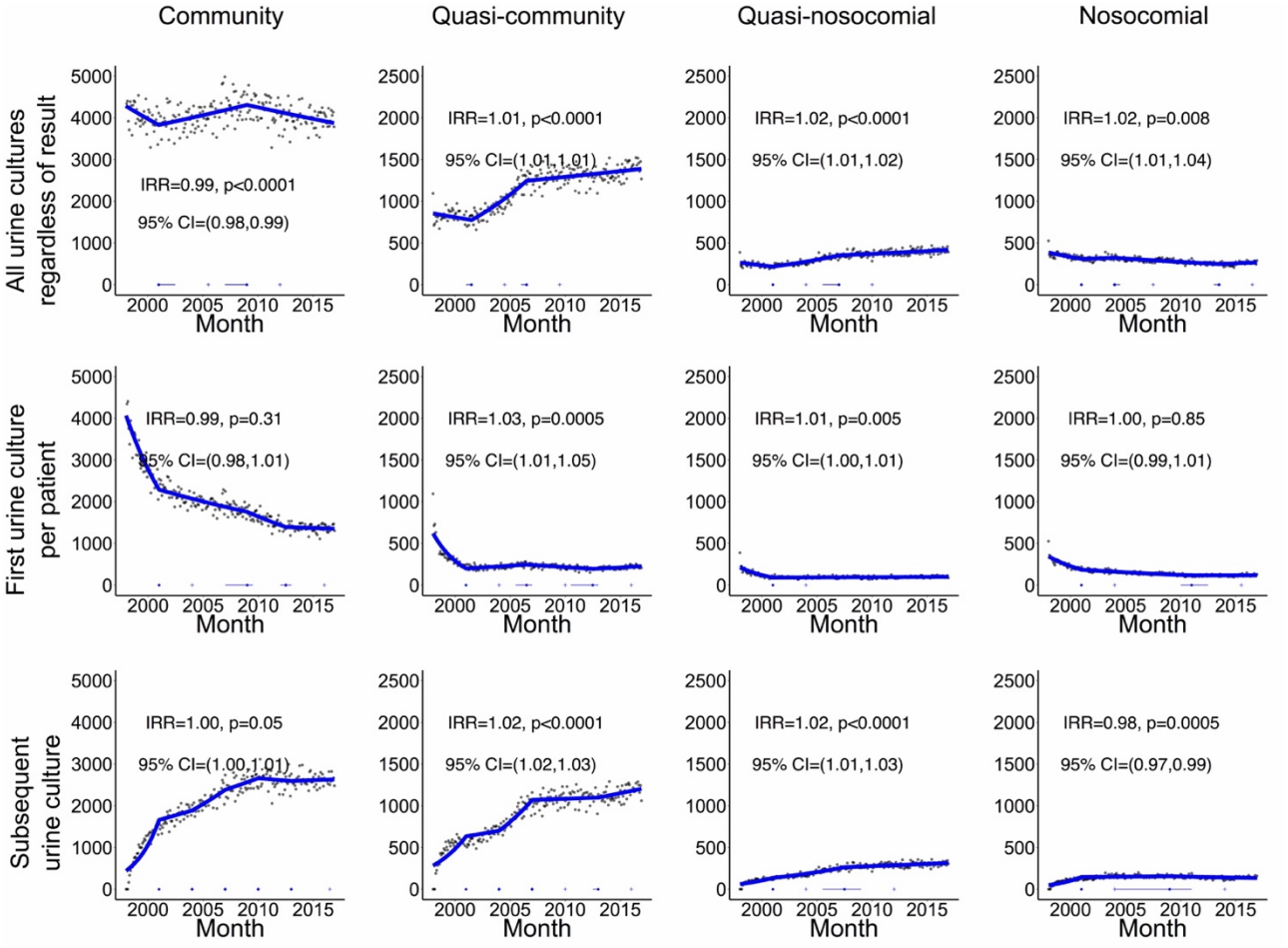


Figure 14 Monthly urine samples submitted for culture regardless of result according to recent hospital-exposure (first and repeat samples per patient)

Note: including repeat samples submitted >90 days after an index sample to match de-deduplication strategy for UTIs. IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016.

Table 4 Overall E. coli bloodstream infection incidence trends in 2016, unadjusted and standardized to the sex and age population of Oxfordshire in 1998

	Community	Quasi-community	Quasi-nosocomial	Nosocomial
	IRR (95% CI) (with breakpoint)	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)
Unstandardized	1.10 (1.04-1.17)	**1.08 (1.06-1.10)**	**1.07 (1.05-1.09)**	**1.03 (1.01-1.05)**
Standardized	1.09 (1.02-1.16)	**1.07 (1.05-1.09)**	**1.06 (1.04-1.08)**	1.02 (1.01-1.04)
Percentage change in regression coefficient*	-10%	-14%	-12%	-23%
Also standardized for number of samples taken per month	1.09 (1.02-1.16)	**1.07 (1.04-1.09)**	**1.06 (1.04-1.08)**	1.02 (1.01-1.04)
Percentage change in regression coefficient*	-9%	-17%	-12%	-23%

* difference in coefficients from standardised and unstandardized estimates expressed as a percentage of the unstandardized estimate
Note: only fitting a single trajectory to incidence for the quasi-nosocomial hospital-exposure group, approximating Figure 1. IRR=annual incidence rate ratio per year in 2016; bold p<0.001, underline p between 0.001-0.05.

Table 5 Incidence trends for first E. coli bloodstream infection per patient, unadjusted and standardized to the sex and age population of Oxfordshire in 1998

	Community	Quasi-community	Quasi-nosocomial	Nosocomial
	IRR (95% CI) (with breakpoint)	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)
Unstandardized	1.10 (1.03-1.17)	**1.07 (1.05-1.10)**	**1.06 (1.04-1.08)**	1.03 (1.01-1.05)
Standardized	1.09 (1.02-1.16)	**1.06 (1.04-1.08)**	**1.05 (1.04-1.07)**	1.02 (1.00-1.04)
Percentage change in regression coefficient*	-10%	-15%	-14%	-26%
Also standardized for samples taken per month	1.09 (1.02-1.16)	**1.06 (1.04-1.08)**	**1.05 (1.04-1.07)**	1.02 (1.00-1.04)
Percentage change in regression coefficient*	-9%	-19%	-13%	-28%

* difference in coefficients from standardised and unstandardized estimates expressed as a percentage of the unstandardized estimate
Note: IRR=annual incidence rate ratio per year in 2016; bold p<0.001, underline p between 0.001-0.05.

2.3.2 Disease severity of *E. coli* bloodstream infections

One possible explanation for increases in incidence of any infection is that more cases who are less unwell are reported over time, meaning that increasing incidence predominantly reflects increased ascertainment of less severe cases. Mortality in the 30 days following *E. coli* bloodstream infection was substantial at 25% (340/1363) in nosocomial cases, 19% (217/1128) in quasi-nosocomial cases, 16% (219/1344) in quasi-community cases and 14% (254/1784) in community cases (**Table 6**). Mortality did decline slightly in the nosocomial (RR=0.98 [95% CI 0.96, 1.00], p=0.03) and quasi-nosocomial (RR=0.98 [0.95, 1.00], p=0.06) groups, potentially consistent with less unwell patients being identified (although other changes in management may also contribute) (**Figure 15**, rates adjusted for both age and sex). However, there was no evidence for changes in 30-day mortality in the quasi-community (RR=0.99 [0.96, 1.01], p=0.32) and community (RR=0.99 [0.96, 1.01], p=0.21) groups, in which incidence was increasing fastest, suggesting that similarly unwell cases were being identified over time, just in far greater numbers.

Changes in haematology/biochemistry test results over time were small and/or non-significant (**Figure 15**), and did not indicate that less severe infections were being identified, or that there were changes in pathogen virulence.

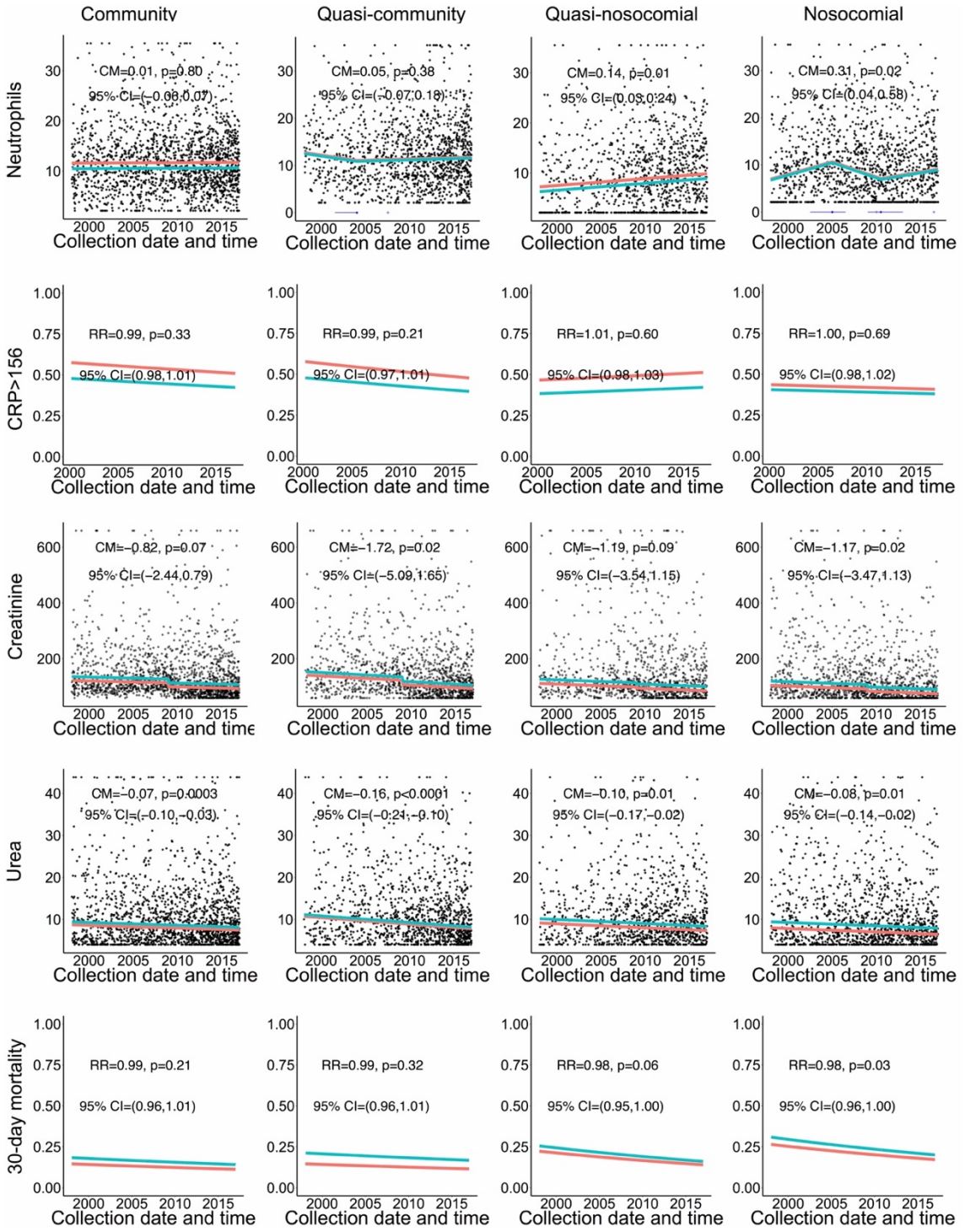


Figure 15 Trends in haematology/biochemistry test results and 30-day mortality following E. coli bloodstream infection according to recent hospital-exposure (first and recurrent infections)

Note: CM=change per year in median value in 2016, that is the relative increase in rate per year as estimated in 2016. Adjusted for age and gender. Fitted lines are for men (blue) and women (red) at mean age 67 years, RR=annual rate ratio in 2016.

Table 6 Yearly numerators/denominators (percentages) for 30-day mortality following E. coli bloodstream infection, co-amoxiclav resistant E. coli bloodstream infections and 30-day mortality following E. coli bloodstream infection in co-amoxiclav resistant versus susceptible cases

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
30-day mortality following E. coli bloodstream infection																				
Community	11/65	9/67	9/79	10/68	7/72	16/80	13/73	10/80	21/81	10/81	12/84	17/88	10/84	13/107	14/112	18/129	18/141	17/133	19/160	254/1858
	(17%)	(13%)	(11%)	(15%)	(10%)	(20%)	(18%)	(12%)	(26%)	(12%)	(14%)	(19%)	(12%)	(12%)	(12%)	(14%)	(13%)	(13%)	(12%)	(14%)
Quasi-community	5/24	8/52	10/41	3/32	9/51	6/52	10/58	4/33	13/49	11/63	10/68	14/69	14/78	14/86	11/90	14/109	13/108	24/117	26/164	219/1346
	(21%)	(15%)	(24%)	(9%)	(18%)	(12%)	(17%)	(12%)	(27%)	(17%)	(15%)	(20%)	(18%)	(16%)	(12%)	(13%)	(12%)	(21%)	(16%)	(16%)
Quasi-nosocomial	5/34	7/36	8/37	5/27	9/38	8/32	10/48	5/32	16/53	15/54	9/62	23/82	16/83	11/82	12/89	14/85	8/76	17/87	19/91	217/1132
	(15%)	(19%)	(22%)	(19%)	(24%)	(25%)	(21%)	(16%)	(30%)	(28%)	(15%)	(28%)	(19%)	(13%)	(13%)	(16%)	(11%)	(20%)	(21%)	(19%)
Nosocomial	18/43	19/60	19/69	14/54	16/54	14/65	14/59	20/73	15/60	13/77	26/85	17/82	20/85	19/83	25/91	18/68	14/71	16/86	23/98	340/1365
	(42%)	(32%)	(28%)	(26%)	(30%)	(22%)	(24%)	(27%)	(25%)	(17%)	(31%)	(21%)	(24%)	(23%)	(27%)	(26%)	(20%)	(19%)	(23%)	(25%)
Numbers of co-amoxiclav resistant E. coli bloodstream infections																				
Community	8/72	3/70	9/87	7/70	6/74	8/85	9/78	7/82	19/86	15/81	16/87	21/93	20/88	22/110	17/113	48/135	44/144	42/136	46/162	367/1853
	(11%)	(4%)	(10%)	(10%)	(8%)	(9%)	(12%)	(9%)	(22%)	(19%)	(18%)	(23%)	(23%)	(20%)	(15%)	(36%)	(31%)	(31%)	(28%)	(20%)
Quasi-community	2/25	7/51	6/32	5/51	8/52	7/58	3/33	18/49	14/63	14/68	14/69	18/77	16/87	11/90	40/109	46/108	43/117	65/164	2/25	337/1343
	(8%)	(14%)	(19%)	(10%)	(15%)	(12%)	(9%)	(37%)	(22%)	(21%)	(20%)	(23%)	(18%)	(12%)	(37%)	(43%)	(37%)	(40%)	(8%)	(25%)
Quasi-nosocomial	5/35	2/36	3/37	1/27	1/38	3/32	8/48	3/32	11/53	18/54	11/63	20/84	36/83	19/81	23/89	41/85	27/76	34/87	47/91	313/1131
	(14%)	(6%)	(8%)	(4%)	(3%)	(9%)	(17%)	(9%)	(21%)	(33%)	(17%)	(24%)	(43%)	(23%)	(26%)	(48%)	(36%)	(39%)	(52%)	(28%)
Nosocomial	6/43	8/61	4/69	8/53	8/54	8/65	8/59	9/73	10/60	23/78	24/85	24/82	29/85	25/83	27/91	40/68	39/71	42/86	54/98	396/1364
	(14%)	(13%)	(6%)	(15%)	(15%)	(12%)	(14%)	(12%)	(17%)	(29%)	(28%)	(29%)	(34%)	(30%)	(30%)	(59%)	(55%)	(49%)	(55%)	(29%)
30-day mortality following E. coli bloodstream infection: co-amoxiclav resistant versus susceptible																				
Community resistant	1/7	0/2	1/7	1/6	1/6	1/8	1/9	0/6	3/18	3/15	2/16	8/19	1/20	4/22	0/17	7/45	6/42	7/41	7/44	54/350
	(14%)	(0%)	(14%)	(17%)	(17%)	(12%)	(11%)	(0%)	(17%)	(20%)	(12%)	(42%)	(5%)	(18%)	(0%)	(16%)	(14%)	(17%)	(16%)	(15%)
Community susceptible	10/58	9/65	8/71	8/61	6/66	15/72	11/63	10/74	18/63	6/65	10/68	9/69	9/64	9/85	13/94	11/84	12/98	9/90	12/115	195/1425
	(17%)	(14%)	(11%)	(13%)	(9%)	(21%)	(17%)	(14%)	(29%)	(9%)	(15%)	(13%)	(14%)	(11%)	(14%)	(13%)	(12%)	(10%)	(10%)	(14%)
Quasi-community resistant	0/1	1/6	0/0	0/6	2/5	1/8	1/7	0/3	5/18	3/14	3/14	3/14	3/18	5/16	3/11	4/40	4/46	8/43	7/65	53/335
	(0%)	(17%)	(-)	(0%)	(40%)	(12%)	(14%)	(0%)	(28%)	(21%)	(21%)	(21%)	(17%)	(31%)	(27%)	(10%)	(9%)	(19%)	(11%)	(16%)
Quasi-community	5/23	7/45	10/40	3/26	7/46	5/44	9/51	4/30	8/31	8/49	7/54	11/55	11/59	9/70	8/79	10/69	9/62	16/74	19/99	166/1006
susceptible	(22%)	(16%)	(25%)	(12%)	(15%)	(11%)	(18%)	(13%)	(26%)	(16%)	(13%)	(20%)	(19%)	(13%)	(10%)	(14%)	(15%)	(22%)	(19%)	(17%)
Quasi-nosocomial resistant	1/5	1/2	0/3	0/1	1/1	2/3	0/8	0/3	3/11	4/18	0/11	5/20	6/36	4/19	3/23	10/41	4/27	5/34	13/47	62/313
	(20%)	(50%)	(0%)	(0%)	(100%)	(67%)	(0%)	(0%)	(27%)	(22%)	(0%)	(25%)	(17%)	(21%)	(13%)	(24%)	(15%)	(15%)	(28%)	(20%)
Quasi-nosocomial	4/29	6/34	8/34	5/26	8/37	6/29	10/40	5/29	13/42	11/36	9/51	18/62	10/47	6/62	9/66	4/44	4/49	12/53	6/44	154/814
susceptible	(14%)	(18%)	(24%)	(19%)	(22%)	(21%)	(25%)	(17%)	(31%)	(31%)	(18%)	(29%)	(21%)	(10%)	(14%)	(9%)	(8%)	(23%)	(14%)	(19%)
Nosocomial resistant	3/6	1/8	1/4	5/8	5/8	1/8	3/8	2/9	4/10	4/22	6/24	10/24	6/29	8/25	9/27	14/40	10/39	13/42	12/54	117/395
	(50%)	(12%)	(25%)	(62%)	(62%)	(12%)	(38%)	(22%)	(40%)	(18%)	(25%)	(42%)	(21%)	(32%)	(33%)	(35%)	(26%)	(31%)	(22%)	(30%)
Nosocomial susceptible	15/37	18/52	18/65	8/45	11/46	13/57	11/51	18/64	11/50	9/55	20/61	7/58	14/56	11/58	16/64	4/28	4/32	3/44	11/44	222/967
	(41%)	(35%)	(28%)	(18%)	(24%)	(23%)	(22%)	(28%)	(22%)	(16%)	(33%)	(12%)	(25%)	(19%)	(25%)	(14%)	(12%)	(7%)	(25%)	(23%)

Note: not adjusted for age and sex, in contrast to model estimates. Mortality based on complete cases (those who could be routinely linked to national information systems).

2.3.3 Impact of previous illness on *E. coli* bloodstream infection

1755 (31%) *E. coli* bloodstream infection occurred in patients with an *E. coli* UTI ≥3 days previously (median (IQR) 213 [43, 918] days previously). However, incidence trends were broadly similar for *E. coli* bloodstream infections with or without *E. coli* UTIs ≥3 days previously. The only difference was that quasi-community *E. coli* bloodstream infections were rising particularly fast in those with previous *E. coli* UTIs (heterogeneity p<0.0001, **Figure 11A**, **Figure 16**).

I next explored whether increases in *E. coli* bloodstream infection were associated with recent symptomatic UTIs, including those without positive urine cultures which **Figure 16** above restricts to. Considering urine samples/results taken within 30 days before the *E. coli* bloodstream infection, and incorporating information on mixed growth and request codes (see Methods), only 760 (13%) *E. coli* bloodstream infections were 'likely urine-associated' (evidence of recent UTI pre-dating the blood culture), with 1613 (28%) being 'urosepsis' (evidence of UTI around the time blood was taken for culture but not before), 1613 (28%) 'unlikely urine-associated' (of which 181 [11%] had a contemporaneous urine specimen positive for another pathogen, and the remainder had no urine specimen), and 1720 (30%) 'unknown' (mixed growth or negative urine culture and either an irrelevant or no request code with the urine sample). However, the relative proportions in these different subgroups did not vary substantially over time (**Figure 17**), suggesting no specific subgroup was associated with increases in incidence. The most notable change was a decrease of over 50% in the proportion of nosocomial cases that were 'likely urine-associated'.

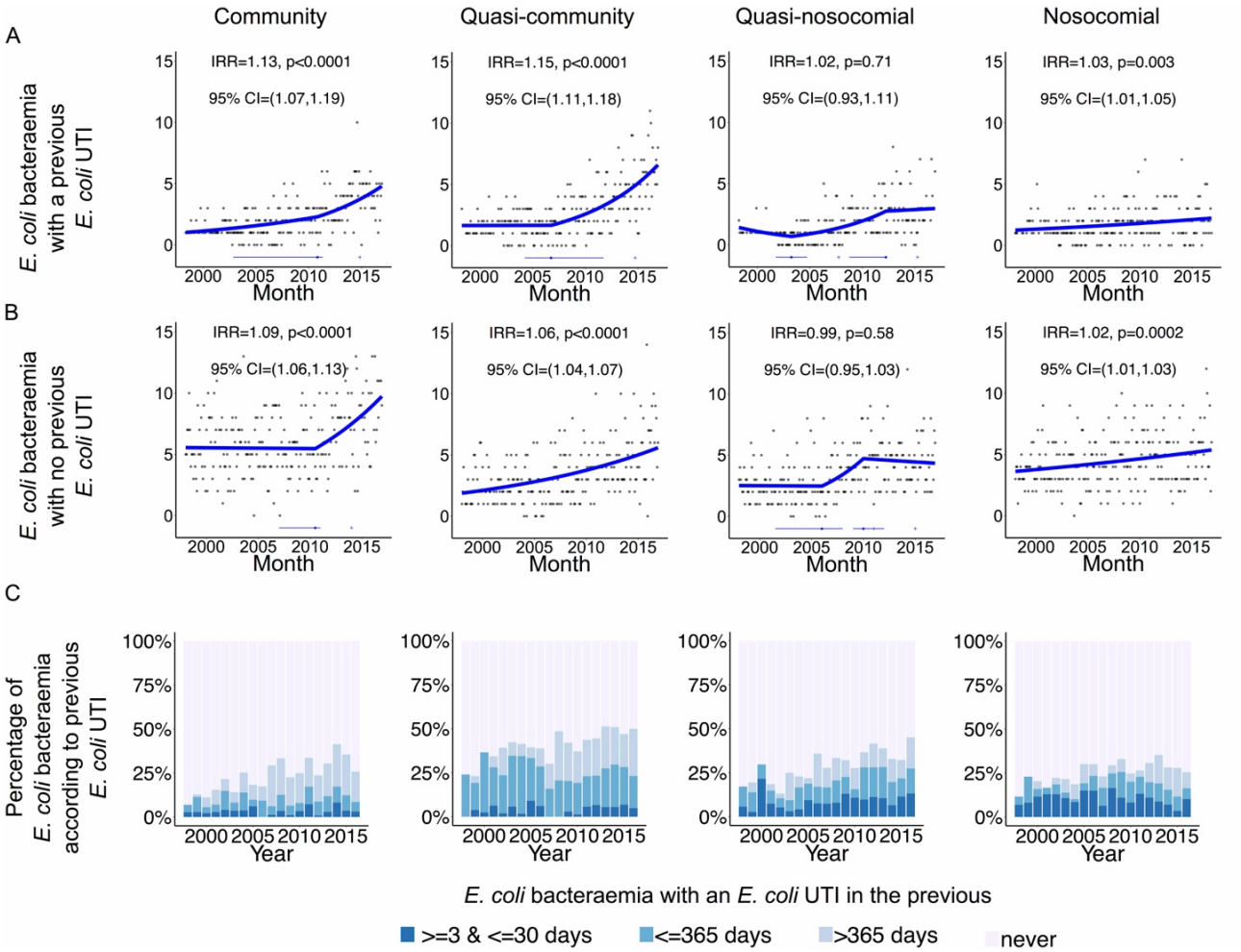


Figure 16 E. coli bloodstream infections according to whether (A) patient had had an E. coli UTI ≥3 days previously (B) patient had not had an E. coli UTI ≥3 days previously (C) E. coli bloodstream infections according to time from previous E. coli UTI

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. See **Table 3** for heterogeneity tests between patients with and without an *E. coli* UTI ≥3 days previously. Results similar restricting to *E. coli* UTIs within the last year or 4 years (data not shown).

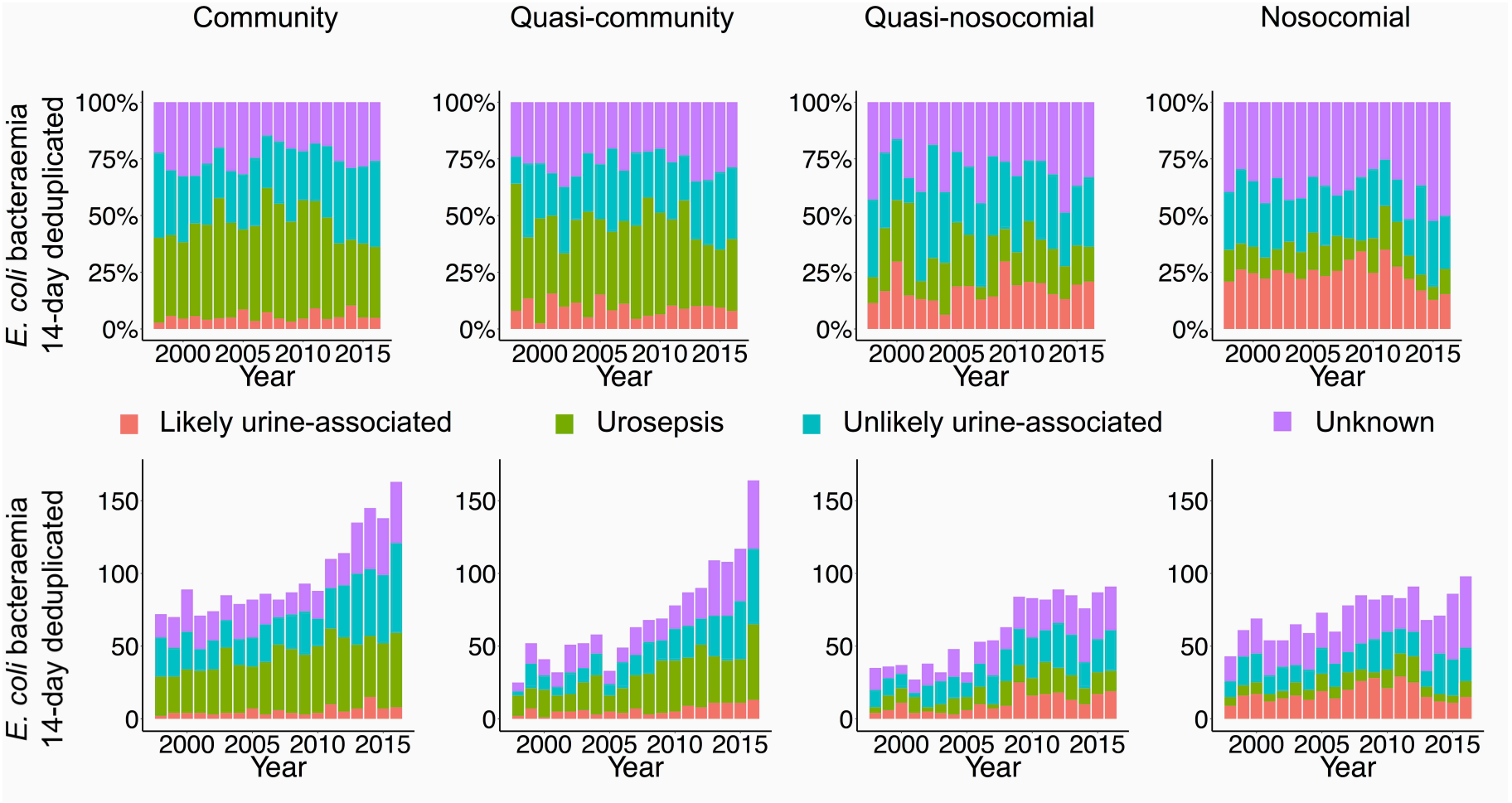


Figure 17 Annual E. coli bloodstream infections according to recent hospital-exposure and urine sample submission/results

Note: 'likely urine-associated' (urine sample taken 3-30 days before *E. coli* bloodstream infection; *E. coli* UTI or mixed growth/negative but UTI suspected clinically from request codes), 'urosepsis' (defined as 'likely urine-associated' but urine samples within (-3,+2] days of the *E. coli* bloodstream infection), 'unlikely urine-associated' (UTI with non-*E. coli* pathogen or no urine sample), or 'unknown' (other).

Next, I explored whether increases in *E. coli* bloodstream infection were associated with a recent indwelling urinary catheter, using catheter urine specimens (CSUs) as a proxy. Percentages of *E. coli* bloodstream infections with a previous CSU increased across hospital-exposure groups, being present in 365 (20%) community, 364 (32%) quasi-community, 541 (40%), quasi-nosocomial, and 584 (43%) nosocomial cases. However, incidence trends were broadly similar for *E. coli* bloodstream infections with or without a previous CSU in the community and quasi-community cases (**Figure 11A**, **Figure 18**, **Table 3**). Quasi-nosocomial *E. coli* bloodstream infections were rising particularly fast in those with a previous CSU (heterogeneity p=0.0002, **Table 3**), while increases in incidence of nosocomial *E. coli* bloodstream infections were restricted to those without a previous CSU (heterogeneity p=0.03, **Table 3**).

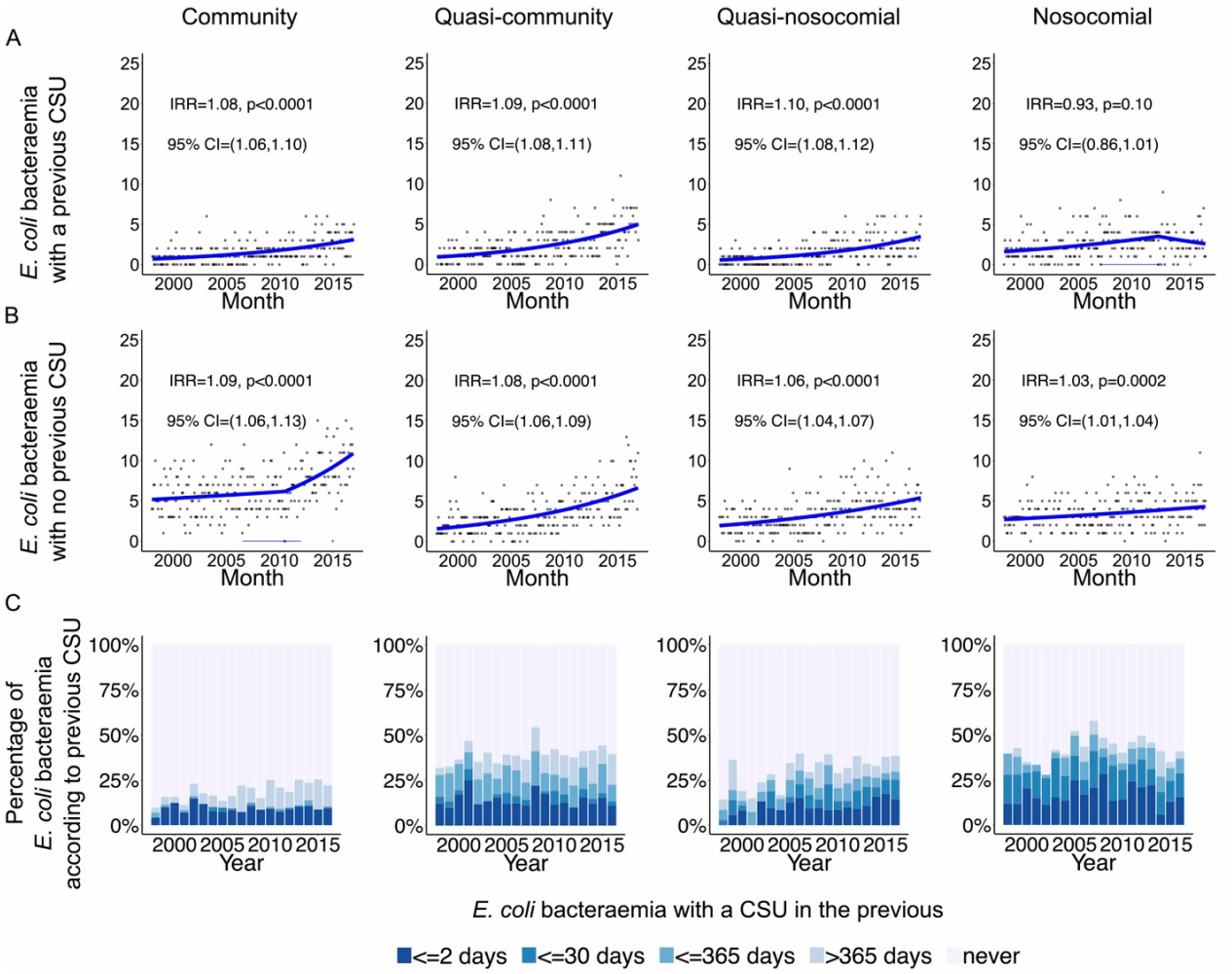


Figure 18 E. coli bloodstream infections according to whether (A) patient had had a catheter urine specimen (CSU) previously (B) patient had not had a CSU previously (C) E. coli bloodstream infections according to time from previous CSU

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. See **Table 3** for heterogeneity tests between patients with and without a CSU previously.

2.3.4 Antimicrobial susceptibility

2.3.4.1 Agreement between susceptibility testing methodologies

There were 2059 *E. coli* bloodstream infections between 31 January 2013, when automated susceptibility testing with the BD Phoenix™, and 31 December 2016 (end of study period). 422 (20%) were tested for co-amoxiclav phenotype by MIC only, 174 (8%) were tested by DD only, with the remaining 1463 (71%) tested by both DD and MIC. Agreement between disk diffusion and Phoenix™ in samples where both were done was reasonable, with a significant increase in yearly percentage of co-amoxiclav resistant *E. coli* bloodstream infections regardless of choice of testing methodology (**Figure 19**), with no evidence that increases in resistance from 2013 were driven by changes in testing methodology. This therefore means that comparisons over time are meaningful, despite the different methodologies.

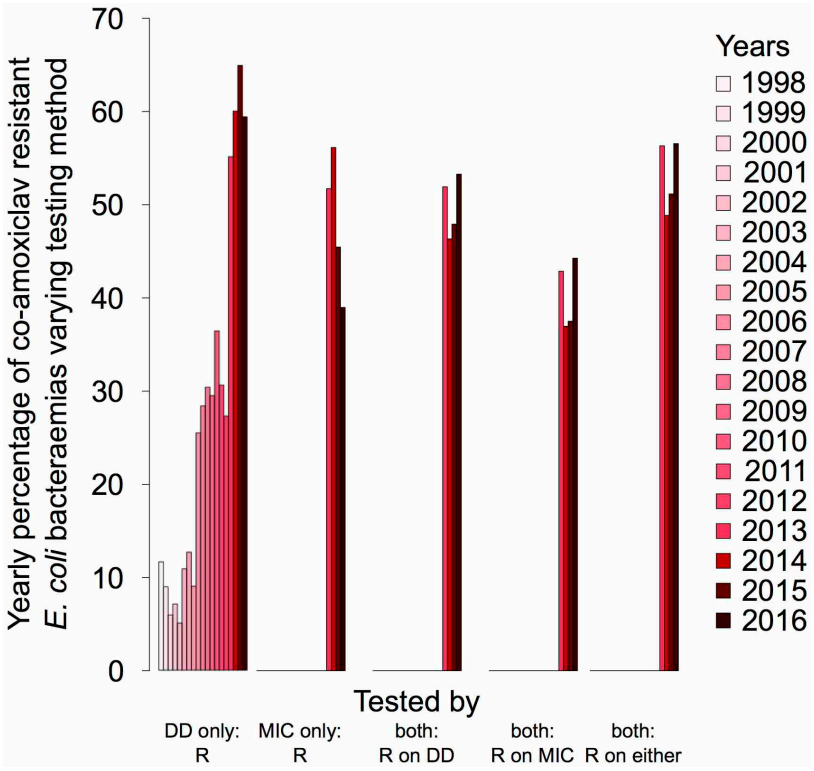


Figure 19 Co-amoxiclav resistance in E. coli bloodstream infections according to different testing methods

Note: DD=disk diffusion. MIC=minimum inhibitory concentration by microbroth dilution (BD Phoenix™).

2.3.4.2 Trends in antibiotic resistant infections

Exploring the possibility that *E. coli* bloodstream infection increases were associated with AMR, the only *E. coli* bloodstream infection antibiotic-resistant phenotype that consistently increased across all hospital-exposure groups was co-amoxiclav ($p<0.0001$; **Figure 11A**, **Figure 20**), with 212 (41%) of 515 *E. coli* bloodstream infections in 2016 being co-amoxiclav resistant (**Table 6, Table 7**). Co-amoxiclav-resistant *E. coli* bloodstream infections increased significantly faster than co-amoxiclav-susceptible *E. coli* bloodstream infections (heterogeneity $p<0.0001$, **Table 3**), but community and quasi-community co-amoxiclav-susceptible *E. coli* bloodstream infections were still increasing significantly in 2016 ($p<0.0001$) (**Figure 20**). Most (942/1412, 67%) co-amoxiclav-resistant *E. coli* bloodstream infections remained susceptible to gentamicin and ciprofloxacin (**Figure 20**).

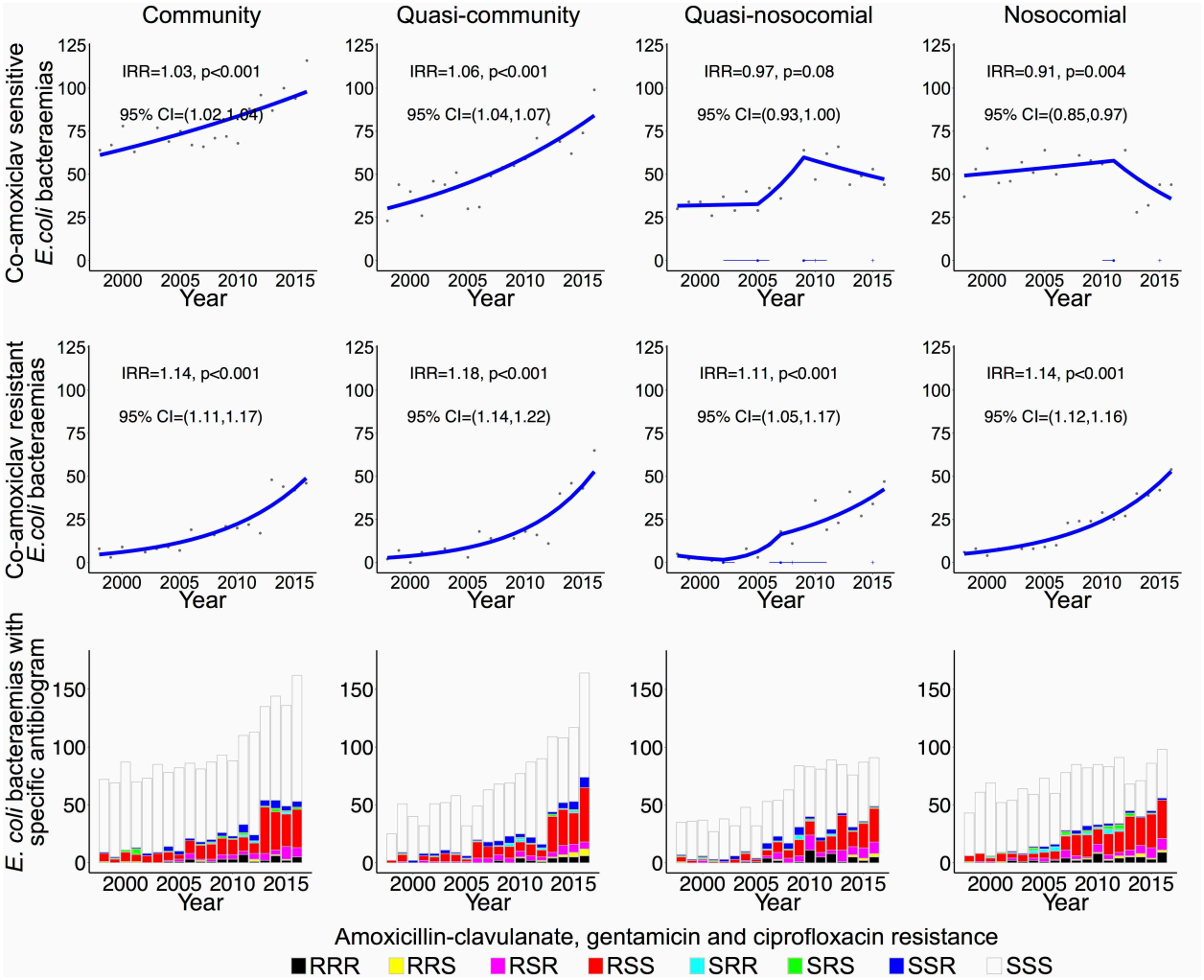


Figure 20 Annual E. coli bloodstream infections that are susceptible and resistant to co-amoxiclav, with and without resistance to gentamicin and ciprofloxacin, according to recent hospital-exposure

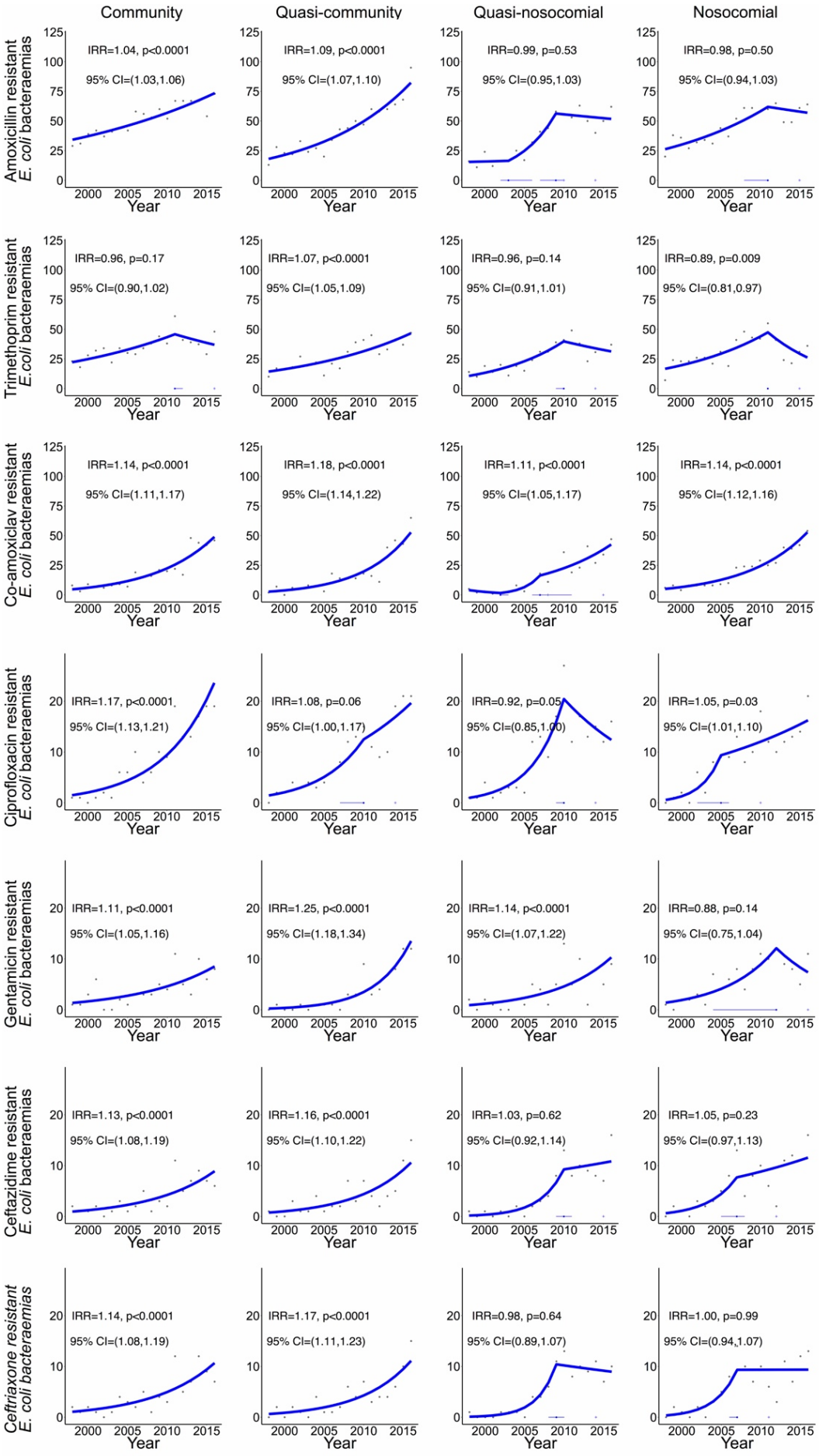
Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016.

Table 7 Total number and percentage of E. coli bloodstream infections and E. coli UTIs tested for each antibiotic and resistant to each antibiotic over the whole period and in 2016

	Tested (%)	Resistant (%)	Tested in 2016 (%)	Resistant in 2016 (%)
***E. coli* bloodstream infections**				
Amoxicillin	5689 (100%)	3357 (50%)	515 (100%)	294 (57%)
Co-amoxiclav	5691 (100%)	1413 (20%)	515 (100%)	212 (41%)
Trimethoprim	5362 (94%)	2230 (35%)	515 (100%)	168 (33%)
Piperacillin-tazobactam	5490 (96%)	434 (7%)	516 (100%)	37 (7%)
Gentamicin	5695 (100%)	327 (5%)	516 (100%)	40 (8%)
Ciprofloxacin	5694 (100%)	672 (10%)	516 (100%)	77 (15%)
Ceftriaxone	5474 (96%)	364 (5%)	516 (100%)	45 (9%)
Ceftazidime	5686 (100%)	352 (5%)	515 (100%)	53 (10%)
Meropenem	5555 (97%)	6 (0%)	516 (100%)	0 (0%)
Amikacin	1003 (18%)	27 (2%)	514 (100%)	12 (2%)
Aztreonam	1703 (30%)	166 (9%)	515 (100%)	54 (10%)
Cefalexin	844 (15%)	211 (22%)	0	0
Cotrimoxazole	1694 (30%)	484 (26%)	512 (99%)	140 (27%)
Ertapenem	2605 (46%)	3 (0%)	515 (100%)	0 (0%)
Fosfomycin	918 (16%)	4 (0%)	512 (99%)	3 (1%)
***E. coli* UTIs**				
Amoxicillin	228183 (100%)	108507 (39%)	13829 (100%)	6329 (46%)
Co-amoxiclav	228054 (100%)	30041 (11%)	13792 (99%)	3921 (28%)
Trimethoprim	228094 (100%)	97281 (35%)	13825 (100%)	4193 (30%)
Piperacillin-tazobactam	59394 (26%)	6098 (8%)	13798 (100%)	366 (3%)
Gentamicin	59917 (26%)	4305 (6%)	13794 (100%)	730 (5%)
Ciprofloxacin	228128 (100%)	14221 (5%)	13826 (100%)	1285 (9%)
Ceftriaxone	55798 (24%)	3830 (6%)	13815 (100%)	720 (5%)
Ceftazidime	59615 (26%)	4098 (6%)	13815 (100%)	683 (5%)
Meropenem	59559 (26%)	103 (0%)	13793 (100%)	6 (0%)
Cefalexin	223197 (98%)	45324 (17%)	13780 (99%)	1932 (14%)
Cotrimoxazole	51033 (22%)	13265 (21%)	13746 (99%)	3552 (26%)
Ertapenem	51837 (23%)	135 (0%)	13787 (99%)	32 (0%)
Fosfomycin	50804 (22%)	499 (1%)	13777 (99%)	90 (1%)
Nitrofurantoin	226236 (99%)	12032 (4%)	13790 (99%)	236 (2%)
Pivmecillinam	28087 (12%)	7514 (22%)	13772 (99%)	1346 (10%)

Increases in other antibiotic-resistant *E. coli* bloodstream infections were most notable in the community and quasi-community groups, with significant year-on-year increments in all but trimethoprim-resistant *E. coli* bloodstream infections, which remained stable in these groups (**Figure 21**).

(A) By antibiotic



(B) summary

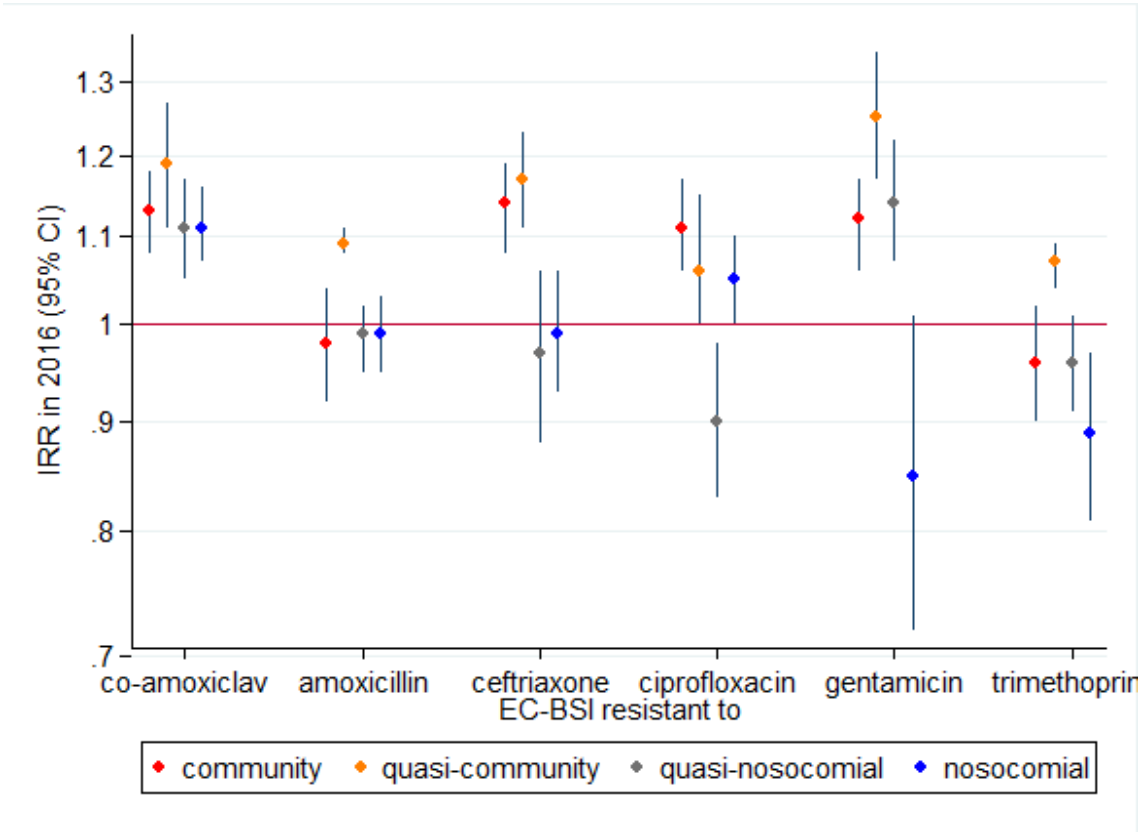


Figure 21 Annual E. coli bloodstream infections according to recent hospital-exposure and antibiotic susceptibility (A) by antibiotic, (B) summary

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. EC-BSI = *E. coli* bloodstream infection.

Co-amoxiclav-resistant *E. coli* UTIs also rose consistently and significantly regardless of healthcare-exposure, but trends were more variable for other antibiotics (**Figure 22**). In 2016, 3921/13792 (28%) *E. coli* UTIs were co-amoxiclav-resistant (**Table 7**).

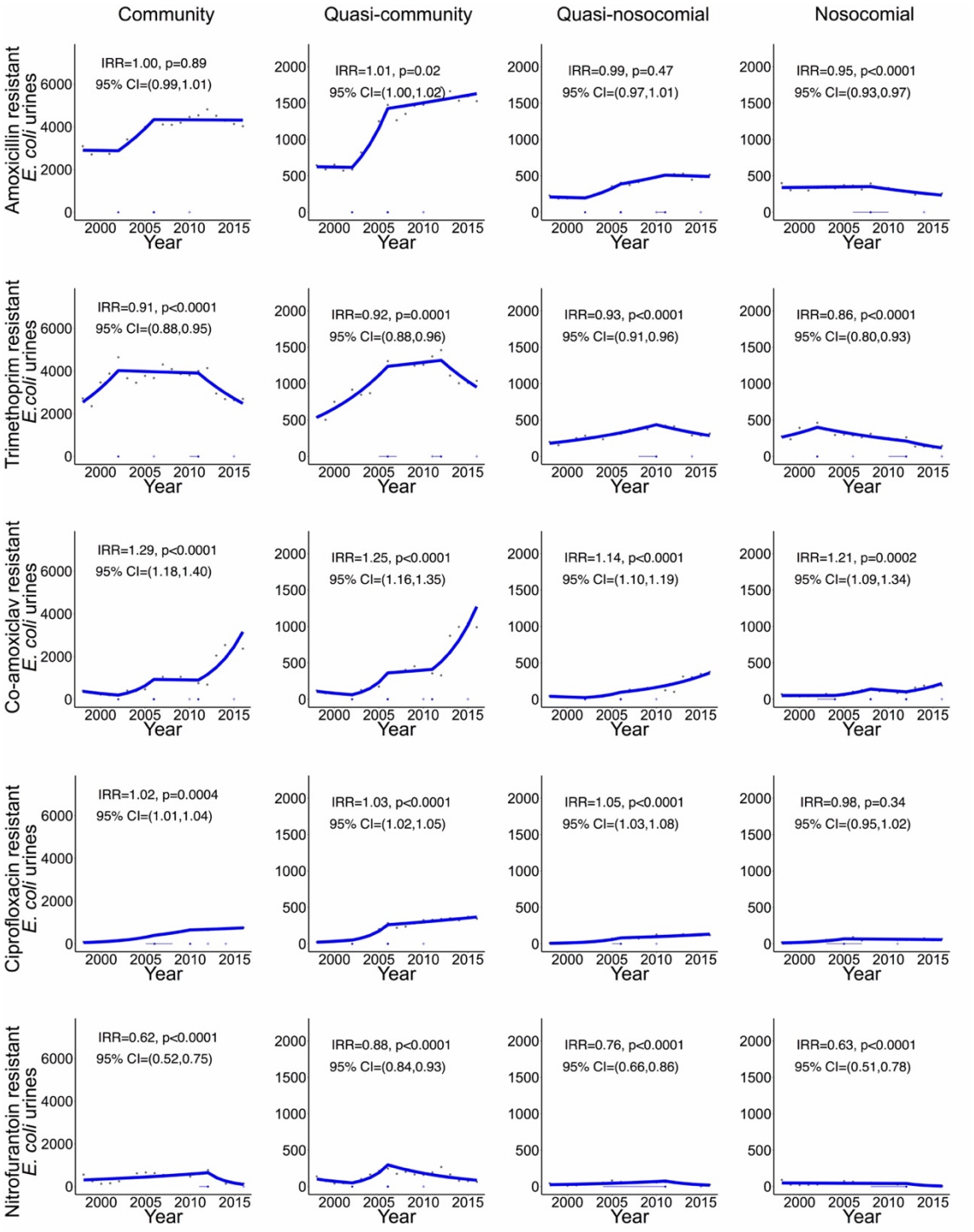


Figure 22 Annual E. coli UTIs according to recent hospital-exposure and antibiotic susceptibility

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016.

2.3.4.3 Severity of co-amoxiclav resistant vs susceptible *E. coli* bloodstream infections

Given the substantial increase in co-amoxiclav resistant *E. coli* bloodstream infections, I investigated whether severity differed in susceptible versus resistant cases. There was no strong evidence that co-amoxiclav-resistant *E. coli* bloodstream infections were associated with higher neutrophil counts in any hospital-exposure group (p>0.04, adjusting for age and sex), or that neutrophil counts were changing differently over time in co-amoxiclav-resistant versus co-amoxiclav-susceptible *E. coli* bloodstream infection (heterogeneity p>0.67; **Figure 23**). Mortality was higher for co-amoxiclav-resistant vs co-amoxiclav-susceptible nosocomial *E. coli* bloodstream infections (unadjusted 30-day mortality 30% (117/395) vs 23% (222/967) respectively; rate ratio adjusting for age and sex=1.32 (1.13-1.46) p=0.002). However, there was no evidence of higher mortality in co-amoxiclav-resistant community/quasi-community/quasi-nosocomial *E. coli* bloodstream infections (p>0.48), and mortality did not change differently over time in any group (heterogeneity p>0.35; **Figure 23**, **Figure 24**, **Table 6**).

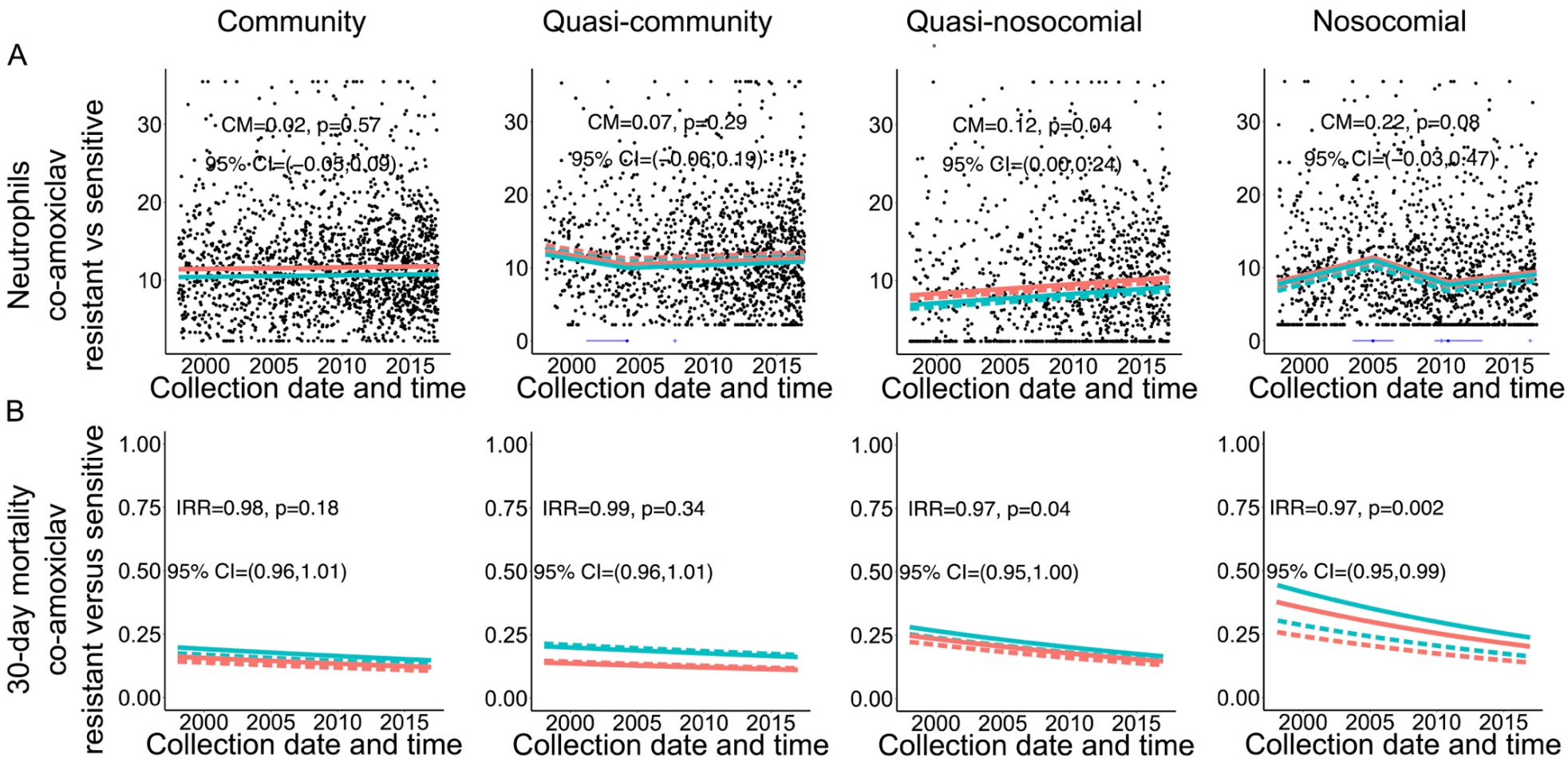


Figure 23 Severity of co-amoxiclav-resistant vs susceptible E. coli bloodstream infections, by (A) neutrophil counts and (B) 30-day mortality across hospital exposure groups

Note: CM=change per year in median value in 2016. Fitted lines are for co-amoxiclav susceptible women (red and dashed), co-amoxiclav susceptible men (blue and dashed), co-amoxiclav resistant women (red and solid), and co-amoxiclav resistant men (blue and solid) at mean age 67 years. IRR=annual rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. Fitted trends in both panels are also adjusted for age and sex.

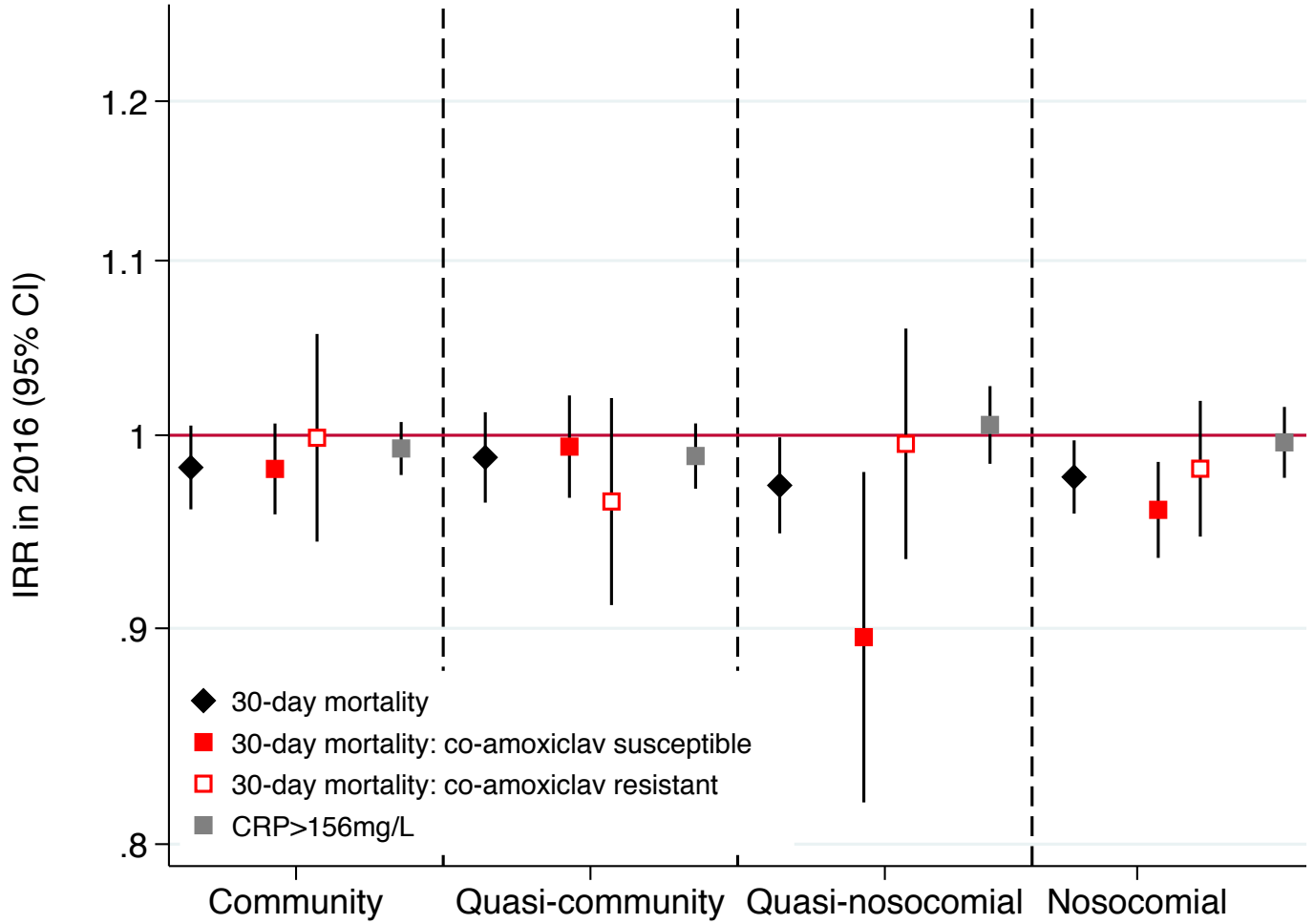


Figure 24 Summary of incidence trends in 2016 for severity of co-amoxiclav resistant and susceptible E. coli bloodstream infections.

Note: IRR=annual incidence rate ratio in 2016, that is the relative increase in rate per year as estimated in 2016. See **Table 3** for numbers and heterogeneity tests.

2.3.4.4 Associations with hospital and community antibiotic use

The strongest associations between any hospital antibiotic use and nosocomial co-amoxiclav-resistant *E. coli* bloodstream infections were with hospital co-amoxiclav (cross-correlation 0.70) and third-generation cephalosporin (0.70) use (**Table 8**; data was only available for financial years 2003 to 2014), with similar results when considering all E. coli bloodstream infections regardless of hospital-exposure (**Table 9**).

Table 8 Cross-correlation between hospital antimicrobial prescribing and co-amoxiclav-resistant nosocomial E. coli bloodstream infection

Antibiotic	Highest absolute cross-correlation [lag (in years) at which this cross-correlation is observed]*
Co-amoxiclav	0.70 [lag 0]
Carbapenem	-0.30 [lag 2]
First generation cephalosporins	-0.41 [lag 1/2]
Second generation cephalosporins	-0.67 [lag 0]
Third generation cephalosporins	0.70 [lag -1/2]
Piptaz	0.53 [lag -1/4]
All cephalosporins	-0.58 [lag 1/4]
Imidazole	-0.58 [lag 1/4]
Lincosamide	0.61 [lag 0]
Macrolide	-0.37 [lag -3]
Beta lactamase resistant penicillins	-0.41 [lag -2 1/4]
Beta lactamase susceptible penicillins	-0.25 [lag 3/4]
Penicillins with extended spectrum	-0.40 [lag 3]
Quinolone	-0.38 [lag -3/4]
Combinations of sulfonamides and trimethoprim, including derivatives	0.33 [lag 2 1/4]

* bivariate cross-correlation between hospital antimicrobial prescribing and nosocomial co-amoxiclav resistant bloodstream infection incidence. For each class of antibiotics, and all antibiotics combined, I considered a time lag of 0 (i.e. same quarter), and all quarters up to -3 and +3, (where -1/4 means antibiotic use in previous quarter against bloodstream infections in current quarter, etc). The cross-correlation of largest absolute magnitude is shown.

Table 9 Cross-correlation between hospital antimicrobial prescribing and co-amoxiclav-resistant E. coli bloodstream infections regardless of hospital-exposure

Antibiotic	Highest absolute cross-correlation [lag (in years) at which this cross-correlation is observed]*
Co-amoxiclav	0.75 [lag 0]
Carbapenem	0.28 [lag 0]
First generation cephalosporins	-0.44 [lag 3/4]
Second generation cephalosporins	-0.71 [lag 0]
Third generation cephalosporins	0.80 [lag 0]
Piptaz	0.62 [lag 0]
All cephalosporins	-0.59 [lag 1/4]
Imidazole	-0.51 [lag 1/4]
Lincosamide	0.69 [lag 0]
Macrolide	-0.31 [lag -3]
Beta lactamase resistant penicillins	-0.49 [lag -2 1/4]
Beta lactamase susceptible penicillins	-0.28 [lag 1 1/4]
Penicillins with extended spectrum	-0.54 [lag 1/4]
Quinolone	-0.45 [lag -2 1/2]
Combinations of sulfonamides and trimethoprim, including derivatives	0.35 [lag 3 1/4]

* bivariate cross-correlation between hospital antimicrobial prescribing and nosocomial co-amoxiclav resistant bloodstream infection incidence. For each class of antibiotics, and all antibiotics combined, I considered a time lag of 0 (i.e. same quarter), and all quarters up to -3 and +3, (where -1/4 means antibiotic use in previous quarter against bloodstream infections in current quarter, etc). The cross-correlation of largest absolute magnitude is shown.

Community prescribing data was only available from 2011, and co-amoxiclav-resistant *E. coli* bloodstream infections were too few to consider relationships with co-amoxiclav use across individual general practices. I therefore considered whether co-amoxiclav resistant *E. coli* UTIs, *E. coli* UTIs (regardless of co-amoxiclav susceptibility), and UTIs (regardless of whether *E. coli* positive) were associated with previous, current or both previous and current co-

amoxiclav usage according to the general practice in which the case was registered and the strength of these associations.

From 2012-2016, general practices prescribing more co-amoxiclav in the previous year had higher rates of subsequent co-amoxiclav-resistant community-*E. coli* UTIs (IRR per 100 DDD higher=1.05 [95% CI 1.02-1.08], p=0.003, **Figure 25**). For the outcome co-amoxiclav-resistant *E. coli* UTI, the best predictor was usage in the previous year rather than usage in the current year and there was no evidence of interactions with the calendar time trend (p=0.22). Co-amoxiclav use in the current year did not add any predictive value (p=0.12 adjusted for previous year vs p=0.64 alone).

In contrast, practices prescribing more co-amoxiclav in the current year had higher rates of community-*E. coli* UTIs overall (regardless of resistance phenotype) and urine specimen submission (IRR=1.02 [1.00, 1.04], p=0.01 and 1.02 [1.01, 1.03], p=0.0001 respectively). For *E. coli* UTIs and all urines, usage in the current year was the better predictor rather than usage in the previous year and there was no evidence of interactions with the calendar time trend (p=0.55). Here co-amoxiclav use in the previous year did not add any predictive value (p=0.58 adjusted for current year vs p=0.11 alone, respectively p=0.21 adjusted for current year vs p=0.006 alone).

The same models were chosen when including all samples regardless of hospital-exposure group and similar results were seen (**Figure 26**). This was also the case when adjusting instead for the proportion aged over 65 and male in 2017 per practice. There was no association between current/prior quinolone use and any of these outcomes (p>0.9).

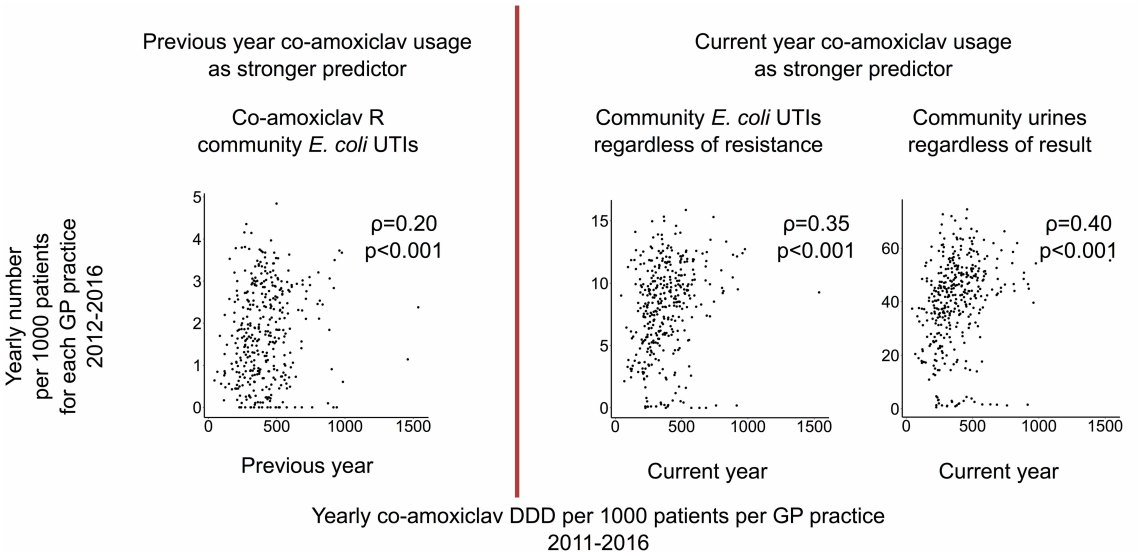


Figure 25 Community co-amoxiclav-resistant E. coli UTIs (A), community E. coli UTIs (B) and community urine samples submitted regardless of result (C) per 1000 patients per general practice 2012-2016 compared with co-amoxiclav usage

Note: showing one point per year per general practice**, restricting to community cases only**. For (A) the strongest predictor was co-amoxiclav DDD per 1000 patients per general practice in the previous year; for (B) and (C) the strongest predictor was co-amoxiclav DDD per 1000 patients per general practice in the current year. Spearman rho (and models) for each panel excludes 5 practices which submitted less than 151 samples over 2011-2016 (all others submitted over 300). Spearman rho for univariable associations with previous vs current co-amoxiclav usage for the 3 outcomes left to right ρ=0.20 vs ρ=0.04, ρ=0.33 vs ρ=0.35, ρ=0.37 vs ρ=0.40 respectively.

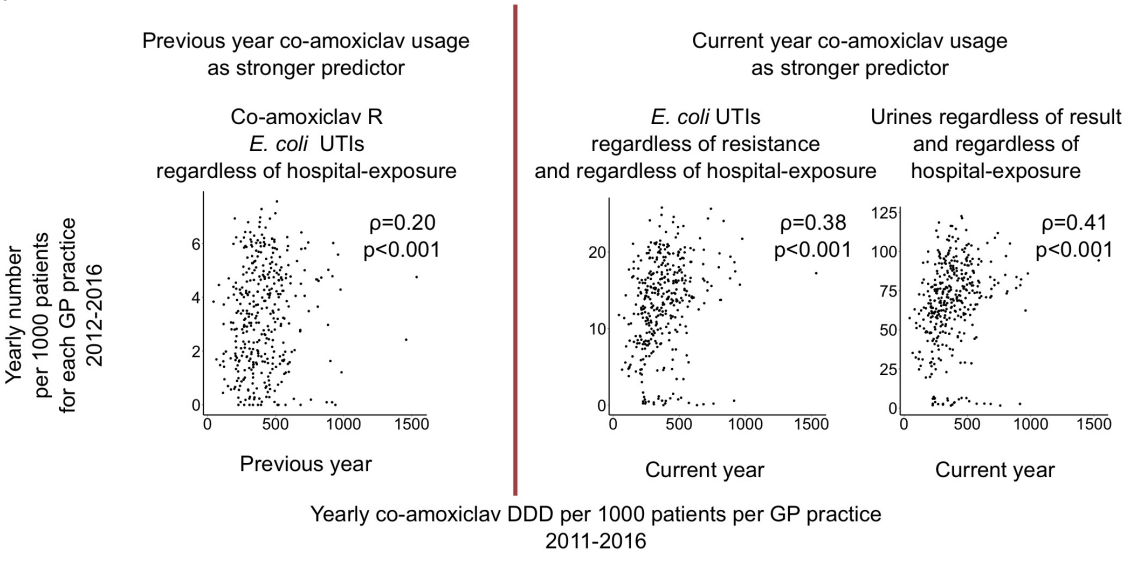


Figure 26 All co-amoxiclav-resistant E. coli UTIs regardless of hospital-exposure (A), E. coli UTIs (B) and urine samples submitted regardless of result (C), per 1000 patients per general practice 2012-2016 compared with co-amoxiclav usage

Note: showing one record per year per general practice, **all cases**. For (A) the strongest predictor was co-amoxiclav DDD per 1000 patients per general practice in the previous year; for (B) and (C) the strongest predictor was co-amoxiclav DDD per 1000 patients per general practice in the current year. Spearman rho (and models) for each panel excludes 5 practices which submitted less than 151 samples over 2011-2016 (all others submitted over 300). Spearman rho univariable associations with previous vs current co-amoxiclav usage for the 3 outcomes left to right ρ=0.2 vs ρ=0.05, ρ=0.36 vs ρ=0.38, ρ=0.38 vs ρ=0.41 respectively.

2.4 Discussion

In this Chapter I have explored potential explanations for continuing increases in *E. coli* bloodstream infections in Oxfordshire over 19 years using extensive, routinely-collected data, including linked laboratory/microbiology results. I found that incidence varied dramatically by hospital-exposure, with overall increases being driven by increases in community/quasi-community cases. This is important given the National Health Service ambition to reduce Gram-negative bloodstream infections by targeting 'healthcare-associated' cases (although this definition incorporates recent community antibiotic use (Public Health England and NHS Improvement 2017)). Previous successful campaigns to reduce methicillin-resistant *Staphylococcus aureus* (MRSA) bloodstream infection and *Clostridium difficile* infections in the UK also focussed on nosocomial risk factors. My findings suggest that defining appropriate strategies aiming to reduce community/quasi-community associated *E. coli* bloodstream infections, such as improved catheter care and improved quality of antibiotic use for UTI management in the community, might have a greater impact in this infection.

In a previous study looking at *E. coli* bloodstream infections occurring in Oxfordshire between 1999 and 2011, cases were split between hospital-associated – defined as having their *E. coli* positive blood culture taken more than 48 hours after being admitted or having been in the hospital in the previous year – and hospital-naïve. They found an increase in the incidence of both groups throughout the decade. One of the strengths of my study was that by exploring my data, rather than arbitrarily categorising groups of data, I was able to show that rates were increasing fastest the further the most recent hospital exposure was. The challenges of grouping data can also be seen in their analysis when they considered

resistance to any of co-amoxiclav, gentamicin, ciprofloxacin and cefotaxime as defining "resistance". This can be problematic given different changes in the prescription of different antibiotics, which are presumed to be the underlying drivers of changes in phenotypes, although arguably is relevant to considering treatment of individual patients.

Co-amoxiclav is one of the most commonly prescribed antibiotics nationally in both the community and hospitals in England (Public Health England 2017; Ashiru-Oredope et al. 2012), and my findings are consistent with this exerting selection pressure for co-amoxiclav resistant *E. coli* UTI and *E. coli* bloodstream infection. Specifically, co-amoxiclav-resistant *E. coli* bloodstream infections rose significantly faster than co-amoxiclav-susceptible *E. coli* bloodstream infections, regardless of hospital-exposure, with the greatest number of co-amoxiclav-resistant *E. coli* bloodstream infections in 2016 being community/quasi-community *E. coli* bloodstream infections. Despite co-amoxiclav being used for empiric bloodstream infection treatment, there were no clinically important changes in mortality. General practices with higher co-amoxiclav prescribing rates in the previous year had more co-amoxiclav resistant *E. coli* UTIs in the subsequent year. Given that recent antibiotic use is the greatest risk factor for subsequent resistant UTIs (Hillier et al. 2007), many people prescribed antibiotics may not have bacterial UTIs (Butler et al. 2017), and many bacterial UTIs may resolve in a similar timeframe without antibiotics (Gágyor et al. 2015), better point-of-care tests that predict benefit from antibiotics may be important to guide decisions to prescribe antibiotics for a suspected UTI or not.

E. coli bloodstream infection is generally considered 'community-acquired' although there are differing definitions of healthcare-associated bloodstream infection (Public Health

England 2016a; Public Health England and NHS Improvement 2017). By linking to previous hospital admissions, one major strength of my study is that I could identify that incidence trends for non-nosocomial *E. coli* bloodstream infections varied significantly by time since discharge. However, an ideal future classification would incorporate information on healthcare-exposure in a hospital, residential/nursing home, but also in the patient's own home. In the present healthcare system, in Oxfordshire, a large number of patients discharged from acute care will go to beds in other healthcare environments (such as intermediate care and community hospitals). This information is however not currently recorded in the electronic health records. Blood sample submission also increased significantly over time, potentially increasing ascertainment of 'mild' cases. However, blood cultures are key to the assessment of unwell patients whenever infection is suspected, and there were no clinically important changes in *E. coli* bloodstream infection-associated severity at presentation (as reflected by results of common blood tests) or mortality, despite substantially increasing incidence, suggesting major ascertainment bias is unlikely. As standardising for age/sex using crude data available had at most modest effects, I did not use this for main analyses.

An ageing population explained between 10-26% of the increase, however the smallest percentage explained was in the community. Even though a high percentage of *E. coli* bloodstream infections had a urinary tract focus, a similarly high percentage have an unknown focus based on the electronic health records available. One interesting hypothesis is that periodontal disease and root canals could represent the focus of a large number of infections. 17% of individuals aged 65 years and over have periodontal disease (National Institute of Dental and Craniofacial Research 2020). Treatment can last up to 2-3 years and

represents both an ideal route for the bacteria to enter the bloodstream, but also for selection of antibiotic resistant pathogens with antibiotics being given both as long-term treatment, but also as prophylaxis during surgery. Unfortunately, over the study period there was no data available on antibiotic use specifically in dentistry.

The increasing trend in nosocomial *E. coli* bloodstream infection incidence was significantly smaller than for community/quasi-community *E. coli* bloodstream infection in Oxfordshire, as observed nationally (Gerver et al. 2014). Multiple infection control interventions were rolled out in UK hospitals from 2005-2010 (David H Wyllie et al. 2011; Duerden et al. 2015) in response to MRSA and *C. difficile*, and some components of these initiatives, particularly relating to general infection prevention and control, could have contributed to lowering nosocomial rates. Consistent with this, increases in hospital-onset Gram-negative bloodstream infection reversed after a MRSA Prevention Initiative was introduced in the US, while community-acquired incidence remained unchanged (Goto et al. 2016).

Whereas MRSA and *C. difficile* are predominantly hospital-associated pathogens, differences in *E. coli* bloodstream infection epidemiology highlight the need for different interventions, particularly in primary care (Public Health England 2016a). In particular, recurrences explain relatively little of the ongoing increases in *E. coli* bloodstream infections. Overall, 42% of *E. coli* bloodstream infection appeared to be more likely amenable to urinary-focussed intervention. Similarly an England-wide study found 51% of *E. coli* bloodstream infections had an underlying urogenital tract focus, with UTI treatment in the prior four weeks the largest independent risk factor (Abernethy et al. 2017). In my study, 13% of *E. coli* bloodstream infections were likely urine-associated and 28% presented as urosepsis; the

first group may be most tractable for prevention but was smallest in both community and quasi-community *E. coli* bloodstream infection, whereas urosepsis was the largest.

One key limitation is lack of data on visits to general practices, meaning my assessment of urinary sources relied on urine samples being submitted for microbiological testing. I may therefore have underestimated the true burden of urinary-associated *E. coli* bloodstream infection, since some patients may have had UTI symptoms and either did not present to their general practitioner or were treated empirically without a urine sample being submitted. Changes in patient health-seeking behaviour or urine sample submission over time could lead to bias in attributing *E. coli* bloodstream infections to different sources. However, successfully treated UTIs, and those resolving without intervention, should not cause bloodstream infection, and bloodstream infections due to UTI treatment failure should be ascertained within my data since guidelines recommend urine samples be submitted from individuals with clinical treatment failure, frequent or recurrent UTI or with a possibly resistant infection (Public Health England 2017).

Much of the burden of *E. coli* bloodstream infections, and especially the rising incidence, is hypothesized to arise from poor urinary catheter care. One key limitation is that I did not have records of the presence of a catheter, but only urine specimens recorded as being taken from a catheter (denoted CSU). I argue that if a catheter was present and causing infection, a specimen would likely have been taken from it at some time. However, only 20% and 30% of the community and quasi-community groups, where incidence is increasing fastest, had a previous CSU, and there was no evidence that incidence was increasing faster in those with a previous CSU versus without.

Similarly, I did not have direct information on the source of each individual *E. coli* bloodstream infection since this is not recorded in routinely available electronic health record data. Interestingly, there was strong evidence that quasi-nosocomial *E. coli* bloodstream infections with UTI or infectious diagnostic codes in the previous admission were rising faster than those without (**Figure 6**). This may reflect underlying predisposition to infection (e.g. chronic illnesses), or that prior antibiotic use adversely affects a patient's microbiota potentially leading to colonisation/overgrowth by more pathogenic *E. coli*, thus predisposing to *E. coli* bloodstream infection.

A limitation of surveillance studies is changes in antimicrobial susceptibility testing methodology (here in February 2013). For example, Leverstein-van Hall et al found a significant increase in the percentage of co-amoxiclav resistant *E. coli* bloodstream infections between 2010 and 2011 from 19% to 31% ($p<0.0001$) only in laboratories which switched from a fixed ratio of amoxicillin/clavulante of 2/1 (CLSI) to a fixed concentration of clavulanate of 2mg/L (EUCAST) (Leverstein-van Hall et al. 2013). Diez-Aguilar et al found discrepancies between results obtained using CLSI versus EUCAST and concluded that further studies are needed in order to investigate which of the two *in vitro* concentrations of clavulanate correlate better with clinical outcome (Díez-Aguilar et al. 2015). Whilst clearly testing protocol can affect results, crucially I showed that changes in co-amoxiclav resistance around this time occurred regardless of method (**Figure 19**). Recent data suggest that broth dilution (BD-Phoenix™) and the gold standard agar dilution have high agreement (T. J. Davies et al. 2019); thus, rising rates of co-amoxiclav-resistant (as defined by EUCAST breakpoints) *E. coli* bloodstream infection/UTI are likely correct.

I also found that general practices with higher co-amoxiclav prescribing rates in the previous year were more likely to have patients diagnosed with co-amoxiclav resistant *E. coli* UTIs in the subsequent year. Similar associations between trimethoprim use and trimethoprim-resistant urine-associated *E. coli* bloodstream infection have been reported in adult women in England (Lishman et al. 2016), and recently more generally across multiple antibiotic classes for *E. coli* UTI (Ironmonger et al. 2018; Cai et al. 2015). Over the period with contemporary prescribing data, co-amoxiclav-resistant *E. coli* bloodstream infections were too few to investigate associations with general practice level antibiotic prescribing within the community. Assessing usage-resistance associations is complicated, since changes in use of one antibiotic are generally accompanied by compensatory prescribing of other antibiotics, and may be compounded by multi-drug resistance.

The key limitation of my study is that comparisons are ecological. My results may also not be generalizable; for example, although the region I studied is sizeable (~1% of the UK), I did not observe a uniform decrease in cephalosporin-resistant and quinolone-resistant *E. coli* bloodstream infections as seen in bloodstream infections caused by *Enterobacteriaceae* across the UK (Livermore et al. 2013). Such differences likely reflect a complex interplay of selection pressures. Another major limitation is that I was unable to assess associations between individual-patient antibiotic use (not available in the research database) and risk of resistant infections or between specific empiric regimens and outcome; these are important future research priorities. However, there were no clinically important changes in mortality overall, by co-amoxiclav-susceptible/resistant phenotype, or by hospital-exposure. Co-amoxiclav remains the recommended first-line empiric treatment for most severe infections

in many hospitals in the UK, so the substantial increase in incidence of co-amoxiclav-resistant bloodstream infections without parallel increases in mortality suggests either that initial inappropriate treatment can be successfully rescued (Fitzpatrick et al. 2016), or that the current definition of co-amoxiclav breakpoints may be suboptimal (European Committee on Antimicrobial Susceptability Testing 2016). Crucially, neither scenario supports a move towards broader empiric antibiotic treatment, consistent with prevailing antimicrobial stewardship messages.

Another inherent limitation is restriction to the routinely collected data available, in particular lack of information on prognostic factors such as illness severity scores, focus of infection, lack of individual antibiotic prescribing data (as above), and having only antibiograms since strain typing was not routinely performed. Representative isolates from 2008-to date have now been whole genome sequenced and their analysis may increase our understanding of the pathogenesis of *E. coli* bloodstream infection. For example, while the distinct change in the monthly incidence of community *E. coli* bloodstream infection in July 2010 might have reflected the introduction and proliferation of a new strain of *E. coli* to the region (Kallonen et al. 2017), there is no evidence of major shifts in strain distribution over the last decade (personal communication, Sam Lipworth).

In summary, on-going increases in *E. coli* bloodstream infection across Oxfordshire over the last 20 years were driven by community and quasi-community cases, and cannot be attributed only to increased recurrences or an aging population. Absence of changes in mortality and severity do not support ascertainment bias playing a major role, although this cannot be completely excluded. Whilst urinary foci are clearly important, at present the

scope for intervening to prevent UTIs progressing to bloodstream infection could be limited. Notably, higher co-amoxiclav use in primary care was associated with higher subsequent rates of co-amoxiclav-resistant *E. coli* UTI, supporting drives to reduce broad-spectrum and inappropriate antibiotic use. However, despite substantial increases in co-amoxiclav-resistant *E. coli* bloodstream infection, evidence that patient clinical outcomes are no worse with these resistant infections does not support broadening empiric antibiotic prescribing from co-amoxiclav (Public Health England 2015c).

3 Predicting antimicrobial resistance in *E. coli* bloodstream infections using machine learning methods and bacterial genome-wide association studies

3.1 Introduction

As explored in the previous Chapter, rates of *E. coli* bloodstream infections have been rising nationally (Public Health England 2019) and regionally (Vihta et al. 2018) and are a major concern in terms of antimicrobial resistance (O'Neill 2014). Over time, there have been multiple changes in susceptibility testing methods (as described in the previous Chapter) and different testing methods are being used across different laboratories. In contrast to phenotypic methods, determining the presence of antimicrobial resistance genomically may be more reproducible, and potentially also more rapid if sequencing can be undertaken without requiring bacterial culture. With whole genome sequencing (WGS) becoming cheaper, using it for routine antimicrobial susceptibility testing is potentially coming within reach.

However, the main challenges in predicting susceptibility from WGS are the very high number of genetic features to be considered together with the constant evolution of the bacteria, particularly under the selective pressure exercised through the use of antibiotics. The problem of predicting antimicrobial resistance from WGS ultimately reduces to identifying patterns within the data and is therefore the ideal setting for machine learning. Machine learning is the scientific study of algorithms and statistical models through which a computer system can learn from training data in order to make predictions on a new set of samples without being explicitly programmed for the task.

In the previous Chapter, I showed that co-amoxiclav resistance in particular has been increasing in Oxfordshire *E. coli* bloodstream infections, with 41% resistance rates in 2016. Co-amoxiclav is still the first-line empiric antibiotic treatment for common infections in the Oxford University Hospitals NHS Foundation Trust (OUH), and many hospitals across England. There is also evidence that different *in vitro* susceptibility methods disagree and it is still unclear which method correlates best with clinical outcome, which is the ultimate goal of susceptibility testing (Díez-Aguilar et al. 2015). The focus of this Chapter is therefore to predict co-amoxiclav resistance from WGS through machine learning algorithms, on the grounds that WGS may ultimately be a more reliable and reproducible way to identify resistance.

3.2 Methods

3.2.1 Study design, population and laboratory phenotype

From IORD, I included all pure and mixed (polymicrobial) *E. coli* bloodstream infections isolated between 1 January 2013, and 31 August 2015, with 90-day deduplication following each index positive culture, i.e. removing repeat blood cultures within 90 days from the same individual. The reason for a longer period of de-duplication in this Chapter compared with Chapter 2 is because here I am aiming to predict phenotypic resistance from WGS data – if a patient has a re-infection or recurrence with the same strain of bacteria, although this is important clinically (Chapter 2), it will have the same phenotype and very similar WGS, and will therefore not really be an independent observation. In this study, I included all isolates from blood cultures with high quality WGS data available from a related study sequencing all *E. coli* blood culture isolates at OUH (n=976) (T. J. Davies et al. 2019). Isolates were only included after automated antimicrobial susceptibility testing (AST) was used in

the OUH routine microbiology laboratory using micro-broth dilution (BD Phoenix™; Beckton, Dickinson and Company) (January 2013). As in the previous Chapter, MICs were interpreted using European Committee on Antimicrobial Susceptibility Testing (EUCAST) breakpoints (European Committee on Antimicrobial Susceptibility Testing 2019). Original MIC reports were censored at both the lower and upper ends, to ≤2/2 mg/L and >8/2 mg/L respectively. Given the right-censoring point was only one doubling dilution above the EUCAST clinical breakpoint (8/2mg/L), all tests originally recorded as >8/2 mg/L were repeated with an updated expanded BD Phoenix™ panel to enable reporting of higher MICs (up to >32/2 mg/L).

3.2.2 DNA extraction and sequencing

Sequencing was undertaken by colleagues from the Modernising Medical Microbiology consortium as follows. Frozen stocks of *E. coli* isolates from blood cultures were stored in nutrient broth plus 10% glycerol at -80^{0}C. Isolates were re-cultured from these frozen stocks and DNA was extracted using the QuickGene DNA Tissue Kit S (Kurabo Industries, Japan) including a chemical lysis step immediately followed by a mechanical lysis step (FastPrep, MP Biomedicals, USA). A combination of standard Illumina and in-house protocols were used to produce multiplexed paired-end libraries which were sequenced on the Illumina HiSeq 2500, generating 151bp paired-end reads. High quality sequences were *de novo* assembled using Velvet (Zerbino and Birney 2008) as previously described (Stoesser et al. 2017; T. J. Davies et al. 2019). *In silico* multi-locus sequence types (MLST) types were defined using ARIBA (Hunt et al. 2017) and the Achtman (Wirth et al. 2006) MLST scheme.

3.2.3 Predicting phenotype using WGS and machine learning algorithms

My primary objective was to predict co-amoxiclav susceptibility using information from WGS and a variety of machine learning algorithms. There were therefore three main avenues to explore:

- input features (which data items I extract from the WGS data as a whole),
- output features (how I define susceptibility, the outcome of interest which I am trying to predict), and
- machine learning algorithms (appropriate for the types of input and output features, exploring from simple to more complicated algorithms, as, depending on the intricacy of the relationships within my data, simple models could perform well or poorly on the given task).

3.2.3.1 Input features

I explored various ways of extracting genetic information for my predictions which can be broadly classified into two groups: mapping WGS onto resistance databases to identify presence of various known resistance-conferring elements and k-mer-based approaches, which do not rely on prior knowledge of the underlying resistance mechanisms.

3.2.3.1.1 Presence and absence of known co-amoxiclav resistance determinants

Sequencing data for each isolate was interrogated by Tim Davies (T. J. Davies et al. 2019) using ARIBA (Hunt et al. 2017) (**Figure 27**) (using default settings) to identify presence/absence of beta-lactamases from the Resfinder database (Zankari et al. 2012) (2000 resistance gene alleles), with the output being the closest matching gene allele from the database and then variation from that reference sequence (e.g. *bla*$_{\text{TEM-156}}$ as the closest

matching allele, and variant mutations such as Ile153Thr from $bla_{\text{TEM-156}}$). Additional searches were also done for the *ampC* promoter sequence ATCC 25922, Lartigue *tem* promoter sequence P3 and porin genes from the reference *E. coli* K-12. Features were deemed present if length adjusted assembly coverage was ≥10, where:

$$length\ adjusted\ assembly\ coverage\ := \frac{\sum contig\ coverage \times contig\ length}{feature\ reference\ length} \geq 10.$$

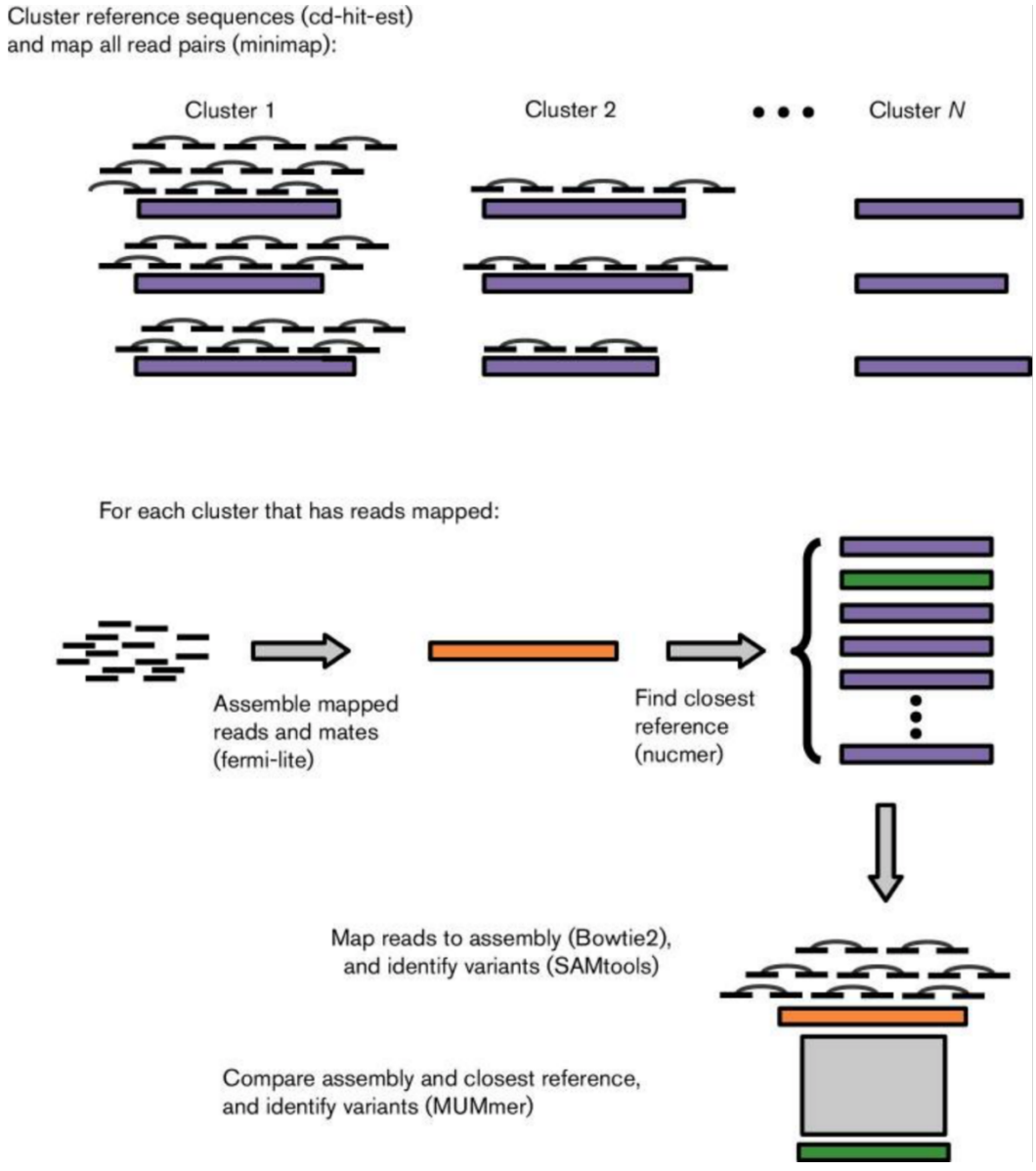


Figure 27 Overview of ARIBA mapping and targeted assembly pipeline

Note: Source: (Hunt et al. 2017). In brief: reference sequences in the Resfinder database are clustered together; reads are mapped onto these clusters; for each cluster that has mapped reads create a local assembly of the reads, called a contig; the closest reference sequence in the Resfinder database is found for this contig; then contig coverage = read depth at each contig position.

In order to create a matrix of features (with a row for each isolate and columns denoting the presence and absence of each feature as 1 or 0 respectively) I started with the closest matching reference name, e.g. ampC_promoter_1_NA together with the variant mutations/insertions/deletions for that allele, e.g. C:79:T | G:124:A | A:142:- | A:143:- | T:144:- | C:145:- | T:146:- | A:147:- | A:148:- | C:149:- | G:150:- | C:151:- | A:152:- | C:168:T | G:230:- | G:233:A | A:245:-. A, C, G and T represent the four types of bases found in DNA: adenine (A), cytosine (C), guanine (G), and thymine (T). C:79:T represents a mutation of C to T at position 79 in the *ampC* promoter sequence (GATCGTTCTGCCGCTGTGGTGTGGTTTACACCGTATGCACCACGCGATGCACGATCTGAAAATCCAC GTACCTGCGGGCAAATGGGTTTTCTACGGTCTGGCTGCTATCCTGACAGTTGTCACGCTGATTGGTG TCGTTACAATCTAACGCATCGCCAATGTAAATCCGGCCCGCCTATGGCGGGCCGTTTTGTATGGAAA CCAGACCCTATGTTCAAAACGACGCTCTGGCGCCTTATTAATTAACCGCCTCTTGCTCCACATTTGCT GCCC). Traditional nomenclature calls position 152 position 0, so 79 would be reported as -73 *ampC* promoter mutation, 124 as -28 *ampC* promoter mutation etc. Dashes represent a deletion if on the right, or an insertion if on the left. I then created a combination of the closest matching reference name and every mutation/insertion/deletion observed in my dataset: so for the example above, ampC_promoter_1_NA_C:79:T | ampC_promoter_1_NA_G:124:A|ampC_promoter_1_NA_A:142:- | ampC_promoter_1_NA_A:143:-|ampC_promoter_1_NA_T:144:- | ampC_promoter_1_NA_C:145:-|ampC_promoter_1_NA_T:146:- | ampC_promoter_1_NA_A:147:-|ampC_promoter_1_NA_A:148:- | ampC_promoter_1_NA_C:149:-|ampC_promoter_1_NA_G:150:- | ampC_promoter_1_NA_C:151:-|ampC_promoter_1_NA_A:152:- | ampC_promoter_1_NA_C:168:T|ampC_promoter_1_NA_G:230:- |

ampC_promoter_1_NA_G:233:A|ampC_promoter_1_NA_A:245:-. I repeated this for every single element present in at least one isolate (beta-lactamases, *ampC* promoter sequence and Lartigue *tem* promoter sequence P3) and their mutations/insertions/deletions and then selected all unique element-mutation/insertion/deletion combinations. I constructed a matrix of presence/absence of these unique element-mutation/insertion/deletion combinations, including the possibility of having the element without any variations as a feature, if this feature was present without any variation in at least one isolate. So if an allele had two mutations, e.g. ampC_promoter_1_NA_C:79:T, ampC_promoter_1_NA_G:124:A, these two would contribute two columns to my matrix of features with entries of 0 for those isolates in which they were absent and 1 for those in which they were present. For the porins *ompF* and *ompC*, mutations/insertions/deletions were not considered, but only presence/absence of a complete gene (ARIBA flag set to 1 if a match exists for the full length of the reference sequence or if the extension to the nearest start and stop codons has the only stop codon at the end of the inferred sequence, and 0 otherwise).

To clean-up the feature list for the machine-learning algorithm, I removed all features which were present in all samples as these will not contain any predictive information (in this case presence of the *ampC* promoter sequence); and set all which were present in more than half the samples to 0 if present (and 1 if absent) in order for all features to be sparse to improve algorithm performance (in this case porins *ompF*, *ompC, ampC* promoter sequence mutations -73 C>T, 81 G>A, and *ampC* promoter sequence deletions 78DelG, 93DelA). The final matrix of presence and absence of known co-amoxiclav resistance determinants consisted of 169 binary features.

3.2.3.1.2 Adding relative coverage

Increases in gene copy number can be a mechanism for resistance, as well as presence-absence/mutation/insertion/deletion. Therefore, relative coverage was estimated from the ratio of gene coverage vs chromosomal coverage, with chromosomal coverage estimated from the mean coverage of the highly conserved MLST genes, i.e.

$$relative\ coverage\ = gene\ coverage \div \frac{\sum MLST\ allele\ coverage}{number\ MLST\ alleles\ found}$$

Gene coverage depth, i.e. number of reads that map to a particular position, was defined as the coverage of the longest contig in the ARIBA assembly mapping to an individual gene (**Figure 27**), not length adjusted as above. This was done based on the assumption that the longest contig would have the highest potential to be the most reliable one to estimate coverage depth from. A DNA relative coverage of 1 would suggest the gene was present in the same quantities as the MLST genes (i.e. 1 per cell) and a DNA relative coverage greater than 1 would suggest the gene was present in higher quantities than the MLST genes (i.e. > 1 per cell, e.g. on more than one plasmid, or on the chromosome and one or more plasmids). I added this relative coverage metric to create a new input feature matrix as it could capture either extra copies of genes or could reflect some issues with sequencing quality if coverage was lower than 1. I therefore constructed another 34 features – corresponding to the unique closest matching beta-lactamases (without considering variations from the sequence), *ampC* promoter sequence ATCC 25922, Lartigue *tem* promoter sequence P3 and porin genes from *E. coli* K-12. Each feature was 0 if the gene was not present in the sample and equalled the DNA relative coverage if the gene was present in the sample. The new matrix consisted of 203 features, 169 binary features and 34 new continuous features presenting gene coverage depth.

3.2.3.1.3 Adding other resistance variant detection methods for all transmissible elements in the Resfinder database, together with quality metrics

ARIBA is one of many methods to try to identify genetic features of interest, each with their own strengths and limitations (Tim Davies, unpublished). Sequencing data for each isolate was interrogated by Tim Davies using four different tools in total (ABRICATE (Seemann 2019), ARIBA, KMERRES (Clausen et al. 2016; Clausen, Aarestrup, and Lund 2018) and SRST2 (Inouye et al. 2014)) with an extended version of the Resfinder database (Hasman et al. 2012) (accessed 16th November, 2017). For all 4 tools, I considered both the presence/absence of each genetic feature as identified by that tool, but also their quality metrics, percentage identity and coverage. I also considered as input features the ARIBA flags present in at least one of the samples ("assembled", "assembled_into_one_contig", "complete_gene", "unique_contig", "has_variant", "hit_both_strands", "scaffold_graph_bad", "variants_suggest_collapsed_repeat", "region_assembled_twice"), or no flag, and the highest minor allele (second most common allele) frequency of variants reported by SRST2 which helps identify mixed/contaminated reads. This value is in the range [0,0.5]; for example, 0 indicates no variation between reads at any aligned base (i.e. at all positions in the alignment, all aligned reads agree on the same base call; although this agreed base may be different from the reference), and 0.25 indicates there is at least one position in the alignment at which all reads do not agree, and the least common variant (either match or mismatch to the reference) is present in 25% of reads (Inouye et al. 2014). Considering a toy gene and 6 aligned reads:

ACGGT

ATGGT

ATGGT

AGGTT

AGGTT

AGGTT

at position 2 G would be the called base and the maximum minor allele frequency would be 2/6, while at position 4 G would be the called base and the maximum minor allele frequency would be 3/6. Overall, the maximum minimum allele frequency reported is the maximum across the gene. When an element was not present according to SRST2, I set the highest minor allele frequency to 0. There were 174 transmissible elements present in at least one of the 976 samples and 24 features for each of these elements, as follows:

- 4 presence/absence results (one per tool), binary;
- 4 percentage identity and 4 coverage values, all 8 scaled to take values in [0,1] by dividing by 100;
- 10 ARIBA flags which take values 1 if present and 0 if absent;
- a binary feature which encodes whether a highest minor allele frequency is reported by SRST2 and if so, what the highest minor allele frequency, a continuous variable in [0,0.5].

The matrix of transmissible elements thus consisted of 4176 features (174 elements * 24 features), and these were added to the previous 203 features to create my new feature matrix. I added the 4176 features to the 203 rather than considering them alone without the 203 features, as the 203 features capture information on the variation from the closest reference sequences from the Resfinder database (variation which is not considered in the creation of these new 4176 features), as well as information on presence/absence and variations from the 2 promoter sequences considered, the *ampC* promoter sequence ATCC 25922 and the Lartigue *tem* promoter sequence P3, information on completeness of porin

genes from *E. coli* K-12, and relative coverage. Further, one advantage of machine learning methods is that they are not sensitive to collinearity in input features (i.e. it is not so important to ensure that input features are not strongly associated with each other) – the algorithms identify the most predictive features as part of their optimisation.

3.2.3.1.4 31-mers

The above matrices of features rely on prior knowledge of resistance elements. One way to avoid relying on this organism and drug specific information is to create matrices of features based on presence and absence of unique words found across the sequences. The assembled sequence of each isolate was split into 31 base k-mers (strings of DNA sequence 31 base-pairs long, moving across one base-pair at a time) and unique 31-mers were counted using DSK (Rizk, Lavenier, and Chikhi 2013). The longer the length of a k-mer the more useful the k-mer is as it is more specific; however, the number of unique k-mers also increases with length – in the case of DNA (with four bases) the total number of possible k-mers is 4^k. A genome-wide association study on Campylobacter published results on 30-mers, but found comparable results with 20-mers and 40-mers when testing the sensitivity of the method to k-mer length (S. K. Sheppard et al. 2013). I chose k=31 as in the middle of this range and an odd number in order to avoid palindromes, i.e. reverse complementary sequences (J. R. Miller, Koren, and Sutton 2010).

Considering a toy example sequence:

ACCGTAGTAGTA

Unique 3-mers and counts would be:

k-mer1 ACC 1

k-mer2 CCG 1
k-mer3 CGT 1
k-mer4 GTA 3
k-mer5 TAG 2
k-mer6 AGT 2

In-house software *gwas_k-mer_pattern* written by Daniel Wilson in C++ was then used to create a deduplicated set of variably present 31-mers across the dataset of isolate sequences, with the presence/absence of each k-mer determined per isolate. As 31-mers were counted from Velvet assemblies, a 31-mer was counted as present if observed at least once in an assembly. The C++ software can be found at https://github.com/jessiewu/bacterialGWAS/blob/master/k-merGWAS/gwas_k-mer_pattern.

Again, considering a toy example of 5 isolates with 6 unique k-mers with its presence/absence matrix:

	Isolate1	Isolate2	Isolate3	Isolate4	Isolate5	Presence and absence pattern
k-mer1	1	0	0	0	0	Pattern1
k-mer2	0	1	1	0	1	Pattern2
k-mer3	0	1	0	1	0	Pattern3
k-mer4	1	0	0	0	0	Pattern1
k-mer5	1	0	0	0	0	Pattern1
k-mer6	0	1	1	0	1	Pattern2

In this case there are 6 k-mers; however, these are present in only 3 unique presence and absence patterns: an algorithm needs only the presence/absence of the patterns not the individual k-mers because the individual k-mers do not add additional information on top of

the pattern for prediction (although they may be used to identify why an individual pattern is predictive). To turn it into a matrix of input features all I need is to transpose it.

	k-mer1	k-mer2	k-mer3	k-mer4	k-mer5	k-mer6
Isolate1	1	0	0	1	1	0
Isolate2	0	1	1	0	0	1
Isolate3	0	1	0	0	0	1
Isolate4	0	0	1	0	0	0
Isolate5	0	1	0	0	0	1

k-mers with the same presence and absence patterns in my 5 isolates are then duplicate features. I therefore drop all duplicates and only keep unique patterns of presence and absence across all my k-mers.

	Pattern1	Pattern2	Pattern3
Isolate1	1	0	0
Isolate2	0	1	1
Isolate3	0	1	0
Isolate4	0	0	1
Isolate5	0	1	0

Among my dataset of 976 isolates there are almost 67 million (66,950,863) unique, variably present 31-mers; however, there are less than 2 million (1,994,299) unique patterns of presence and absence. Even once I only consider the patterns, rather than the 31-mers, as features, the number of features is still much higher than the number of isolates, and I therefore carried out feature selection. I used the R package *bugwas* (Earle et al. 2016) to

test the genome-wide association of the patterns with binary co-amoxiclav resistant/susceptible phenotype using a linear mixed model to adjust for population structure (Zhou and Stephens 2012). I then carried out feature selection by choosing the patterns with $p<0.05/1994299$ ($-\log_{10}(p)>-\log_{10}(0.05/1994299)$ for the first feature set, and the ones that have $p<0.01$ ($-\log_{10}(p)>2$) for the second feature set. The former threshold is the Bonferroni corrected significance threshold to maintain overall 5% Type I error, that is each individual hypothesis is tested at a corrected significance level of 0.05/total number of hypotheses tested (the number of unique patterns of presence and absence). More details with regards to genome-wide association studies (GWAS) can be found in Section 3.2.4.

3.2.3.2 Output features

Given the importance of co-amoxiclav to hospital prescribing, and that incidence of co-amoxiclav resistant bloodstream infections is rising faster than co-amoxiclav susceptible infections (Chapter 2), my primary objective was to predict co-amoxiclav susceptibility in *E. coli*. However, initial work by Tim Davies demonstrated that simple rule-based prediction of phenotype from genotype had very poor performance for this bug-drug combination (T. J. Davies et al. 2019). Therefore, I also investigated prediction of amoxicillin susceptibility as a simpler test case, since this should be more straightforward since amoxicillin is only one drug, rather than a combination of a beta-lactamase and a beta-lactamase inhibitor (co-amoxiclav) and is arguably the simplest beta-lactamase for which resistance could be predicted. I first considered the binary phenotype obtained by MICs being interpreted using EUCAST breakpoint MICs, that is 8 mg/L for both amoxicillin and amoxicillin-clavulante. MICs are both left and right censored, therefore I explored the impact of this on performance, by replacing left censored values with the value itself, 1/2 and 1/4 of the left-censored value,

and right censored values with 2, 4 and 8 times the right-censored value (Croghan and Egeghy 2003). All MICs were modelled on the $\log_2$ scale so this corresponds to moving censored values 1, 2 or 3 doubling dilutions to left or right respectively, that is my output features took values in the following 3 sets: {1,2,3,4,5,6}, {0,2,3,4,5,7} and {-1,2,3,4,5,8}.

3.2.3.3 Machine learning algorithms

State-of-the-art machine learning algorithms were implemented in Python using the Scikit-learn module (Pedregosa, Weiss, and Brucher 2011). The collection of ordinal regression algorithms were implemented in Python using the mord module (Pedregosa-Izquierdo 2015), which follows a Scikit-learn compatible API and for which the source code can be found at http://github.com/fabianp/mord. Algorithms were chosen empirically based on their performance. However, further details of the mathematics behind each method are now discussed for the interested reader particularly as these can provide insight into why certain methods outperform others. All algorithms can however be easily implemented without knowledge of the underlying mathematics.

Different machine learning algorithms need to be considered depending on the type of the output feature: binary (resistant vs susceptible) or continuous ($\log_2$ MIC).

To describe the machine learning approaches taken it is necessary to first define some notation. Let X_{nxp} be the matrix of p input features for my n isolates, denoting the i^{th} row of X by X_i, and let y be a column vector with n rows recording the phenotype of my n isolates. Given my isolates have labels, I am in a supervised learning framework. I therefore try to learn a function $f: \mathcal{X} \to \mathcal{Y}$ from my input space $\mathcal{X} = \mathbb{R}^p$ (reflecting the genotypes X_i, the

input features) to my output space $\mathcal{Y}$ (reflecting my phenotypes y_i, the output features) based on my set of paired observations $(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$ and a loss function $L: \mathcal{Y} \times \mathcal{Y} \times \mathcal{X} \to \mathbb{R}_+$ which measures the discrepancy between the predicted phenotype $f(X_i)$ and the true phenotype y_i with input features X_i. I assume that $(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)$ are independent identically distributed random samples coming from an unknown joint probability distribution and I try to find the function which minimizes the expectation of the loss, called risk $R(f) = \mathbb{E}_{X,y}[L(y, f(X), X)]$, an expectation of a function of random variables X and y with some joint probability distribution. However this function R(f) is unknown since the joint probability distribution of X,y is unknown. Therefore, in practice I minimize the empirical risk $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i), x_i)$ over the hypothesis class of functions f, that is the set of functions which I am considering. The empirical risk is commonly referred to as the cost function and it is just the mean of the loss function evaluated at all data points and their predictions. I consider all linear functions $f(x) = w^T x + b$, parameterized by weights $w \in \mathbb{R}^p$ and bias $b \in \mathbb{R}$, with parameters to be estimated from the data.

3.2.3.3.1 Binary phenotype

For the binary phenotype, following (Murphy 2012; Hastie, Tibshirani, and Friedman 2009; Bishop 2007), I considered:

- Base classifiers

Logistic regression focuses on maximizing the probability of observing the data assuming a linear log-odds model, or equivalently minimizing the empirical risk. However, what I am actually interested in is minimizing the true risk. Empirical risk minimisation risks overfitting

if the function I am trying to estimate is allowed to be too complex: one can imagine finding a function which simply memorizes the data, but such a function would not generalize well. Regularization is a technique that penalizes model complexity in order to avoid overfitting and improve generalization performance. The difference between logistic l1 and logistic l2 (two different types of regularisation) is the penalty function.

- Logistic l1

Logistic regression with l1-regularization minimises the following cost function over $w \in \mathbb{R}^p$ and bias $b \in \mathbb{R}$

$$\min_{w,b} \|w\|_1 + C \sum_{i=1}^{n} \log\left(1 + e^{-y_i(X_i^T w + b)}\right)$$

The penalty function (first term) is just the sum of the absolute values of the weights w_i, i=1,2, …,p corresponding to the p features. Typically there is no closed solution, that is the weights and bias which minimize the cost function cannot be analytically calculated. However, the function is convex and differentiable and therefore there exists a unique minimum and iterative optimisation techniques like gradient descent can be used (**Figure 28**).

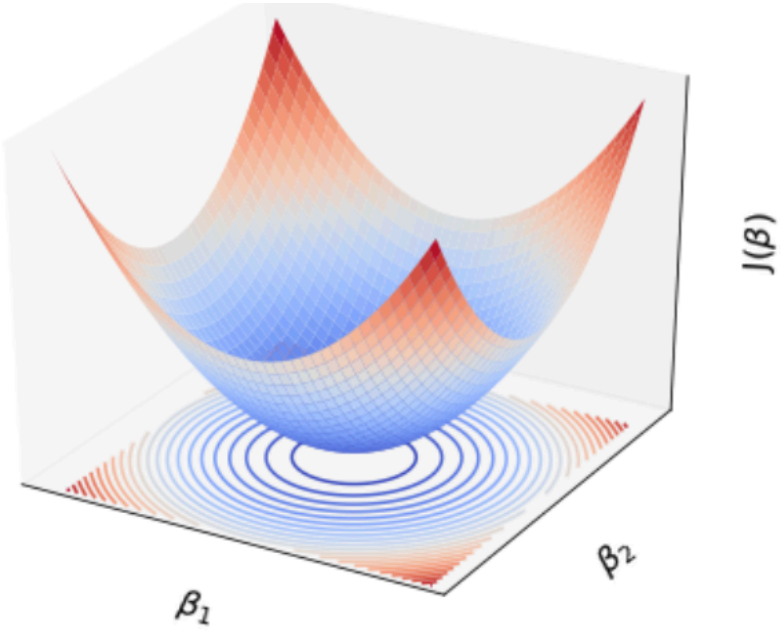


Figure 28 Gradient descent for optimization of the cost function

Note: I make the simplifying assumption that p=2 in order to be able to plot the cost function J(β) as I vary weights β_1 and β_2. The weights would be those at which the cost J(β) is minimum.
Source:https://towardsdatascience.com/regularization-in-machine-learning-connecting-the-dots-c6e030bfaddd

I have added a parameter C which controls the strength of the penalty. This parameter needs to be chosen by the modeller and is called a hyperparameter. Details of how these were chosen will follow when the machine learning procedure is described below. Smaller values of C will favour sparser models, i.e. models where many of the weights are 0 and these input features therefore have no effect on the prediction.

- Logistic l2

Logistic regression with l2-regularization minimises a slightly different cost function

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \log\left(1 + e^{-y_i(X_i^T w + b)}\right)$$

The difference is that the weights are squared which means that larger values of weights w will be penalized more than smaller ones. Moreover, l1-regularization can yield sparse models, while l2-regularization cannot.

- Support vector machines (SVM) with linear kernel

For classification, support vector machines try to find the separating hyperplane (a subspace whose dimension is one less than that of its main space; in 3 dimensions this would be a standard plane, while in 2 dimensions this would be a line) that maximizes the distance between the closest point to the margin and observations classified into different groups, called the support vectors. This is equivalent to minimizing the cost function

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i(X_i^T w + b)\}$$

The hyperparameter C controls the simplicity of the decision function, with a low C favouring a smooth decision function, while a high C aims to classify all isolates correctly at the cost of a more complex function which might not generalize well.

- SVM with radial basis function (RBF) kernel

One problem is that a dataset may not be linearly separable, that is there may not exist a hyperplane in the multi-dimensional feature space which separates my isolates into susceptible/resistant. In a 2-dimensional space, that would mean there is no line which separates my isolates into susceptible/resistant. Kernel methods allow the creation of non-linear combinations of my features which uplift my samples onto a higher-dimensional feature space where they may become linearly separable (see **Figure 29** for a toy example) and where I can therefore then use a linear decision boundary to separate the classes, just as above, thus creating a non-linear machine learning algorithm.

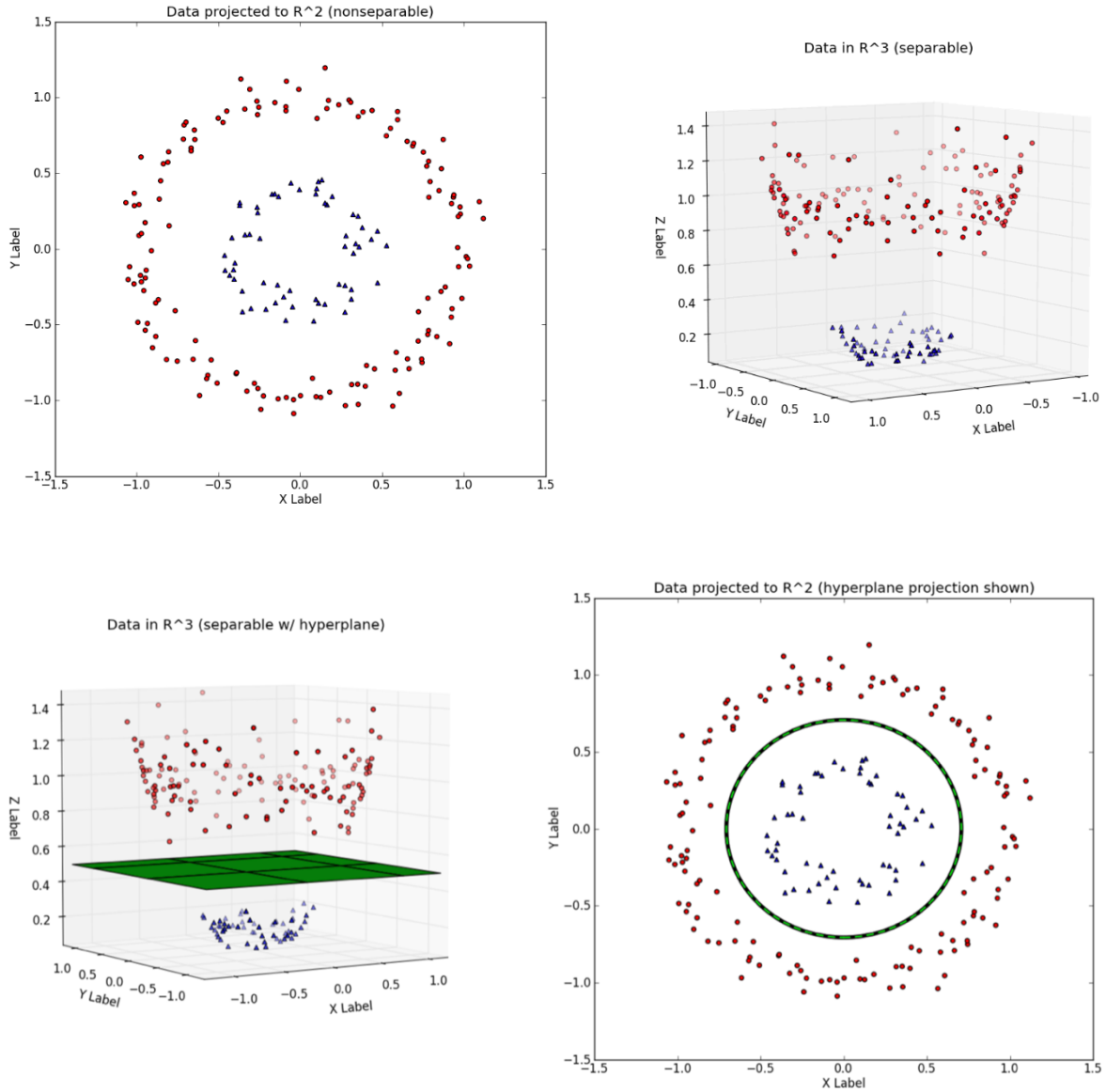


Figure 29 Toy example illustrating the power of kernel methods

Note: data becomes linearly separable in a higher dimensional space (top row), and by projecting the linear decision boundary from this higher dimensional space onto the original space, a non-linear decision boundary is obtained in the original space. Top Left: a dataset in $\mathbb{R}^2$ not linearly separable (a circular decision boundary is appropriate in this case), Top Right: the same dataset uplifted onto $\mathbb{R}^3$by the transformation: $[x_1, x_2] = [x_1, x_2, x_1^2 + x_2^2]$. Source: (E. Kim 2015).

The power of kernel methods is in the fact that an explicit mapping is not required in order to get the linear learning algorithm to learn the non-linear decision boundary. The radial basis function (RBF) is $e^{-\gamma\|x-x'\|_2^2}$ and γ defines how much influence a training data has, and

γ can therefore be seen can be seen as the inverse of the radius of influence of the isolates selected by the model as support vectors.

- Ensemble methods

Ensemble methods use the simple, yet very powerful, idea of combining a number of different machine learning algorithms in order to improve predictive performance. This is also known as meta-learning. I specifically considered decision trees as the machine algorithms to be combined. A decision tree asks a series of sequential questions; the answers determine which question is asked next (equivalent to any individual isolate going down one of 2 "branches" coming from each question). If a decision tree only asks a few questions (only has a few nodes), then it might have a high bias, i.e. underfit; however, if the tree is too deep (has many questions/nodes before it makes a final decision on the output), then it might have a high variance, i.e. it will overfit. One way to achieve a good bias-variance trade-off (**Figure 30**) is to perform hyperparameter tuning. However, even if hyperparameter tuning is carried out, decision trees are unstable, that is small changes in the input data lead to different trees. Ensemble methods help mitigate this and produce low bias, low variance models.

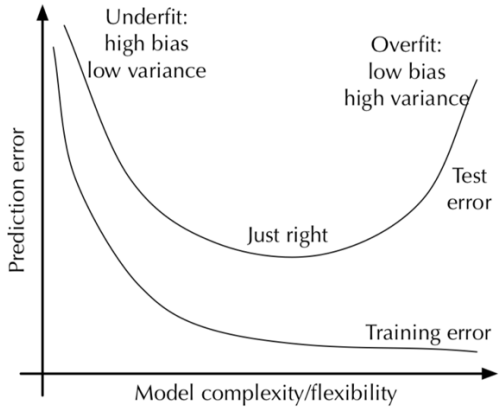


Figure 30 Bias-variance trade-off

Note: Source: http://www.stats.ox.ac.uk/~palamara/teaching/SML19/HT19_lecture7.pdf.

Formally, a decision tree recursively partitions the feature space $\mathbb{R}^p$ such that isolates from the same class are eventually grouped together. Suppose there are k classes (where a binary output corresponds to k=2). Then for node m representing a subspace R_m of the feature space $\mathbb{R}^p$ with N_m observations, define

$$p_{m,l} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = l), l = 1, \dots, k$$

i.e. the proportion of isolates from class l in node m. In order to obtain a tree which differentiates well between the classes, I want to minimize the "impurity" of the nodes (that is, the proportion of observations assigned to the node from a different class to the majority class) which I chose to define using the Giny impurity function

$$\sum_{l=1}^{k} p_{m,l}(1 - p_{m,l})$$

Considering a split based on feature j which splits the data from node m into a left node and a right node (for example left node will be all the isolates which have a feature present, while the right node will be all the isolates which have this feature absent), then the impurity of node m is

$$\frac{N_{left}}{N_m} \sum_{l=1}^{k} p_{left,l}(1 - p_{left,l}) + \frac{N_{right}}{N_m} \sum_{l=1}^{k} p_{right,l}(1 - p_{right,l})$$

- Random forest

In a random forest classifier, each tree in the ensemble is built from a sample drawn with replacement from the training set to reduce over-fitting to a specific dataset, by assuming it represents a larger population which can be repeatedly sampled using replacement. Furthermore, in constructing each tree, when I split a node I choose the best split from a random subset of features instead of choosing from all the features in order to reduce over-

fitting, by avoiding a few strong predictors being consistently chosen for the top levels of the trees, that is very similar structured and hence highly correlated trees. Nodes are split until all isolates reaching a node are classified as either susceptible or resistant. The overall prediction for an individual isolate from a random forest classifier is the class with the majority number of votes across all the classifiers (trees), therefore the method is known as bagging. The strength of a random forest lies in its ability to prevent overfitting through the random choice of features, which improves its generalisability to unseen data by comparison to a simple decision tree. Since all the trees are allowed to grow until all isolates in a leaf (i.e. a node at the end of a branch) are all classified either susceptible or resistant, the averaging is done over a number of low bias trees, therefore obtaining a low bias, low variance tree.

- Adaboost

Adaboost stands for adaptive boosting. All isolates are initially assigned weights of 1/N, where N is the number of isolates, and a small decision tree is trained on the original data. Here I chose to use trees with maximum depth of 3, ie a maximum of three features to classify each isolate in each model (tree). Because of this depth restriction, this model is only slightly better than random guessing and for that reason it is called a weak learner. I then iteratively modified the isolate weights by increasing the weights of the isolates which were incorrectly predicted by the weak learner and decreasing the weights of the isolates which were correctly predicted, and then training a new tree. With each iteration, the weights force the model to focus on those isolates which were misclassified by the previous models. The final prediction is a weighted majority vote of the predictions from all the models. I chose the number of weak learners to be 100.

- Gradient boosted trees

Just like other boosting methods, this method creates a sequence of weak learners and the ultimate prediction is a weighted majority vote of the predictions from all the models. At each stage, a regression tree is fit to the negative gradient of the binomial deviance loss function, that is every tree minimizes the residual errors made by the previous tree. I again chose the maximum depth of the tree to be 3 and the number of weak learners to be 100.

In summary, Adaboost and Gradient boosted trees both work by sequentially adding trees to the ensemble which focus on the better prediction of the previously misclassified isolates. The difference between the two lies in how they treat the misclassified isolates: at every iteration Adaboost increases the weights of the misclassified isolates, while Gradient boosted trees fits a new model to the residual errors made by the previous predictor.

3.2.3.3.2 MIC

Whilst MIC may be thought of as a continuous outcome, in practice the way it is measured by microbroth dilution, and the relatively small number of discrete values that it can take, make it more like an ordered categorical outcome, formally called an “ordinal” outcome. For the MIC phenotype I therefore considered the following methods:

- Ignoring order in MICs (treating the output as multiclass)

Multinomial logistic regression is a generalization of logistic regression for a multiclass problem, where the different MIC categories are simply considered as different unrelated labels (that is, the ordering of MICs is not taken into account). Suppose there are k classes.

- Multinomial l1

$$\min_{w,b} \|w\|_1 + C \sum_{i=1}^{n} \sum_{j=1}^{k} 1\{y_i = j\} \log \left(\frac{e^{-y_i(X_i^T w+b)}}{\sum_{l=1}^{k} e^{-y_l(X_i^T w+b)}} \right)$$

- Multinomial l2

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \sum_{j=1}^{k} 1\{y_i = j\} \log \left(\frac{e^{-y_i(X_i^T w+b)}}{\sum_{l=1}^{k} e^{-y_l(X_i^T w+b)}} \right)$$

- Random forest classification

Decision trees for binary phenotype were described above for the general case of classification for k classes. With a multinomial learner, nodes are split until all isolates reaching a node are classified as belonging to one of the k classes, rather than just one of two classes. The overall prediction for an individual isolate from a random forest classifier is the class with the majority number of votes across all the classifiers (trees).

- Ordinal models or threshold-based models

Ordinal models or threshold-based models are an extension of logistic regression. The extension from binary to ordinal labels is done by using k-1 thresholds $\theta_1, \theta_2, \ldots, \theta_{k-1}$ to split the real-line into k segments. For each isolate a predicted value f(x_i) lies on the real-line – if this falls between two thresholds (θ_{j-1}, θ_j) then this would predict sample i belonging to class j. One crucial detail when considering an ordinal outcome is that, when the true MIC is 16/2 mg/L, predicting an MIC of 2/2 mg/L is worse than predicting an MIC of 8/2 mg/L. The difference between the Ordinal Immediate-Threshold and the Ordinal All-Threshold models is how they penalize for the number of thresholds between the true category and the predicted one, that is through the choice of loss function.

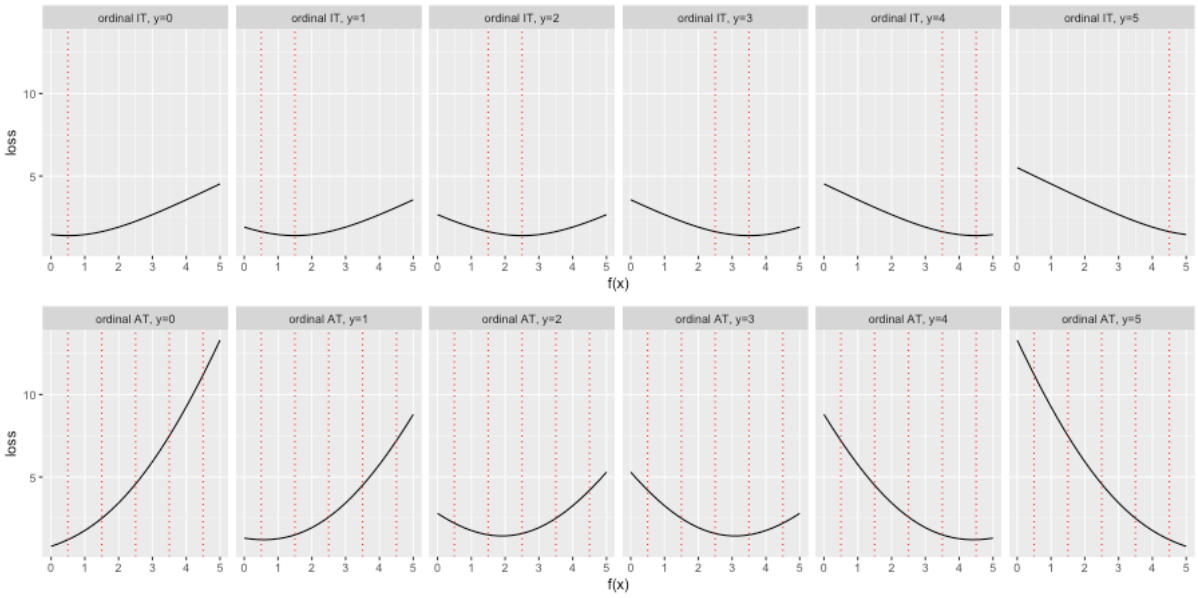


Figure 31 Loss functions from Top: Ordinal Immediate-Threshold (IT), Bottom: Ordinal All-Threshold (AT)

Note: considering an ordinal label with 6 categories (0,1,2,3,4,5). Predictions assumed to vary between 0 and 5 and choosing (0.5,1.5,2.5,3.5 and 4.5) as the 5 thresholds.

- Ordinal Immediate-Threshold (IT) (Rennie and Srebro 2005; Pedregosa, Bach, and Gramfort 2014; Pedregosa-Izquierdo 2015)

For the immediate-threshold variant, for isolate i with true label y_i and predicted label $f(x_i)$ the loss function is of the form

$$L(y_i, f(x_i), x_i) = l\big(f(x_i) - \theta_{y_i-1}\big) + l(\theta_{y_i} - f(x_i))$$

where $l(x)$ is the logit function

$$l(x) = -\log\left(\frac{1}{1+e^{-x}}\right)$$

Therefore, I am trying to maximize the accuracy, i.e. the number of correctly classified observations as a proportion of all observations.

- Ordinal All-Threshold (AT) (Rennie and Srebro 2005; Pedregosa, Bach, and Gramfort 2014; Pedregosa-Izquierdo 2015)

For the all-threshold variant, I consider all threshold violations and sum over all the penalties.

Define $s(m, y_i) := \begin{cases} -1 \; if \; m < y_i \\ +1 \; if \; m \geq y_i \end{cases}$. Then the all-threshold loss is:

$$L(y_i, f(x_i), x_i) = \sum_{m=1}^{k-1} l(s(m, y_i) * (\theta_m - f(x_i)))$$

This variant minimizes the mean absolute error, where the error is the difference between the model prediction and the correct label.

- Ordinal Squared Error (SE) (Pedregosa-Izquierdo, 2015; Pedregosa et al., 2014)

Ordinal SE uses minimizes the mean squared error, rather than the mean absolute error as in ordinal AT above.

In summary, Ordinal IT, AT and SE all model the ordering of the MICs and penalize for the number of thresholds between the true label and the predicted label. All algorithms therefore favour solutions that minimize the number of thresholds that are crossed. However, **Figure 31** illustrates how, even though the penalty increases the further away the predicted label is from the true label for both Ordinal IT and Ordinal AT, the increase is not very big for the former; the latter was designed so that with each threshold crossed the penalty increases more steeply. Ordinal SE takes this one step further and squares the error in order to increase the penalty even more.

- Regression-based models (treating the output as continuous)

Rather than treating the MICs as ordinal or categorical labels, they can instead be considered as truly continuous variables.

- Ordinal Ridge (Pedregosa-Izquierdo, 2015; Pedregosa et al., 2014)

This method fits an l2-regularized linear least squares model (i.e. minimizes the mean squared error), rounds the predicted MIC and sets all values which are below the minimum to the minimum and all which are above the maximum to the maximum.

- Least absolute deviation (LAD) (Pedregosa-Izquierdo, 2015; Pedregosa et al., 2014)

SVM can be extended to regression problems, i.e. to a continuous prediction rather than a dichotomous class prediction. Analogously to how the binary classification model produced by an SVM effectively disregards any observations from the training data that lie close to the hyperplane separating classes, the model produced by Support Vector Regression (SVR) also ignores the data close to the prediction during the model building, that is, it will disregard observations that are less than some value ε from the prediction (Paisitkriangkrai 2012).

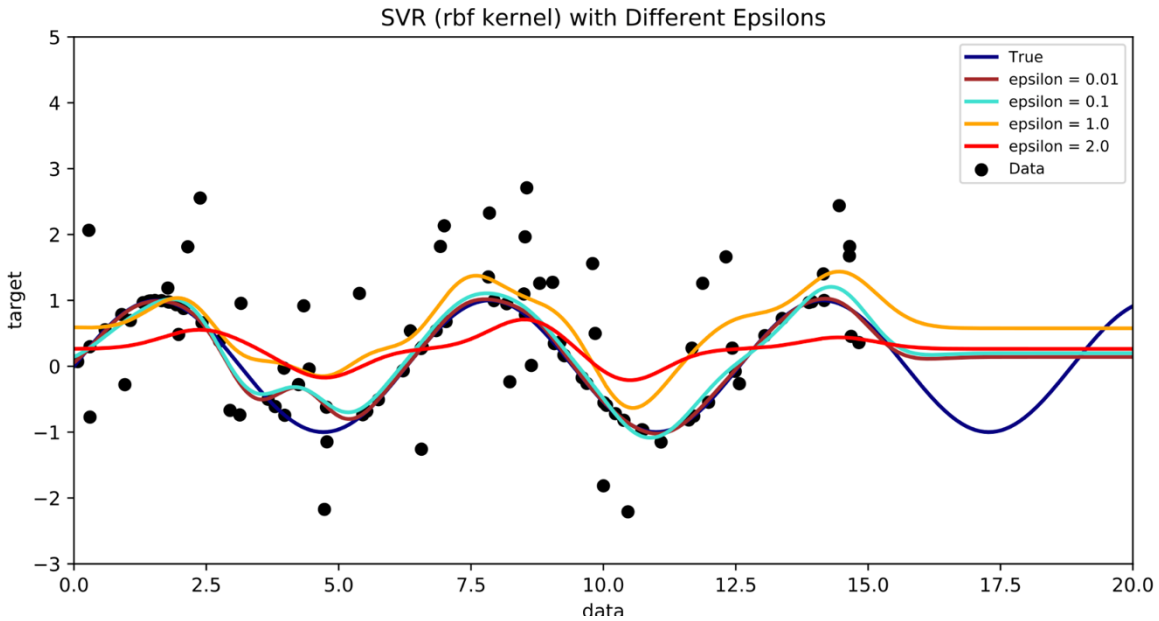


Figure 32 SVR prediction varying ε.

Note: As ε increases the predictions become less sensitive to errors. Source: https://en.wikipedia.org/wiki/Support-vector_machine#/media/File:Svr_epsilons_demo.svg.

3.2.3.3.3 Feature importance

In traditional statistical models, regression coefficients correspond to the importance of the features in the model prediction. This can be done similarly for machine learning regression models, but alternative approaches can also be considered. Focusing on the Ordinal SE model, the relative importance of features to the prediction can be retrieved from the weights corresponding to each feature, with higher squared weights leading to a higher ranking of importance.

An alternative approach applicable to all methods using weights is recursive feature elimination. This starts with a model which includes all features. It then drops the feature with the lowest weight in the model and the new model is trained on this new set of features. The process is recursively repeated until all features are dropped. The relative importance of features is then the reverse order in which they were dropped. This method could also be used as a feature selection technique by iteratively removing the least important features. This can lead to improved performance as noise is removed, while also decreasing training time.

3.2.3.4 Pipeline for testing classifiers

I start with 5-fold stratified cross-validation on the 976 samples, i.e. I randomly divided my dataset into five (folds), but kept the susceptible/resistant ratio equal across the folds. I trained on 80% of the data in each fold (yellow boxes) and evaluated performance on the remaining 20% in each fold (green boxes) as illustrated below (**Figure 33**). This gives a total of 5 iterations from which one value for each performance metric is obtained from each iteration (based on performance of the model trained on the yellow boxes and evaluated on

each green box). I also repeat this process n=10 times. I assume performance metrics $P_1, P_2, \ldots, P_{n*5}$ are independent identically distributed normal random variables with unknown mean μ and unknown variance σ^2. Then considering the sample mean $\bar{P} = \frac{\sum_{i=1}^{n*5} P_i}{n*5}$ and the sample variance $S^2 = \frac{1}{n*5-1}\sum_{i=1}^{n*5}(P_i - \bar{P})^2$, $\frac{\bar{P}-\mu}{\frac{S}{\sqrt{n*5}}} \sim t_{n*5-1}$, $\left(\bar{P} \pm t_{n*5-1}\left(\frac{\alpha}{2}\right)\frac{S}{\sqrt{n*5}}\right)$ is a $1-\alpha$ confidence interval for μ, choosing α=0.05 The higher the number of repeats n, the tighter the confidence intervals, but an increase in number of repeats also means an increase in computation time. I therefore chose n=10 repeats as a compromise; that is, I repeat the division of the data into folds 10 times, from which I compute the mean and standard deviation of each performance metric (Vanwinckelen and Blockeel 2012).

Iteration	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
1	Train	Train	Train	Train	Test
2	Train	Train	Train	Test	Train
3	Train	Train	Test	Train	Train
4	Train	Test	Train	Train	Train
5	Test	Train	Train	Train	Train

Figure 33 Splitting of the data during cross-validation

For every single training dataset (4 of the 5 folds, yellow boxes within each row), I balance the training dataset, i.e. subsample from the isolates in the class with higher numbers, susceptible for co-amoxiclav, resistant for amoxicillin, in order to give equal numbers of resistant and susceptible isolates in the training dataset. Balancing should be done if a dataset is highly imbalanced: for example, 95% accuracy could be obtained on a dataset with 95% susceptible samples by calling everything susceptible, but will of course be a useless model. I only balance the training dataset, not the test dataset, as if I were to apply

the model in a clinical setting I would of course have no control over the distribution of the phenotypes.

3.2.3.4.1 Setting the threshold for calling resistance

The optimal balance between false resistant and false susceptible calls is to some extent a subjective judgement. I therefore chose to produce an algorithm that balanced both errors by minimising the absolute difference between the $very\ major\ error\ rate = \frac{false\ susceptible}{true\ resistant + false\ susceptible}$ and the $major\ error\ rate = \frac{false\ resistant}{true\ susceptible + false\ resistant}$.

All of the binary classification models which I consider make a probability prediction between 0 and 1 of the isolate belonging to a class. I varied the probability required for a resistant call, to balance the false negative and positive rates. The default threshold is 0.5, so a predicted probability of >0.5 would mean the isolate was predicted to be resistant. However, this may not be optimal. Therefore, before predicting in the test dataset (green above), I first split the balanced training dataset (yellow above) 80%/20% into a nested train/validate set (**Figure 34**) in order to choose the probability threshold defining resistant (vs susceptible) from {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}.

Iteration	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5
1	Train	Train	Train	Train	Test
2	Train	Train	Train	Test	Train
3	Train	Train	Test	Train	Train
4	Train	Test	Train	Train	Train
5	Test	Train	Train	Train	Train

Figure 34 Additional splitting of the data for internal validation of the resistance calling threshold during cross-validation

Note: Showing validation 20% of the training dataset in orange: validation is a random 20% sample across all folds in the training dataset, not necessarily balanced across folds.

3.2.3.4.2 Hyperparameter tuning

Tuning a threshold is generally considered different from tuning hyperparameters, since tuning a threshold does not change the model parameters or the flexibility of the model. Instead hyperparameters are used in the process of estimating model parameters, helping this process by controlling the complexity of the model.

I only carried out hyperparameter tuning for SVM and did this using library Optunity (Claesen et al. 2014). Before choosing the probability threshold, I carried out 2-fold stratified cross-validation on the remaining 80% of the balanced training dataset (excluding the validation set) to find the hyperparameters which maximize the area under the curve (AUC) (defined below). For the linear kernel, only the regularization parameter C needs to be tuned, penalizing the complexity of the decision surface. For the RBF kernel, both C and γ need to be tuned, where γ can be seen as the inverse of the radius of influence of isolates selected by the model as support vectors. I used the recommended hyperparameter search strategy, called particle swarm optimization (Clere M and Kennedy J 2002; Kennedy 2010), where the hyperparameters are randomly initialised and their movement after that is influenced by the value of the cost function which I am trying to minimize. I chose 10 evaluations in the space $C \times \log_{10} \gamma = [0,10] \times [-5,0]$. Once the optimal hyperparameters were selected, I trained again on the full 80% of the balanced training dataset.

For the MIC prediction (rather than resistant/susceptible prediction), I repeated the stratified 5-fold cross-validation 10 times as described above; however, I only ran these with the default hyperparameters, and did no hyperparameter tuning, therefore the validation step was no longer carried out.

3.2.3.5 Performance metrics

For binary outputs, accuracy, sensitivity, specificity and AUC are considered when evaluating models, defined as below:

$$accuracy := \frac{true\ resistant + true\ susceptible}{true\ resistant + true\ susceptible + false\ resistant + false\ susceptible}$$

$$sensitivity := \frac{true\ resistant}{true\ resistant + false\ susceptible}$$

$$specificity := \frac{true\ susceptible}{true\ susceptible + false\ resistant}$$

$$AUC = \int_0^1 ROC(t)dt, where\ t = 1 - specificity\ and\ ROC(t) = sensitivity$$

- The receiving operating characteristic (ROC) curve is a summary performance measurement averaged over different threshold settings.
- AUC is the area under the ROC curve and is a measure of separability between susceptible and resistant isolates.
- For all of these metrics, higher values mean better performance, that is, a better capability of distinguishing between susceptible and resistant isolates.
- When sensitivity increases, specificity decreases and vice-versa.
- If I decrease the probability threshold defining resistance within the models, I increase the number of predicted resistant isolates and hence increase the sensitivity and decrease the specificity; conversely, if I increase the probability threshold, I increase the number of predicted susceptible isolates and hence increase the specificity and decrease the sensitivity.

For MIC as the output, I considered essential agreement and accuracy (also called categorical agreement) when evaluating models, defined as below:

$$essential\ agreement := \frac{true\ MIC + MIC\ within\ a\ doubling\ dilution\ of\ the\ true\ MIC}{number\ of\ predictions}$$

$$accuracy := \frac{true\ MIC}{number\ of\ predictions}$$

Using essential agreement as a performance metric is consistent with current FDA standards for antimicrobial susceptibility test systems (Devices and Services 2003; J. H. Jorgensen 1993) and with established clinical microbiology practices (James H Jorgensen and Ferraro 2009). An acceptable essential agreement would be above 89.9%, however, accuracy would also generally be expected to be above 89.9% (Devices and Services 2003). Sensitivity and specificity are defined as above and calculated for MIC predictions dichotomised using the EUCAST breakpoint (so if predicted MIC>8 mg/L then resistant, if predicted MIC≤8 mg/L then susceptible).

3.2.4 Bacterial genome-wide association study

Bacterial genome-wide association studies are a well-established method for investigating associations between genetic variants and a specific trait (Earle et al., 2016). As the name indicates, variants are investigated across the entire genome, in contrast to gene-specific candidate-driven studies. Studies based on single nucleotide polymorphisms rely on mapping sequence data to a highly conserved reference genome and only capture variation within the core genome. I therefore chose to capture variation using k-mers which overcomes this limitation and are able to capture single nucleotide polymorphisms, gene presence and absence, short insertions and deletions, therefore including both the core genome common to all organisms with a species and the accessory genome, containing additional genetic material that may or may not be present.

As described in section 3.2.3.1.4:

- I split the sequencing data for each isolate into 31-mers
- I created a list of unique variably present 31-mers
- I created a list of presence and absence patterns within my 976 samples and a list which recorded the index of the pattern corresponding to each 31-mer
- I calculated a relatedness matrix or Kinship matrix K of size $n \times n$ defining how related the samples were to each other based on presence and absence patterns, $K = \frac{1}{p}\sum_{j=1}^{p}(X_j - \frac{1}{p}\sum_{k=1}^{p} X_k) \times (X_j - \frac{1}{p}\sum_{k=1}^{p} X_k)^T$, where X_j is the j[th] column of the $X_{n\times p}$ is the matrix of p features for my n isolates, where the p features are in this case the presence and absence patterns. This was done using the software GEMMA (Zhou and Stephens 2012)
- I ran the R package *bugwas* which tests both patterns of presence and absence of 31-mers and strain lineages (defined by principal components) for their association with the phenotype.

The R package *bugwas* calls the software GEMMA (Zhou and Stephens 2012) which implements the Genome-wide Efficient Mixed Model Association algorithm. It is used to test for association of patterns of presence and absence of 31-mers with the phenotype, adjusting for population structure which can have strong effects in bacterial GWAS, reflected by the effects of the principal components. This is done by fitting a univariate linear mixed model (LMM) in the following form:

$$y = W\alpha + x\beta + u + \epsilon,$$

with

$$u = X\gamma,$$

and assuming

$$\gamma \sim MVN_p\left(0, \lambda\tau^{-1}I_p\right),$$

$$\epsilon \sim MVN_n(0, \tau^{-1}I_n),$$

and so

$$u \sim MVN_n(0, \lambda\tau^{-1}K),$$

- y is an n-vector of phenotypes for my n=976 isolates,
- W_{nxc} is a matrix of fixed effects, which in my case is just a column of 1s,
- α is a c-vector of the corresponding coefficients including the intercept, which in my case includes just the intercept and nothing else,
- x is an n-vector of the marker genotypes, which in my case is the pattern of presence and absence of the 31-mers which is being tested,
- β is the effect size of these genotype patterns,
- u is an n-vector of random effects,
- ε is an n-vector of errors,
- λ is the ratio between the two variance components,
- τ^{-1} is the variance of the residual errors,
- K is the $n \times n$ relatedness matrix (reflecting population structure).

MVN_n is standard notation for an n-dimensional multivariate normal distribution.

GEMMA tests the null hypothesis $H_0: \beta = 0$ against the alternative hypothesis $H_1: \beta \neq 0$ using a likelihood ratio test. This is done for each pattern of presence and absence of the 31-mers, one at a time.

For prediction of phenotype in an individual isolate, the approach to prediction in *bugwas* is first to estimate the individual isolate "background" resistance, the random effects γ. To do this, set α=0 and β=0. Then the posterior distribution of γ is:

$$\gamma|X, y, \lambda, \tau^{-1} \sim MVN_p\left(\left(X^TX + \frac{1}{\lambda}I_p\right)^{-1}X^Ty, \left(\tau X^TX + \frac{1}{\lambda}I_p\right)^{-1}\right)$$

λ and τ were estimated using restricted maximum likelihood (REML).

The posterior mean of γ is given by $\left(X^TX + \frac{1}{\lambda}I_p\right)^{-1}X^Ty$, also known as the best linear unbiased prediction (BLUP) of the random effects γ.

The effect sizes (i.e. the posterior mean estimates of random effects) of each of the patterns present within each isolate are then added (Earle et al., 2016), that is the predicted phenotypes are X*γ, and the continuous outcome dichotomised by calling all negative predictions susceptible and all positive predictions resistant.

Focusing on predicting $\log_2$(MIC), I chose to deal with left and right censoring of the MIC values by:

- setting all values which were below the lower limit to the lower limit itself and those which were greater than the upper limit to double the upper limit (x2) (as observed values were ≤2/2 mg/L, but strictly greater than 32/2 mg/L),
- setting all values which were below the lower limit to half the lower limit (/2) and those which were greater than the upper limit to four times the upper limit (x4) (i.e. move these values to one doubling dilution more extreme on the $\log_2$(MIC) scale),
- setting all values which were below the lower limit to quarter the lower limit (/4) and those which were greater than the upper limit to eight times the upper limit (x8)

(i.e. move these values to two doubling dilutions more extreme on the $\log_2$(MIC) scale),

which is standard practice for statistical analyses of left and right censored data where this cannot straightforwardly be incorporated into the distribution likelihood, as here (Croghan and Egeghy 2003).

LMMs were chosen in order to control for population structure. However, this can significantly reduce power to detect associations of population-stratified variants (in my case patterns of presence and absence of 31-mers where these are both strongly associated with phenotype and with specific strains or sub-strains). Principal component analysis (PCA) is a dimensionality reduction technique which aims to preserve the variance in the data. It has been observed that leading principal components (PCs) of k-mer patterns trace paths through the deepest branches of the corresponding phylogenetic tree which reflect an underlying relationship between principal component analysis and genealogical history (Earle et al., 2016; McVean, 2009). I therefore defined lineages by principal components of my k-mer patterns. Principal components are uncorrelated by definition and their construction, therefore I expect my power to detect lineage associations to increase by using PCs compared to defining lineages using genetic clustering approaches (e.g. multilocus sequence types (STs)) where these could be correlated. I assessed the significance of the association between each PC and the phenotype using a Wald test (Wald 1943). Each individual hypothesis is tested at a corrected significance level of 0.05/total number of hypotheses tested, that is the number of PCs which is equal to 976, the number of isolates.

In order to visualise the results genome-wide using Manhattan plots, I used Bowtie2 (Langmead and Salzberg 2012) to map my 31-mers against complete reference genomes of *E. coli* CFT073, and two examples of the common, multi-drug resistant lineage ST131, specifically: Ecol_244, SAMN05511157 - fimH22 and Ecol_AZ146, SAMN05511152, fimH30-Rx, with data for their complete genomes being available at https://www.ncbi.nlm.nih.gov/bioproject/?term=316786 (Stoesser et al. 2017).

I used KMA (Clausen, Aarestrup, and Lund 2018) to map my 31-mers onto the Resfinder database (Hasman et al. 2012) and onto 11144 curated plasmids from NCBI, downloaded in August 2018. The curation methods used were similar to those published in (Orlek et al. 2017) and aimed to identify genuine complete plasmid sequences. The plasmids are from all bacteria, i.e. taxonomy query term: bacteria[porgn:__txid2], and all identical plasmid sequences were deduplicated.

The R package *bugwas* was developed by Sarah Earle to test for associations between variants and the phenotype whilst controlling for population structure and identify lineage effects and can be found at the following github page: http://github.com/sgearle/bugwas.

3.2.5 *Bugwas* prediction in relation to machine learning algorithms

The solution of the *bugwas* prediction is the same as that of an l2-regularized linear regression, the machine learning algorithm Ordinal Ridge in Section 3.2.3.3.2, when considering MIC as a truly continuous variable. I did not treat the outcome as continuous when considering a binary phenotype when using machine learning algorithms.

As the predictions from *bugwas* are actually continuous values, the resistance calling threshold could be tuned, in the same way in which it was done for the machine learning algorithms. I did however not pursue this as the main use of *bugwas* in this Chapter was for feature selection.

3.3 Results

3.3.1 Routine laboratory phenotypes and genotypes

Between January 2013 and August 2015, there were 1039 *E. coli* bloodstream infections after 90 day de-duplication. 1008 (97%) had complete amoxicillin and co-amoxiclav susceptibility results, and 976/1008 (97%) sequences passed quality control (T. J. Davies et al. 2019). 531/976 (54%) had amoxicillin MIC >8 mg/L by BD Phoenix™, i.e. were classified amoxicillin-resistant using EUCAST breakpoints. 355/976 (36%) had co-amoxiclav MIC > 8/2 mg/L by BD Phoenix™, i.e. were classified co-amoxiclav-resistant using EUCAST breakpoints. When the 355 isolates with co-amoxiclav MIC > 8/2 mg/L were retested, one (0.3%) MIC changed to ≤2/2 mg/L, one (0.3%) to 4/2 mg/L and 14 (4%) to =8/2 mg/L; that is, a total of 16/355 (5%) changed from resistant to susceptible, leaving 339/976 (35%) with a co-amoxiclav MIC > 8/2 mg/L (but now identified as 16/2, 32/2 or >32/2 mg/L in all cases).

Table 10 Distribution of co-amoxiclav MICs within 976 E. coli bloodstream infections

Co-amoxiclav MIC (mg/L)	≤2/2	4/2	8/2	16/2	32/2	>32/2
Number of samples	165	304	168	89	59	191
(%)	(17%)	(31%)	(17%)	(9%)	(6%)	(20%)

Figure 35 illustrates a bimodal phenotype for amoxicillin and piperacillin **(a)** and **(c)** versus a continuous phenotype for co-amoxiclav and piperacillin-tazobactam **(b)** and **(d).** Whilst both are traditionally dichotomised through cut-off points, for the latter these are somewhat

arbitrary given the lack of bimodality in the distributions.

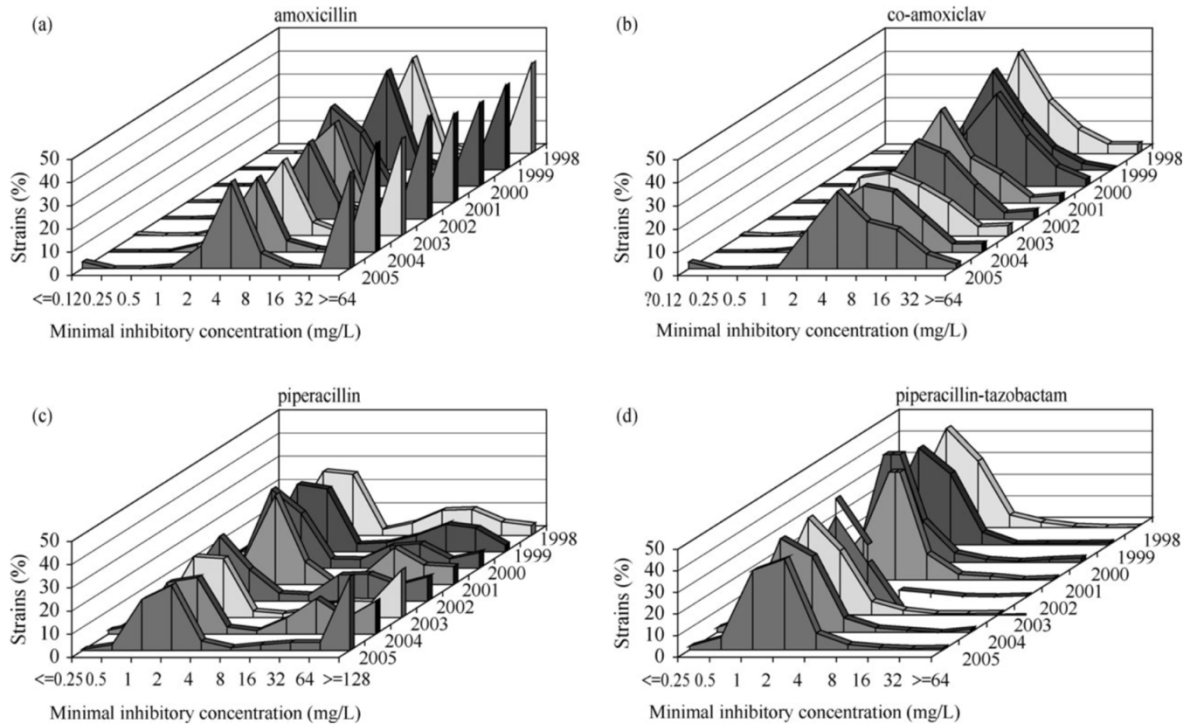


Figure 35 MIC distributions of E. coli isolated from urology services in the Netherlands (1998–2005)

Note: Source: (Nys et al. 2008).

3.3.2 Binary phenotype: comparing machine learning algorithms

Focusing first on amoxicillin WGS resistance prediction with a binary phenotype as the outcome and input features constructed from genetic features identified from mapping onto resistance databases, **Figure 36** compares machine learning algorithms in terms of accuracy, sensitivity, specificity and AUC considering input features 'known' (known co-amoxiclav resistance determinants, p=169, Section 3.2.3.1.1), 'known + relative coverage' (p=203, Section 3.2.3.1.2) and 'known + rc + resfinder' (known co-amoxiclav resistance determinants plus relative coverage (rc) plus all transmissible elements and quality metrics by the 4 different tools) (p=4379, Section 3.2.3.1.3). Performance was excellent across the input feature sets, regardless of the machine learning algorithm, with means >95% across all

metrics, and relatively little difference between the feature sets, with at most 'known + rc + resfinder' having slightly lower specificity than the other two feature sets.

Focusing on the 'known + rc + resfinder' feature set, gradient boosted trees perform slightly better than the other models in terms of specificity, while performing as well as the rest in terms of sensitivity, and were therefore slightly better than other methods in terms of accuracy. However, SVM with RBF kernel outperformed the other algorithms in terms of AUC across all the input feature sets, despite not outperforming them in terms of the other 3 performance metrics. However, even though AUC is widely used in the academic literature for comparison of classifiers, studies (Hand 2013, 2009) have shown that the AUC is equivalent to assuming different cost distributions for different classifiers and therefore it should not be used to compare classifiers, but only to evaluate their performance against random classification. For this reason, overall gradient boosted trees can be considered to offer the best overall performance for prediction of amoxicillin resistance.

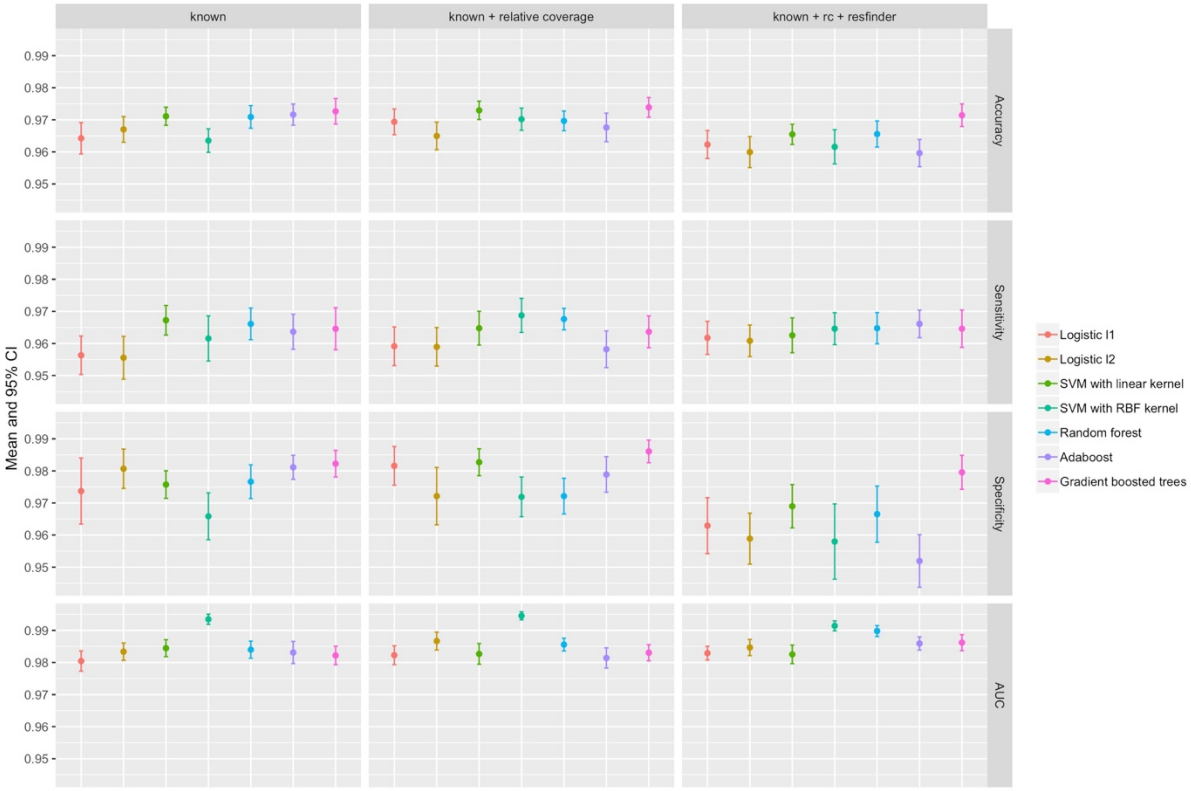


Figure 36 Amoxicillin resistance prediction performance of various machine learning classifiers applied to the first 3 input feature sets and a binary phenotype

Focusing on co-amoxiclav WGS resistance prediction with a binary phenotype as the outcome and input features constructed from mapping onto resistance databases, **Figure 37** compares the same machine learning algorithms in terms of the same performance metrics across the same three feature sets. Performance is significantly worse than that for amoxicillin (note change in scale on **Figure 37** (y-axis minimum 0.75) compared with **Figure 36** (y-axis minimum 0.95)). Adding relative coverage to the 'known' input feature set improved specificity, without a corresponding decrease in sensitivity, and therefore improved accuracy (and AUC) across all machine learning models – however, sensitivity and specificity were still only 80-85%. Adding all transmissible elements in Resfinder to 'known + relative coverage' did not affect performance when looking across all algorithms, despite adding an extra 4176 features to only 203 in the 'known + relative coverage' feature set.

This is probably as these new features added little information. However, a decrease in performance might be expected, given the number of isolates n=976 is more than 4 times smaller than the number of features (p=4379) in this case. Across the 3 input feature sets adaboost achieved higher sensitivities (~85%), but at the cost of lower specificity (~75%) and therefore lower accuracy (~80%) too. The best performance across the 4 metrics was achieved with the input feature set 'known + relative coverage' with models SVM with RBF kernel and random forest. These models had better performance across the four metrics on the 'known + relative coverage' (p=203) versus 'known + rc + resfinder' (p=4379) feature sets, illustrating how feature selection can improve performance.

Focusing on the 'known + rc + resfinder' input feature set, for amoxicillin adaboost performed as well or outperformed all other models for accuracy, sensitivity and specificity. However, looking at co-amoxiclav this was no longer the case, illustrating that there is not one best model for all questions.

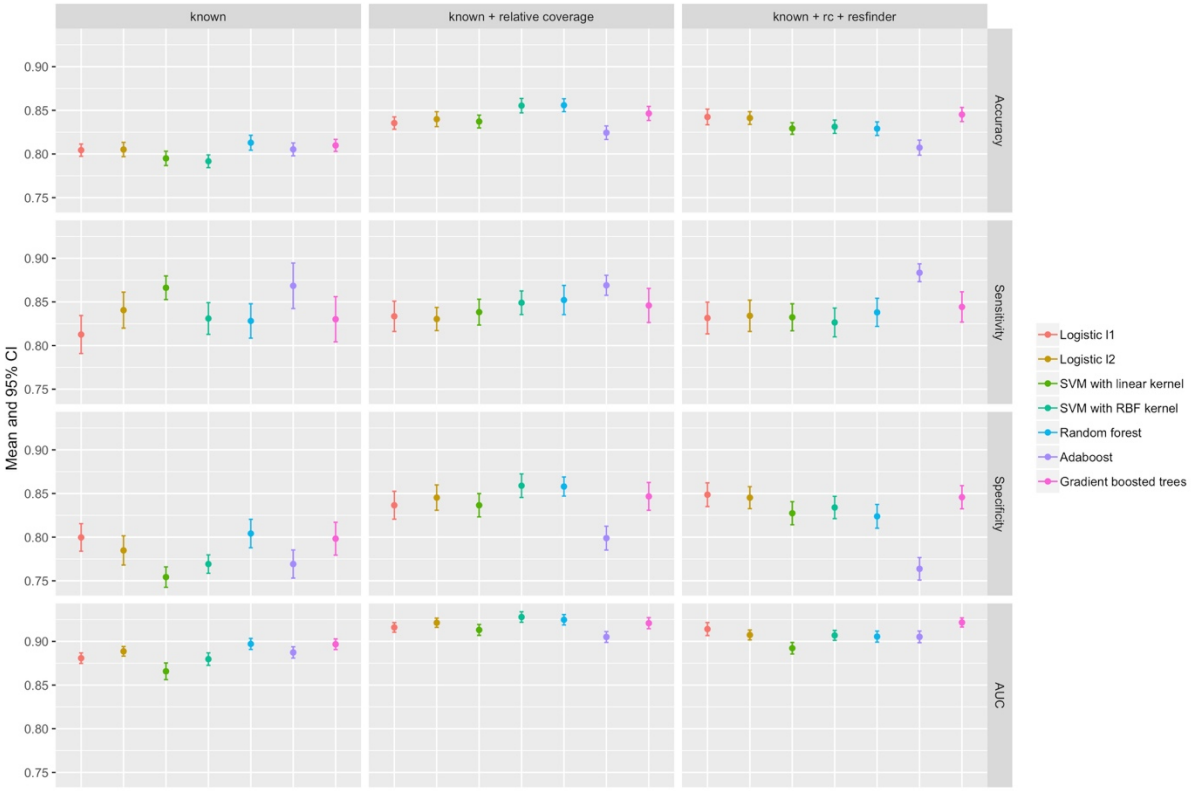


Figure 37 Co-amoxiclav resistance prediction performance of various machine learning classifiers applied to the first 3 input feature sets and a binary phenotype

3.3.3 Bacterial genome-wide association study

I first performed PCA on the patterns of presence and absence of the 31-mers. **Figure 38** depicts the leading 3 principal components on a phylogeny generated from core single-nucleotide polymorphisms identified from mapping sequences to the CFT073 genome. These correspond to major lineages in the phylogeny, as has been previously observed (Earle et al. 2016; McVean 2009).

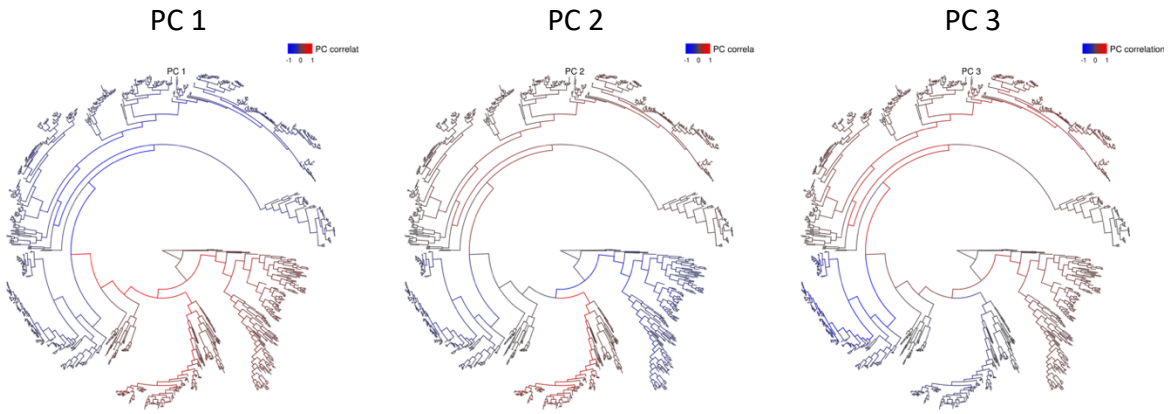


Figure 38 Leading principal components correspond to major lineages in bacterial phylogenies

Note: Phylogenetic tree of the 976 samples based on a core SNP alignment from mapped genomes corrected for recombination using ClonalFrameML (Didelot and Wilson 2015); units are substitutions per genome. Tree was built by Sam Lipworth. Branches of the tree are coloured according to their correlation with PC 1 (left), PC 2 (middle) and PC 3 (right). Positive correlations are shown in red, while negative correlations are shown in blue, with brighter colours representing stronger correlations. Each principal component appears to trace a path through the phylogenetic tree, separating one lineage (red branches) from another (blue branches). A square root transform was applied to the branch lengths to visualise fine-scale structure at the tips of the phylogenies.

Figure 39 presents the significance of lineage associations of *E. coli* principal components with co-amoxiclav resistance for the top 20 most significant PCs according to my Wald test. Only PCs 32 and 3 were below the Bonferroni-corrected significance threshold of 0.05/976. However, no branch in the phylogeny was most strongly correlated with PC 32. **Figure 40** illustrates the *E. coli* phylogeny annotated by co-amoxiclav resistance and principal component projections of PC 3. The branches of the tree with the strongest correlations reveal that PC 3 represented the split between ST131 and other ST types. ST131 was the second most common sequence type in my dataset; it had the highest percentage of phenotypically co-amoxiclav-resistant isolates and is a well-recognised drug-resistant lineage. However, co-amoxiclav resistance was otherwise distributed throughout the phylogenetic tree, indicating that acquisition would have occurred in the past in highly drug-

resistant lineages such as ST131 but also through horizontal gene transfer between lineages where resistance was more variably present.

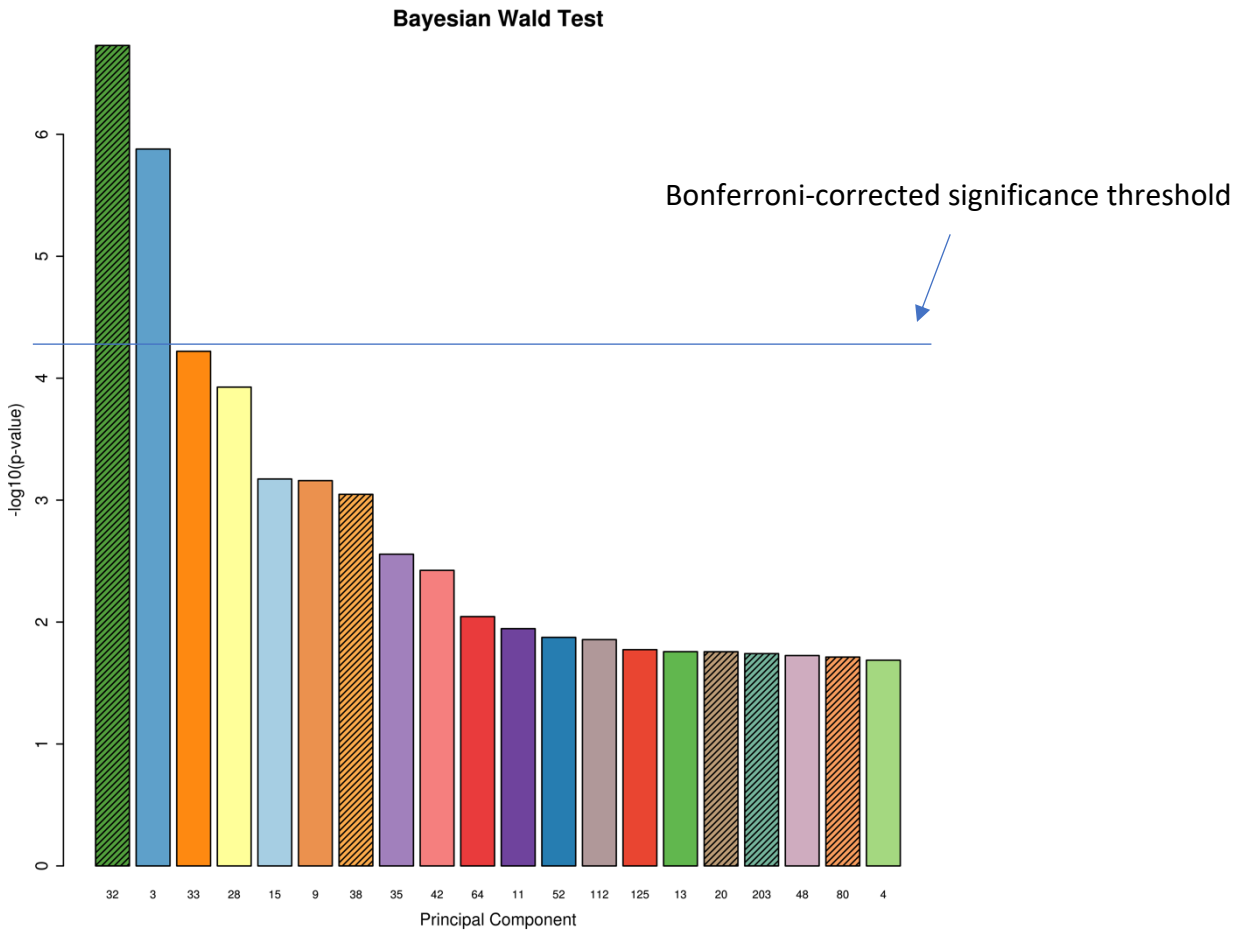


Figure 39 Wald test of significance of lineage associations of E. coli principal components with co-amoxiclav resistance

Note: The first 20 most significant PCs are shown. PCs 32, 38, 20, 203 and 80 are hashed to indicate that no branch in the phylogeny was most strongly correlated with this PC. Only PC-32 and PC-3 were significantly associated with co-amoxiclav resistance after correcting for multiple testing, Bonferroni-corrected significance threshold 0.05/976.

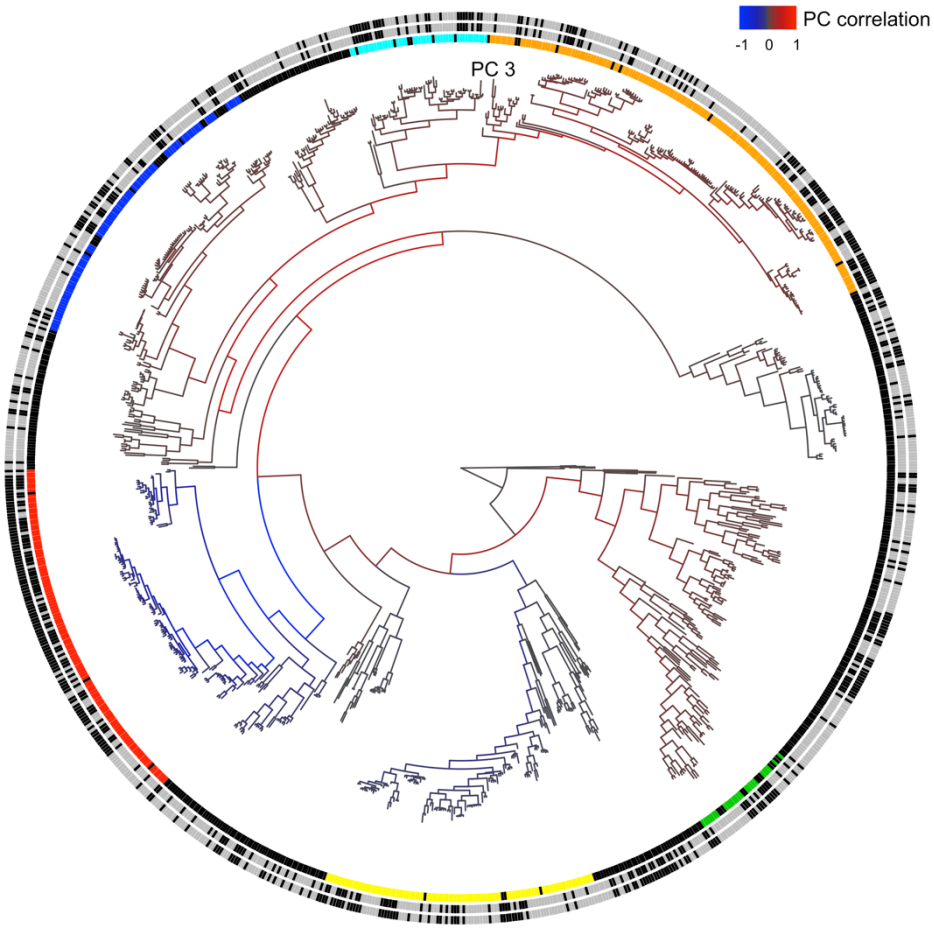


Figure 40 The E. coli phylogeny annotated by co-amoxiclav resistance and principal component projections of PC 3

Note: PC 3 is the one PC which was above the Bonferroni-corrected significance threshold and was also most strongly correlated with a branch in the phylogeny. Tips are annotated from inner circle to outer circle by: ST type with ST131 in red, ST73 in orange, ST69 in yellow, ST95 in blue, ST12 in cyan, ST10 in green and all other ST types in black; co-amoxiclav resistance laboratory phenotype, grey for susceptible and black for resistant; co-amoxiclav resistance predicted phenotype (same colours). The branches of the tree are coloured according to their correlation with the projections of the isolates onto PC 3. Positive correlations are coloured red and negative correlations blue, with brighter colours representing stronger correlations. As above, PC 3 appears to trace a path through the clonal genealogy, contrasting one lineage (red branches) with another (blue branches) enabling the interpretation of what the significant lineage effect represents.

3.3.3.1 Binary phenotype prediction using GWAS

Figure 41 depicts the maximum likelihood phylogeny annotated with the first 20 most significant PCs and the true and predicted phenotypes, focusing on a binary phenotype. The LMM had 90% accuracy, 95% sensitivity, 88% specificity, where the LMM coefficients were estimated based on all 976 isolates and predictions were made on the same 976 isolates. In

contrast, using cross-validation the machine learner/feature set combination with the highest accuracy, random forest on the 'known + relative coverage' feature set had accuracy 86% [95% CI (85%, 86%)] sensitivity 85% [95% CI (84%, 87%)], specificity 86% [95% CI (85%, 87%)] and AUC 92% [95% CI (92%, 93%)].

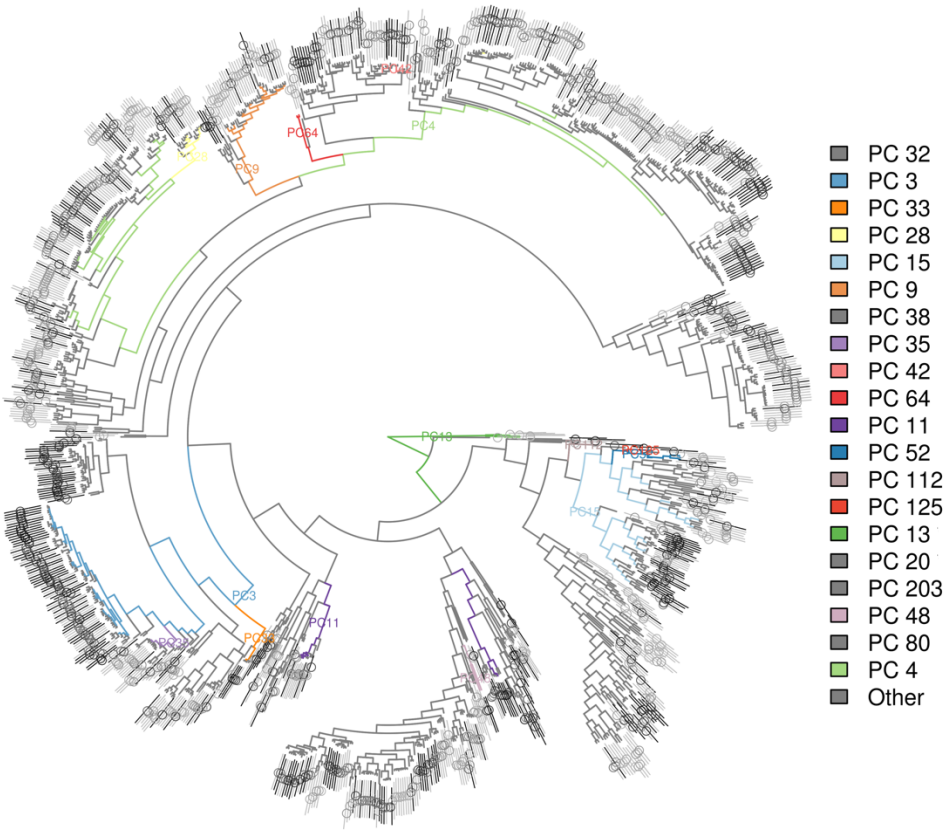


Figure 41 Phylogenetic phenotype prediction distribution

Note: Maximum likelihood phylogeny annotated with the first 20 most significant PCs and the true and predicted phenotypes. Branches are coloured by the PC they were most strongly correlated with. Lines represent true phenotype, circles represent LMM predicted phenotype, with light grey as susceptible and dark grey as resistant.

Figure 42 represents the mapping of the 31-mers to reference genomes *E. coli* CFT073, and two examples of ST131 genomes, specifically: Ecol_244, SAMN05511157 - fimH22 and

Ecol_AZ146, SAMN05511152, fimH30-Rx from a genomic epidemiological study on carbapenemase-producing *E. coli* (Stoesser et al. 2017). Considering the *E. coli* CFT073 complete reference genome, all 31-mers which were above the Bonferroni-corrected significance threshold were unmapped. Looking across all three reference genomes, 31-mers which were just below the Bonferroni-corrected significance threshold mapped onto the *ampC* promoter region of each reference genome. The two ST131 reference genomes contain mobile genetic elements and multiple resistance genes. Therefore, as expected, **Figure 42** indicates multiple 31-mers above the significance threshold mapping in different places to the two ST131 reference genomes which contain these mobile genetic elements, reducing the number of unmapped reads (far left "peak"). Further Ecol_AZ146 is known to be multi-drug resistant and therefore the number of unmapped 31-mers is much lower for this genome (far right panel), and also the significance level of these unmapped k-mers is decreased.

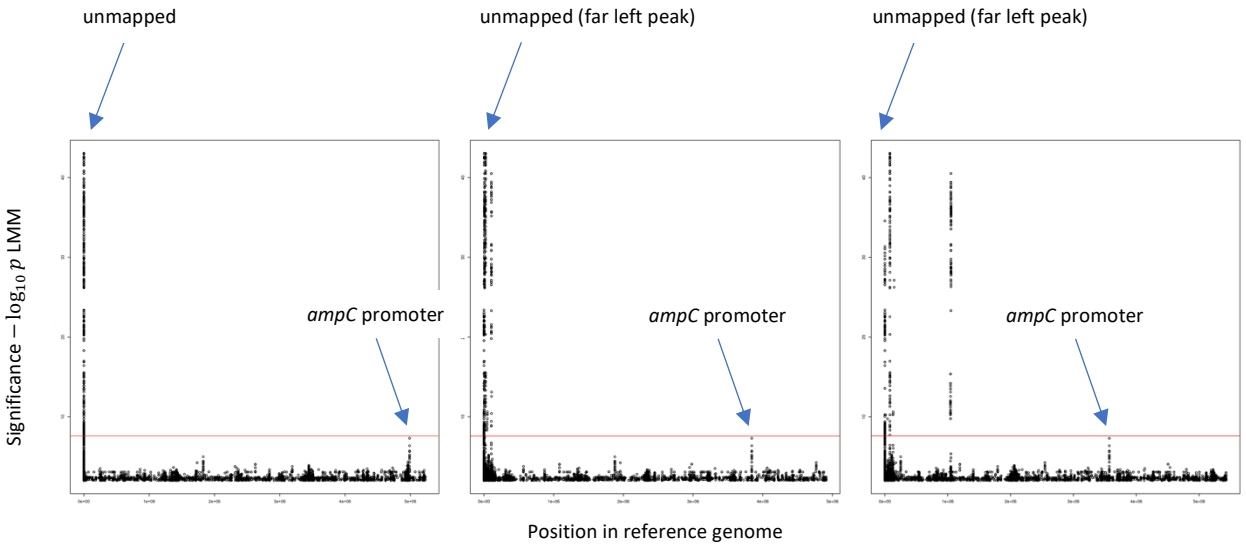


Figure 42 Significance of association of the presence/absence of 31-mers with co-amoxiclav resistance after controlling for population structure using LMM, mapped to E. coli CFT073 complete genome, and two ST131 candidates

Note: The two ST131 candidates are: Ecol_244, SAMN05511157 - fimH22 and Ecol_AZ146, SAMN05511152, fimH30-Rx, with data for their complete genomes being available at https://www.ncbi.nlm.nih.gov/bioproject/?term=316786. All 31-mers from patterns with significance below

10^{-2} are plotted. The red line denotes the Bonferroni-corrected significance threshold accounting for the number of 31-mer patterns of presence and absence tested. The x-axis is the position in the reference genome (ie differs in each panel), while the y-axis is the $-\log_{10}$ p of each individual 31-mer.

3.3.3.2 MIC prediction using GWAS

Using GWAS to predict $\log_2$(MIC), **Table 11A**, **Table 12A**, **Table 13A** compare observed co-amoxiclav laboratory phenotype with the MIC predicted from genotype, dealing with left and right censoring in the three different ways described in methods, while **Table 11B**, **Table 12B**, **Table 13B** present the respective number of isolates with all possible $\log_2$(MIC) differences between predicted genotype and laboratory phenotype. When MIC values of ≤2/2 and >32/2 mg/L were set to 2 and 64 respectively in the GWAS model (first method), the essential agreement was 875/976 (90%), while the accuracy was only 332/976 (34%). When MIC values of ≤2/2 and >32/2 mg/L were set to 1 and 128 respectively (second method), there was a 13% increase in accuracy (to 460/976 (47%)), but no increase in essential agreement (874/976 (90%)). When MIC values of ≤2/2 and >32/2 mg/L were set to 1/2 and 256 respectively (third method), there was a further 5% increase in accuracy (507/976 (52%)), but a 2% decrease in essential agreement (857/976 (88%)).

I also reran *bugwas* excluding isolates with sequence type ST131, the second most common strain type, but with the highest percentage of phenotypically-resistant isolates, to explore whether this would impact prediction performance. Results are presented in **Table 14A&B**, setting MIC values <2/2 and >32/2 mg/L to 1/2 and 256 mg/L respectively (third method). Essential agreement was essentially the same (756/853 (89%)); however, accuracy dropped (343/853 (40%)) despite choosing to deal with the left and right censoring with the method which obtained the highest accuracy when using at the whole dataset.

Table 11 Co-amoxiclav phenotype and predicted genotype from GWAS using log_2(MIC) as the outcome and first strategy for interval censoring

(A) Concordance

Phenotype (mg/L)	Genotype prediction (mg/L) ≤2/2	4/2	8/2	16/2	32/2	>32/2
≤2/2	10	119	36			
4/2		114	166	20		
8/2		1	71	85	11	
16/2			12	55	22	
32/2			1	23	25	10
>32/2			6	27	101	57

Note: MICs ≤2/2 and >32/2 mg/L set to 2/2 and 64/2 mg/L respectively.

(B) Number of isolates with log_2(MIC) difference between predicted genotype from GWAS and phenotype

Log_2(MIC) difference	-3	-2	-1	0	1	2	3
Number of isolates	6 (0.6%)	28 (3%)	137 (14%)	332 (34%)	402 (41%)	67 (7%)	4 (0.4%)

Note: Essential agreement 875/976 (90%), accuracy 332/976 (34%).

Table 12 Co-amoxiclav phenotype versus predicted genotype from GWAS using log_2(MIC) as the outcome and second strategy for interval censoring

(A) Concordance

	Genotype prediction (mg/L)					
Phenotype (mg/L)	≤2/2	4/2	8/2	16/2	32/2	≥64/2
≤2/2	84	60	20	1		
4/2	2	134	137	23	8	
8/2		2	68	77	20	1
16/2		2	13	45	29	
32/2			1	23	21	14
≥64/2			5	21	57	108

Note: MICs ≤2/2 and >32/2 are set to 1/2 and 128/2 mg/L respectively.

(B) Number of isolates with log_2(MIC) difference between predicted genotype from GWAS and phenotype

Log_2(MIC) difference	-3	-2	-1	0	1	2	3
Number of isolates	5 (0.5%)	24 (3%)	97 (10%)	460 (47%)	317 (32%)	63 (6%)	10 (1%)

Note: Essential agreement 874/976 (90%), accuracy 460/976 (47%).

Table 13 Co-amoxiclav phenotype versus predicted genotype from GWAS using log_2(MIC) as the outcome and third strategy for interval censoring

(A) Concordance

Phenotype (mg/L)	Genotype prediction (mg/L) ≤2/2	4/2	8/2	16/2	32/2	≥64/2
≤2/2	110	38	16	1		
4/2	16	145	105	28	7	3
8/2		7	64	67	26	4
16/2	2	2	13	39	27	6
32/2			2	21	16	20
≥64/2			4	18	36	133

Note: MICs ≤2/2 and >32/2 are set to 1/2 and 256 mg/L respectively.

(B) Number of isolates with log_2(MIC) difference between predicted genotype from GWAS and phenotype

Log_2(MIC) difference	-3	-2	-1	0	1	2	3	4
Number of isolates	6 (0.6%)	22 (2%)	93 (10%)	507 (52%)	257 (26%)	76 (8%)	12 (1%)	3 (0.3%)

Note: Essential agreement 857/976 (88%), accuracy 507/976 (52%).

Table 14 Co-amoxiclav phenotype versus predicted genotype from GWAS using log_2(MIC) as the outcome (third interval censoring strategy): excluding ST131 E. coli bloodstream infections

(A) Concordance

Phenotype (mg/L)	Genotype prediction (mg/L) ≤2/2	4/2	8/2	16/2	32/2	≥64/2
≤2/2	68	64	27	2		
4/2	1	81	172	23	2	
8/2		1	52	73	19	2
16/2		2	9	36	26	2
32/2			1	17	15	10
≥64/2			4	13	40	91

Note: MICs ≤2/2 and >32/2 are set to 1/2 and 256 mg/L respectively.

(B) Number of isolates with log_2(MIC) difference between predicted genotype from GWAS and phenotype

Log_2(MIC) difference	-3	-2	-1	0	1	2	3
Number of isolates	4 (0.4%)	16 (2%)	68 (7%)	343 (35%)	345 (35%)	71 (7%)	6 (0.6%)

Note: Essential agreement 756/853 (89%), accuracy 343/853 (40%).

3.3.4 Bacterial genome-wide association study as a feature selection tool for binary prediction

Focusing on co-amoxiclav WGS resistance prediction with a binary phenotype as the outcome, **Figure 43** compares machine learning algorithms in terms of accuracy, sensitivity, specificity and AUC considering input features 'known' (p=179), 'known + relative coverage' (p=203), 'known + rc + resfinder' (p=4379) as presented in Section 3.3.2, but also alongside k-mer based input features, specifically the k-mer patterns with p<Bonferroni corrected significance threshold (p=326, denoted 'k-mer326'), or $p<10^{-2}$ (p=5561, denoted 'k-mer5561'). The input feature set 'k-mer326' had a similar performance to 'known' input features across the machine learning algorithms; whilst they had similar sizes, one was obtained from mapping WGS onto resistance databases, while the other only relied on 31-mers from assemblies significantly associated with outcomes in (univariate) GWAS.

Copy number is not accounted for in the 31-mers input features, since these are based only on patterns from assemblies. Despite this, looking across all input features, input feature set 'k-mer5561' had higher prediction performance across all metrics across all individual machine learning algorithms, despite being the largest input feature set which could decrease performance given the number of isolates (only 976). Overall, the best performing model on the 'k-mer5561' feature set was l2-regularised logistic regression with accuracy 89% [95% CI (88%, 89%)], sensitivity 88% [95% CI (87%, 90%)], specificity 89% [95% CI (87%, 90%)] and AUC 96% [95% CI (95%, 96%)] (with errors estimated out-of-sample using cross-validation), compared to the original GWAS with accuracy 90%, sensitivity 95% and specificity 88% which was still therefore better overall (although errors were estimated in-sample).

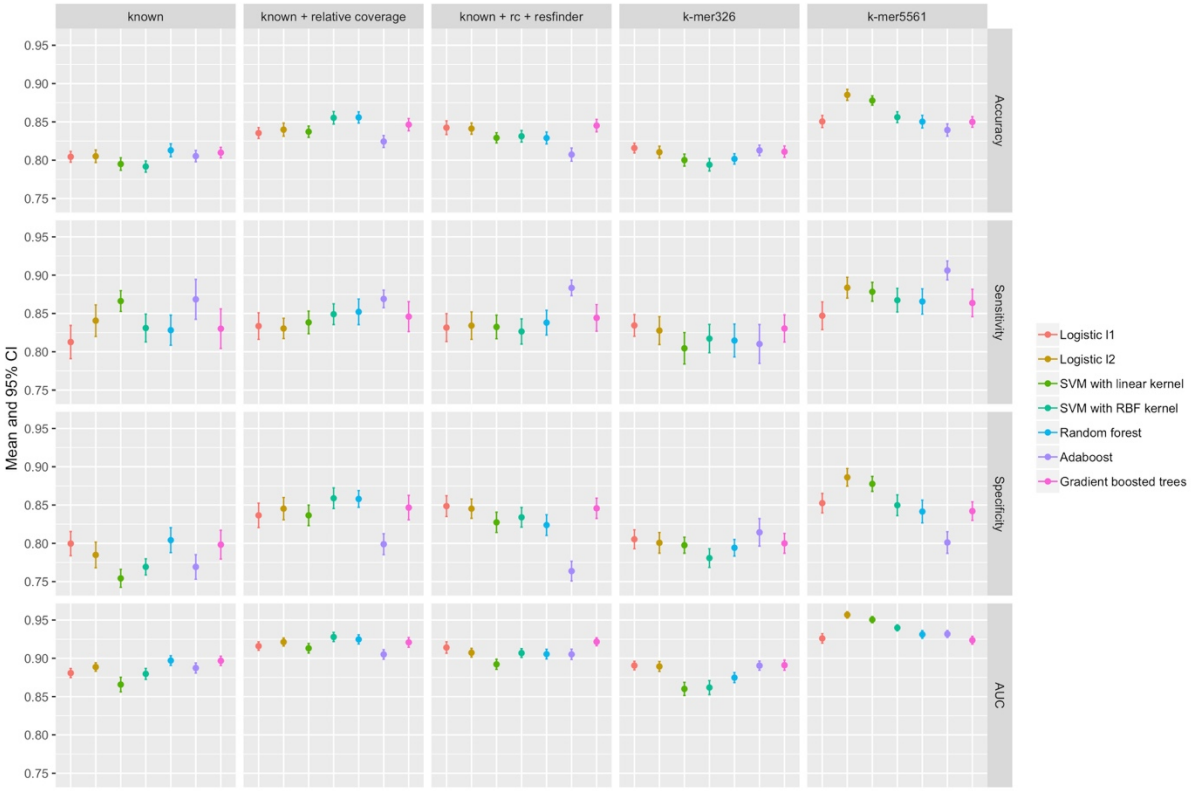


Figure 43 Co-amoxiclav resistance prediction performance of various machine learning classifiers applied to all 5 input feature sets and a binary phenotype

3.3.5 Prediction of minimum inhibitory concentration

Focusing on co-amoxiclav WGS MIC prediction considering all 5 input feature sets, that is the input features constructed from mapping onto resistance databases, but also input features constructed from 31-mers, **Figure 44** compares machine learning algorithms in terms of essential agreement and accuracy, and **Figure 45** in terms of sensitivity and specificity. The specific input feature sets were 'known' (p=179), 'known + relative coverage' (p=203), 'known + rc + resfinder' (p=4379), but also 'k-mer280' and 'k-mer5088', specifically the k-mer patterns with p<Bonferroni-corrected significance threshold and p<0.01 (-$\log_{10}$(p)>2)respectively. The size of the input feature sets differs from the preceding section as these are from GWAS with MIC as the response variable (section 3.2.4).

Similarly to the performance for the binary output, when considering the MIC, input feature set ‘k-mer280’ had a similar performance to the ‘known’ input feature set across the machine learning algorithms. Focusing on essential agreement as the performance metric, considering relative coverage in addition to presence/absence considerably improved performance compared with presence/absence alone across all the machine learning algorithms. In particular, the lowest performing algorithm when considering ‘known + relative coverage’, Ordinal IT, still had higher mean essential agreement, 88% [95% CI (88%,89%)], than the highest performing one when considering only the ‘known’ feature set, Ordinal Ridge 87% [95% CI (86%,88%)] (essential agreement was almost identical for Ordinal AT and Ordinal SE).

Essential agreement across the machine learning algorithms using the ‘known + relative coverage’ feature set illustrates clearly how very informative features can considerably improve performance regardless of the choice of model, although different models can still increase/decrease performance (e.g. Ordinal SE has higher accuracy (92% [95% CI 92%,93%]) than Ordinal IT (88% [88%,89%]) and Multinomial l1 (90% [89%,90%])).

Adding all the transmissible elements in Resfinder with their quality metrics led to a considerable decrease in essential agreement only for the Multinomial l2 and the random forest classifiers, while the performance of the other models was not affected by the considerably higher number of features.

Input feature set ‘k-mer5088’ illustrates how model choice can increase performance, with Ordinal AT, IT and SE having considerably higher essential agreement compared to all other

models for this feature set, whereas differences between models were less extreme with other feature sets.

Overall, the best performing model in terms of essential agreement was Ordinal SE (which penalises predictions further away from the truth in the most extreme manner) with the 'known + relative coverage' feature set. This classifier had essential agreement 92% [95% CI 92%,93%], accuracy 58% [57%,59%] (categorical agreement), sensitivity 74% [72%, 75%] and specificity 98% [97%, 98%] (the latter two measures against the predicted binary phenotype). Ordinal SE with the 'k-mer5088' feature set had fairly similar performance, with essential agreement 92% [91%, 92%], accuracy 55% [54%, 55%], sensitivity 78% [76%, 79%] and specificity 94% [94%, 95%]. In comparison, using GWAS essential agreement was 90% with an accuracy of only 47%. In addition, using GWAS the model fitting and predictions were both done on the whole dataset and performance would therefore be expected to decrease on an independent dataset of isolates, whereas the machine learners estimated performance metrics based on cross-validation (i.e. in a different sub-sample to those used to fit the models).

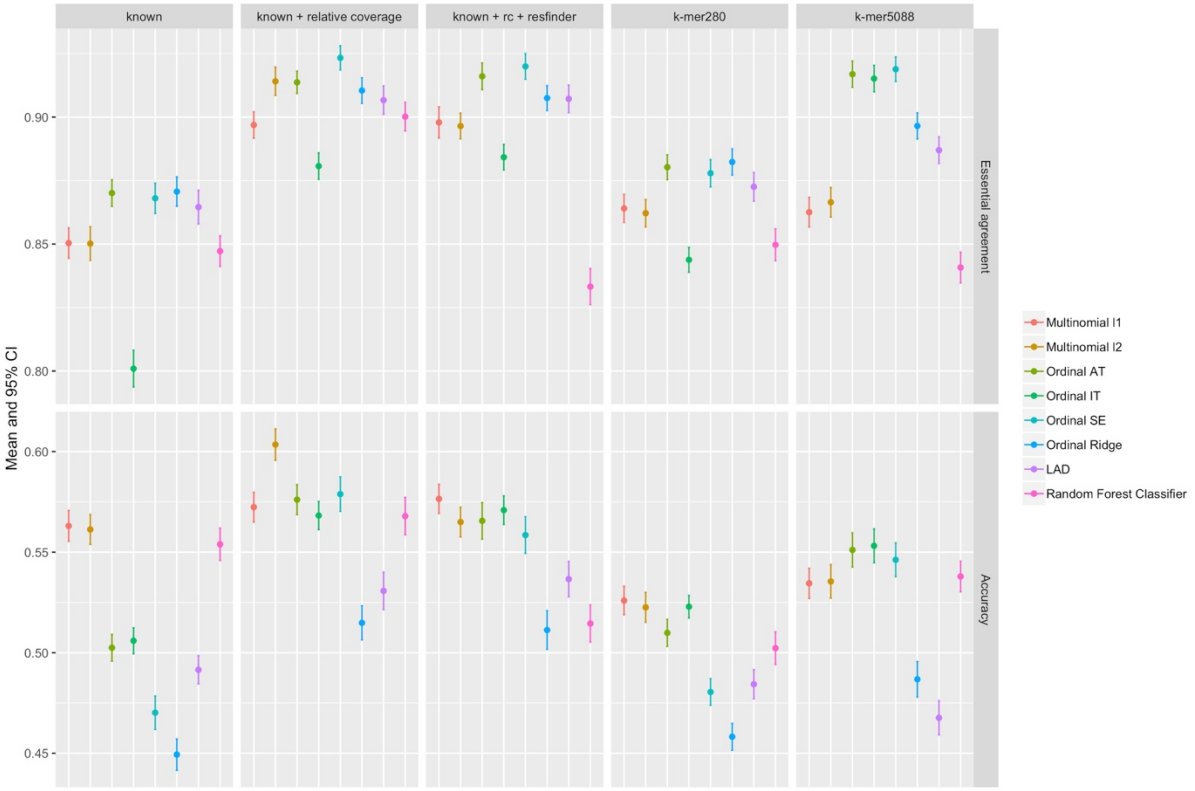


Figure 44 Co-amoxiclav MIC prediction performance of various machine learning classifiers applied to all 5 input feature sets: essential agreement and accuracy (categorical agreement)

Note: Showing mean and 95% confidence intervals.

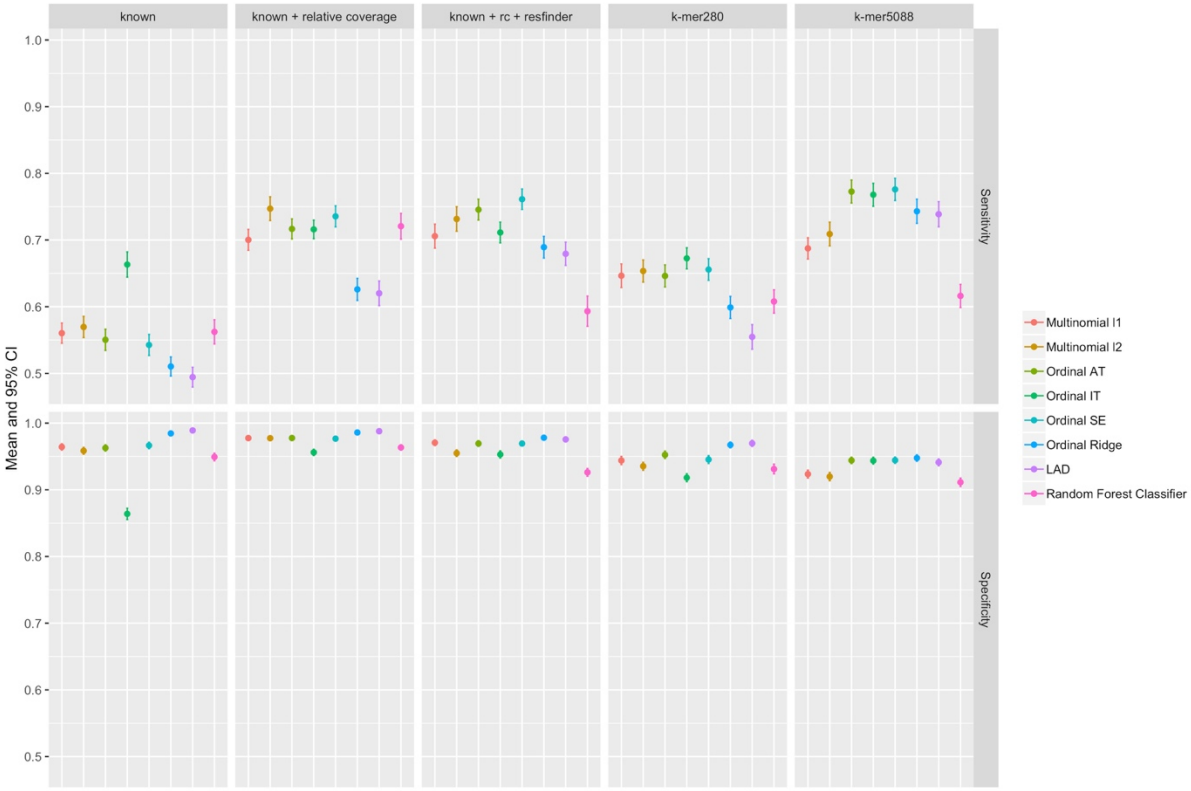


Figure 45 Co-amoxiclav MIC prediction performance of various machine learning classifiers applied to all 5 input feature sets: sensitivity and specificity

Note: Showing mean and 95% confidence intervals. Sensitivity and specificity obtained by dichotomising the outcome using the EUCAST breakpoint, i.e. predicted MIC>8/2 mg/L resistant, predicted MIC≤8/2 mg/L susceptible.

3.3.6 Feature importance

When considering relative feature importance, I chose to focus on predicting MIC and on the top performing model Ordinal SE. **Table 15** and **Table 16** present the top 20 features from the 'known + relative coverage' feature set obtained using Ordinal SE as my classifier and either the squared weights of the features or recursive feature elimination for ranking, respectively. The 2 ranking methods agreed on 17 out of the top 20 features.

Table 15 Top 20 features for co-amoxiclav MIC prediction from 'known + relative coverage' feature sets using the Ordinal SE classifier and the squared weights of the features for ranking

Feature	Squared weight	Type of feature
ampC promoter mutation -32 T>A	56.1	mutation
ampC promoter insertion -15_-14 T	41.6	insertion
ampC promoter mutation – 42 C>T	32.5	mutation
ampC promoter insertion -16_-15InsG	32.3	insertion
TEM promoter mutation C37T	31.4	mutation
ampC promoter mutation – 11 C>T	28.0	mutation
$bla_{CTX\text{-}M\text{-}15}$ mutation Ser133Gly	24.3	mutation
$bla_{OXA\text{-}1}$ presence/absence	18.5	presence/absence
$bla_{CMY\text{-}2}$ presence/absence	18.0	presence/absence
$bla_{TEM\text{-}30}$ relative coverage	14.6	relative coverage
ampC promoter insertion -24InsT	11.1	insertion
$bla_{TEM\text{-}1D}$ presence/absence	10.7	presence/absence
***TEM* promoter mutation G167T**	9.6	mutation
$bla_{OXA\text{-}1}$ relative coverage	9.5	relative coverage
ampC promoter insertion -17_-16InsC	8.0	insertion
ampC promoter mutation -17 T>C	8.0	insertion
$bla_{TEM\text{-}34}$ relative coverage	7.9	relative coverage
$bla_{CTX\text{-}M\text{-}55}$ relative coverage	6.9	relative coverage
$bla_{TEM\text{-}1C}$ presence/absence	6.9	presence/absence
$bla_{CMY\text{-}2}$ mutation Cys127Thr	6.0	mutation

Note: **Bold** means feature in top 20 using the squared weights of the features for ranking, but not when using recursive feature elimination for ranking, see **Table 16**.

Table 16 Top 20 features for co-amoxiclav MIC prediction from 'known + relative coverage' using the Ordinal SE classifier and recursive feature elimination for ranking

TEM promoter mutation C37T

ampC promoter mutation -32 T>A

ampC promoter mutation -42 C>T

bla_{OXA-1} relative coverage

ampC promoter mutation -11 C>T

ampC promoter insertion -15_-14InsT

ampC promoter insertion -16_-15InsG

ampC promoter insertion -17_-16InsC

$bla_{CTX-M-15}$ mutation Ser133Gly

***ampC* promoter deletion 93DelA**

bla_{TEM-1D} presence/absence

bla_{TEM-30} relative coverage

bla_{CMY-2} presence/absence

$bla_{CTX-M-55}$ relative coverage

bla_{TEM-34} relative coverage

ampC promoter mutation -17 T>C

ampC promoter insertion -24InsT

***TEM* promoter deletion 38DelT**

bla_{TEM-33} relative coverage

bla_{OXA-1} presence/absence

Note: **Bold** means feature in top 20 using recursive feature elimination for ranking, but not when using the squared weights of the features for ranking, see **Table 15**.

3.4 Discussion

In this Chapter I aimed to explore various machine learning algorithms in order to predict antimicrobial susceptibility of *E. coli* bloodstream infections from WGS. I focused on co-amoxiclav for two reasons. Firstly, because even though it is one of the most commonly used antibiotics in England (Public Health England 2017), I have shown in Chapter 2 that it is the only antibiotic for which the number of resistant isolates is rising regardless of previous hospital-exposure. Secondly, defining clinical susceptibility is particularly challenging for this beta-lactam/beta-lactamase inhibitor combination with different methods yielding different results (Díez-Aguilar et al. 2015; Leverstein-van Hall et al. 2013). Looking at the co-amoxiclav MIC distribution within my dataset, the breakpoint of 8/2 mg/L seems to have been somewhat arbitrarily chosen, as in contrast to other simpler antibiotics whose MIC distributions can clearly be dichotomised with many isolates being observed on either side of an obvious breakpoint, but not at or around the breakpoint itself, the co-amoxiclav MIC distribution is instead more similar to a bell curve. I therefore explored whether WGS is a more reliable and reproducible way of defining co-amoxiclav susceptibility.

3.4.1 Choice of input features

One of the largest studies considering prediction of antimicrobial susceptibility from WGS using machine learning methods looked at 13402 *Mycobacterium tuberculosis* isolates for 11 antibiotics (Kouchaki et al. 2018). Their matrix of features was created by aligning each isolate to a reference strain and considering SNPs in genes known to be associated with resistance of one or more drugs because *Mycobacterium tuberculosis* does not present the challenge of an accessory genome and therefore gene presence/absence effects are not relevant. However, this approach was not suitable when considering *E. coli* due to the

accessory genome being known to carry a large number of elements and genes with an important role in resistance. Therefore, instead of mapping onto a reference genome I considered two different approaches for the construction of my matrix of features: mapping onto databases of known resistance elements and creating agnostic patterns of presence/absence of 31-mers.

Feature sets based on the reference database performed similarly to those based on 31-mers; however, the advantage of the k-mer approach is that it does not rely on a curated database of genes which is time consuming to maintain and keep up to date, and highly prone to errors. Keeping a resistance database up to date is particularly challenging for *E. coli* because single point mutations in a gene can change phenotypic expression and because new resistance genes are being continuously imported into the species from other Enterobacteriaceae due to horizontal gene transfer. Models based on 31-mers can also provide guidance on novel sequence features for subsequent experimental testing of resistance mechanisms.

When constructing my matrix of features, I considered the presence/absence of specific elements even if the element had a mutation – that is if a mutated *bla-TEM*$_{123}$ was present, I considered both the original element as a feature and the mutation as a feature. My models therefore effectively considered whether the mutation added any additional predictive power over the main element. An alternative would be to create my matrix only from elements which help differentiate the isolates; presence/absence of specific elements could be set to 1 if the element was present without any mutations and 0 if any mutations were observed. There are advantages and disadvantages to both approaches: where mutations

are rare and do not change the expressed phenotype, my approach should be more powerful as I would generally incorporate the effect of the gene being present/absent regardless of mutation. However, if small changes drastically effect phenotype, I might miss these effects.

Feature sets 'k-mer326' (binary phenotype) and 'k-mer280' (MIC) which only included patterns of presence and absence above the Bonferroni-adjusted significance threshold from GWAS had low performance, similar to that of the 'known' feature set which only included presence/absence of elements without relative coverage. However, as 31-mers mapping onto the *ampC* promoter region were just below the Bonferroni-corrected significance threshold, this low performance might be expected given the known importance of mutations within the *ampC* promoter region for resistance (Richet et al. 1999). Performance was considerably improved by considering larger k-mer feature sets, with predictions performing similarly to the 'known + relative coverage' feature set with MIC as the outcome and having the best performance with a binary phenotype as the outcome. My work shows that bigger feature sets may be a way to get good performance without having to rely on curated databases.

Focusing on the k-mer input features, one of the main limitations in terms of the sequencing data is working with k-mers taken from the assemblies, rather than from the reads directly. This is because assemblies are known to make certain decisions and simplifying assumptions in collapsing reads and therefore might be dropping crucial information in terms of resistance prediction. For example, there may be multiple copies of closely related beta-lactamase genes with different effects on resistance in an isolate which are collapsed in the

assembly into one gene. However, working with the reads directly comes with its own challenges, namely computational requirements and more errors. Specifically, the error profile of reads is known to vary substantially across them. Low-quality regions at end of reads would have to be dropped, and low frequency k-mers would have to be dropped as these are more likely to represent errors. When counting k-mers from assemblies, these were deemed to be present if they had a count of one or higher. However, if I were to count k-mers from reads, then errors in the reads could be accounted for by only deeming k-mers to be present if the count was above a certain threshold. If this was to be done, the counts would likely have to be normalised, for example by the total k-mer count per sample, and multiplied by 10^x, with x to be identified empirically across the isolates, and then dropped if the rate of k-mer observation per total-k-mers observed in sample was below some threshold, for comparability across samples. Including depth/copy number of k-mers in this way may however lead to a significant improvement in performance, because it would incorporate the effects of multiple similar resistance-conferring genes within a sample, similar to the one obtained when adding relative coverage as features in addition to the simple presence/absence features. The fact that hyperexpression of beta-lactamases plays an important role in co-amoxiclav resistance has been shown in previous studies (T. J. Davies et al. 2019); it has also been shown that it can increase MIC for another beta-lactam/beta-lactamase combination piperacillin-tazobactam (J. Kim et al. 2019). Hyperexpression of certain genes (e.g. ampC) can be indirectly represented by copy number (which I should be able to detect with DNA sequencing, at least approximately). However, other mechanisms of resistance such as efflux pumps represent a response to the antibiotic being in the cell that is dependent on the expression of RNA or other messenger molecules. Whilst these could be measured by RNA-seq, proxies for them are not present in genomic

DNA and I could therefore not assess them. I found one study that had investigated predicting antimicrobial susceptibility in *Pseudomonas aeruginosa* using genomic and transcriptomic markers; it found that for 3 out of 4 antibiotics, gene expression values improved susceptibility predictive performance (Khaledi et al. 2019). Among their most predictive features were also presence of multidrug efflux pumps such as *amrB (mexY)*. However, RNA-seq is at the moment too expensive for routine clinical practice.

Performing feature selection can reduce overfitting and improve accuracy due to less noise guiding the decisions made in the prediction algorithms. Moreover, it can also decrease training times, which is particularly useful when considering very large feature sets such as those constructed from k-mers (even more so k-mers from reads). In this work I chose one feature selection technique when looking at 31-mers based on GWAS, but there are many more I could have considered (Saeys, Inza, and Larrañaga 2007). Given biological knowledge, it is worth focusing particularly on those which could account for interactions, that is when the presence of one gene can lead to one phenotype, which however in the presence of another gene can lead to the suppression of the effect or they combine and produce a different phenotype. Future work should therefore explore various feature selection methods first when working with k-mers, but also when looking at the 'known + relative coverage' feature set.

A major choice in feature selection when considering k-mers is the length of a k-mer. A k-mer length of 31 has been chosen for bacterial GWAS in the past based on performance following a sensitivity analysis (S. K. Sheppard et al. 2013); however, it is possible that my

model could improve if a sensitivity analysis for best k-mer length was carried out for my specific task.

3.4.2 Choice of output features

In my analysis the greatest improvement in performance was obtained by moving away from a binary phenotype as the outcome to considering the minimum inhibitory concentration as the outcome: this may be expected as the distribution of MICs of co-amoxiclav for *E. coli* bloodstream infections is clearly more normal than bimodal, and therefore the breakpoint of 8 mg/L arbitrarily splits a single distribution. Davies et al performed EUCAST agar dilution and CLSI agar dilution on a subset of 261 isolates from my dataset (160 [61%] co-amoxiclav resistant) and obtained an overall agreement of only 85% between these two reference phenotyping methods considering a binary susceptible/non-susceptible split categorisation (T. J. Davies et al. 2019). They also showed that the reproducibility of the laboratory results was very poor, with 61% (158/261) of isolates having a change of at least one doubling dilution on repeating the EUCAST agar dilution test and 28% (73/261) on repeating the CLSI agar dilution test. This, together with my results from machine learning, emphasises the need to move away from a "resistant" vs "susceptible" paradigm for co-amoxiclav. EUCAST, for example, have introduced a "susceptible under increased exposure category" for certain antibiotics such as piperacillin-tazobactam when considering *Enterobacterales*, making an isolate with an MIC of 8/4 mg/L or less susceptible, with an MIC of 16/4 mg/L susceptible under increased exposure, and an isolate with an MIC greater than 16/4 mg/L resistant (European Committee on Antimicrobial Susceptibility Testing 2019). Even though the MIC distribution is continuous, the final decision is a dichotomous one: to use a specific antibiotic or not. In reality, the question that

needs to be answered is whether the antibiotic will treat the infection in the patient or not and training algorithms with outcome of successful treatment as the output features could be one way of moving away from the traditional “susceptible” vs “resistant” paradigm. Moreover, patient information which is generally used by infectious disease specialists in order to decide on treatment (e.g. previous antibiotic exposure, previous antibiograms, other risk factors) could be incorporated into the input features.

3.4.3 Choice of methods

I saw that for predicting co-amoxiclav MIC, Ordinal SE with the “known + relative coverage” feature set performed best with essential agreement 92% [95% CI 92%, 93%], accuracy 58% [95% CI 57%, 59%], sensitivity 74% [95% CI 72%, 75%] and specificity 98% [95% CI 97%, 98%]. Interestingly, this method is the one with the most extreme penalties the further the incorrect prediction is from the truth. Given the immediate clinical cost of predicting an isolate to be susceptible when it is truly resistant is higher than predicting an isolate to be resistant when it is truly susceptible (the former could lead to inappropriate treatment, while the latter could select for resistance), future work could consider minimizing the absolute difference between the false negative rate and the false positive rate when predicting MIC, and/or even prioritizing higher sensitivity.

One of the main challenges of machine learning algorithms is interpretability. Focusing on my best performing model, Ordinal SE, I considered two different methods in order to rank my features based on their importance for the model prediction. The first one was considering the weights of the final model as I would expect the highest weights to identify the features which are very powerful predictors of MIC. The second one was recursive

feature elimination as I would expect this to pick up those features which are most important in terms of model fit. Focusing on the top 20 features, 17 out of the 20 features were actually identical between these two approaches. Interestingly, the highest ranking *ampC* promoter mutations from the larger 'known + relative coverage' feature set, at positions -42 and -32, have been shown to directly enhance transcription of *ampC* (Richet et al. 1999; Olsson et al. 1983).

Focusing on feature importance, one of the main limitations of machine learning methods is that from highly correlated features, one will arbitrarily be chosen and then the rest discarded, as machine learning algorithms are designed to improve prediction performance rather than identify independent effects. Therefore, had I looked at this in other feature sets, e.g. that including all transmissible elements not just those associated with beta-lactamase resistance, I would have expected to find some associated with cross-resistance (Kouchaki et al. 2018; Yang et al. 2019) since machine learning is known to leverage confounding for better prediction. For example, looking at the multivariable model in (T. J. Davies et al. 2019), bla_{TEM} relative coverage is independently associated with changes in doubling dilution MIC, however, this is only ranked 101 by recursive feature elimination possibly due to $bla_{TEM\text{-}1D}$ presence/absence being ranked in the top 20.

In terms of the machine algorithms considered, hyperparameter tuning is known to be able to lead to increased performance. In this Chapter I only carried this out for SVM since this is computationally intensive and time consuming. However, now that the top performing sets ('known+relative coverage' and 'k-mer5088' with MIC as the outcome) and models (Ordinal

SE) have been identified, further work should focus on hyperparameter tuning of the regularization parameter, which may improve performance further.

Other machine learning algorithms I did not consider include extreme gradient boosting and deep learning: these have been shown to have a very good performance in many different classification problems and further work could compare their performance to that of the models considered in this Chapter. I did train some deep neural networks; I have not included their results here as their performance was disappointing. I consider this likely to be due to not enough time being spent on designing the architecture (number of layers, number of nodes, connections) and potentially due to the size of the dataset as deep neural networks are known to require massive amounts of data in order to be able to learn the thousands to millions of weights needed depending on the complexity of the architecture. Considering phenotype prediction from genetic sequences, at least one other study has also shown that 3-layer neural networks with equal number of nodes in each layer have poorer performance than classical machine learning algorithms such as extreme gradient boosting (J. Kim et al. 2019), supporting my experience. However, deep neural network often outperform other methods in terms of image classification, for example, where they have much more complex architectures with a much higher number of layers, varying number of nodes and different connections between the different layers (Krizhevsky, Sutskever, and Hinton 2012). These deep neural networks are generally trained on millions of images and in practice people will simply finetune the weights of these networks to improve performance on their specific research questions, rather than design new architectures from scratch as would be needed here for phenotype prediction. These architectures built for image classification exploit adjacency being informative (e.g. visually – lines and edges, or in time –

before and after). One of the challenges is therefore representing the information in WGS through images; but one concern is that adjacency is not found within this type of data because of the mobility of different resistance elements e.g. through transposition, plasmid exchange, horizontal gene transfer etc.

One limitation of my study is that the performance and generalisability of my results would ideally have been tested on a completely independent test set. However, in practice a separate independent test set is often not available. During cross-validation, the 5 folds are independent, but they will of course not be independent across the 10 repeats which is an assumption made for the calculation of my confidence intervals. This is also what is currently being done in the literature but is not ideal. Further sequencing has been done on *E. coli* isolates from bloodstream infections from 1 September 2015 through 31 December 2018, so this would be one possibility in future. Future work should include obtaining an independent test set from a different region/part of the world in order to evaluate generalisability. I would expect performance to decrease; however, if enough data is available from other regions, I could split and combine with Oxfordshire data instead.

When comparing the performance of the machine learning algorithms with that of the bacterial genome-wide association study, it is important to note that the performance on the latter was evaluated on the same set of isolates on which the model was fitted. I would therefore expect these error estimates to be optimistic and for performance to be lower on new data. Lack of an independent test set is therefore a major limitation of the GWAS component of this study, a limitation that I avoided for the machine learning algorithms by using cross-validation.

Similarly to previous studies (Earle et al. 2016), in my bacterial genome-wide association study I retained power to identify associations with population stratified variants by defining lineages through the PCs and testing for associations between PCs and the phenotype. PC 3 was the only one which was significantly associated with phenotype and also correlated with a branch in the phylogeny. This corresponded to ST131, which was also identified by (T. J. Davies et al. 2019) as the only one sequence type which was associated with co-amoxiclav resistance (chi-squared $p<0.0001$ compared with $p>0.16$ for all other STs).

3.4.4 Comparison with other studies

In a parallel analysis within the group, (T. J. Davies et al. 2019) constructed a separate LMM which did not account for population structure (unlike my GWAS LMM model), but did account for multiple and varying effects of the different known genetic features using multivariable models with careful variable selection and with reference standard agar dilution as the outcome. This model produced essential agreement of 97% (681/704) and accuracy of 78% (548/704) on an independent test set of 704 isolates. However, the model was developed including model selection for relationships between copy number and MIC as well as presence/absence effects, grouped beta-lactam resistance elements based on known effects (so-called Bush-Jacoby classes) and also incorporated interactions between different elements (reflecting saturation effects on MIC).

Such a model in which very careful variable selection is carried out by an expert based on specialist knowledge should contain only necessary and sufficient variables and is therefore

expected to be the best model for the analysed data. However, with the increase in the size of available datasets through the:

- number of sequences as prices are decreasing and they are accumulating over time
- number of bacteria of interest
- number of antibiotics of interest
- number of resistance elements

having an expert dedicated to building such models is becoming increasingly difficult. However, as the sizes of the datasets increase, these become ideal candidates for machine learning algorithms to learn from especially as resistance determinants may change over time. Increasing the size of a dataset is also known to be the most effective way of improving the performance of a machine learning model.

The main appeal of machine learning algorithms is that very large numbers of input features can be considered and good prediction performance can be obtained without the requirement of careful feature selection or expert knowledge. For example, adding another approximately 4000 features to the 'known + relative coverage' input features improved performance when considering the binary phenotype as the outcome and did not decrease performance when considering MIC as the outcome, despite only having 976 isolates.

The GWAS LMM is equivalent to an l2-regularized linear regression, a machine learning algorithm, and it was given a matrix of almost 2 million patterns of presence and absence as input. Every individual pattern of presence and absence was included as a fixed effect in the GWAS LMM accounting for population structure, and feature selection was carried out based on the significance of these individual patterns for the machine learning algorithms

considered in this Chapter. Interestingly, l2-regularised logistic regression was the best performing model when considering a binary phenotype as the outcome, however, its performance was lower than the GWAS l2-regularised linear regression. This may be as a dichotomous split is not appropriate for co-amoxiclav and therefore treating it as a truly continuous variable leads to better performance. However, when considering the MIC as the outcome the best performing model, Ordinal SE, outperformed the GWAS l2-regularised linear regression and this may be due to carrying out feature selection for the Ordinal SE model features.

One other reason why performance may have been higher in the previous multivariable model than with machine learning/GWAS is because the training dataset used for the former over-selected for co-amoxiclav resistant isolates (as additional phenotyping was performed on the subset used to train the data), therefore leaving dataset for prediction with a lower percentage of resistance, which may have been easier to predict. MIC prediction was also only done on isolates which did not contain any new genetic features compared with the training dataset (704/715 (98%)). Ideally, one would test a classifier's performance on the whole dataset as new genetic features will occur and prediction will have to be carried out if these methods are to be implemented routinely.

In terms of other organisms, Kouchaki et al looked at 13402 *Mycobacterium tuberculosis* infections for 11 antibiotics with a susceptible/resistant phenotype (Kouchaki et al. 2018). The same machine learning algorithms as those considered in this Chapter for the binary phenotype were considered plus a product of marginals model, which however performed worst for all 11 antibiotics. For 9 out of the 11 antibiotics they also found the best

performing model to be l2-regularized logistic regression, with gradient boosted trees being the best performing model for the remaining 2. However, performance was generally very similar across all algorithms considered (with the exception of the product of marginals model), with the main difference in performance being seen between antibiotics, with performance ranging from 76% sensitivity and 87% specificity for moxifloxacin to 92% sensitivity and 98% specificity for isoniazid.

Extreme gradient boosting (XGBoost) (Chen and Guestrin 2016) is an algorithm designed for boosting trees – adding new models to correct errors made by previous models – with a focus on execution speed and model performance. Nguyen et al looked at 5278 nontyphoidal *Salmonella* genomes collected over 15 years in the United States and used extreme gradient boosting to predict MICs for 15 antibiotics using 10-mers from assemblies as input features (Nguyen, Long, et al. 2018). Overall (across 15 antibiotics) they achieved essential agreement of 95% [95% CI (95%, 95%)], very major error of 2.7% [95% CI (2.4%, 3%)] (inverse of sensitivity) and major error 0.1% [95% CI (0.1%, 0.2%)] (inverse of specificity). Their analysis illustrates the strength of machine learning algorithms, specifically in the large amounts of data on which these algorithms are trained, with performance increasing from 88.5% to 91.4% as the size of the training set increased from 250 to 1000 genomes, and further increasing to 95.2% as the training set increased to 4500 genomes. Co-amoxiclav was one of the 15 antibiotics with 18% resistance prevalence (746/4095) and essential agreement for this antibiotic was 93% [95% CI (93%, 94%)] (based on Figure 2) when the model was trained on the 4500 genomes, higher than the 92% [95% CI (92%, 93%)] I achieved in *E. coli* with best algorithm Ordinal SE. However, this raises the question of how increasing my dataset might affect performance, given the 5% increase seen with

the XGBoost model going from 1000 to 4500 isolates. They conducted an equivalent analysis on 1668 *Klebsiella pneumoniae* clinical isolates with an overall 92% essential agreement (Nguyen, Brettin, et al. 2018). However, isolates in this dataset were highly resistant with 87% (1455/1668) having a co-amoxiclav MIC of 32/2 mg/L and 98% (1635/1668) an amoxicillin MIC of 32 mg/L. They also considered 10-mers from whole genome assemblies, 10-mers from AMR genes only and also 10-mers from non-AMR genes only with all giving similar accuracies. They argued that this shows there is enough information within the non-AMR genes to predict resistance; however, this probably illustrates the fact that what the machine learning model is doing is learning the structure within their data, picking up more drug resistant "clades", such as ST131 in *E. coli,* which could lead to very poor generalisability in future.

Focusing on the length chosen for the k-mers, Nguyen et al considered k-mers of length 10, much shorter than my length of 31 (Nguyen, Long, et al. 2018). The advantage of using a shorter k-mer length is having a smaller number of features; however, this also makes interpreting the results more challenging as a k-mer of shorter length will map to multiple regions. This is also noted by the authors in one of their following papers which focuses on predicting MIC for nontyphoidal *Salmonella*, but also on identifying which genomic features were most predictive. In this second study, in order to reduce the number of redundant k-mers within each genome and to allow analysis with the BLAST program (Camacho et al. 2009), they had to increase the k-mer length in their feature set from 10 to 15 (Nguyen, Long, et al. 2018).

Drouin et al looked at phenotypes of *C. difficile*, *M. tuberculosis*, *P. aeruginosa* and *S. pneumoniae* for 17 antibiotics and worked with 31-mers from assembled genomes (Drouin et al. 2016). 111 to 556 genomes were included for each pathogen-antibiotic combination, with the number of 31-mers varying from 10 to 123 million. Their goal was to have an easily interpretable model and therefore they considered a different algorithm, the Set Covering Machine algorithm. This produces sparse rules-based models and grows linearly with the number of genomes and 31-mers and is therefore able to learn from all the 31-mers. In contrast, the other algorithms which they considered (decision trees, regularised logistic regression) required an initial feature selection step, which they carried out through chi-squared tests. They did not however consider co-amoxiclav. The average number of features used by the Set Covering Machine for prediction varied from 1.0 to 4.9, supporting their main conclusion that their method was able to generate interpretable and accurate models. However, I feel that this means the set covering machine may not be appropriate, as it produced accurate prediction models of phenotypes which were nevertheless determined by a small set of genomic variations which may not be generalisable. In a further study, Drouin and Marchand also considered 1524 *E. coli* bloodstream infections with binary results for co-amoxiclav susceptibility, with 30% resistance prevalence (464/1524) and ~48.5 million k-mers (Drouin and Marchand 2018). The highest accuracy obtained was 81.8% from the Set Covering Machine algorithm with an average of 4.6 features used for each rule. They also considered decision trees, regularised logistic regression, SVM and Naïve Bayes; however, all had a lower performance compared to my equivalent models.

3.4.5 General issues

One of the main strengths of my dataset is that it comes from an unselected population with a great variety of MICs by comparison to many studies which only considered highly resistant isolates. For example, Stoesser et al selected all retrievable extended-spectrum cephalosporin-resistant isolates (commonly representative of multidrug-resistant phenotypes) for study using heuristic rules, comprising 74 *E. coli* and 69 *K. pneumoniae* (Stoesser et al. 2013); Peseky et al considered 78 highly resistant clinical *Enterobacteriaceae* isolates and used only logistic regression (Pesesky et al. 2016). However, more recent studies (described above) have used larger datasets of less highly selected isolates.

Other future work could also consider predicting genotype for other antibiotics, first individually, but also as a multi-label learning problem (Yang et al. 2019) as this would allow resistance co-occurrence to be exploited for increased predictive performance. Based on my and others' experience (Drouin and Marchand 2018; Nguyen, Brettin, et al. 2018), I would expect all methods to perform better with more bimodal resistance distributions – that is, I expect co-amoxiclav phenotype to be one of the hardest to predict.

Another natural extension is evaluating how well these models trained on *E. coli* bloodstream infections would perform on predicting the genotype for other Enterobacteriaceae which have similar enough resistance mechanisms such as *K. pneumoniae*. This question could also be approached by testing the predictive performance of the extreme gradient boosted tree fitted and tuned on *K. pneumoniae* (Nguyen, Brettin, et al. 2018) on my *E. coli* isolates.

In summary, when considering approximately 1000 *E.* coli bloodstream infections the best performing machine learning model was Ordinal SE considering MIC as the outcome of interest. Performance was very similar considering known co-amoxiclav resistance elements and relative coverage as input features versus the most significant approximately 5000 k-mers, with significance determined by GWAS. Therefore, the agnostic k-mer approach is more appealing as it does not rely on a heavily curated database which requires constant effort in order to be kept up to date. Counting k-mers directly from reads in order to be able to incorporate information on coverage is the next step in order to further improve my model, as shown by the significant improvement in performance when relative coverage was added to the known co-amoxiclav feature set which contained only presence and absence. Further work should also consider methods of k-mer selection from GWAS which could identify/incorporate significant interactions, for example the interaction between bla_{TEM} genes and TEM promoters, rather than just considering each k-mer (pattern) individually as done in this Chapter. An example of such a tool is the harmonic mean p-value (Wilson 2019) which was able to bring together information from two individually non-significant features into one significant feature in a human GWAS example for neuroticism and a joint human-pathogen GWAS for hepatitis C viral load. The algorithms should be trained again as more isolates become available and tuning of the regularisation parameter should be carried out. Extreme gradient boosting as well as neural networks should also be further explored as considering all these further steps could lead to an increase in performance which would then be acceptable for routine clinical practice.

4 Can resistance in bloodstream infections be estimated from resistance in other types of specimen? – a new method for low-income countries to monitor antimicrobial resistance

4.1 Introduction

In the previous Chapter I investigated whether WGS could replace laboratory susceptibility testing. However, in low-income countries, even standard testing of blood cultures is difficult, if not impossible with the existing facilities. In this Chapter I therefore aimed to build a tool for the surveillance of antimicrobial resistance in low-income countries. Specifically, I investigated if it is possible to estimate resistance rates in bloodstream infection by using microbiology results from cultures of other anatomical sites, which are typically easier to obtain and process than blood cultures.

Antimicrobial resistance is one of the top 10 global health threats (World Health Organisation 2019). From the limited data that has been analysed, the problem is even more severe in low- and middle-income countries (Vernet et al. 2014; Laxminarayan et al. 2003; Byarugaba 2004). With increasing travel, emergence and dissemination of highly resistant strains is a global concern. The importance of surveillance of antimicrobial resistance as a tool to fight the rising rates of antimicrobial resistance has been highlighted in the “Essential medicines and health products” by the World Health Organization (World Health Organization 2012). The monitoring and detection of threats allows rational and informed prevention and intervention actions to be taken (Johnson 2015). Antimicrobial resistance surveillance is particularly important in low-income settings where lack of

laboratory capacity prevents routine patient-level antimicrobial susceptibility testing (Ayukekbong, Ntemgwa, and Atabe 2017). Therefore, antimicrobial surveillance data representative of the local population, and which is carried out over time in order to capture changes in susceptibility patterns, can help healthcare practitioners choose antimicrobials to which patients are more likely to respond (Ayukekbong, Ntemgwa, and Atabe 2017). The lack of surveillance data on antimicrobial resistance development was identified as one of the key contributors to global antibiotic resistance (Chokshi, A., Sifri, Z., Cennimo, D., & Horng 2019). Without an understanding of the size of the problem of antimicrobial resistance, and the early detection of its emergence, the pragmatic approach which has to be taken in hospitals with limited laboratory facilities generally results in overtreatment (Hart and Kariuki 1998). The importance of mapping antimicrobial drug resistance has therefore been underlined in numerous papers, particularly as it could guide empirical antibiotic treatment for which very little data is available at the moment in low-income countries.

One reason why surveillance is generally poorer in low income countries is lack of well-functioning laboratories (Vlieghe, Bal, and Gould 2010; Vernet et al. 2014). For this reason, laboratory capacity strengthening has been a major focus of the Fleming Fund for the last few years https://www.flemingfund.org/grants-funding/. The Fleming Fund is aligned with the Global Antimicrobial Resistance Surveillance System (GLASS) (World Health Organisation 2015) launched by the World Health Organization in 2015 which encourages the collection, analysis and sharing of antimicrobial resistance data according to well established standards (which are to be improved as the action plan is carried out and weaknesses of the current standards are brought to light). The goal of setting up such a surveillance system is the

ability to develop targeted and informed programmes for prevention and control of antimicrobial resistance as well as to be able to assess the impact of interventions. However, greater efforts need to be made in order to carry out antimicrobial resistance surveillance in low and middle-income countries at the standards required by GLASS and this is underlined by studies focusing on the implementation in these specific settings (Seale et al. 2017).

In Chapter 2, I conducted a longitudinal analysis of *E. coli* bloodstream and urine infections in a large UK teaching hospital over a period of 19 years, particularly focussing on co-amoxiclav resistant isolates. I found that co-amoxiclav resistant bloodstream and urine infections were both increasing. This raises the broader question as to whether, for each pathogen-drug combination, the proportion of resistant bloodstream infections could be estimated by the proportion of resistant isolates from other sites. This would mean that the latter has the potential to be used as a surveillance method for the former. This is attractive as, in low-income countries, setting up a laboratory and providing the ongoing infrastructure (e.g. reagents, personnel, electricity etc) to conduct blood culture surveillance compared to urine culture surveillance would approximately double the costs (London School of Hygiene & Tropical Medicine 2016). From discussions with Nicola C. Gordon based on a manuscript in progress on costs for bacterial identification, a standard, safe bacteriology laboratory would cost approximately £50,000 to equip from scratch (crude estimates based on (London School of Hygiene & Tropical Medicine 2016), assuming a building is already available. Some laboratories from low-income settings are already processing urine samples, however, but without the standard requirements which include:

1. Infrastructure and utilities – e.g. certain type of flooring, benches, clean water, reliable electricity (mains power, backup generator, and battery standby to ensure constant power supply to incubators, freezers, fridges),
2. Basic biosafety aspects – controlled access, screened windows, handwashing facilities, and facilities for waste disposal (i.e. a dedicated autoclave for contaminated waste, and an incinerator),
3. Basic equipment for making agar (hotplate, clean autoclave), an incubator, a fridge.

Processing urine cultures (or cultures of other anatomical sites) without autoclaving waste, properly sterilising equipment and having access to running water is considered bad practice and should not be encouraged. However, even after considering the costs of bringing these laboratories to accepted quality standards in terms of processing urine samples safely, the main additional cost incurred in order to process a blood culture sample versus a urine sample consists of the cost of the blood culture machine (approximately £30,000) together with the sampling bottles (the cost of a blood culture bottle varies between £2.5 and £5 and ideally two bottles should be taken from an adult, i.e. an additional cost of £5-10 per sampling timepoint). In contrast, the cost of a urine culture is only 50p at one timepoint.

The other factor to be considered in the surveillance of antimicrobial resistance, is the low expected positivity rate of blood cultures of approximately 10%, meaning that the cost per blood culture on which any drug susceptibility results can be obtained (because an organism is identified at all) is 10-fold higher. Further, of positive cultures only about 9% grow *E. coli* (estimates based on surveillance of bloodstream infection isolates at a large teaching hospital in Malawi, 1998-2016 (Musicha et al. 2017)), meaning the cost per *E. coli* positive

blood culture (on which drug susceptibility to probably the most important pathogen can be obtained) is around 100-fold higher than the cost per blood culture. In work carried out by Nicola C. Gordon for the Fleming Fund, making the assumption that in order to obtain 100 *E. coli* bloodstream infections, blood from at least 11000 patients will have to be sampled, at the lowest cost for the bottles, and if only a single bottle per patient is taken, the additional cost for blood culture equipment and consumables for a laboratory that already has all the equipment and reagents necessary to do urine cultures is approximately £66,300. This factors in the cost of the instrument, bottles and the cost of identifying positive samples, but does not take into account the additional training, staff costs and waste disposal costs, which are fairly significant for blood cultures. Even with this investment, uncertainty around prevalence estimates based on only 100 observations in the denominator will be substantial, limiting utility.

In this study, I considered both *E. coli* and *Staphylococcus aureus* as these are two of the most common bacteria causing bloodstream infection (Edelsberg et al. 2014; Poolman and Anderson 2018; Vernet et al. 2014), and are often found in urine or on the skin, respectively. As a result, a large collection of these two pathogens isolated from both blood and urine/skin is already available. Taking urine cultures and skin swabs is less invasive than taking blood cultures, and these samples are more frequently culture positive, reducing the costs per drug phenotype obtained. The fact that these kinds of isolates are therefore easily available in large numbers therefore makes them ideal for surveillance purposes which has the benefits both of identifying emerging resistance, and also providing data-driven guidance for choice of empirical antimicrobial treatment.

Related to this, I also explored whether in high-income countries, antibiograms of concurrent or previous *E. coli* urinary tract infections had the potential to be used as guidance for empirical antimicrobial treatment of a new bloodstream infection. Within high-income countries, current microbiological methods take around 48 hours for susceptibility results to become available, during which time infections are treated empirically. If the infection was resistant to the empirical antibiotic, this could lead to poorer outcomes and also contribute to the development of further antibiotic resistance (Laxminarayan et al. 2003).

4.2 Methods

I obtained data on all *E. coli* and *S. aureus* isolates from IORD from 1 January 1998 to 31 December 2018. This extract was prepared by Phuong Quan. I selected all *E. coli* bloodstream infections and urinary tract infections. For *S. aureus* isolates, following (S. Walker et al. 2008), I excluded cultures from vascular tips due to their relationship with blood cultures (as in our laboratory there is a policy that a line tip culture is only done if there is a positive blood culture), and methicillin-resistant *S. aureus* (MRSA) screening cultures because only MRSA could be isolated from these. I then categorised the remaining *S. aureus* cultures into bloodstream infections and *S. aureus* non-blood cultures. Non-blood cultures were further subdivided into:

- sterile sites (clinical samples cultured from respiratory and normally sterile sites taken under aseptic technique, such ascites, joint aspirates, cerebrospinal fluid, pre-prosthetic material, and collections of pus) which one could hypothesise would be more similar to blood cultures in terms of resistance since they also reflect invasive

infection. They are still potentially harder to obtain (e.g. needing sampling during operations, etc), but once obtained easier to culture.

- non-sterile sites (everything else, typically surface swabs, skin surface, genital, urine) which are easier to obtain and therefore more likely to be available, but may reflect colonisation with a commensal rather than infection with a pathogen.

I identified all antimicrobials for which antibiotic susceptibilities had been tested in >80% of samples throughout 1998-2018, that is,

- in both blood and urine cultures for *E. coli*: amoxicillin, co-amoxiclav, ciprofloxacin and trimethoprim,
- in blood and non-blood cultures for *S. aureus*: ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin and tetracycline. Susceptibility of *S. aureus* to penicillin stopped being routinely tested for in 2013, and I therefore only considered years 1998-2012 for this drug.

I used the same breakpoints regardless of type of isolate, namely the ones recommended by EUCAST for bloodstream infections (European Committee on Antimicrobial Susceptibility Testing 2019).

For each relevant pathogen and drug combination, I compared the proportion (and 95% confidence interval (CI)) resistant in blood versus the proportion (and 95% CI) resistant in the other sample type in each calendar year. Due to the longitudinal nature of the data, in order to identify the lag at which the correlation between the two time-series is strongest, I considered time-series cross-correlation functions between the proportion resistant in blood over calendar years and the proportion resistant in the other sample type over calendar years. Following the statistical methods for assessing agreement between two

methods of clinical measurement described in (Bland and Altman 1986), I analysed the agreement between the proportion resistant in blood versus the proportion resistant in the other sample type, using blood cultures as the reference standard, using limits of agreement and plotted the difference between the proportion resistant per year in blood and the other sample type. The number of comparisons outside 5% or 10% variation were calculated to determine its use as a surveillance tool. For surveillance, allowing for 5%-10% variation seems sensible given these could influence decisions on empiric antibiotic treatment, and so I also considered whether the 95% CIs lay within these boundaries. As the last simple summary, I calculated Lin's concordance correlation coefficient (CCC) (Lin 1989) defined by the formula:

$$CCC := 1 - \frac{\mathbb{E}[(Y-X)^2]}{\mathbb{E}[(Y-X)^2|\delta=0]} = 1 - \frac{\mathbb{E}[(Y-X)^2]}{(\mu_X - \mu_Y)^2 + \sigma_X^2 + \sigma_Y^2}$$

$$= 1 - \frac{expected\ squared\ perpendicular\ deviation\ from\ the\ line\ y=x}{expected\ squared\ perpendicular\ deviation\ from\ the\ line\ y=x,\ when\ X\ and\ Y\ are\ uncorrelated}$$

This quantifies the agreement between the proportion of resistant urines (respectively swabs) and the gold standard, that is the proportion of resistant blood cultures. This is analogous to the Spearman's correlation coefficient which tests for an association between two different variables; however, the difference is that Lin's concordance correlation coefficient takes into account the fact that the expectation is that the two variables being compared should be the same. Similarly to a standard correlation coefficient, the CCC takes values between -1 and 1 and is equal to 1 when the two methods of measuring proportion of resistant isolates are in perfect agreement, -1 when they are in perfect disagreement, and 0 if there is no association. It combines measures of precision and accuracy in order to determine how far the observed values deviate from the line of perfect concordance,

overcoming some of the limitations of correlation coefficients in general, paired t-tests and least squares approaches (Lin 1989). However, I also plotted the ordinary least squares regression of the proportion resistant in the other sample type against the proportion resistant in blood cultures in order to visualise this compared to the line of perfect concordance.

The methods outlined above analyse the data without modelling it. In order to model the data, I considered three linear mixed models (or special cases of these), an extension of simple linear models which allow both fixed and random effects. I chose these in order to incorporate the standard errors in the estimated proportion resistant blood cultures, and/or infections of other sites.

First, I conducted a standard meta-analysis: in traditional meta-analysis terms, I treated every calendar year as an independent study, resistant vs susceptible as the outcome, the blood cultures as the treatment group and the other sites of infection as the control group. I considered the log odds ratio as a measure of the difference between these two groups. I considered both fixed and random effect models. The later assumes that each year comes from a normal distribution centred on the average difference between the two sample types across the years. Pooling of effect estimates across the calendar years was done by inverse variance weighting. I also considered the degree of inconsistency across the years and the heterogeneity across years using I^2 statistics (Higgins et al. 2003). Similar to the Bland-Altman methods above, these models estimate an overall fixed difference between the two sample types over all calendar years, but do allow individual years to vary around this.

I also conducted two different types of meta-regression, allowing for uncertainty in the response variable, but not the explanatory variable, in contrast to standard regression models which assume both are known exactly without error, and incorporating calendar year as a random effect (as for the meta-analysis above). The first meta-regression was a standard extension of the meta-analysis above, modelling the log odds ratio for resistance in blood vs other sample types (with its standard error) as the response variable, and including year as an explanatory variable. The second directly modelled the log odds of resistance in bloodstream infections as the response variable, against the log odds of resistance in other sample types as the explanatory variable. The reason for not incorporating error in the estimate of resistance in the other sample types is that the numbers of other sample types per year are very high, meaning that the error in the estimate of the odds (or proportion) resistant is much smaller, in contrast to estimates of resistance in blood cultures. The reason for modelling the log odds is mathematical convenience, since the log odds ranges from minus to plus infinity, whereas proportions are bounded by zero and 1. The advantage is that, rather than assuming the same fixed difference between the proportion resistant in blood and other sample types as the standard meta-analysis above, meta-regression allows for the relationship between these proportions to vary. For example, the difference between them could be smaller (or larger) at high (or low) resistance prevalence, or over calendar time. Prediction intervals were calculated as they indicate where a prediction made from the proportion of resistant other sample types would be expected to fall (Riley, Higging, and Deeks 2011)

$$\hat{\mu} \pm 1.96\sqrt{standard\ error(\hat{\mu})^2 + \hat{\tau}^2}$$

where $\hat{\mu}$ is the estimated effect size and $\hat{\tau}^2$ is the estimated amount of heterogeneity, i.e. the estimated variance in the true outcome across the years. These were then transformed back onto the original proportion scale using

$$proportion\ resistant = \frac{\exp\left(predicted \log(odds)\right)}{1 + \exp\left(predicted \log(odds)\right)}$$

One important caveat of the analyses above is that the underlying populations with bloodstream infections and infections in other sites may genuinely differ, that is different resistance rates may be observed simply because of differences in the underlying populations being sampled. For a secondary analysis, I therefore considered whether using the susceptibility of isolates from current infections in other sites could be used to predict the susceptibility results for isolates from blood, on an individual basis. For all blood isolates, I selected the closest culture from a different site which was taken less than 3 days before or up to 2 days after (if available) (or within [-72 hours, +48 hours] instead in sensitivity analysis). Only matched pairs were analysed, ensuring that estimates of proportions resistant in the different sample types came from the same population.

I also considered the ability of susceptibility of a previous *E. coli* urinary tract infection to predict susceptibility of a subsequent *E. coli* bloodstream infection. For all blood isolates, I therefore selected the closest culture from a different site which was taken between 3 and 90 days before the bloodstream infection (if available).

For these matched pairs of isolates, I assessed concordance using McNemar's test for marginal homogeneity, that is testing whether bloodstream infections and the other type of

infection disagreed randomly or not in terms of their susceptibility. I also computed the following of performance metrics:

$$accuracy := \frac{true\ resistant + true\ susceptible}{true\ resistant + true\ susceptible + false\ resistant + false\ susceptible}$$

$$very\ major\ error\ (VME) := \frac{false\ susceptible}{true\ resistant + false\ susceptible} = 1 - sensitivity$$

$$major\ error\ (ME) := \frac{false\ resistant}{true\ susceptible + false\ resistant} = 1 - specificity$$

$$positive\ predictive\ value\ (PPV) := \frac{true\ resistant}{true\ resistant + false\ resistant}$$

$$negative\ predictive\ value\ (NPV) := \frac{true\ susceptible}{true\ susceptible + false\ susceptible}$$

I conducted all analyses using R 3.5.1. Meta-analyses were conducted using the metafor package (Viechtbauer 2010).

4.3 Results

4.3.1 Population

4.3.1.1 *E. coli* infections

This study was carried out on isolates cultured from the population of Oxfordshire whose size increased from just under 600,000 in 1998 to approximately 680,000 in 2018 (mid-year population estimates from the Office for National Statistics https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates). A total of 8080 *E. coli* bloodstream infections and 322071 *E. coli* urinary tract infections were included. **Figure 46** illustrates the number of *E. coli* bloodstream infections (**A**), respectively *E. coli* urinary tract infections (**B**), per year, between 1998 and 2018, split

by susceptibility to amoxicillin, co-amoxiclav, ciprofloxacin, trimethoprim. Other antibiotics such as cephalosporins are not considered here because very few *E. coli* urine cultures were tested for susceptibility to these drugs prior to the introduction of automated testing in 2013. A relationship between the proportion of resistant bloodstream infections and proportion of resistant urinary tract infections seemed plausible and I therefore analysed this formally. Work carried out in the Modernising Medical Microbiology group considering ST types and whole genome sequences, as well as the prevalence of known resistance genes found no evidence of major shifts in strain distribution over the last decade or major changes in proportions of known resistance genes (personal communication, Sam Lipworth).

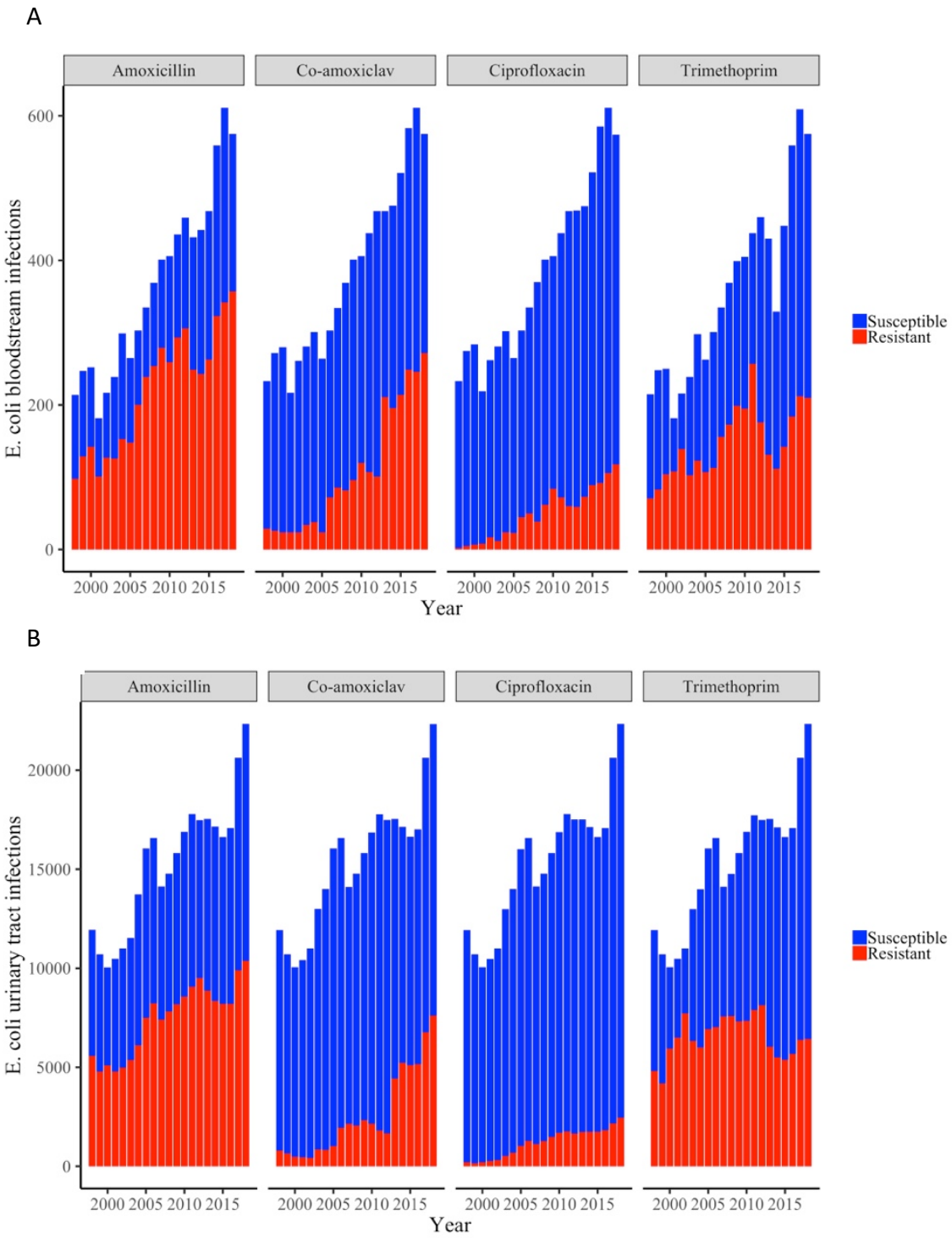


Figure 46 Number of (A) E. coli bloodstream infections and (B) E. coli urinary tract infections split by susceptibility to amoxicillin, co-amoxiclav, ciprofloxacin, trimethoprim, 1998-2018

Figure 47 illustrates the full analyses assessing the proportion co-amoxiclav resistant *E. coli* urinary tract infections as a measure of the proportion of co-amoxiclav resistant *E. coli* bloodstream infections. **Panel A** illustrates trends in proportions of co-amoxiclav-resistant *E. coli* bloodstream infections and urinary tract infections per year, 1998-2018. These follow very similar trends, with large changes in proportion occurring in both types of infections in 2006 and 2013. **Panel B** illustrates the cross-correlation function between the two time series shown in **Panel A**, allowing for different lags between the two; for example, a lag of 1 year is the cross-correlation between the proportion of resistant urines in one year and proportion of resistant blood cultures in the following year. The maximum cross-correlation between the proportion of resistant bloodstream infections and the proportion of resistant urinary tract infections was 0.96 and occurred at lag 0, that is, no displacement. As the displacement increased, the cross-correlation decreased, but positive displacement, that is resistance being determined earlier in urines than bloodstream infections led to slightly higher cross-correlations than the corresponding negative displacement.

For **Panels C**, **D** and **G** I shifted the focus from trends over time to agreement at different levels of resistance. For this pathogen-drug combination, three levels of resistance can be clearly identified, corresponding to periods between large increases in resistance, that is 1998-2005, 2006-2012, respectively 2013-2018. **Panels C** and **D** test the original hypothesis that the proportion of resistant bloodstream infections could be measured by the proportion of resistant urinary tract infections. **Panel C**, similar to the Bland-Altman plot traditionally used to assess the suitability of a replacement measurement tool, clearly shows the difference between the two proportions was smaller at lower than higher levels of resistance, and the limits of agreement do not include 0, which would be perfect

agreement, and are wide. However, expecting perfect agreement could be argued to be unrealistic, and considering agreement in terms of levels of absolute levels of resistance within certain thresholds may still be clinically helpful (similar to for example a +/-log2 dilution accepted in the prediction of a minimum inhibitory concentration). **Panel D** shows where resistance proportions lie compared with what might be considered clinically acceptable error (±5% and ±10%): whilst for 10 of the 21 years, the proportion of resistant bloodstream infections differed from the proportion of resistant urine infections by more than 10%, in only 3 years did the 95% CI of the proportion of resistant bloodstream infections not cross this boundary. The ordinary least squares regression showed generally greater underestimation of resistance as prevalence increased. For co-amoxiclav, Lin's concordance correlation (CCC) was 0.71 [95% CI 0.54, 0.83], which was again far from perfect concordance ($CCC = \pm 1$), but also from no concordance ($CCC = 0$).

Panel E is a forest plot of the meta-analysis of the log odds ratio for resistance in bloodstream vs urine cultures. The model estimates a pooled odds ratio of 1.99 [95% CI 1.79, 2.21] under the assumption of a random effects model (fixed effect estimate similar), but illustrates how far the individual years deviate from this, represented by the 74% I^2 statistic ($p<0.01$). Again, for perfect agreement (original hypothesis) the 95% CI in **Panel E** would have contained 1; in fact it shows that the odds of resistance are approximately doubled in blood cultures compared with urine cultures. The meta-regression in **Panel F** shows that there is no strong evidence that this odds ratio is varying linearly over time (per year effect on log odds ratio -0.01 [95% CI -0.03,0.01], p=0.23); the fact that the relationship between the two resistance prevalences is relatively constant means it may still be of clinical use. **Panel G** shows this overall association between co-amoxiclav resistance

prevalence in bloodstream and urine infections from my final meta-regression. As the relationship was modelled as linear in the log odds, transforming back to the proportion scale leads to a slight curvature in the association. Overall the association was very strong (log odds of resistance in blood cultures increasing by 0.86 [95% CI 0.75,0.97] for every unit increase in the log odds of resistance in urine cultures (perfect agreement being indicated by a coefficient of 1), $p<0.001$). All but two of the proportions of resistant bloodstream infections lay within the 95% prediction interval based on the proportion of resistant urines (which also takes into account variability in this estimate), and all 95% CI around these proportions of resistant bloodstream infections substantially overlapped with the prediction interval.

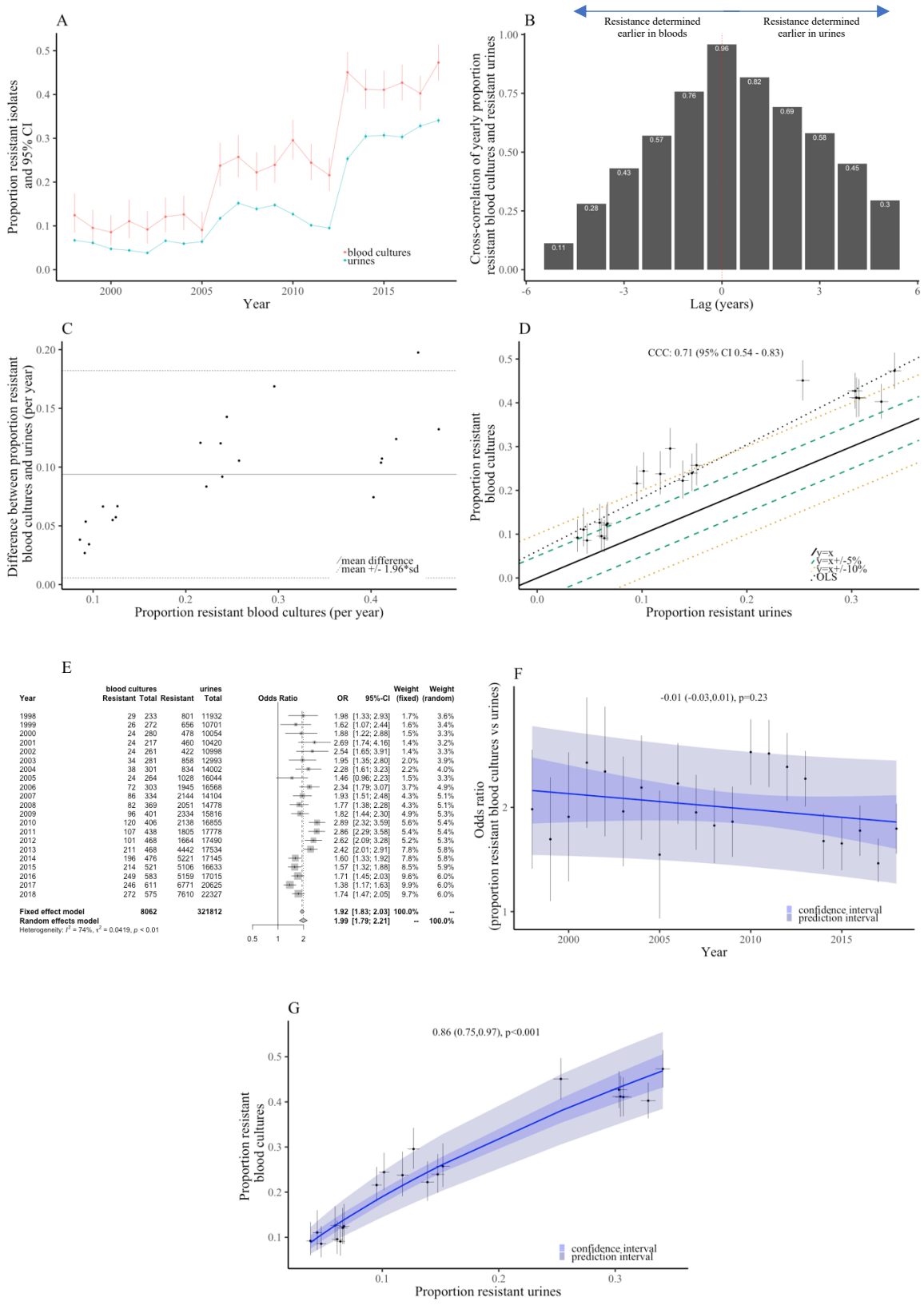


Figure 47 Analyses of agreement between proportion of co-amoxiclav-resistant E. coli bloodstream infections and proportion of co-amoxiclav-resistant E. coli urinary tract infections

Note: A – proportion and 95% CI of co-amoxiclav-resistant *E. coli* bloodstream infections and proportion and 95% CI of co-amoxiclav-resistant *E. coli* urinary tract infections per year, 1998-2018.
B – cross-correlation function of yearly proportion resistant bloods and urines. A lag of 1 year is the cross-correlation between the proportion resistant urines and the proportion resistant bloods in the following year.
C– plot (similar to Bland-Altman) comparing the difference between the proportion of resistant bloods and urines per year against the gold standard, that is the proportion resistant blood cultures per year, with the limits of agreement (95% CI for the difference).
D – Proportion resistant urines (95% CI) vs proportion resistant bloods (95% CI) per year, with one point per year, and ordinary least squares regression (OLS) predicted association. The line y=x represents perfect agreement, while y=x+/-5% and y=x+/-10% represent what could be considered a clinically acceptable error.
E – Meta-analysis of the log odds ratio for resistance in bloodstream vs urine cultures per year.
F – odds ratio for resistance in blood vs urine samples on calendar year and predicted association, its 95% CI and the 95% prediction interval from meta-regression modelling the log odds ratio for resistance in blood vs urine samples with calendar year as an explanatory variable; impact of calendar time on log odds ratio (95% CI) and p-value.
G – proportion resistant in blood vs urine samples per year and the predicted association, its 95% CI and the 95% prediction interval from meta-regression of the log-odds of resistant bloods on the log-odds of resistant urines.
Please see electronic version for easier visualisation and **Table 17** for summary of key results.

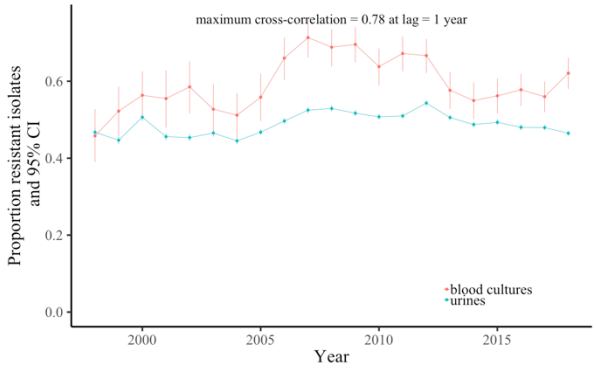
The same analyses are included for all the remaining pathogen-drug combinations in **Figure 58** – **Figure 66** (Supplementary Material). In order to allow comparison of the characteristics of each antibiotic, in **Figure 48** only **Panels A**, **G** and **E** are presented, including the summary of **Panel B** by adding the maximum cross-correlation and the lag at which this occurred onto **Panel A** and adding Lin's concordance correlation coefficient onto **Panel G**.

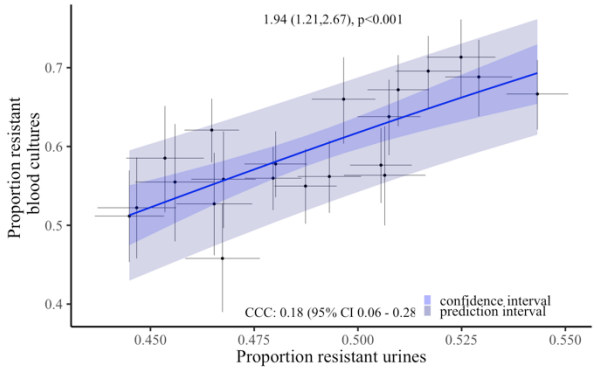
Figure 48 illustrates the analyses of agreement between proportion of resistant *E. coli* bloodstream infections and proportion of resistant *E. coli* urinary tract infections, for amoxicillin, co-amoxiclav, ciprofloxacin and trimethoprim. For amoxicillin, co-amoxiclav and ciprofloxacin there were similar trends in the proportion of resistant bloodstream infections to urinary tract infections, but with proportions consistently higher in blood cultures than in urines. Trends were similar when considering trimethoprim as well, despite considerable changes over time; however, proportions were no longer consistently higher in blood cultures than in urines throughout the whole time period. The maximum cross-correlation occurred at lag 0 for all drugs except amoxicillin, for which it occurred at lag 1 year with a

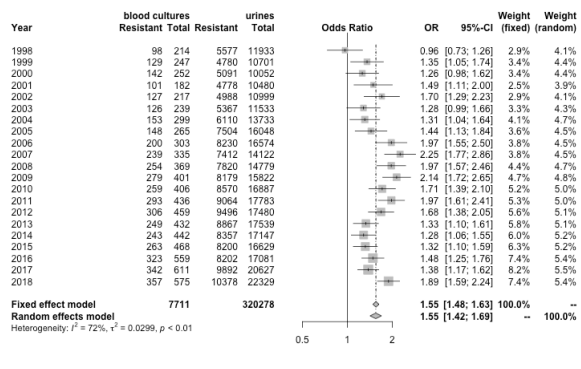
value of 0.78; however, at lag 0 the cross-correlation was very similar (0.77) and this was the lowest maximum cross-correlation achieved across the antibiotics for *E. coli*. The proportion of resistant urines was highly associated with the proportion of resistant bloodstream infections ($p<0.001$ for all four antibiotics) with the majority of observations of resistant proportions in bloodstream infections within the 95% prediction intervals based on the model and the observed resistance prevalence in urine infections. Considering the meta-analysis, for amoxicillin and ciprofloxacin results were similar to those for co-amoxiclav, with greater resistance prevalence in bloodstream infections compared with urine infections but for trimethoprim there was no overall evidence of differential prevalence (pooled odds ratio (random effects model) 0.93 [95% CI 0.82, 1.05]). However, considering the meta-regression for variation in this odds ratio over calendar year, trimethoprim was the only antibiotic for which there was evidence for a trend over time ($p<0.001$, **Figure 60**) with $p>0.2$ for the other 3 drugs (**Panel F** in **Figure 58** and **Figure 59, Table 17**). For trimethoprim, resistance was initially slightly higher in urine samples, and then became if anything slightly greater in blood, but differences in resistance in the two sample types for trimethoprim were relatively small overall.

Figure 48 Analyses of agreement between proportion of resistant E. coli bloodstream infections and proportion of resistant E. coli urinary tract infections, for amoxicillin, co-amoxiclav, ciprofloxacin and trimethoprim

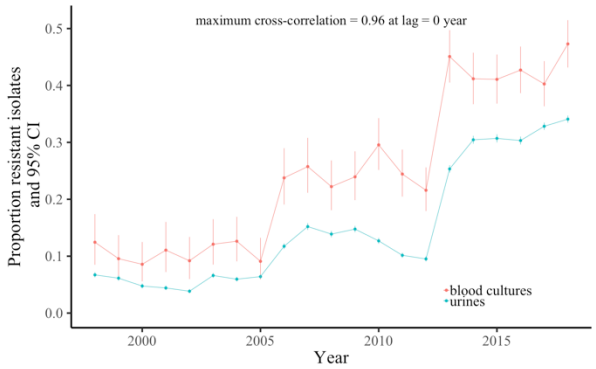
Amoxicillin-resistant

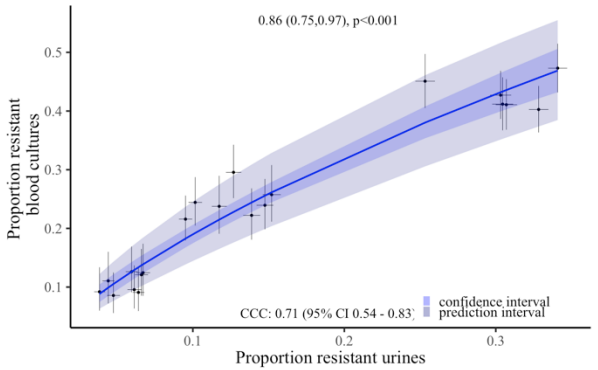


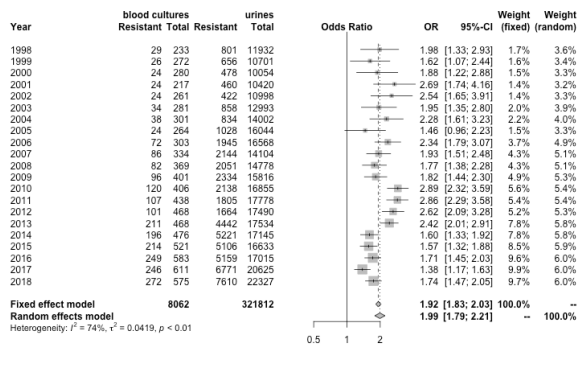




Co-amoxiclav-resistant

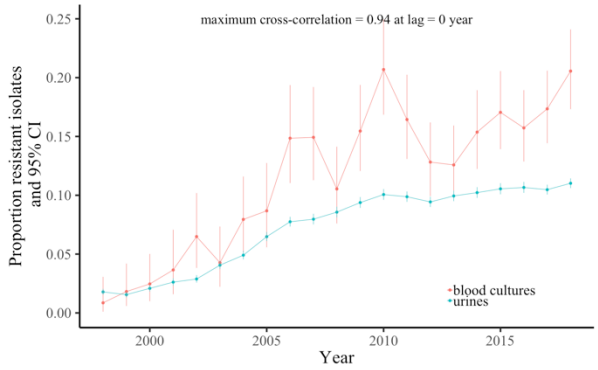


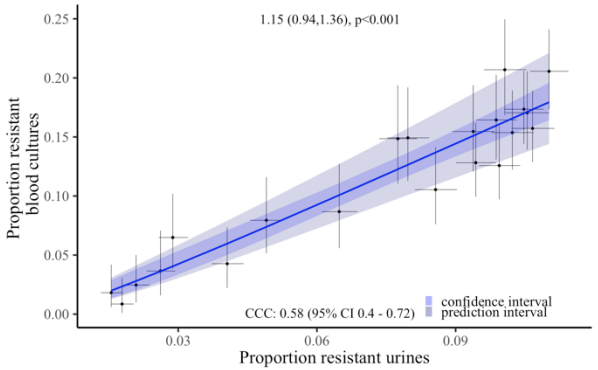


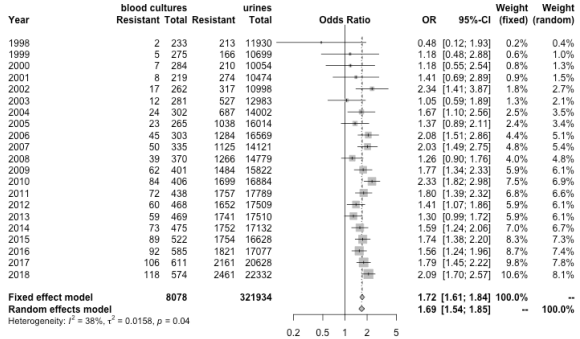


Note: In the middle plot the x and y axes are on different scales. Please see electronic version for easier visualisation and **Table 17** for summary of key results.

Ciprofloxacin-resistant





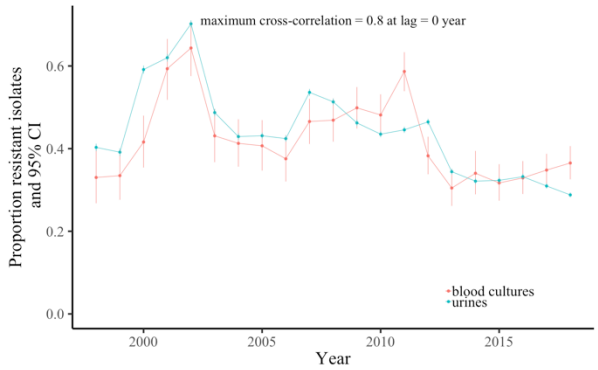


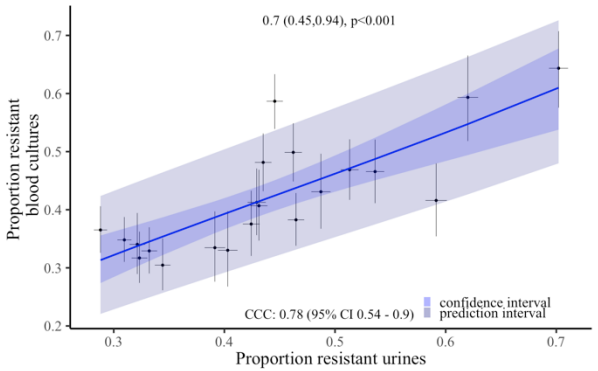
Year	blood cultures Resistant	blood cultures Total	urines Resistant	urines Total	OR	95%-CI	Weight (fixed)	Weight (random)
1998	2	233	213	11930	0.48	[0.12; 1.93]	0.2%	0.4%
1999	5	275	166	10699	1.18	[0.48; 2.88]	0.6%	1.0%
2000	7	284	210	10054	1.18	[0.55; 2.54]	0.8%	1.3%
2001	8	219	274	10474	1.41	[0.69; 2.89]	0.9%	1.5%
2002	17	262	317	10998	2.34	[1.41; 3.87]	1.8%	2.7%
2003	12	281	527	12983	1.05	[0.59; 1.89]	1.3%	2.1%
2004	24	302	687	14002	1.67	[1.10; 2.56]	2.5%	3.5%
2005	23	265	1038	16014	1.37	[0.89; 2.11]	2.4%	3.4%
2006	45	303	1284	16569	2.08	[1.51; 2.86]	4.4%	5.1%
2007	50	335	1125	14121	2.03	[1.49; 2.75]	4.8%	5.4%
2008	39	370	1266	14779	1.26	[0.90; 1.76]	4.0%	4.8%
2009	62	401	1484	15822	1.77	[1.34; 2.33]	5.9%	6.1%
2010	84	406	1699	16884	2.33	[1.82; 2.98]	7.5%	6.9%
2011	72	438	1757	17789	1.80	[1.39; 2.32]	6.8%	6.6%
2012	60	468	1652	17509	1.41	[1.07; 1.86]	5.9%	6.1%
2013	59	469	1741	17510	1.30	[0.99; 1.72]	5.9%	6.1%
2014	73	475	1752	17132	1.59	[1.24; 2.06]	7.0%	6.7%
2015	89	522	1754	16628	1.74	[1.38; 2.20]	8.3%	7.3%
2016	92	585	1821	17077	1.56	[1.24; 1.96]	8.7%	7.4%
2017	106	611	2161	20628	1.79	[1.45; 2.22]	9.8%	7.8%
2018	118	574	2461	22332	2.09	[1.70; 2.57]	10.6%	8.1%
Fixed effect model	**8078**		**321934**		**1.72**	**[1.61; 1.84]**	**100.0%**	--
Random effects model					**1.69**	**[1.54; 1.85]**	--	**100.0%**

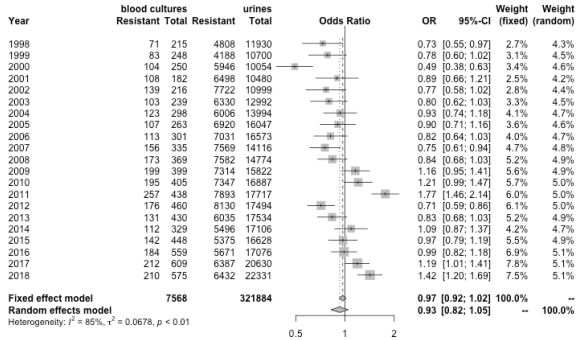
Heterogeneity: I^2 = 38%, τ^2 = 0.0158, p = 0.04

Odds Ratio

Trimethoprim-resistant







Year	blood cultures Resistant	blood cultures Total	urines Resistant	urines Total	OR	95%-CI	Weight (fixed)	Weight (random)
1998	71	215	4808	11930	0.73	[0.55; 0.97]	2.7%	4.3%
1999	83	248	4188	10700	0.78	[0.60; 1.02]	3.1%	4.5%
2000	104	250	5946	10054	0.49	[0.38; 0.63]	3.4%	4.6%
2001	108	182	6498	10480	0.89	[0.66; 1.21]	2.5%	4.2%
2002	139	216	7722	10999	0.77	[0.58; 1.02]	2.8%	4.4%
2003	103	239	6330	12992	0.80	[0.62; 1.03]	3.3%	4.5%
2004	123	298	6006	13994	0.93	[0.74; 1.18]	4.1%	4.7%
2005	107	263	6920	16047	0.90	[0.71; 1.16]	3.6%	4.6%
2006	113	301	7031	16573	0.82	[0.64; 1.03]	4.0%	4.7%
2007	156	335	7569	14116	0.75	[0.61; 0.94]	4.7%	4.8%
2008	173	369	7582	14774	0.84	[0.68; 1.03]	5.2%	4.9%
2009	199	399	7314	15822	1.16	[0.95; 1.41]	5.6%	4.9%
2010	195	405	7347	16887	1.21	[0.99; 1.47]	5.7%	5.0%
2011	257	438	7893	17717	1.77	[1.46; 2.14]	6.0%	5.0%
2012	176	460	8130	17494	0.71	[0.59; 0.86]	6.1%	5.0%
2013	131	430	6035	17534	0.83	[0.68; 1.03]	5.2%	4.9%
2014	112	329	5496	17106	1.09	[0.87; 1.37]	4.2%	4.7%
2015	142	448	5375	16628	0.97	[0.79; 1.19]	5.5%	4.9%
2016	184	559	5671	17076	0.99	[0.82; 1.18]	6.9%	5.1%
2017	212	609	6387	20630	1.19	[1.01; 1.41]	7.8%	5.1%
2018	210	575	6432	22331	1.42	[1.20; 1.69]	7.5%	5.1%
Fixed effect model	**7568**		**321884**		**0.97**	**[0.92; 1.02]**	**100.0%**	--
Random effects model					**0.93**	**[0.82; 1.05]**	--	**100.0%**

Heterogeneity: I^2 = 85%, τ^2 = 0.0678, p < 0.01

Odds Ratio

Note: In the middle plot the x and y axes are on different scales. Please see electronic version for easier visualisation and **Table 17** for summary of key results.

4.3.1.2 *S. aureus* infections

There were a total of 6926 *S. aureus* bloodstream infections and 165775 *S. aureus* non-blood cultures: 51612 (31%) sterile site cultures and 114163 (69%) non-sterile site cultures. **Figure 49** illustrates the number of *S. aureus* bloodstream infections (**A**), respectively *S. aureus* non-blood cultures (**B**) split by susceptibility to ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin, tetracycline, between 1998 and 2018. In **Figure 50** *S. aureus* non-blood cultures are split into sterile site cultures (**A**) and non-sterile site cultures (**B**). Penicillin resistance was analysed only between 1998 and 2012 as routine testing was stopped in 2013. Ciprofloxacin, erythromycin and flucloxacillin presented strikingly similar trends in proportions resistant over time in the different types of samples, again supporting the plausibility of using non-blood cultures as a proxy for blood culture surveillance. The majority of cultures were also tested for vancomycin resistance, but there were no resistant bloodstream infections, only 2 resistant sterile site cultures and 4 resistant non-sterile site cultures. I therefore did not consider vancomycin resistance further.

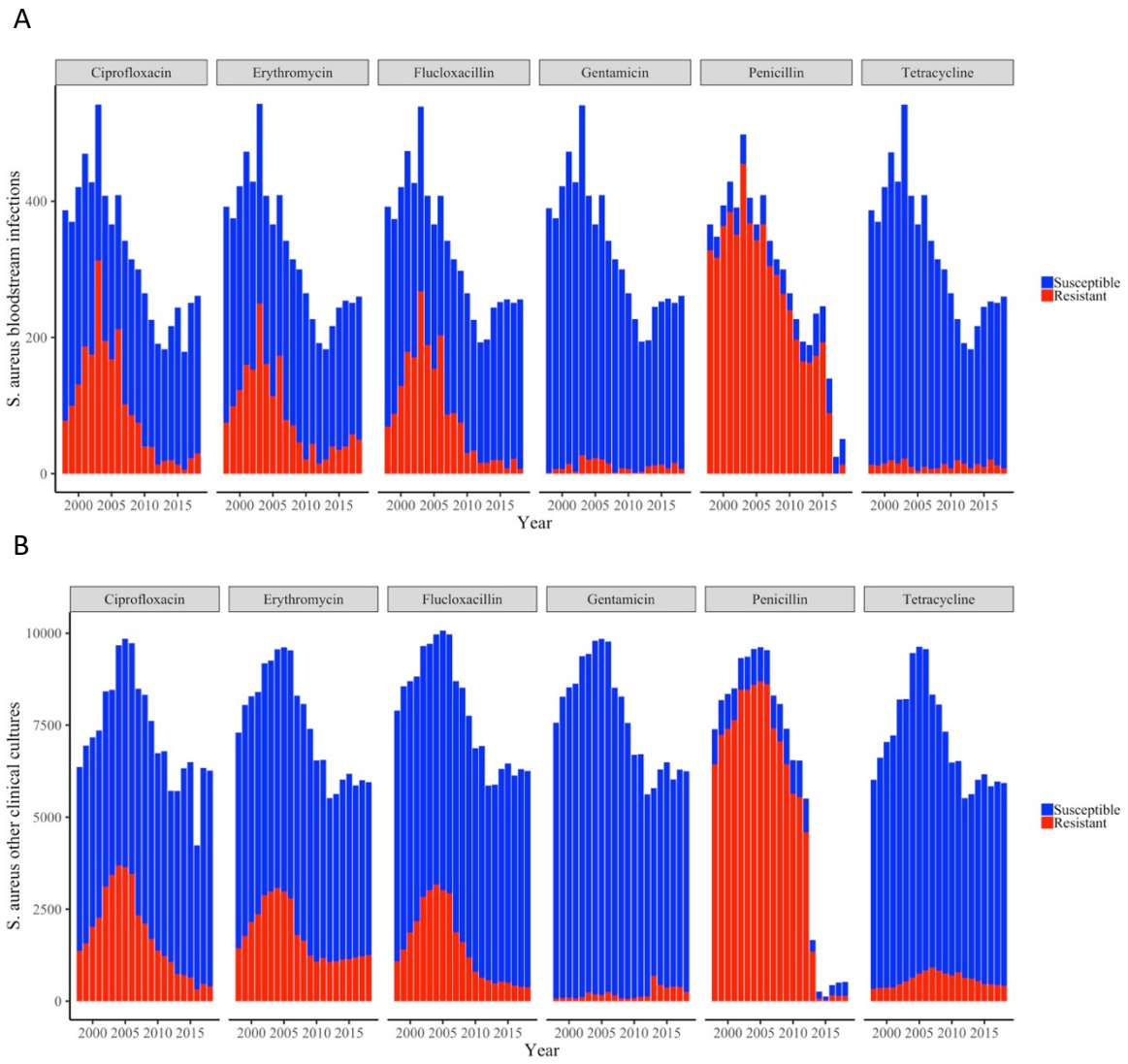


Figure 49 Number of (A) S. aureus bloodstream infections (B) S. aureus non-blood cultures split by susceptibility to ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin, tetracycline, 1998-2018

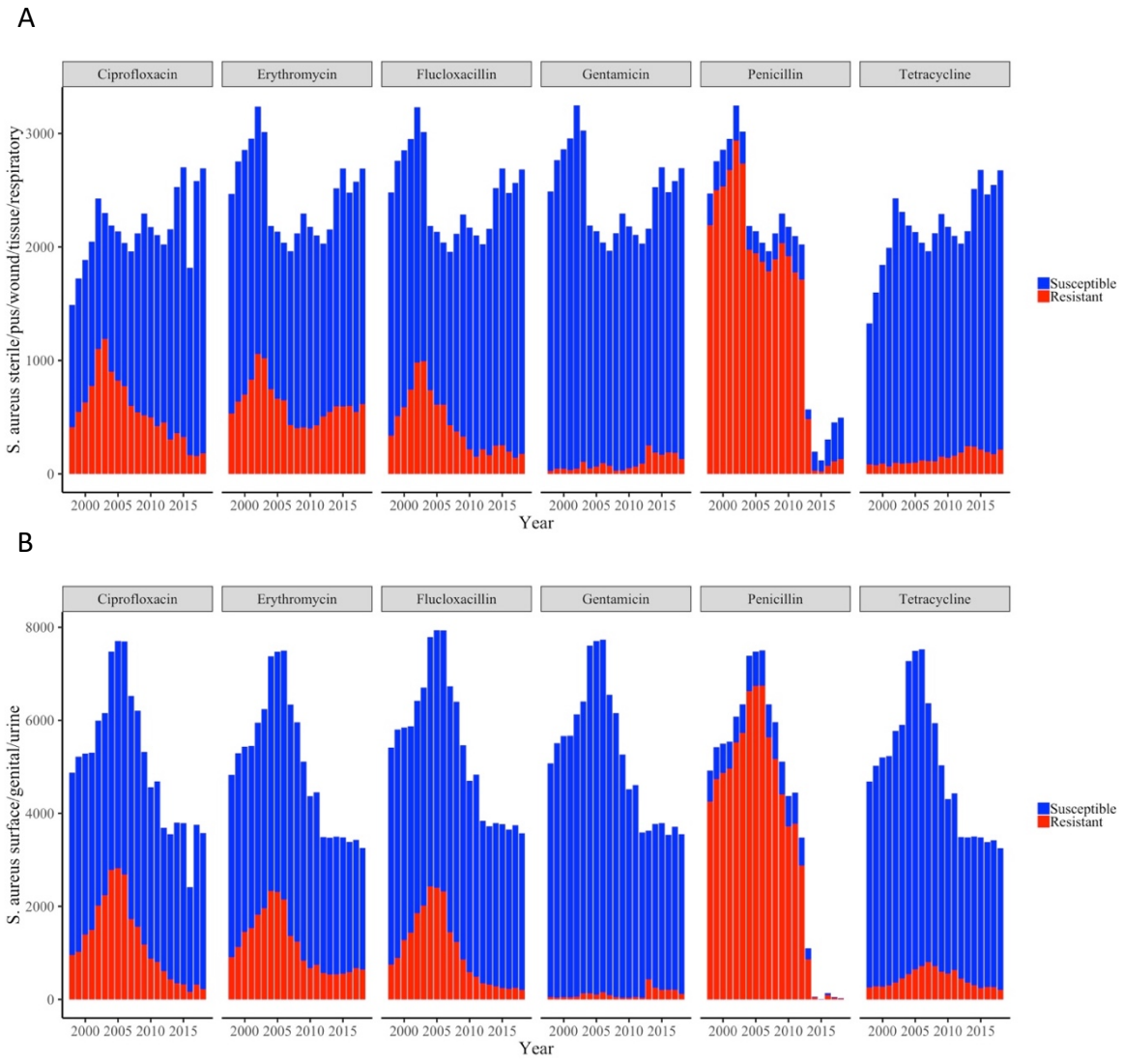


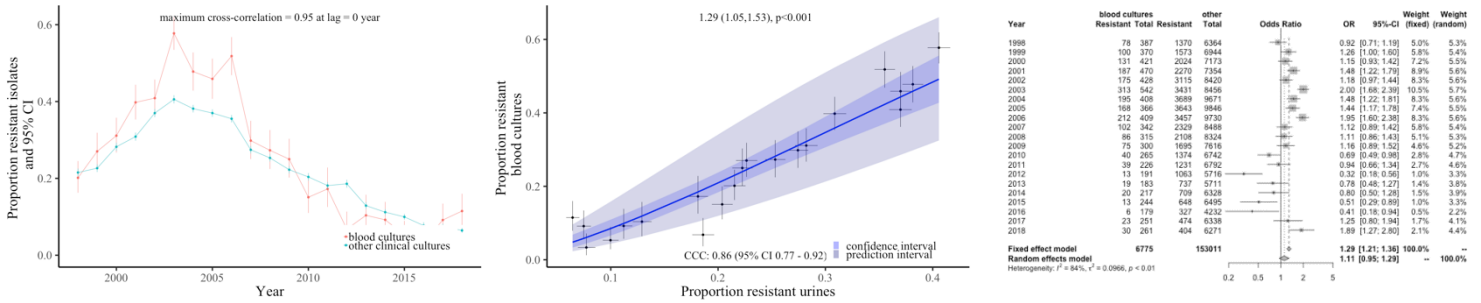
Figure 50 Number of (A) S. aureus sterile site cultures (excluding blood cultures) (B) S. aureus non-sterile site cultures split by susceptibility to ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin, tetracycline, 1998-2018

Figure 51 and **Figure 52** illustrate results for non-blood cultures, sterile site cultures and non-sterile site cultures for ciprofloxacin and erythromycin. Results are similar between the three groups of infection sites and resemble those seen for trimethoprim when focusing on *E. coli*. As expected, agreement is slightly better when considering resistance in sterile site cultures versus bloodstream infections, but differences are smaller than for *E. coli*. As in a low-income country, non-sterile site cultures would be more likely to be taken and hence have susceptibility tests I chose to present the results for all six antibiotics comparing this

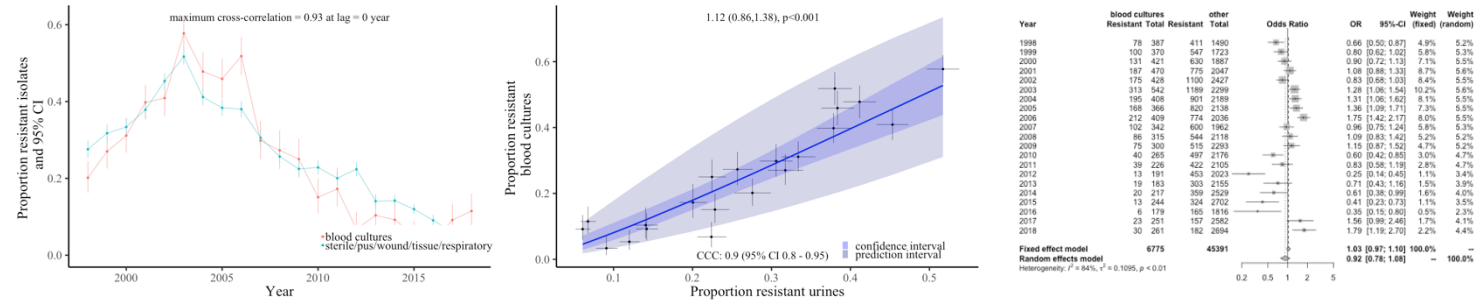
group of infection sites against bloodstream isolates. Results were very similar between ciprofloxacin, erythromycin and flucloxacillin, and bloodstream infections were much more similar to other sites of infections when considering pooled odds ratios compared to *E. coli* (**Table 17**). They were also similar for penicillin, however, >80% of isolates were resistant in every year where testing was done, as expected.

For gentamicin and tetracycline resistance there was a lot more uncertainty in the fitted models due to the low levels of resistance (<15%). However, it is notable that tetracycline was the one drug where there was little similarity between the time series of the proportion resistant over time in the different types of samples. Overall, there were also stronger effects of calendar time for resistance in *S. aureus*, reflecting generally higher rates of resistance in blood compared with non-sterile site cultures during the MRSA epidemic in the mid-2000s which then became more similar (**Figure 61**-**Figure 66**).

S. aureus
ciprofloxacin-resistant
blood cultures
vs
other clinical cultures



S. aureus
ciprofloxacin-resistant blood
cultures
vs
sterile/pus/wound/
tissue/respiratory



S. aureus
ciprofloxacin-resistant
blood cultures
vs
skin
surface/genital/urine
cultures

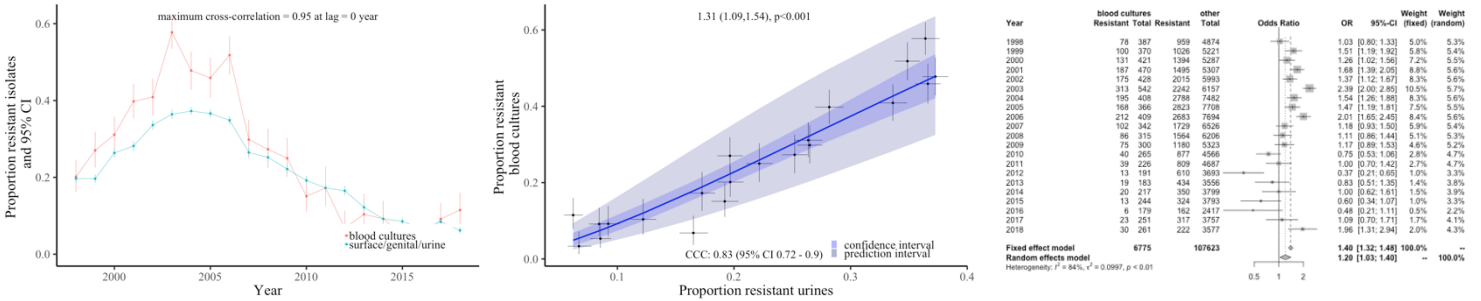
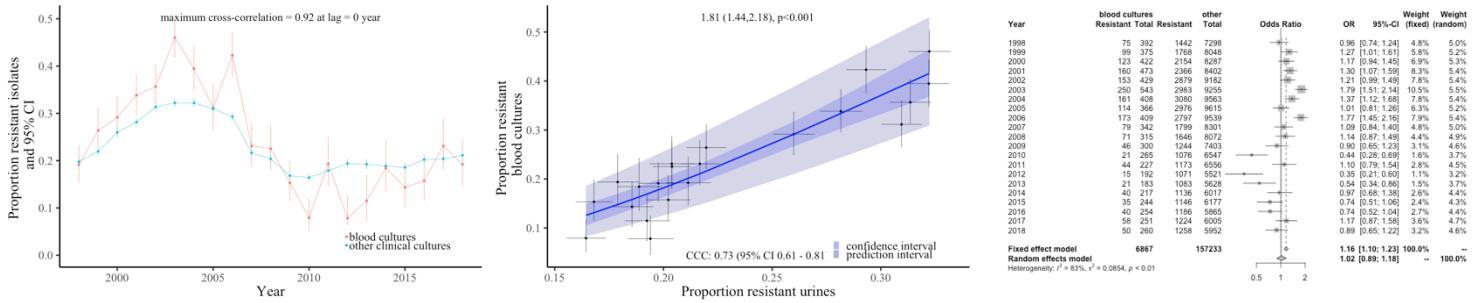
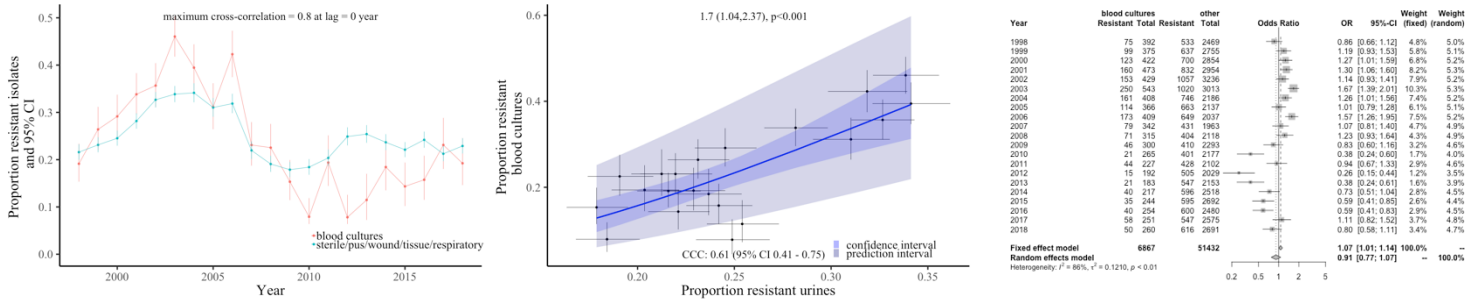


Figure 51 Proportion ciprofloxacin-resistant blood cultures vs non-blood cultures, sterile site cultures and non-sterile site cultures
Note: In the middle plot the x and y axes are on different scales. Please see electronic version for easier visualisation and **Table 17** for summary of key results.

S. aureus
erythromycin-resistant
blood culture
vs
other clinical cultures



S. aureus
erythromycin-resistant
blood culture
vs
sterile/pus/wound/
tissue/respiratory



S. aureus
erythromycin-resistant
blood culture
vs
skin
surface/genital/urine

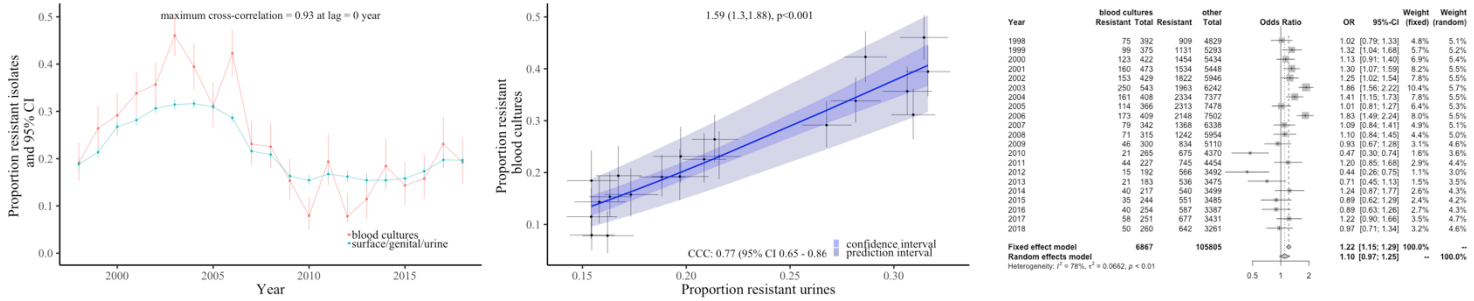
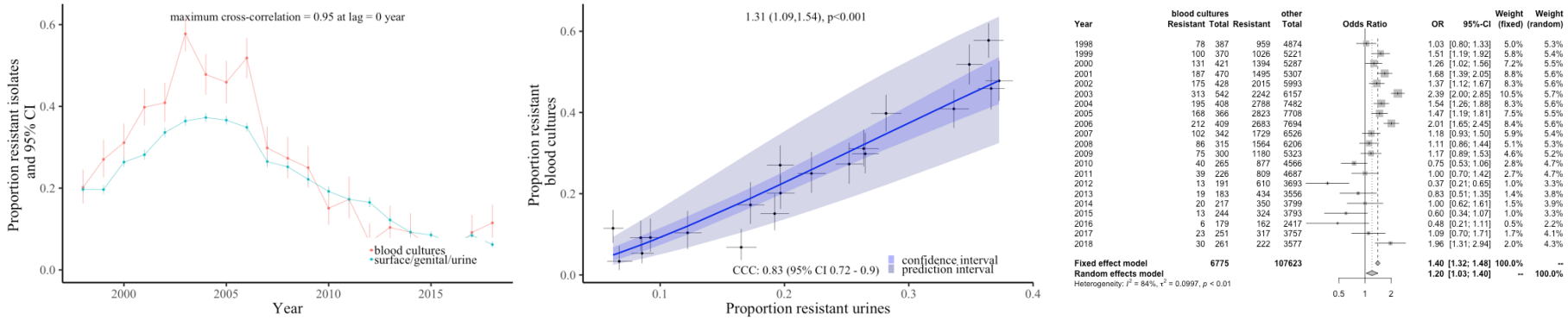


Figure 52 Proportion erythromycin-resistant blood cultures vs non-blood cultures, sterile site cultures and non-sterile site cultures

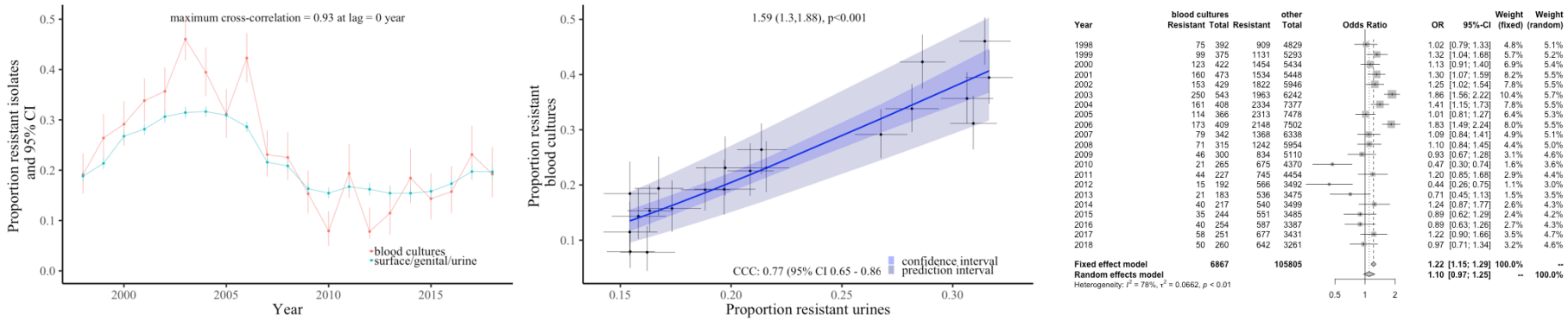
Note: In the middle plot the x and y axes are on different scales. Please see electronic version for easier visualisation and **Table 17** for summary of key results.

Figure 53 Agreement between proportion of resistant S. aureus bloodstream infections and proportion of resistant S. aureus non-sterile site cultures, for ciprofloxacin, erythromycin, flucloxacillin, gentamicin, penicillin, tetracycline

Ciprofloxacin-resistant

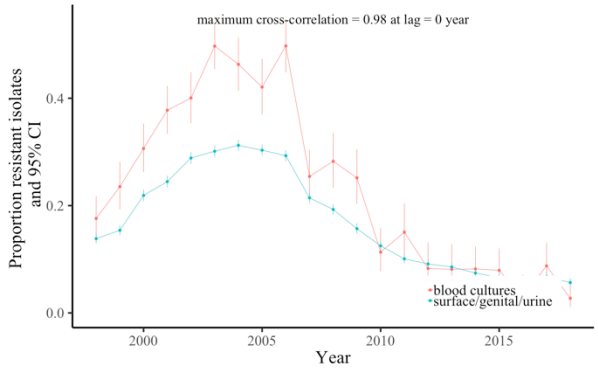


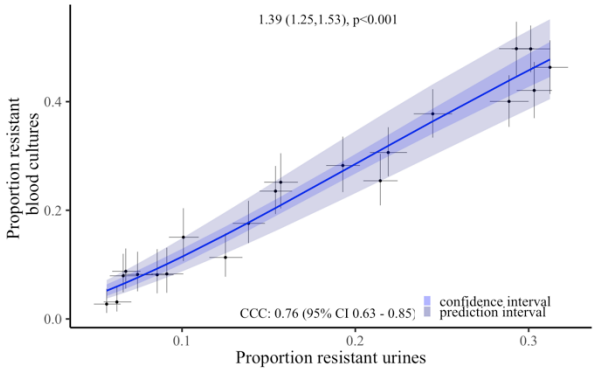
Erythromycin-resistant

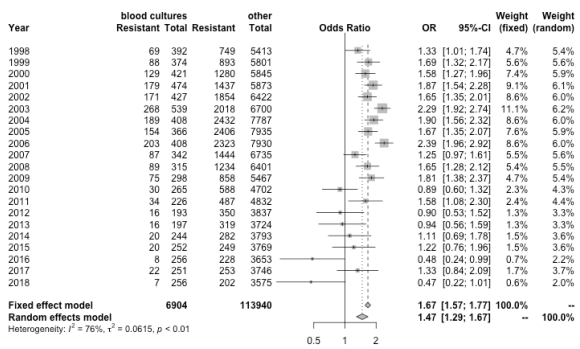


Note: In the middle plot the x and y axes are on different scales. Please see electronic version for easier visualisation and **Table 17** for summary of key results.

Flucloxacillin-resistant



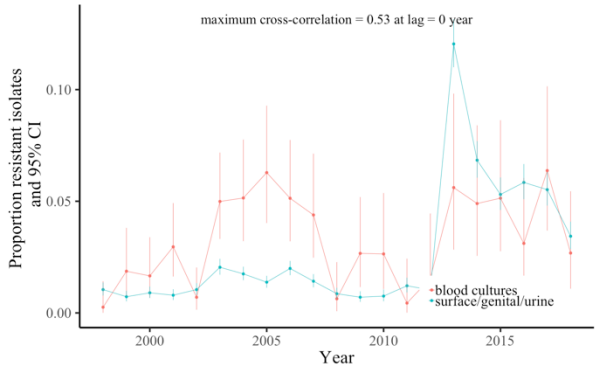


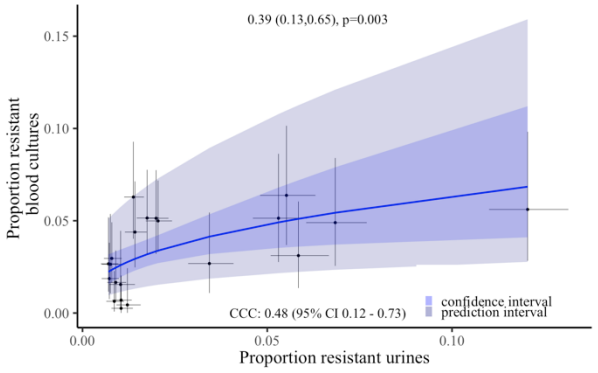


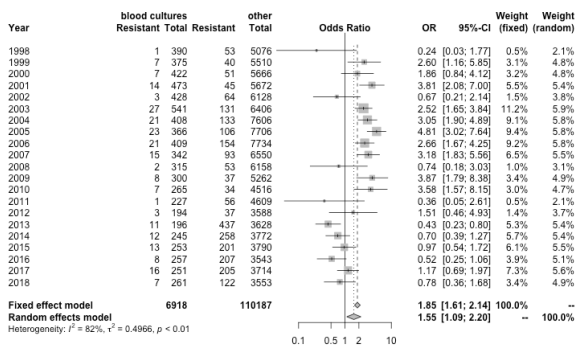
Year	blood cultures Resistant	Total	other Resistant	Total	Odds Ratio	OR	95%-CI	Weight (fixed)	Weight (random)
1998	69	392	749	5413		1.33	[1.01; 1.74]	4.7%	5.4%
1999	88	374	893	5801		1.69	[1.32; 2.17]	5.6%	5.6%
2000	129	421	1280	5845		1.58	[1.27; 1.96]	7.4%	5.9%
2001	179	474	1437	5873		1.87	[1.54; 2.28]	9.1%	6.1%
2002	171	427	1854	6422		1.65	[1.35; 2.01]	8.6%	6.0%
2003	268	539	2018	6700		2.29	[1.92; 2.74]	11.1%	6.2%
2004	189	408	2432	7787		1.90	[1.56; 2.32]	8.6%	6.0%
2005	154	366	2406	7935		1.67	[1.35; 2.07]	7.6%	5.9%
2006	203	408	2323	7930		2.39	[1.96; 2.92]	8.6%	6.0%
2007	87	342	1444	6735		1.25	[0.97; 1.61]	5.5%	5.6%
2008	89	315	1234	6401		1.65	[1.28; 2.12]	5.4%	5.5%
2009	75	298	858	5467		1.81	[1.38; 2.37]	4.7%	5.4%
2010	30	265	588	4702		0.89	[0.60; 1.32]	2.3%	4.3%
2011	34	226	487	4832		1.58	[1.08; 2.30]	2.4%	4.4%
2012	16	193	350	3837		0.90	[0.53; 1.52]	1.3%	3.3%
2013	16	197	319	3724		0.94	[0.56; 1.59]	1.3%	3.3%
2014	20	244	282	3793		1.11	[0.69; 1.78]	1.5%	3.6%
2015	20	252	249	3769		1.22	[0.76; 1.96]	1.5%	3.6%
2016	8	256	228	3653		0.48	[0.24; 0.99]	0.7%	2.2%
2017	22	251	253	3746		1.33	[0.84; 2.09]	1.7%	3.7%
2018	7	256	202	3575		0.47	[0.22; 1.01]	0.6%	2.0%
Fixed effect model	**6904**		**113940**			**1.67**	**[1.57; 1.77]**	**100.0%**	--
Random effects model						**1.47**	**[1.29; 1.67]**	--	**100.0%**

Heterogeneity: I^2 = 76%, τ^2 = 0.0615, p < 0.01

Gentamicin-resistant





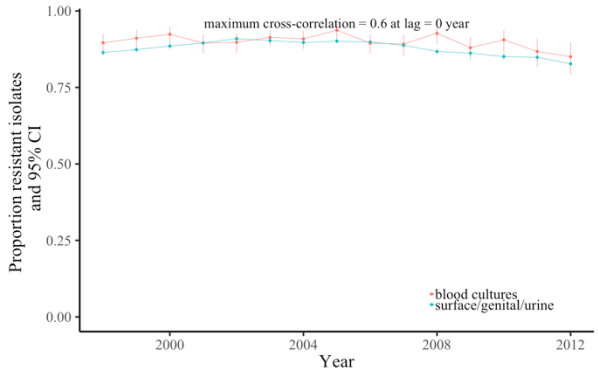


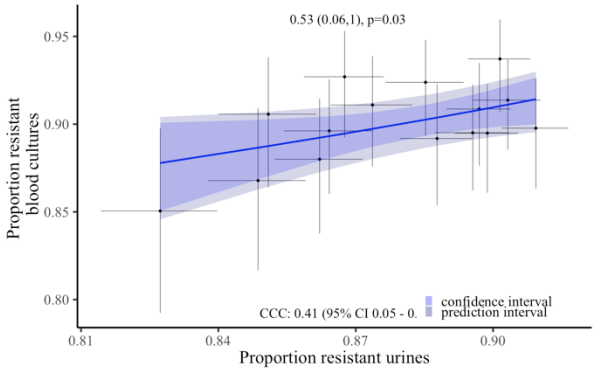
Year	blood cultures Resistant	Total	other Resistant	Total	Odds Ratio	OR	95%-CI	Weight (fixed)	Weight (random)
1998	1	390	53	5076		0.24	[0.03; 1.77]	0.5%	2.1%
1999	7	375	40	5510		2.60	[1.16; 5.85]	3.1%	4.8%
2000	7	422	51	5666		1.86	[0.84; 4.12]	3.2%	4.8%
2001	14	473	45	5672		3.81	[2.08; 7.00]	5.5%	5.4%
2002	3	428	64	6128		0.67	[0.21; 2.14]	1.5%	3.8%
2003	27	541	131	6406		2.52	[1.65; 3.84]	11.2%	5.9%
2004	21	408	133	7606		3.05	[1.90; 4.89]	9.1%	5.8%
2005	23	366	106	7706		4.81	[3.02; 7.64]	9.4%	5.8%
2006	21	409	154	7734		2.66	[1.67; 4.25]	9.2%	5.8%
2007	15	342	93	6550		3.18	[1.83; 5.56]	6.5%	5.5%
2008	2	315	53	6158		0.74	[0.18; 3.03]	1.0%	3.1%
2009	8	300	37	5262		3.87	[1.79; 8.38]	3.4%	4.9%
2010	7	265	34	4516		3.58	[1.57; 8.15]	3.0%	4.7%
2011	1	227	56	4609		0.36	[0.05; 2.61]	0.5%	2.1%
2012	3	194	37	3588		1.51	[0.46; 4.93]	1.4%	3.7%
2013	11	196	437	3628		0.43	[0.23; 0.80]	5.3%	5.4%
2014	12	245	258	3772		0.70	[0.39; 1.27]	5.7%	5.4%
2015	13	253	201	3790		0.97	[0.54; 1.72]	6.1%	5.5%
2016	8	257	207	3543		0.52	[0.25; 1.06]	3.9%	5.1%
2017	16	251	205	3714		1.17	[0.69; 1.97]	7.3%	5.6%
2018	7	261	122	3553		0.78	[0.36; 1.68]	3.4%	4.9%
Fixed effect model	**6918**		**110187**			**1.85**	**[1.61; 2.14]**	**100.0%**	--
Random effects model						**1.55**	**[1.09; 2.20]**	--	**100.0%**

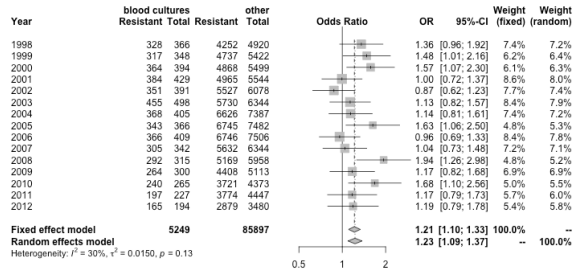
Heterogeneity: I^2 = 82%, τ^2 = 0.4966, p < 0.01

Note: In the middle plot the x and y axes are on different scales. Please see electronic version for easier visualisation and **Table 17** for summary of key results.

Penicillin-resistant



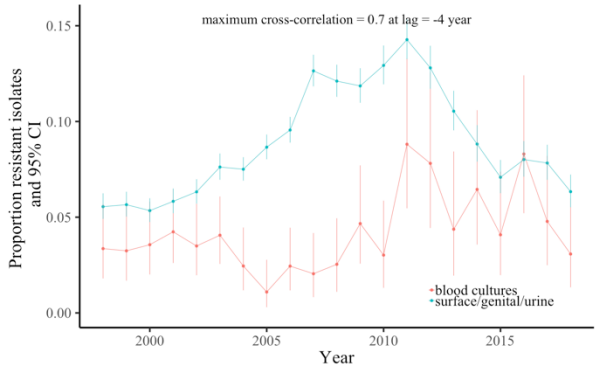


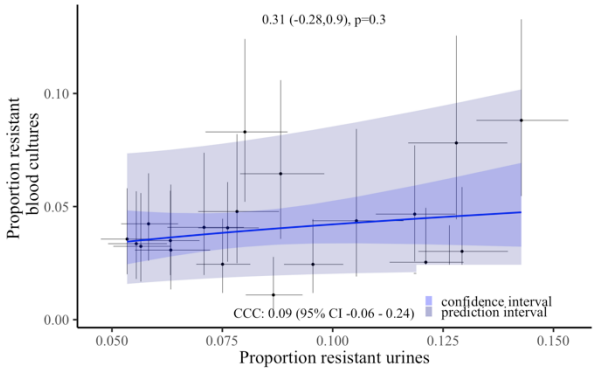


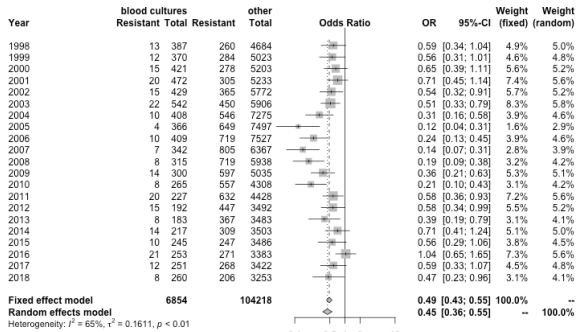
Year	blood cultures Resistant	Total	other Resistant	Total	Odds Ratio	OR	95%-CI	Weight (fixed)	Weight (random)
1998	328	366	4252	4920		1.36	[0.96; 1.92]	7.4%	7.2%
1999	317	348	4737	5422		1.48	[1.01; 2.16]	6.2%	6.4%
2000	364	394	4868	5499		1.57	[1.07; 2.30]	6.1%	6.3%
2001	384	429	4965	5544		1.00	[0.72; 1.37]	8.6%	8.0%
2002	351	391	5527	6078		0.87	[0.62; 1.23]	7.7%	7.4%
2003	455	498	5730	6344		1.13	[0.82; 1.57]	8.4%	7.9%
2004	368	405	6626	7387		1.14	[0.81; 1.61]	7.4%	7.2%
2005	343	366	6745	7482		1.63	[1.06; 2.50]	4.8%	5.3%
2006	366	409	6746	7506		0.96	[0.69; 1.33]	8.4%	7.8%
2007	305	342	5632	6344		1.04	[0.73; 1.48]	7.2%	7.1%
2008	292	315	5169	5958		1.94	[1.26; 2.98]	4.8%	5.2%
2009	264	300	4408	5113		1.17	[0.82; 1.68]	6.9%	6.9%
2010	240	265	3721	4373		1.68	[1.10; 2.56]	5.0%	5.5%
2011	197	227	3774	4447		1.17	[0.79; 1.73]	5.7%	6.0%
2012	165	194	2879	3480		1.19	[0.79; 1.78]	5.4%	5.8%
Fixed effect model	**5249**		**85897**			**1.21**	**[1.10; 1.33]**	**100.0%**	--
Random effects model						**1.23**	**[1.09; 1.37]**	--	**100.0%**

Heterogeneity: I^2 = 30%, τ^2 = 0.0150, p = 0.13

Tetracycline-resistant







Year	blood cultures Resistant	Total	other Resistant	Total	Odds Ratio	OR	95%-CI	Weight (fixed)	Weight (random)
1998	13	387	260	4684		0.59	[0.34; 1.04]	4.9%	5.0%
1999	12	370	284	5023		0.56	[0.31; 1.01]	4.6%	4.8%
2000	15	421	278	5203		0.65	[0.39; 1.11]	5.6%	5.2%
2001	20	472	305	5233		0.71	[0.45; 1.14]	7.4%	5.6%
2002	15	429	365	5772		0.54	[0.32; 0.91]	5.7%	5.2%
2003	22	542	450	5906		0.51	[0.33; 0.79]	8.3%	5.8%
2004	10	408	546	7275		0.31	[0.16; 0.58]	3.9%	4.6%
2005	4	366	649	7497		0.12	[0.04; 0.31]	1.6%	2.9%
2006	10	409	719	7527		0.24	[0.13; 0.45]	3.9%	4.6%
2007	7	342	805	6367		0.14	[0.07; 0.31]	2.8%	3.9%
2008	8	315	719	5938		0.19	[0.09; 0.38]	3.2%	4.2%
2009	14	300	597	5035		0.36	[0.21; 0.63]	5.3%	5.1%
2010	8	265	557	4308		0.21	[0.10; 0.43]	3.1%	4.2%
2011	20	227	632	4428		0.58	[0.36; 0.93]	7.2%	5.6%
2012	15	192	447	3492		0.58	[0.34; 0.99]	5.5%	5.2%
2013	8	183	367	3483		0.39	[0.19; 0.79]	3.1%	4.1%
2014	14	217	309	3503		0.71	[0.41; 1.24]	5.1%	5.0%
2015	10	245	247	3486		0.56	[0.29; 1.06]	3.8%	4.5%
2016	21	253	271	3383		1.04	[0.65; 1.65]	7.3%	5.6%
2017	12	251	268	3422		0.59	[0.33; 1.07]	4.5%	4.8%
2018	8	260	206	3253		0.47	[0.23; 0.96]	3.1%	4.1%
Fixed effect model	**6854**		**104218**			**0.49**	**[0.43; 0.55]**	**100.0%**	--
Random effects model						**0.45**	**[0.36; 0.55]**	--	**100.0%**

Heterogeneity: I^2 = 65%, τ^2 = 0.1611, p < 0.01

Note: In the middle plot the x and y axes are on different scales. Please see electronic version for easier visualisation and **Table 17** for summary of key results.

Table 17 Summary of key results for agreement between proportion resistant in blood cultures vs non-sterile site cultures

	Maximum cross-correlation (panel B)	**Lin's concordance coefficient (95% CI)**	**Pooled OR for resistance in blood vs other samples* (95% CI) (panel E)**	**Heterogeneity I^2, p (panel E)**	**Impact of calendar time on log OR (95% CI) p (panel F)**	**Impact of log odds resistant other samples on log odds resistant bloods (95% CI) p (slope of model in panel G)**
E. coli						
Co-amoxiclav	0.96 at lag=0 year	+0.71 (+0.54,+0.83)	1.99 (1.79,2.21)	74%, p<0.01	-0.01 (-0.03,+0.01), p=0.23	+0.86 (+0.75,+0.97), p<0.001
Amoxicillin	0.78 at lag=1 year	+0.18 (+0.06,+0.28)	1.55 (1.42,1.69)	72%, p<0.01	+0.01 (-0.01,+0.02), p=0.27	+1.94 (+1.21,+2.67), p<0.001
Ciprofloxacin	0.94 at lag=0 year	+0.58 (+0.40,+0.72)	1.69 (1.54,1.85)	38%, p=0.04	+0.01 (-0.01,+0.03), p=0.32	+1.15 (+0.94,+1.36), p<0.001
Trimethoprim	0.80 at lag=0 year	+0.78 (+0.54,+0.90)	0.93 (0.82,1.05)	85%, p<0.01	+0.03 (+0.01,+0.04), p<0.001	+0.70 (+0.45,+0.94), p<0.001
S. aureus						
Ciprofloxacin	0.95 at lag=0 year	+0.83 (+0.72,+0.90)	1.20 (1.03,1.40)	84%, p<0.01	-0.03 (-0.06,+0.00), p=0.03	+1.31 (+1.09,+1.54), p<0.001
Erythromycin	0.93 at lag=0 year	+0.77 (+0.65,+0.86)	1.10 (0.97,1.25)	78%, p<0.01	-0.02 (-0.04,+0.00), p=0.05	+1.59 (+1.30,+1.88), p<0.001
Flucloxacillin	0.98 at lag=0 year	+0.76 (+0.63,+0.85)	1.47 (1.29,1.67)	76%, p<0.01	-0.04 (-0.06,-0.02), p<0.001	+1.39 (+1.25,+1.53), p<0.001
Gentamicin	0.53 at lag=0 year	+0.48 (+0.12,+0.73)	1.55 (1.09,2.20)	82%, p<0.01	-0.07 (-0.12,-0.02), p=0.004	+0.39 (+0.13,+0.65), p=0.003
Penicillin	0.60 at lag=0 year	+0.41 (+0.05,+0.67)	1.23 (1.09,1.37)	30%, p=0.13	+0.00 (-0.03,+0.03), p=0.90	+0.53 (+0.06,+1.00), p=0.03
Tetracycline	0.70 at lag=-4 year	+0.09 (-0.06,+0.24)	0.45 (0.36,0.55)	65%, p<0.01	+0.01 (-0.03,+0.05), p=0.63	+0.31 (-0.28,+0.90), p=0.30

* from random effects meta-analysis treating each calendar year as an independent "study" vs urines for *E. coli*, vs non-sterile site cultures for *S. aureus*

4.3.1.3 Pooling all antibiotics

As results within the same pathogen were similar across all drugs, I also analysed results for all drugs pooled. One could argue that many clinicians would prescribe an antibiotic if resistance prevalence is less than 5% and not prescribe an antibiotic if resistance prevalence is greater than 20%. For mild infections an antibiotic could possibly still be administered if resistance prevalence was between 10% and 20%, while if resistance prevalence was between 5% and 10%, decision would be unclear. Whilst all these thresholds are arbitrary, **Figure 54** illustrates however how results would be very similar if these thresholds were to be shifted. I have therefore analysed the proportion of drug-years proportion resistant blood cultures depending on proportion resistant other cultures, using the 4 levels of resistance (<5%, 5-10%, 10-20%, >20%) (**Figure 55**). Pooling together all the drugs allowed a higher number of proportion resistant blood cultures to be available for analysing associations when resistance was less than 20%. Allowing for the uncertainty when proportion resistant was between 5% and 10%, proportion resistant in other cultures would have been able to inform appropriateness of empirical antimicrobial therapy with a specific drug for bloodstream infections in Oxfordshire (**Figure 55**). 142/204 (70%) drug-years had the same resistance category for both bloods and other types of culture (categorical agreement), 197/204 (97%) had either the same resistance category or within ± 1, i.e. adjacent category (essential agreement). The remaining discrepancies were observed in 5 years looking at tetracycline resistance categories in *S. aureus* cultures, with 10-20% resistance prevalence in non-sterile cultures, and <5% in blood cultures, and in 2 years looking at co-amoxiclav resistance categories in *E. coli* with <5% in urines vs 10-20% in blood cultures (year 2001), respectively 5-10% in urines and >20% in blood cultures (year 2012).

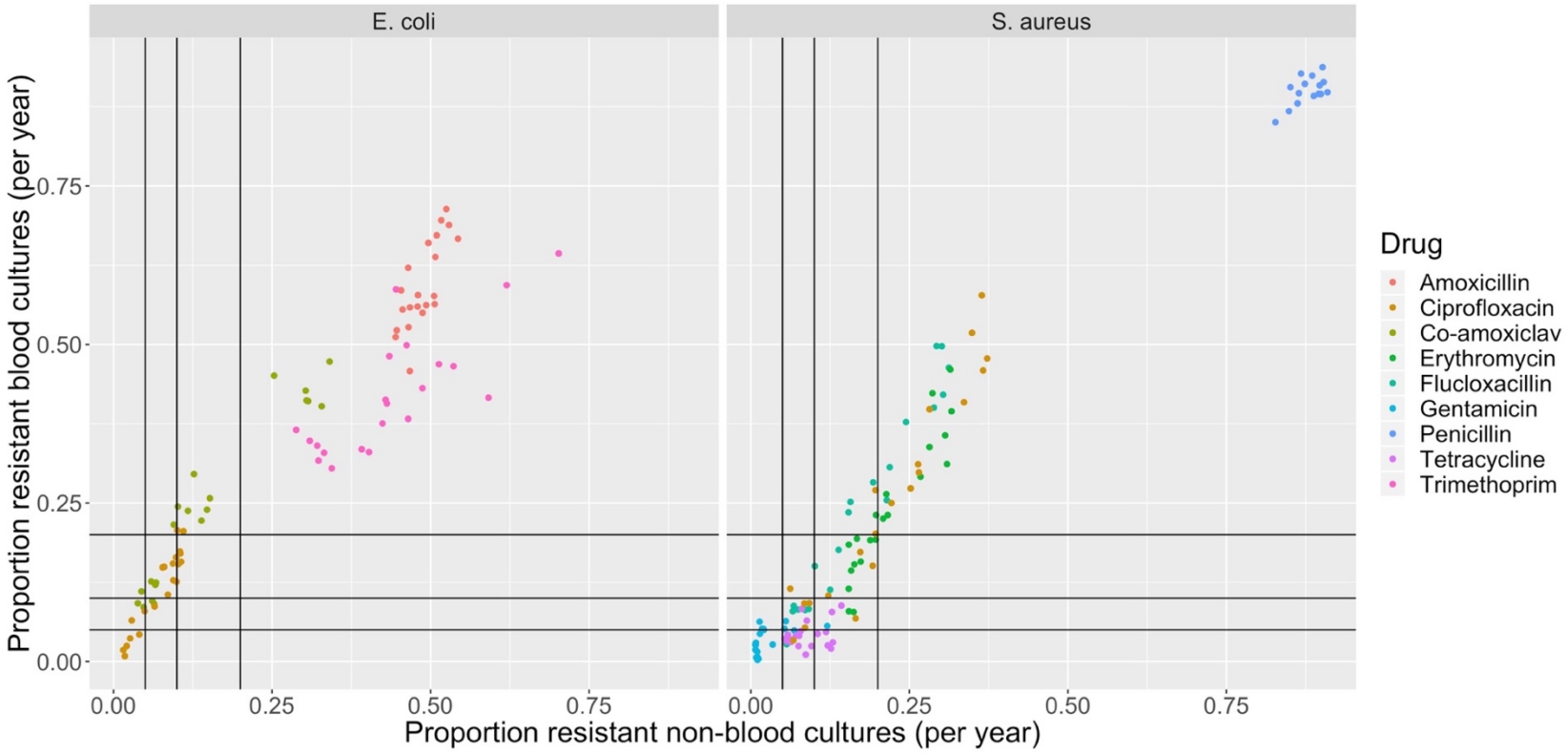


Figure 54 Proportion resistant non-blood cultures versus proportion resistant blood cultures per year for all antibiotics

Note: Each dot reflects resistance prevalence for one antibiotic in one specific year, with 21 dots per antibiotic (one for each year 1998-2018). Lines are drawn at 0.05 (5%), 0.1 (10%) and 0.2 (20%) resistance prevalence in each of blood vs non-blood cultures for ease of visualisation of agreement between clinically meaningful resistance categories in these two types of samples.

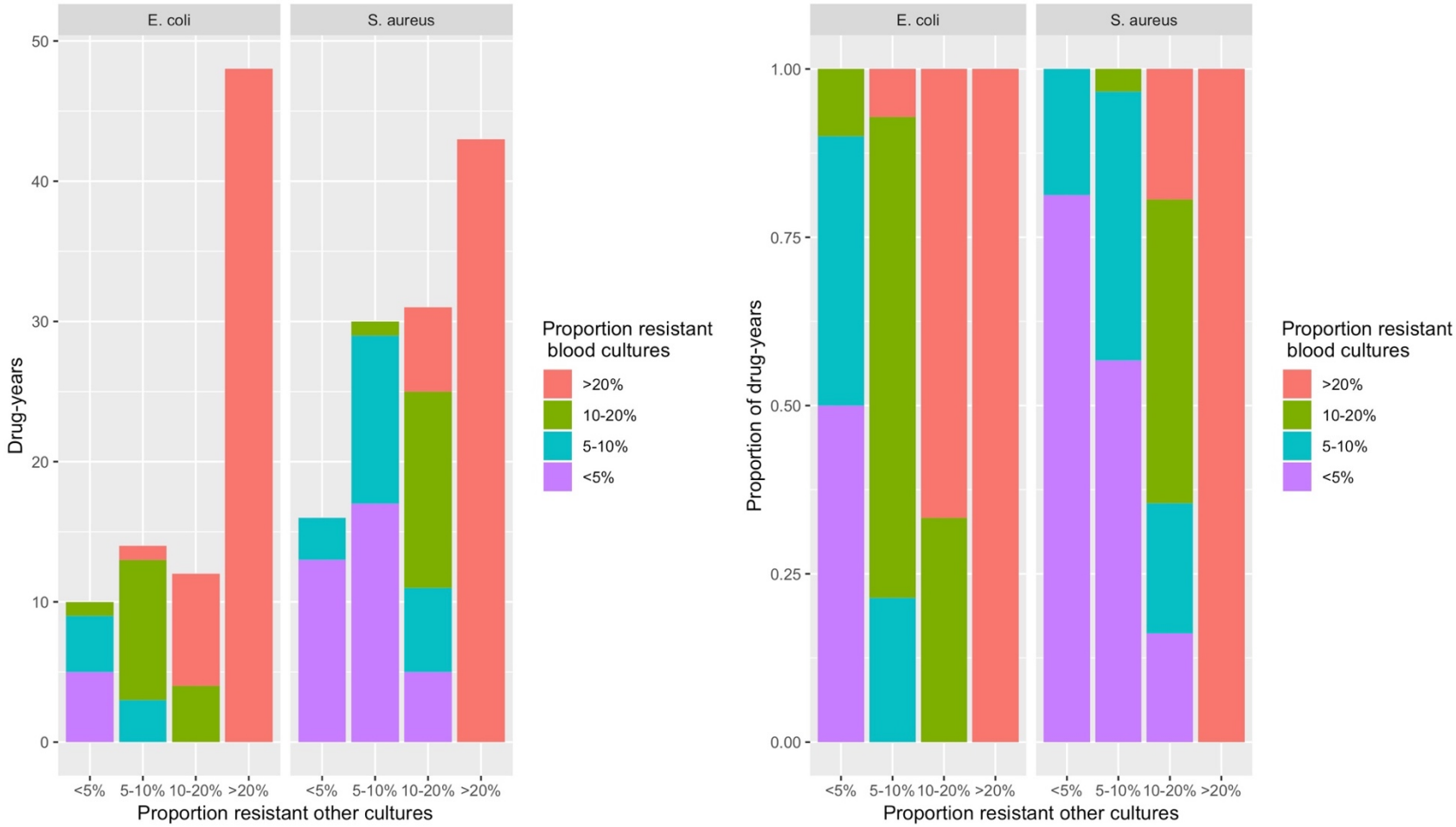


Figure 55 Counts (left) and proportions (right) of drug-years with varying proportions of resistant blood cultures classified by the proportion with resistant urine/non-sterile site cultures

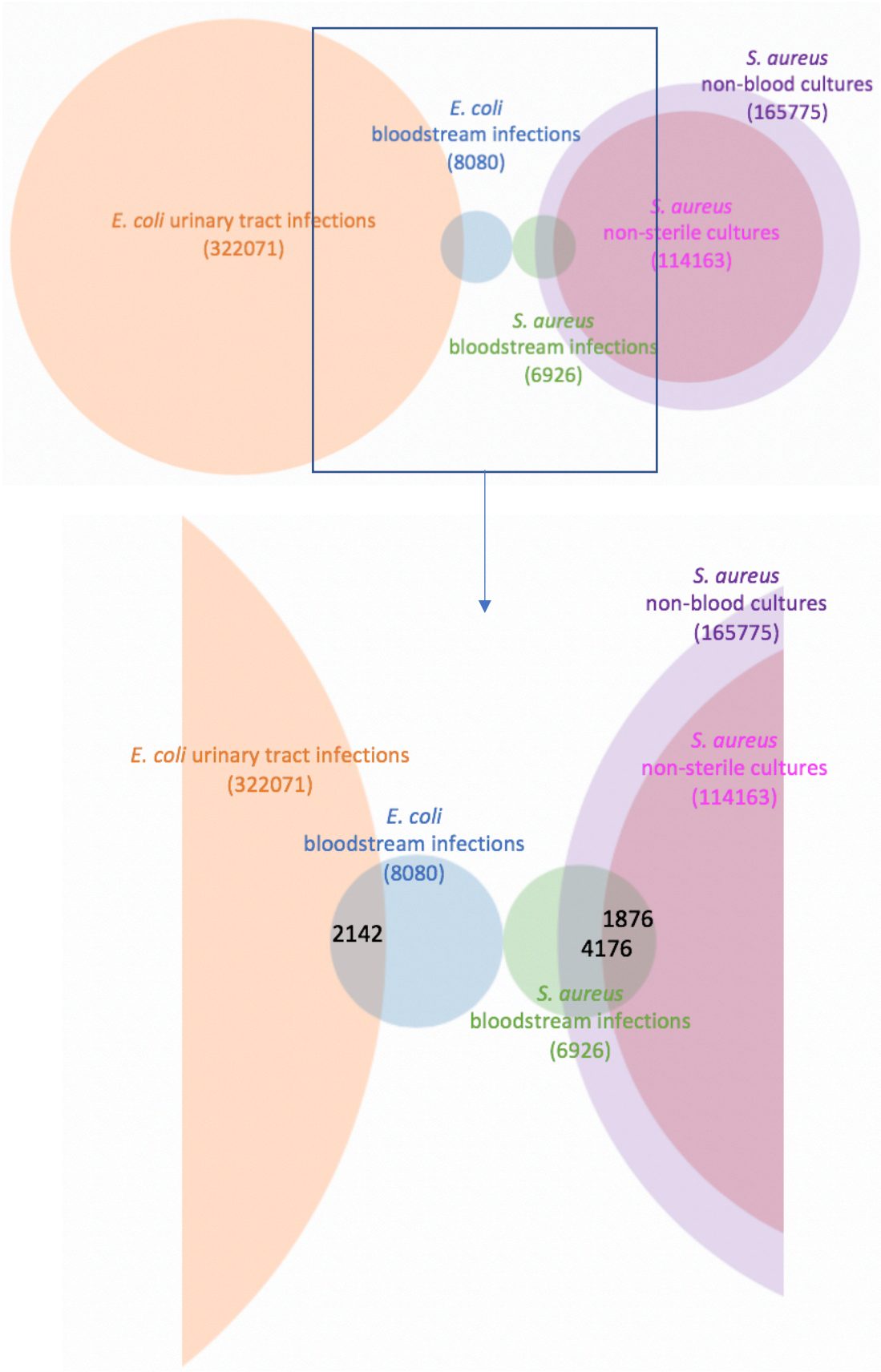
Note: Grouping points from Figure 54 into clinically meaningful resistance categories. For example, in the leftmost column there are 10 drug-years with <5% resistance prevalence in E. coli urine cultures, 5 years (5/10=0.5 on the right) for which resistance prevalence in blood cultures was <5%, 4 (4/10=0.4) for which it was 5-10% and 1 (1/10=0.1) for which it was 10-20%.

4.3.2 Individual-level analyses

One reason for the less than perfect agreement between resistance prevalence in blood and other cultures in the analyses above is that different patients are included in the two estimates, which are simply based on all bloodstream isolates and all isolates from other samples, regardless of patient population (as this total dataset would be simple to obtain from a microbiology laboratory). In contrast, in this section I focus on using susceptibility results in other sites of infections to predict resistance in bloodstream infections within the same patient. Of the total 8080 *E. coli* bloodstream infections I was able to analyse in the first part of this Chapter (not all with susceptibility test results for all antibiotics, hence denominators are slightly smaller in individual figures), only 2142 (27%) had a *E. coli* urinary tract infection within [-3,+2] days of the blood culture **(Figure 56**), and only 1137(14%) had one within [-90,-3] days. Similarly of 6926 *S. aureus* bloodstream infections, only 1876 (27%) had a *S. aureus* skin surface/genital/urine culture within [-3,+2] days of the blood culture (**Figure 56**). The percentages were much higher when considering *S. aureus* from other clinical cultures with 4176 (60%) of the *S. aureus* bloodstream infections having one within [-3,+2] days (**Figure 56**). Therefore, even though there is overlap between bloodstream infections and non-blood cultures in both *E. coli* and *S. aureus*, there is also a large proportion of bloodstream infections with no matching positive non-blood culture, indicating that it could have developed from some other unobserved factors.

In **Table 18** I present contingency tables for all antibiotics for *E. coli* and *S. aureus*, together with performance metrics which are classically used when a replacement diagnostic tool is tested on the same sample (that is, paired data).

Figure 56 Venn diagram of E. coli urinary tract infections, E. coli bloodstream infections, S. aureus non-sterile site cultures, S. aureus non-blood cultures and S. aureus bloodstream infections



Focusing on the performance of using susceptibility results of the closest *E. coli* urinary tract infection taken at any time from 3 days before up to 2 days after the bloodstream infection isolate within the same patient, interestingly, even on this paired data, resistance to trimethoprim was higher in urinary than blood isolates for trimethoprim (McNemar $p<0.001$), but was lower in urinary than blood isolates for amoxicillin ($p=0.02$) and co-amoxiclav ($p<0.001$) (similarly to results in all isolates described above). There was no asymmetry in resistance for ciprofloxacin ($p=0.26$) with high overall agreement (99%). Guidance on susceptibility testing tools (Devices and Services 2003) require a major error (ME) of less than 3% and a very major error (VME) of less than 7.5% (upper 95% confidence limit for the true VME) which were only satisfied for ciprofloxacin when considering *E. coli*. But, for *E. coli* at least, there is no evidence that the modestly different resistance rates seen in blood and urine isolates in the overall analysis were due to differential sampling of intrinsically different underlying populations.

I next considered the concordance of susceptibility results from the closest prior *E. coli* urinary tract infection taken between 3 and 90 days before the bloodstream infection within a patient, since this could be a potential indicator of inappropriate empiric therapy within a high-income country. Whilst resistance to amoxicillin and co-amoxiclav was still lower in urinary than blood isolates (McNemar $p<0.001$), interestingly there was no difference in trimethoprim resistance between prior urinary isolates and current blood isolates ($p=0.75$) potentially implicating inadequate treatment of a prior urinary tract infection with trimethoprim as contributing to bloodstream infection. Diagnostic performance was lower for all antibiotics compared with concurrent isolates described above; however, positive predictive value (PPV) was 80% or higher for all drugs including co-amoxiclav, meaning that

four out of five patients with a prior co-amoxiclav resistant urinary tract infection would turn out to have a co-amoxiclav resistant bloodstream infection.

For *S. aureus*, unlike *E. coli*, overall accuracy between resistance from concurrent blood and non-blood cultures was very high (97-99%) with ciprofloxacin, erythromycin and flucloxacillin having VME under 3% and ME under 2% when looking at other clinical cultures, and only slightly higher values when considering only non-sterile site cultures (VME under 5% and ME under 3%). For these three drugs there was some evidence of asymmetry (McNemar p-values between 0.04-0.07), but numbers were small, and if anything resistance was higher in non-blood cultures. The ME for gentamicin and tetracycline were only 1%, however the VME were 30% and 8% respectively. The ME for penicillin was 8%, however the VME was only 2%; for penicillin there were twice as many patients with bloodstream-resistant non-blood culture-susceptible isolates than the converse.

Table 18 Performance of using susceptibility results within the same patient in other sites of infections as a marker for bloodstream infection susceptibility

	Susceptible bloodstream infections		Resistant bloodstream infections		ME	VME	McNemar	Accuracy	PPV	NPV
	Susceptible other site	Resistant other site	Susceptible other site	Resistant other site						
Concurrent E. coli										
Co-amoxiclav	1303	72	127	373	5%	25%	<0.001	89%	84%	91%
Amoxicillin	640	55	84	1071	8%	7%	0.018	92%	95%	88%
Ciprofloxacin	1600	13	7	257	1%	3%	0.26	99%	95%	100%
Trimethoprim	837	171	88	728	17%	11%	<0.001	96%	81%	90%
Previous *E. coli*										
Co-amoxiclav	479	47	103	194	9%	35%	<0.001	82%	80%	82%
Amoxicillin	217	33	94	461	13%	17%	<0.001	84%	93%	70%
Ciprofloxacin	622	18	33	151	3%	18%	0.05	94%	89%	95%
Trimethoprim	290	83	78	338	22%	19%	0.75	80%	80%	79%
Non-blood *S. aureus*										
Ciprofloxacin	1450	34	18	707	2%	2%	0.04	98%	95%	99%
Erythromycin	1505	35	17	556	2%	3%	0.02	98%	94%	99%
Flucloxacillin	1567	21	10	674	1%	1%	0.07	99%	97%	99%
Gentamicin	2074	18	26	60	1%	30%	0.29	98%	77%	99%
Penicillin	154	16	35	1530	9%	2%	0.01	97%	99%	81%
Tetracycline	1987	10	6	69	1%	8%	0.45	99%	87%	100%
Non-sterile site *S. aureus*										
Ciprofloxacin	896	29	12	473	3%	2%	0.01	97%	94%	99%
Erythromycin	877	29	16	335	3%	5%	0.07	96%	92%	98%
Flucloxacillin	950	20	12	451	2%	3%	0.21	98%	96%	99%
Gentamicin	1244	14	17	42	1%	29%	0.72	98%	77%	99%
Penicillin	77	12	24	982	13%	2%	0.07	97%	99%	76%
Tetracycline	1199	4	3	35	0%	8%	1.00	99%	90%	100%

4.3.3 *E. coli* co-amoxiclav breakpoint

In the work above, I have used common breakpoints for *E. coli* co-amoxiclav resistance in blood and urine cultures. This is consistent with the goal of the Chapter of predicting resistance in blood cultures using resistance in urine cultures. However, interestingly, in 2014, EUCAST introduced a different breakpoint for co-amoxiclav for uncomplicated UTIs, of 32/2 mg/L rather than 8/2 mg/L, the recommended one for all other cultures. As Phoenix™ only allows one breakpoint per antibiotic per pathogen, UTIs with an MIC >8/2 mg/L were reported as resistant to prescribers. It is notable that a large proportion of urines with an MIC >8/2 mg/L are actually ≤32/2 mg/L. This raises important questions about whether these UTIs really are complicated/uncomplicated and whether co-amoxiclav is still an appropriate choice.

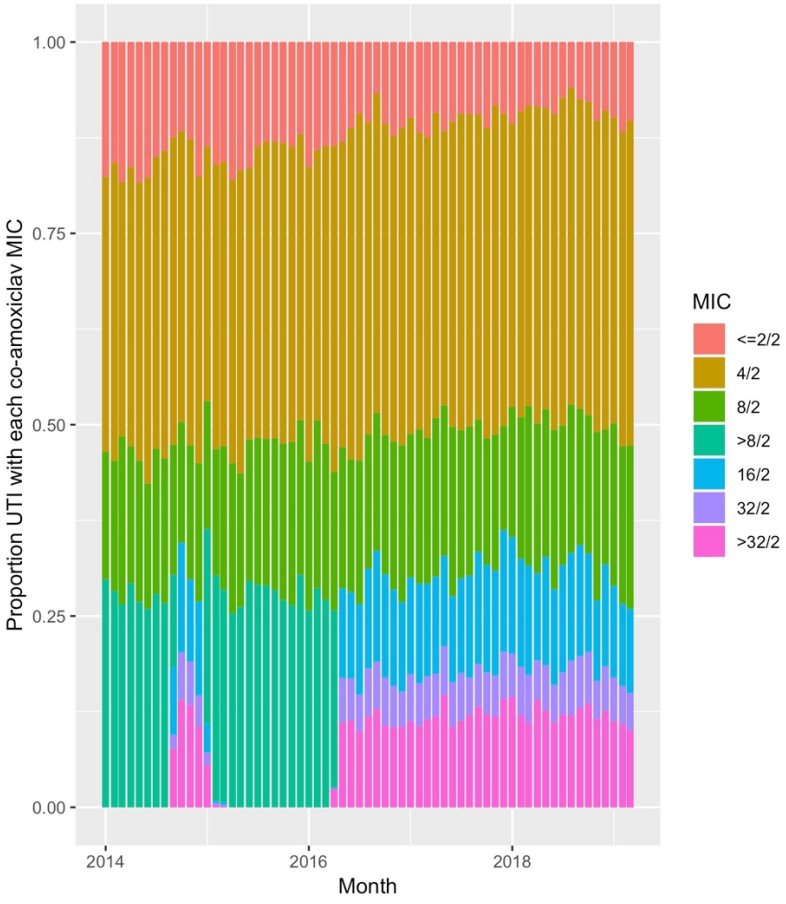


Figure 57 Co-amoxiclav MIC distribution per month post EUCAST breakpoint change

4.4 Discussion

In this Chapter I showed that, on balance, antimicrobial resistance results from infections in sites other than blood could potentially be used for the surveillance of antimicrobial resistance in bloodstream infections in low-income countries. Focusing on two of the most common bacteria which cause infectious diseases, namely *E. coli* and *S. aureus*, I showed that even though there is not complete agreement between the proportion of resistant blood cultures and the proportion of resistant other clinical cultures for the majority of antibiotics, for all antibiotics there is a strong association between the two, both in terms of actual values, but also – potentially more importantly - in terms of trends over time. This suggests that monitoring changes in resistance over time could be done using samples that are easier to collect than blood.

From my search of the published literature, I have not found any studies directly looking at associations between resistance in blood versus other sites of infection. In 2011, Health Protection Scotland and National Services Scotland published a protocol for the surveillance of antimicrobial resistance in urinary tract infections in Scotland motivated by the fact that even though antimicrobial resistance surveillance is currently focusing on bloodstream infections, monitoring resistance in other sites of infection where the infection initially starts before entering the bloodstream would be more suitable for the identification of emergence and spread of resistance (Health Protection Scotland and National Services Scotland 2011). However, the hypothesised relationship was not analysed using the collected and published data (Health Protection Scotland and Information Services Division 2016). One study from the Netherlands considered antimicrobial resistance trends between 1998 and 2016 from livestock in commensal *E. coli*, as an animal reservoir for antimicrobial

resistance in humans (Hesp et al. 2019). However, the study did not compare trends seen in commensal *E. coli* from livestock to those seen in *E. coli* bloodstream infections. One other study considered resistance in blood, sputum, urine and wound cultures (Xie et al. 2012), however, their focus was on comparing resistance patterns in elderly nursing home and general community patients, rather than between the different sites of infection.

Current efforts in the direction of antimicrobial resistance surveillance in low-income countries are being directed towards building laboratories and educating staff in order to carry out antimicrobial susceptibility testing for blood cultures, as if this can be achieved, then other types of culture will become possible too. However, setting up such a system is very expensive, and it will be many years until they will be able to provide enough results on blood cultures for surveillance purposes. Further, a study by Vernet et al draws attention to the importance of building an infrastructure which is able to capture unbiased antimicrobial susceptibility trends. Existing infrastructures in low-income countries are often only available to wealthy patients leading to biased results (Vernet et al. 2014). Bloodstream infections are also a lot less common than less serious infections in other body sites; therefore, even once a number of laboratories have been set up to carry out routine clinical testing, there will be a period of time when not enough samples will be available to provide surveillance results with high enough confidence. Taking a culture from urine/skin surface/genital sites is a lot less invasive and therefore much easier to do on a large number of people from different communities. Being non-sterile sites, growth is much easier to achieve than from a sterile site like blood, and therefore many more of the samples taken would be useful for antimicrobial resistance surveillance. Even when considering Oxfordshire, a region in a high-income country, the advantage of looking at samples from

non-sterile sites for surveillance of antimicrobial resistance over blood cultures or samples from sterile sites can be seen in the narrower confidence intervals due to the larger number of isolates available.

One of the most important limitations of this work is that the associations between resistance proportions in other clinical cultures and blood cultures relies on the assumption that most bloodstream infections have had the source of infection from the other body site, that is, infection is driven from commensal colonising organisms that become pathogenic, rather than organisms being newly acquired relatively directly into the blood. Even though this may hold within the population I studied, it is possible, although fairly unlikely, that this may not generalise to a low-income country. One other limitation is that I have only considered surveillance of the proportion of resistant infections, not numbers or rates of resistant infections; however, in terms of empiric treatment recommendations, it is the proportion rather than absolute numbers that is of interest.

One question raised by my analyses is how good one would require the agreement with resistance from blood cultures to be in order for surveillance of resistance in samples from other sites to be clinically useful. In my dataset, significantly different resistance proportions were observed in the different sample types for many pathogen-drug combinations with some variation over time and less than perfect agreement. However, the cross-correlations were high, and overall the figures suggest that considering other sites of infection than blood for surveillance can be useful, even though traditional criteria for performance measures (like VME/ME) are difficult to apply as I expect the agreement to not be perfect. For example, one might expect a pooled odds ratio to not contain 1 based on current

epidemiological knowledge; specifically, a large proportion of *E. coli* bloodstream infections are known to be of urinary focus, but if the *E. coli* urinary tract infection was susceptible I would expect it to be treated and not lead to a bloodstream infection, which is what is hypothesised to happen with resistant poorly treated urinary tract infections. Hence one might expect the proportion of resistant bloodstream infections to be higher due to inadequate treatment, but underlying time trends to track each other, as I observed. Considering all pooled drug-years and 5%, 10% and 20% as arbitrary thresholds for resistance prevalence I showed that proportion resistant in other cultures could be used to inform empirical antimicrobial therapy.

One other limitation could have been the fact that different people were sampled for blood and other specimens, therefore differences might be due to different populations. However, my paired analysis suggests it is unlikely for this to be the case.

I have not considered surveillance of antimicrobial resistance by focusing on the entire antibiogram, that is patterns of resistance within one isolate across different antibiotics; rather I have analysed one antibiotic at a time, focusing in particular on pathogen-drug combinations which would be considered for treatment in clinical practice, such as ciprofloxacin and erythromycin for *S. aureus* bloodstream infections. I also pooled all samples together: in future work I could consider whether relationships are generalizable between different groups of patients. For instance, results might theoretically differ between children vs adults, community vs hospital infections, method of resistance testing, and across different countries (high- vs low- and middle-income countries). However, unlike in low-income countries where a large number of infections are seen in children, in the

Oxfordshire dataset that I have explored the number of bloodstream infections occurring in children is very small (<5% of *E. coli* bloodstream infections and <10% of *S. aureus* bloodstream infections).

In a high-income country, susceptibility testing is still not instantaneous, with microbiology results taking at least 48 hours to become available to clinicians. During this time, empiric antimicrobials are administered; however, increasing use of electronic health records means that it is becoming much simpler to check the history of the patient, for example looking for a previous urinary tract infection and if one is available, guiding their prescription based on this antibiogram, rather than simply treating with the guideline recommended empiric antibiotic. My dataset allowed me to explore whether a previous resistant *E. coli* urinary tract infection was associated with a subsequent resistant *E. coli* bloodstream infection. I looked at an arbitrarily chosen interval of 3 months – 3 days, guided by the analyses in Chapter 2; however, future work could consider a sensitivity analysis where I could investigate how the strength of this association changes as I looked further back in time. Along these lines, Yelin et al looked at predicting antimicrobial resistance for an individual community-acquired urinary tract infection in an individual patient by using personal clinical history details including records of previous resistant or susceptible previous urinary tract infections (Yelin et al. 2019). They found that incorporating demographic information together with susceptibility results of previous urinary tract infections and antibiotic purchases in a machine learning model allowed them to improve guidance of empirical treatment.

Asking clinicians to reduce antibiotic use and running successful stewardship programmes cannot be done in a complete information vacuum. Two studies from India identified the lack of microbiology facilities and the unwillingness of patients to have a sample taken as some of the main factors influencing antibiotic use (Kotwani et al. 2010; Ganguly et al. 2011). Antimicrobial resistance surveillance plays a key role in the optimal use of antibiotics and hence in the reduction of resistance. Low- and middle-income countries face numerous challenges in setting up a surveillance system similar to those in high-income countries including the lack of coherent governance, budget, technical expertise, information technology systems and co-ordination. It would be very expensive to set up laboratories that can handle blood cultures, and a lot of the blood culture results would be negative, severely restricting the information on antimicrobial susceptibility that can be obtained from them. However, for example *E. coli* is often found in urine, which is much easier to sample, much cheaper, faster and easier to grow from. The relationship between proportion resistant bloodstream infections and that from other sites of infection is therefore of great clinical significance as it can help low-income countries set up an appropriate strategy for antimicrobial resistance surveillance which in turn can provide evidence for decisions such as empiric antibiotic treatment. However, using population level surveillance as a substitute for individual patient testing should only be considered a stepping stone, i.e. a pragmatic step, and should not interfere with efforts to improve diagnostic capacity for individual patient management.

The next steps would be to identify some middle-income settings from which similar data can be obtained, even though the numbers are expected to be much smaller and the periods of time much shorter. This will allow me to test how well the results observed in

Oxfordshire generalise. Local antimicrobial susceptibility patterns will vary largely based on local antimicrobial prescribing practices. Implementing this surveillance strategy would involve sampling from non-sterile sites of infection, perhaps every 6 months from specific regions of interest. Depending on the resistance levels, both the width of the 95% confidence intervals around the estimated resistance prevalence would vary, but also their clinical significance. The higher the number of positive samples, the tighter the confidence intervals will be, however, obtaining more samples would increase the cost of surveillance. In designing such a surveillance strategy, estimates for the positivity rate of specific cultures from non-sterile sites would also be required in order to choose the number of cultures which should be taken.

4.5 Supplementary material

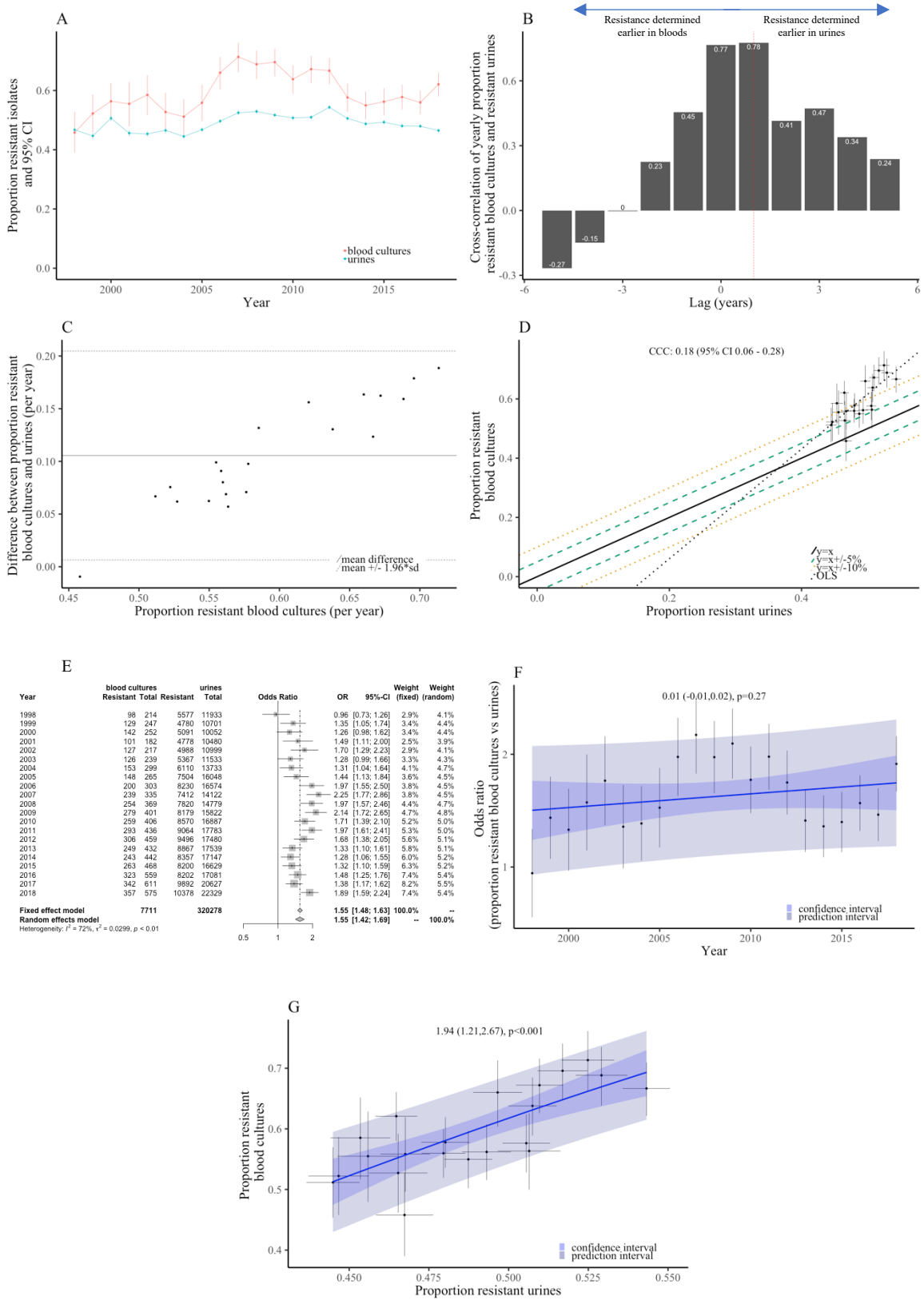


Figure 58 Agreement between proportion of amoxicillin-resistant E. coli bloodstream infections and proportion of amoxicillin-resistant E. coli urinary tract infections

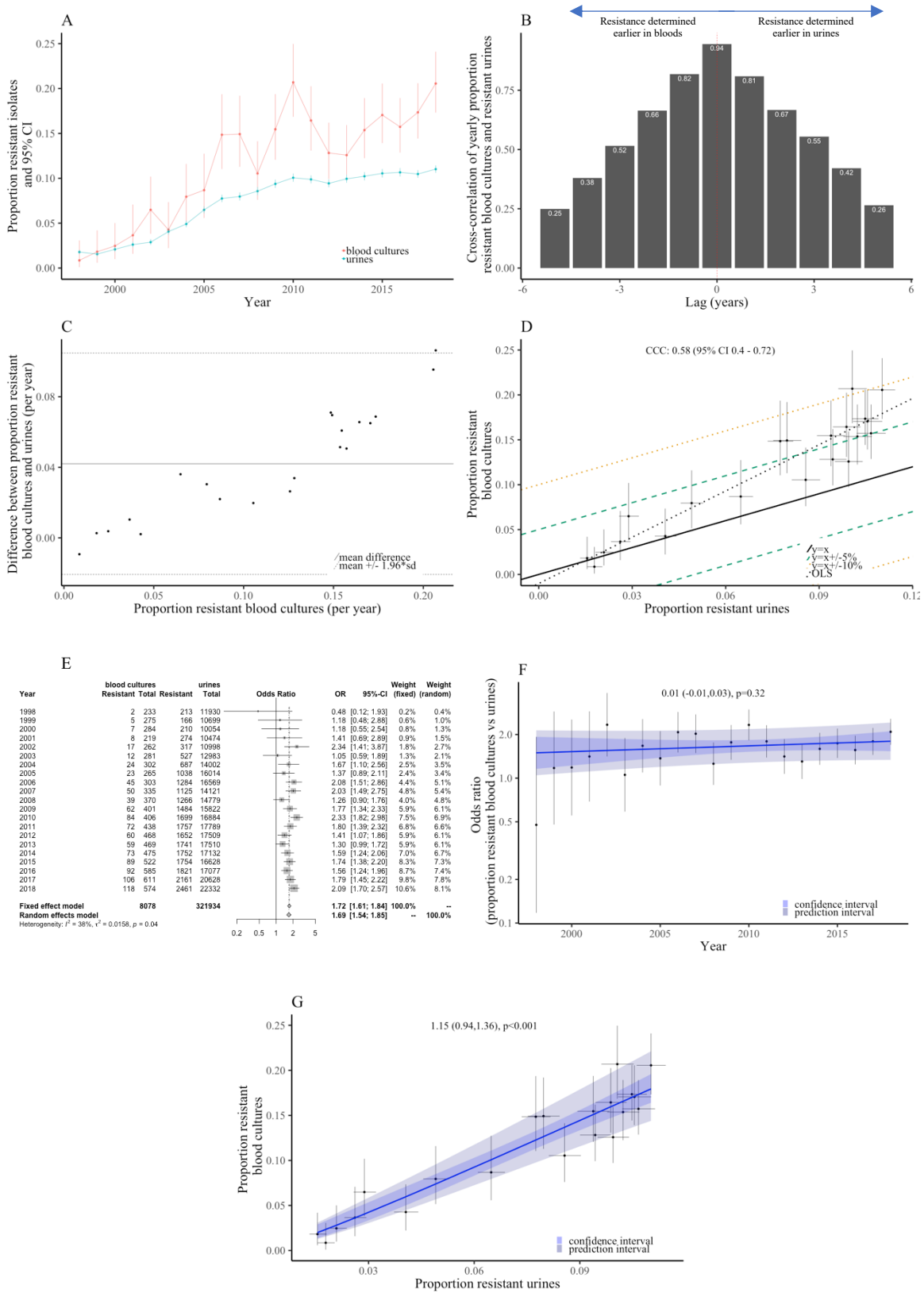


Figure 59 Agreement between proportion of ciprofloxacin-resistant E. coli bloodstream infections and proportion of ciprofloxacin-resistant E. coli urinary tract infections

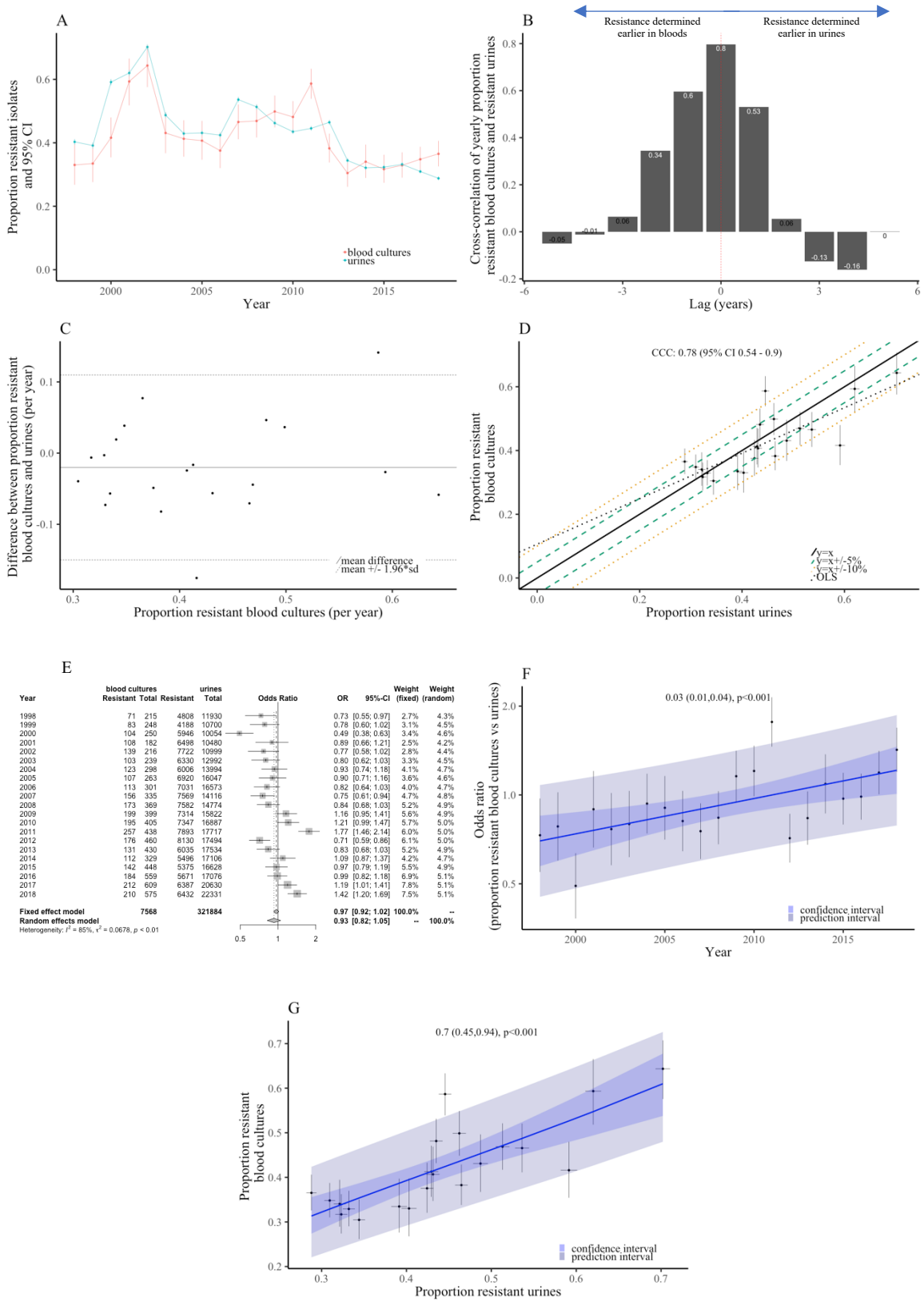


Figure 60 Agreement between proportion of trimethoprim-resistant E. coli bloodstream infections and proportion of trimethoprim-resistant E. coli urinary tract infection

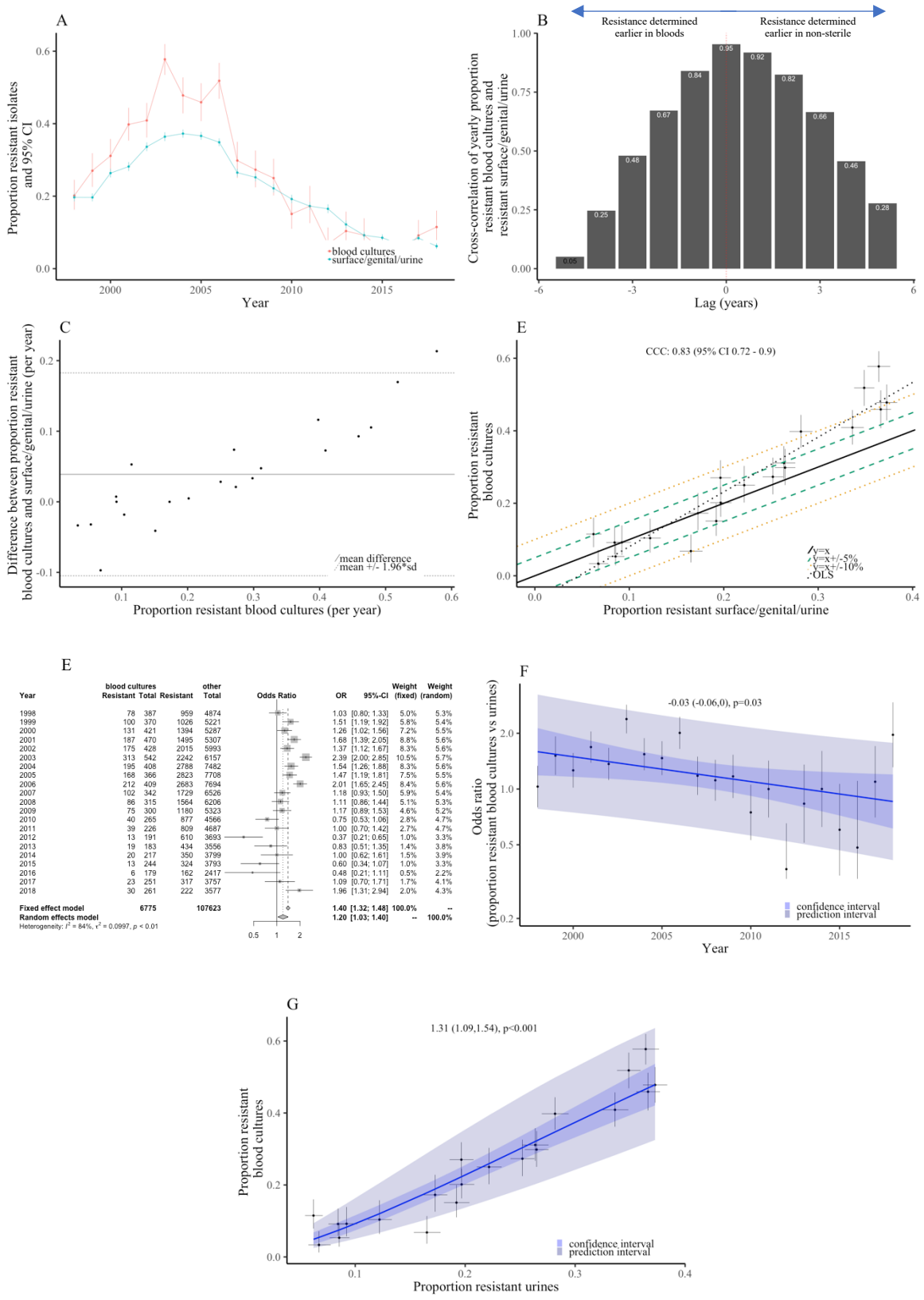


Figure 61 Agreement between proportion of ciprofloxacin-resistant S. aureus bloodstream infections and proportion of ciprofloxacin-resistant S. aureus non-sterile site cultures

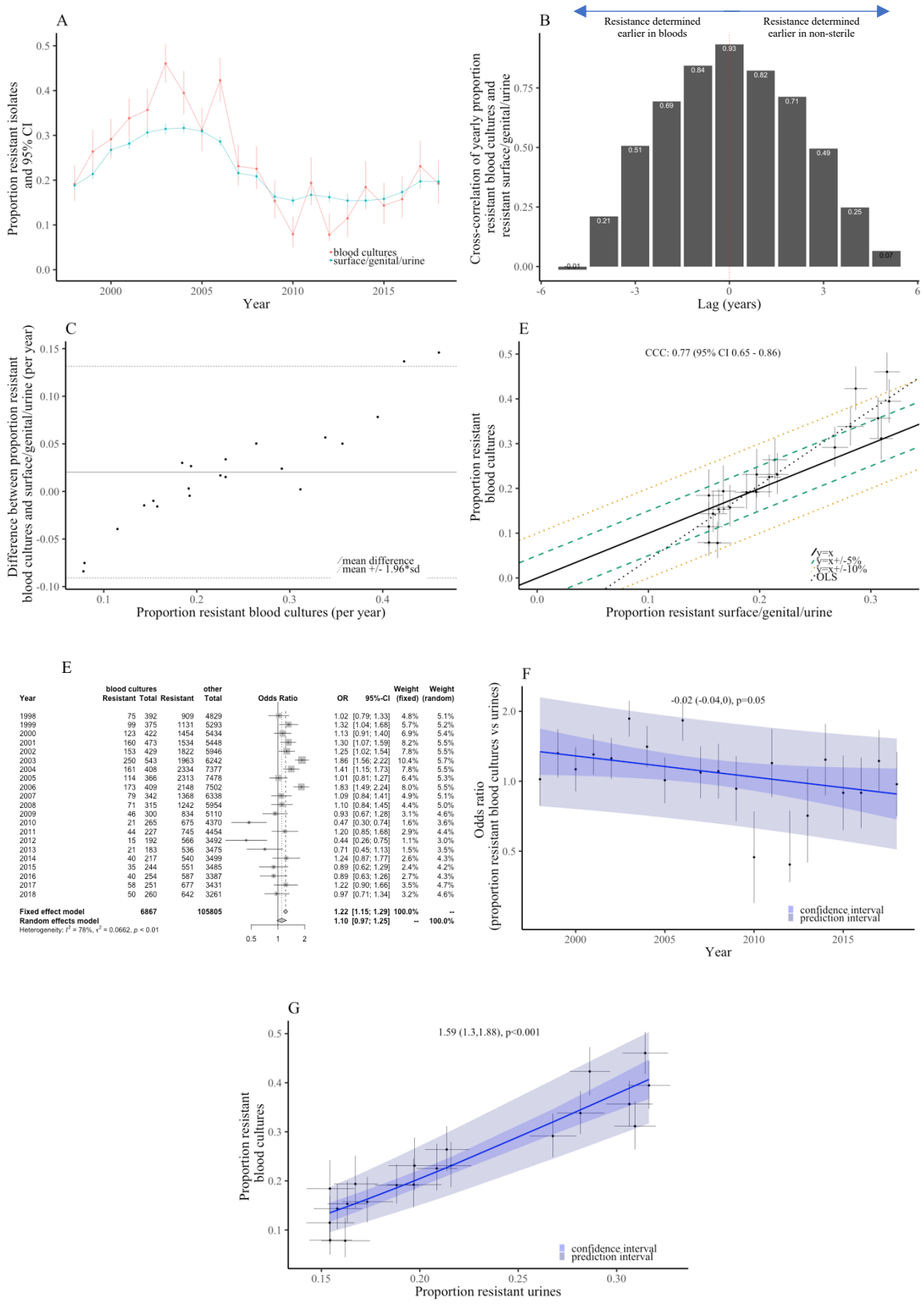


Figure 62 Agreement between proportion of erythromycin-resistant S. aureus bloodstream infections and proportion of erythromycin-resistant S. aureus non-sterile site cultures

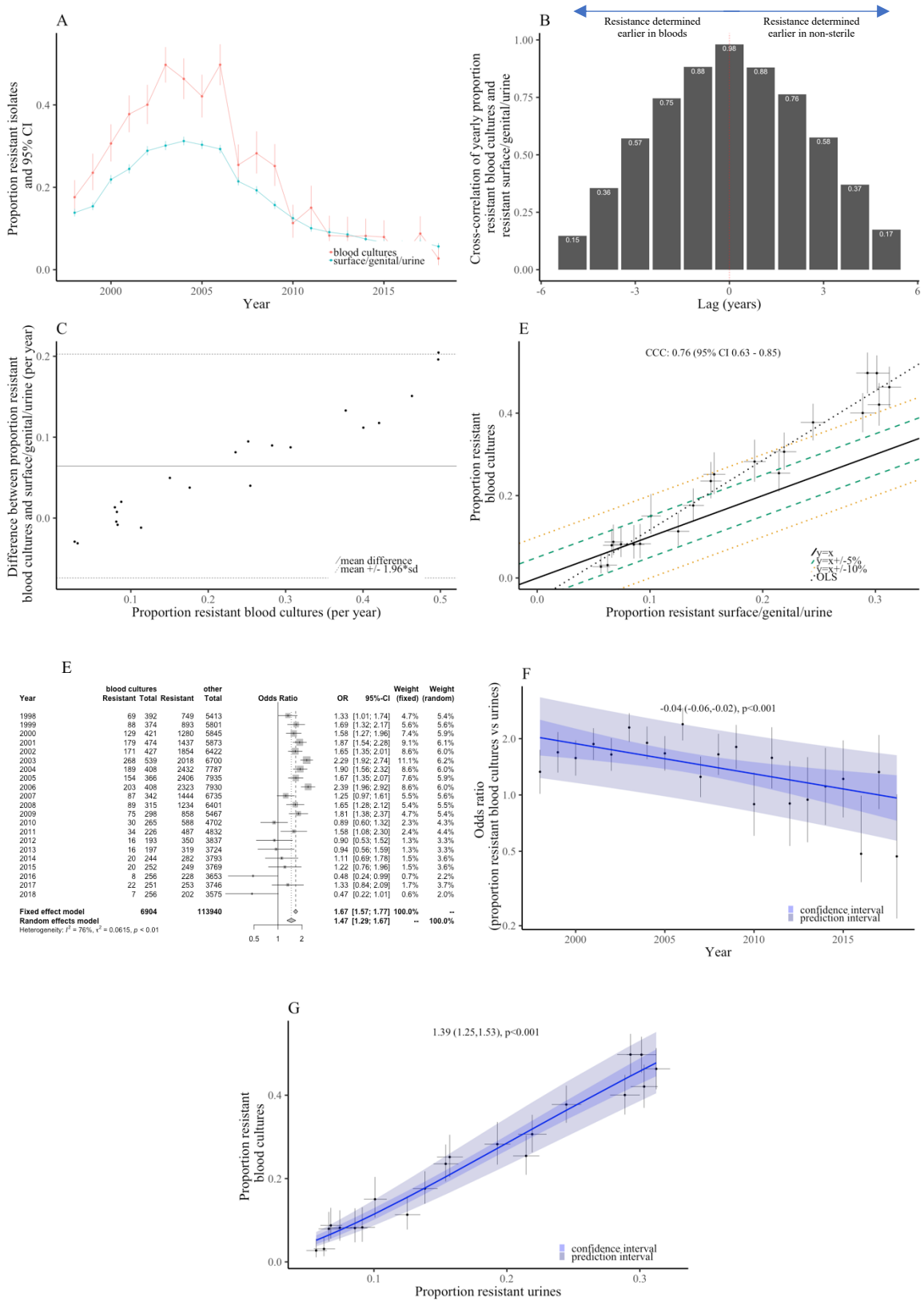


Figure 63 Agreement between proportion of flucloxacillin-resistant S. aureus bloodstream infections and proportion of flucloxacillin-resistant S. aureus non-sterile site cultures

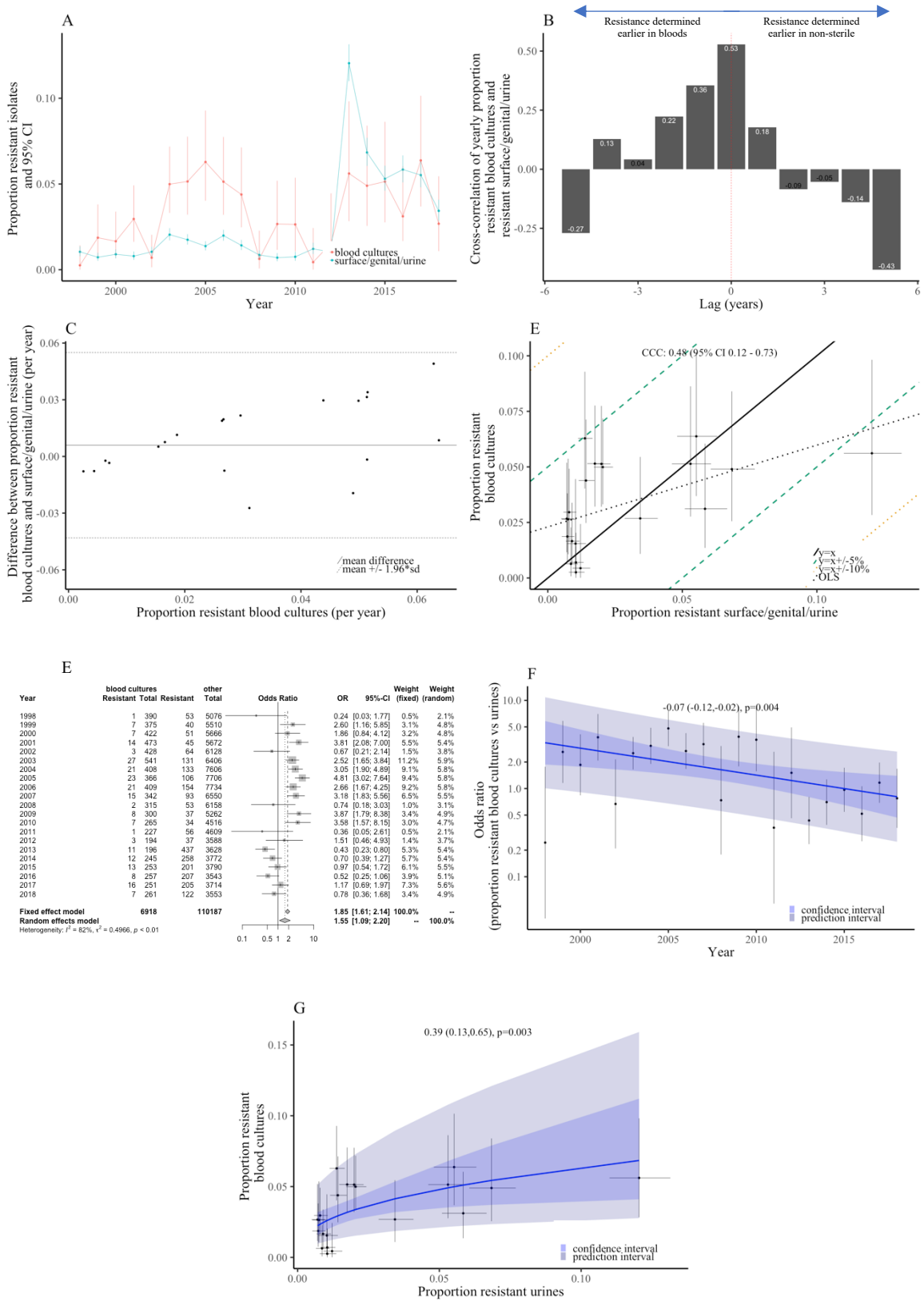


Figure 64 Agreement between proportion of gentamicin-resistant S. aureus bloodstream infections and proportion of gentamicin-resistant S. aureus non-sterile site cultures

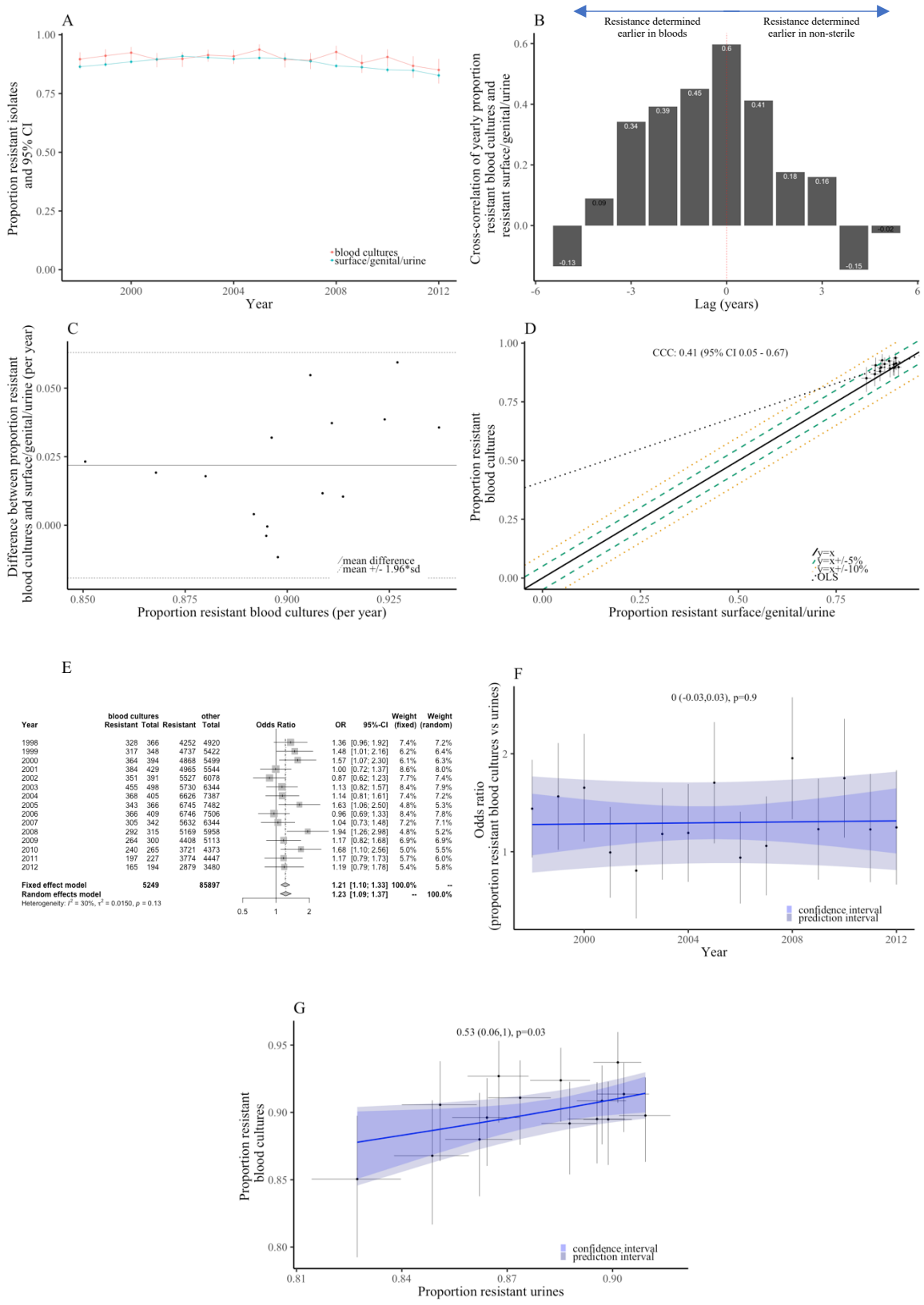


Figure 65 Agreement between proportion of penicillin-resistant S. aureus bloodstream infections and proportion of penicillin-resistant S. aureus non-sterile site cultures

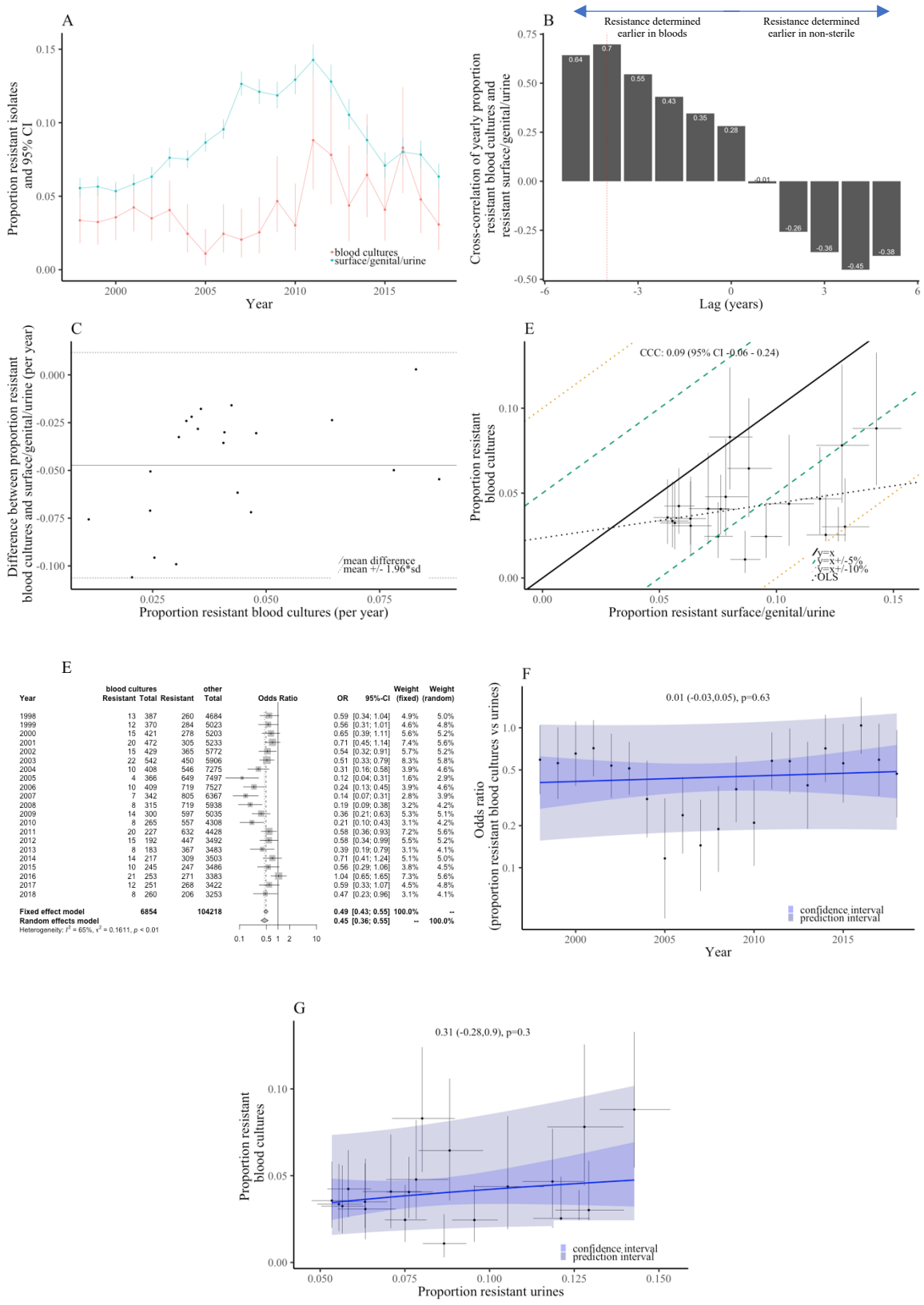


Figure 66 Agreement between proportion of tetracycline-resistant S. aureus bloodstream infections and proportion of tetracycline-resistant S. aureus non-sterile site cultures

5 Conclusions and future work

During preliminary work carried out during an internship with the Modernising Medical Microbiology group, I investigated all bloodstream infections occurring in Oxfordshire over time. I chose bloodstream infections as ascertainment bias is expected to be minimal due to their severity. *E. coli* was the most common pathogen causing bloodstream infections and their number had been continuously rising, with an accelerated increase since 2002.

The main goal of this thesis was to explore possible reasons for the rise in *E. coli* bloodstream infections and to determine possible methods to control the rise and manage the infections.

In my first results Chapter, I aimed to understand what was driving this increase by exploring various different hypotheses using electronic health record data available within IORD. I was able to show, for the first time, that the rise in *E. coli* bloodstream infections accelerated with increasing, rather than decreasing, time since the last hospital admission. This is a major contrast to other healthcare-associated infections such as MRSA and *C. difficile*. However, despite exploiting the different types of data available, I was not able to identify specific driver for most of the increase. A small percentage of the increase could be explained by an aging population. However, the percentage explained by ageing actually decreased with increasing time since the last hospital admission. The differential rise in community onset disease could not be attributed to increased ascertainment. Even though there was an increase in the number of blood cultures taken for processing regardless of result, this increase was significantly smaller in patients who had not been admitted in the

hospital in the year before the blood culture being taken. Moreover, due to the severity of bloodstream infections one would expect a blood culture to be taken if a bloodstream infection occurred. This lack of change in severity was also supported by no clinically significant changes in the results of blood tests taken and bloodstream infections, or mortality following it, over time, overall or in different groups, with community cases still having a 15% 30-day mortality. This suggests that ascertainment was also random – as more blood cultures were sent, it was not the case that more were sent from more well patients. Antecedent UTIs or presence of a catheter (as evidenced by a previous CSU labelled sample) did not appear to be driving the increase. Importantly, supporting generalisability of my study, the rise in *E. coli* bloodstream infections was also observed at a national level for the English national surveillance programme (Public Health England 2019). However, information on previous hospital-exposure, previous UTI, previous presence of a catheter was not available at a national level through mandatory reporting. Instead, Public Health England had conducted a 3-month enhanced sentinel surveillance study, collecting data on previous healthcare exposure in 1688 patients with *E. coli* bloodstream infection in 35 NHS hospitals. The proportion of cases which appeared to be amenable to a urinary-focused intervention, as well as the proportion of cases which appeared to be amenable to a urinary catheter intervention, were similar to those seen in Oxfordshire, suggesting that results seen in Chapter 2 could be generalisable to the whole of the UK. Even though the rise was not restricted to those with a previous UTI or a previous CSU, the high proportion of such cases in Oxfordshire suggests prevention strategies targeting UTI management and catheter care could at least slow down the rise in *E. coli* bloodstream infections.

I classified isolates based on time of *E. coli* bloodstream infection after admission and previous hospital-exposure. Future classifications would ideally also incorporate information on nursing homes or community care. Previously, some effort has been made in order to incorporate nursing homes in the NHS data sources underlying IORD using postcodes of patients and matching them with those of registered nursing homes in Oxfordshire. However, at the time, an unrealistically small percentage of patients could be matched directly to a nursing home so this was not taken forwards.

The rise in incidence in *E. coli* bloodstream infections was occurring in both co-amoxiclav susceptible and co-amoxiclav resistant isolates. However, the increase was significantly faster in the latter group. I was able to show that general practices prescribing more co-amoxiclav were associated with higher rates of co-amoxiclav resistant *E. coli* UTIs in the subsequent year. Bloodstream infections were much less common than UTIs (2.5%) and therefore there was insufficient power to look at associations with bloodstream infections as there were too few per general practice. However, it is plausible to think that faster increases in co-amoxiclav resistant bloodstream infections could also be driven by increased co-amoxiclav use in the community, since the source of the bloodstream infection is often considered to be urinary (although, interestingly, I did not find evidence that urinary infections appeared to be driving increases overall). There are two mechanisms by which this could occur, both equally important. Firstly, higher population level antibiotic use could increase the frequency of resistant organisms in the community. This means that individuals are more likely to acquire resistant organisms. Secondly, resistant pathogens could be selected within the individual through individual antibiotic use and poorly treated *E. coli* UTIs. This means that UTIs which are treated with an inappropriate antibiotic could develop

into *E. coli* bloodstream infections. In the future, linkage of individual patient community with hospital antibiotic usage records would allow exploration of the latter hypothesis.

This work suggests strategies targeting UTIs in order to minimize antibiotic usage could lead to a decrease in resistant infections. Such strategies could focus on identifying the severity of the infection in order to guide treatment. If the infection is mild enough or in a patient with no comorbidity, natural treatments such as drinking plenty of fluids could first be recommended. Empirical antibiotic treatment for UTIs could then be reserved for severe cases. If the infection does not clear on its own or through natural remedies, if possible, antibiotics to which the pathogen is known to be susceptible should be chosen. However, this requires a pathogen to be isolated. The large scale move to nitrofurantoin as first line antibiotic across England may help as resistance prevalence is still very low (Public Health England 2017). In practice, I showed that submission of urine samples for testing is highly variable. There are multiple diagnostic companies currently trying to develop point of care tests for UTIs to address this challenge. These strategies could reduce unnecessary antibiotic use, could prevent the selective pressure of using antibiotics to which a pathogen was resistant, and could help with prolonging the usefulness of broad-spectrum antibiotics.

The rise in *E. coli* bloodstream infection was slowest in the nosocomial group. This was potentially a result of previous successful prevention and intervention strategies targeting healthcare-associated infections. This was also suggested by a decrease of over 50% in the proportion of nosocomial cases that were 'likely urine-associated' as well as the increase being confined to those without a previous CSU. However, the percentages of *E. coli* bloodstream infections with a previous CSU increased with increasing proximity to hospital-

exposure, from 20% in the community to 43% in the nosocomial cases, highlighting the burden of catheterisation in this group. Similarly, 30-day mortality increased with increasing proximity to hospital-exposure from 14% in the community to 25% in the nosocomial group. Therefore, the management of nosocomial infections remains of great importance given their substantial mortality, which may reflect in part that those acquiring an infection in hospital are sicker.

I was not able to identify a clear driver of the rise in *E. coli* bloodstream infections. However, good choice of treatment is required in order to manage the rise. Very few new antibiotics have been developed and deployed in the past few decades. Therefore, electronic health records provide an opportunity to guide the exploitation of those antibiotics which are already available in order to use them as efficiently as possible, prolonging their utility. I found no evidence for increased severity or mortality in bloodstream infections caused by resistant versus susceptible isolates. Two possible hypotheses for this are that: mortality is too crude a marker of poorer outcome, and/or current susceptibility testing methods do not correlate well with the effectiveness of the antimicrobial therapy in the patient. Alternatively, rapid diagnostics and rapid switches to effective antibiotics might mitigate not receiving the correct initial treatment, and/or the sickest patients get escalated to broader antibiotics empirically. In Oxfordshire, hospital prescribing records per individual are available since October 2015 and these will allow for the latter two hypotheses to be explored.

Before I started this work, the general feeling was that increased co-amoxiclav usage was due to changing testing methodology. I was able to confidently refute this and show instead

that there was a genuine change in resistance around the time testing methods changed. In future, considering both treatment and surveillance, changes in susceptibility testing method, or changes in any other method in general, I would recommend both the old and the new methods to be carried out for a period of time in order to allow comparisons of the two different methods within the laboratory, to provide a better chance of disentangling true changes in the epidemiology from changes which are due to testing alone.

Work carried out in parallel in the group showed that a binary categorisation of susceptibility may not be appropriate for co-amoxiclav due to the unimodal distribution of MICs that is split by the breakpoint, with disagreement between repeat MIC results also being observed. With WGS becoming cheaper, using sequence data routinely is potentially coming within reach and it may be able to produce more reliable predictions of the effectiveness of antibiotic treatment. Moving towards the routine use of WGS would have the added benefit of being able to track spread of infection (Quainoo et al. 2017; Dingle et al. 2017; Kong et al. 2018; Eyre et al. 2013). I chose to investigate machine learning algorithms in order to predict genotypic "resistance" as these can produce very good models without requiring expert knowledge (Chapter 3). The best performing machine learning models were able to produce good predictions of co-amoxiclav MICs, within a doubling dilution. However, this still led to relatively poor prediction of resistance vs susceptibility primarily because the breakpoint defining resistance effectively lies within the distribution of observed MICs.

The challenges of choosing breakpoints in order to interpret MICs before reporting resistance to prescribers is also suggested by the continuous updates of EUCAST

recommendations. In 2014, EUCAST introduced a different breakpoint for co-amoxiclav for uncomplicated UTIs, of 32/2 mg/L rather than 8/2 mg/L. However, Phoenix^{TM} only allows the selection of one breakpoint for *E. coli* positive cultures for each drug, including co-amoxiclav, and therefore the lower one was chosen and is reported, i.e. susceptible versus resistant using 8/2 mg/L as the breakpoint for both blood and urine cultures. This means resistance is overcalled by the routine hospital laboratory for some isolates from urine. However, guidelines recommend urines to be sent for culture only in complicated UTI cases, so in theory co-amoxiclav resistance should be overcalled only in a small number of cases. Information on UTI complexity would also not be routinely available in the laboratory in order to apply a different breakpoint for samples from uncomplicated cases. One option for *E. coli* positive urine cultures would be to directly report MICs together with breakpoints for complicated and uncomplicated UTIs to prescribers in order to allow them to review their decision-making in light of the clinical syndrome of the patient. This sort of approach would also have benefits when there is not a clear breakpoint to separate susceptible from resistant isolates, as for co-amoxiclav. Along these lines, in 2019, EUCAST made the decision to change definitions to: “susceptible, standard dosing regime”, “susceptible, increased exposure”, “resistant”. This can be seen as a step towards personalised medicine, where information on patient (e.g. weight, site of infection, co-morbidities) and bacteria (MIC, AMR genes) could be used to better predict risk and outcome. However, this still essentially draws arbitrary thresholds across a continuous distribution of MICs, just two thresholds rather than one, and moreover one measured with error.

At the moment susceptibility testing is carried out in vitro, however, the final goal is understanding how well an antibiotic will perform in vivo, i.e. in the patient. In this thesis I

focused on using bacterial DNA to predict in vitro susceptibility testing results. It may be that a shift in focus is required from “resistance”-specific genes to genetic features associated with a poor outcome. In reality, the elements influencing the effectiveness of antimicrobial therapy can be categorised into three groups:

- Host (comorbidities, past illnesses, genetics, microbiome),
- Pathogen (virulence, AMR genes),
- Antibiotics and other medical treatments (e.g. fluids).

Given concerns about traditional susceptibility testing methods, in my machine learning algorithms, training labels represented currently by in vitro susceptibility testing results could be replaced by measures of the efficacy of treatment. These could be metrics derived based on the evolution of physiological measurements (e.g. temperature, blood pressure), evolution of blood test results (e.g. white cell counts, CRP level). Time to hospital discharge and 30-day mortality could also be used, but these are quite crude measurements. This would also allow the models to continuously be updated as more data is gathered without the need for traditional microbiology susceptibility methods to be carried out.

One major advantage of machine learning algorithms is their potential to be applied across multiple pathogen-drug combinations. Moreover, they can be trained on updated datasets by simply rerunning them, and therefore new important genetic resistance mechanisms which are known to arise in the population can be learnt by the model. At present, I do not consider that the performance I was able to achieve is sufficient to recommend their routine use in resistance prediction. However, bigger datasets are also a known method of improving the performance of machine learning algorithms. Therefore, as datasets increase

and the models are retrained, their performance is expected to improve. Training algorithms to predict "poor outcome" would also be one mechanism to address an often repeated criticism of resistance prediction from genotype, rather than observing phenotype, namely that it would be impossible to identify a genuinely novel resistance gene from the sequence alone.

In all countries, regardless of income, an initial empiric antibiotic treatment has to be administered to patients. In high-income countries, this is in the first few days of treatment before results from a culture from the site of infection are available, the bacteria causing the infection is identified and antimicrobial susceptibility testing is carried out, potentially leading to a change in antibiotic treatment and explaining the lack of difference in mortality I observed between infections with resistant vs susceptible isolates. However, in low- and middle-income countries it is generally the exception rather than the norm that carrying out such tests is possible, and so patients will remain on a potentially ineffective empiric regimen for longer. Therefore, using samples of urine or from non-sterile sites, which could be available in far greater numbers, could be a much more efficient method for antimicrobial resistance surveillance which could guide empiric antibiotic treatment. Changes in the proportion of resistant bloodstream infections of a size which would inform change of treatment could be identified by looking at the proportion resistant in non-blood cultures. I also found that there could be potential value in considering susceptibility results of a prior *E. coli* UTI in order to make a personalised empiric treatment choice in a high-income country before susceptibility results of *E. coli* bloodstream infections become available. However, unfortunately given its widespread use as empiric therapy, among the four antibiotics I considered, the poorest performance of using susceptibility results of prior

E. coli UTI as a marker of resistance in *E. coli* bloodstream infection was for co-amoxiclav. This was likely due to the challenges of a binary phenotype when considering co-amoxiclav. However, as the positive predictive value was 80% or higher for all drugs including co-amoxiclav – meaning that four out of five patients with a prior co-amoxiclav resistant urinary tract infection would turn out to have a co-amoxiclav resistant *E. coli* bloodstream infection – in the absence of an antibiogram, this could still be useful.

The largest foci of *E. coli* bloodstream infections is considered to be the urinary tract. However, my work in Chapter 2 showed that even though these represent a large proportion of the bloodstream infections, there are many with a different or unknown focus and there was no evidence for a faster increase in those of urinary tract infection origin. This work is further supported by the results seen in Chapter 4, where the agreement between the antibiograms of the UTIs taken within [-3,+2] days of the bloodstream infection was lower than might be expected if such a high percentage of bloodstream infections had been of urinary tract origin.

The complexities of antimicrobial stewardship are numerous. I will present a few which were illustrated within my dataset. In the hospital context, moving away from one antibiotic means introducing another, since antimicrobial therapy still needs to be provided. In Oxfordshire, in 2006 there was a shift in prescribing from second- and third-generation cephalosporins, which cause selective pressure for *C. difficile*, towards empiric treatment with co-amoxiclav and gentamicin. This led to a reduction in the number of *C. difficile* infections (Dingle et al. 2017), but plausibly also to an increase in the number of co-amoxiclav resistant *E. coli* bloodstream infections.

Drugs from the same antibiotic classes have similar genetic resistance mechanisms, with resistance to one often also causing resistance to the other. More importantly, and particularly for gram-negative pathogens, different resistance-conferring mechanisms are commonly carried together, because of horizontal gene transfer and co-selection. In Chapter 4, ciprofloxacin, erythromycin and flucloxacillin presented strikingly similar trends in the proportions of resistant *S. aureus* infections over time. Resistance to these drugs in *S. aureus* is known to be caused by different gene mechanisms which are also independent biologically. This suggests that a large proportion of genetic mechanisms of resistance to these drugs are co-carried, even though not co-located or mechanistically linked. Thus, one explanation for the correlation between trends in ciprofloxacin, erythromycin and flucloxacillin resistant cases is that MRSA strains that were observed in Oxfordshire were also ciprofloxacin and erythromycin resistant. There was epidemiological evidence that the MRSA strain in which the largest decrease was observed in 2006, ST36, was also erythromycin resistant in Oxfordshire (David H Wyllie et al. 2011). A case-control study on isolates from 2007 found that almost all MRSA cases were also ciprofloxacin resistant (100% of hospital-exposed cases, and 98% of community cases) (R. Miller et al. 2010). This also supports the plausibility of using non-blood cultures as a proxy for blood culture surveillance.

In the future, advances in sequencing directly from samples and decreases in the cost of sequencing and data processing/storage could make microbiology laboratories less important for routine clinical practice. However, at the moment sequencing is still significantly more expensive than the very cheap traditional microbiology tests. But if the

exponential progress that has been seen in storage capacity over the last decade, for example, were to continue, sequencing could be a faster and cheaper method than building microbiology laboratories in low-income countries.

One challenging question is at what proportion of resistance should empiric antimicrobial treatment be switched. There is certainly not one universal value than can be used. The main challenge is that using an antibiotic inevitably creates resistance to it. Logically, were resistance not a concern, the most sensible strategy would be to use the most effective, broadest spectrum antibiotic, e.g. meropenem, in every person immediately after an infection was suspected. However, resistance means such a strategy would inevitably render the antibiotic useless over a period of time. Therefore, in considering when an empiric antimicrobial recommendation should be changed, morbidity and mortality associated with a particular infection within the population needs to be considered, before antibiotic resistance needs are analysed. The consequences of a resistant infection are one of the crucial determinants of making the decision to move away from recommending a specific antibiotic as empiric treatment. Considering a resistant *Mycoplasma pneumoniae* infection, which generally only causes a mild illness in older children and young adults, one could easily allow a very high level of resistance before switching empirical antibiotic treatment (or not using an antibiotic at all). However, considering infectious diseases affecting children in low- and middle-income countries with very high mortality rates, particularly among babies (e.g. neonatal sepsis), even a 5-10% resistance prevalence within the population of infections could be extremely serious and immediate measures might need to be taken. This may depend on whether changing empiric antibiotic treatment having observed poor response can “rescue” the initial sub-optimal treatment or not.

In this thesis, I have focused on resistance in specific isolated pathogens. However, for empiric treatment we are interested in the effectiveness of antibiotics without knowing the bacteria causing the infection, since antibiotics are administered before a culture is available. Moreover, a culture can present no growth or mixed growth. Therefore, by only considering a limited number of individual pathogens, I may only be accounting for a modest proportion of disease and, as a result, could recommend moving away from an empiric antibiotic with a high prevalence of resistance to my individual pathogens to a different drug which is ineffective against a greater proportion of other pathogens.

To address this, the weighted incidence syndromic combination antibiogram (WISCA) is designed to guide empiric antibiotic treatment per syndrome (Hebert et al. 2012). It calculates the likelihood that different antibiotics and combinations of antibiotics will treat a particular syndrome, taking into account the distribution of pathogens causing that particular syndrome, including the possibility of infections caused by multiple pathogens. In comparison to a traditional antibiogram which gives susceptibilities to individual antibiotics for a specific pathogen, WISCA was shown to improve the usefulness of the information presented for the treatment of 8232 UTIs and 997 abdominal-biliary infections. *E. coli* was the causal organism for 49% of UTIs, but only for 11% of abdominal-biliary infections. Therefore, considering for example fluoroquinolone resistance, even though 84% of *E. coli* infections were susceptible to a fluoroquinolone, only 62% of UTIs were susceptible to one, and an even lower percentage of 37% of abdominal-biliary infections because many more were caused by other pathogens. This clearly illustrated the usefulness of considering WISCAs rather than traditional antibiograms. Comparing standard empiric antimicrobial

treatment received by 163 patients with critical care infections, to a choice of dual therapies which would have been guided by WISCAs, there was a 56%, 42% and 18% increase in the 12h, 24h and 48h coverage of infections (Randhawa et al. 2014).

In conclusion, in this thesis I have shown how, through the use of linked routinely collected electronic health records, the epidemiology of *E. coli* bloodstream infections in Oxfordshire could be better understood without the need for expensive and time-consuming manual data collection, which often also suffers from large amounts of missing data. As more types of records, such as general practice prescribing, are linked to databases like IORD, more hypotheses will be developed and tested. I have shown that machine learning algorithms have the potential to predict antimicrobial susceptibility from WGS without the need for expert knowledge. In future, information on hospital antimicrobial prescribing and patient outcome could be used to train machine learning models which could predict effectiveness of therapy. As ease of travel makes transmission of pathogens and AMR genes much more common, managing antimicrobial resistance requires a global approach. Using Oxfordshire data, I have identified a tool which can help with the surveillance of antimicrobial resistance in low- and middle-income countries. I showed that monitoring could be done by using proportion resistant non-blood cultures as these are easier to sample and can therefore be collected in much larger numbers. Epidemiology, resistance prediction and surveillance will all continue to be essential tools in combatting the threat posed by antimicrobial resistance.

6 Bibliography

Abernethy, J, R Guy, E A Sheridan, S Hopkins, M Kiernan, M H Wilcox, A P Johnson, and R Hope. 2017. “Epidemiology of Escherichia Coli Bacteraemia in England : Results of an Enhanced Sentinel Surveillance Programme.” *Journal of Hospital Infection* 95: 365–67.

Acheson, E. D. 1964. “Oxford Record Linkage Study. a Central File of Morbidity and Mortality Records for a Pilot Population.” *British Journal of Preventive & Social Medicine* 18: 8–13.

Ashiru-Oredope, Diane, Mike Sharland, Esmita Charani, Cliodna Mcnulty, and Jonathan Cooke. 2012. “Improving the Quality of Antibiotic Prescribing in the NHS by Developing a New Antimicrobial Stewardship Programme : Start Smart — Then Focus.” *Journal of Antimicrobial Chemotherapy* 67 (Suppl 1): i51–63. https://doi.org/10.1093/jac/dks202.

Ayukekbong, James A., Michel Ntemgwa, and Andrew N. Atabe. 2017. “The Threat of Antimicrobial Resistance in Developing Countries: Causes and Control Strategies.” *Antimicrobial Resistance and Infection Control* 6 (1): 1–8. https://doi.org/10.1186/s13756-017-0208-x.

Bishop, C. 2007. *Pattern Recognition and Machine Learning*. Springer.

Bland, J M, and D. G. Altman. 1986. “Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement.” *Lancet* 1 (8476): 307–10.

Buetti, Niccolò, Andrew Atkinson, Jonas Marschall, and Andreas Kronenberg. 2017. “Incidence of Bloodstream Infections : A Nationwide Surveillance of Acute Care Hospitals in Switzerland 2008 – 2014.” *BMJ Open* 7: 1–5. https://doi.org/10.1136/bmjopen-2016-013665.

Butler, Christopher, Frank Dunstan, Margaret Heginbothom, Brendan Mason, Zoë Roberts,

Sharon Hillier, Robin Howe, Stephen Palmer, and Anthony Howard. 2007. “Containing Antibiotic Resistance: Decreased Antibiotic-Resistant Coliform Urinary Tract Infections with Reduction in Antibiotic Prescribing by General Practices.” *British Journal of General Practice*, no. October.

Butler, Christopher, Nick Francis, Emma Thomas-Jones, Carl Llor, Emily Bongard, Michael Moore, Paul Little, et al. 2017. “Variations in Presentation, Management, and Patient Outcomes of Urinary Tract Infection: A Prospective Four-Country Primary Care Observational Cohort Study.” *British Journal of General Practice* 67 (665): e830–41. http://bjgp.org/content/67/665/e830.abstract.

Byarugaba, D. K. 2004. “Antimicrobial Resistance in Developing Countries and Responsible Risk Factors.” *International Journal of Antimicrobial Agents* 24 (2): 105–10. https://doi.org/10.1016/j.ijantimicag.2004.02.015.

Cai, Tommaso, Gabriella Nesi, Sandra Mazzoli, Francesca Meacci, Paolo Lanzafame, Patrizio Caciagli, Liliana Mereu, et al. 2015. “Asymptomatic Bacteriuria Treatment Is Associated With a Higher Prevalence of Antibiotic Resistant Strains in Women With Urinary Tract Infections.” *Clinical Infectious Diseases* 61: 1655–61. https://doi.org/10.1093/cid/civ696.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 9: 1–9. https://doi.org/10.1186/1471-2105-10-421.

Cardoso, Teresa, Mónica Almeida, N Deborah Friedman, Irene Aragão, and Altamiro Costa-pereira. 2014. “Classification of Healthcare-Associated Infection : A Systematic Review 10 Years after the First Proposal.” *BMC Medicine* 12 (40). https://doi.org/10.1186/1741-7015-12-40.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. San Francisco, California, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Chokshi, A., Sifri, Z., Cennimo, D., & Horng, H. 2019. “Global Contributors to Antibiotic Resistance.” *Journal of Global Infectious Diseases*, no. 11(1): 36–42. https://doi.org/doi:10.4103/jgid.jgid_110_18.

Claesen, Marc, Jaak Simm, Dusan Popovic, Yves Moreau, and Bart De Moor. 2014. “Easy Hyperparameter Search Using Optunity.” http://arxiv.org/abs/1412.1114.

Clatworthy, Anne E, Emily Pierson, and Deborah T Hung. 2007. “Targeting Virulence: A New Paradigm for Antimicrobial Therapy.” *Nature Chemical Biology* 3 (9): 541–48.

Clausen, Philip T.L.C., Frank M. Aarestrup, and Ole Lund. 2018. “Rapid and Precise Alignment of Raw Reads against Redundant Databases with KMA.” *BMC Bioinformatics* 19 (1): 1–8. https://doi.org/10.1186/s12859-018-2336-6.

Clausen, Philip T.L.C., Ea Zankari, Frank M. Aarestrup, and Ole Lund. 2016. “Benchmarking of Methods for Identification of Antimicrobial Resistance Genes in Bacterial Whole Genome Data.” *The Journal of Antimicrobial Chemotherapy* 71 (9): 2484–88. https://doi.org/10.1093/jac/dkw184.

Clere M, and Kennedy J. 2002. “The Particie Swarm: Explosion, Stability, and Convergence in a Multi-Dimensional Complex Space.” *IEEE Transactions on Evolutionary Computation* 6 (1): 58–73.

Cohen, Mitchell L. 2000. “Changing Patterns of Infectious Disease.” *Nature* 406 (August): 762–67.

Croghan, Carry W, and Peter P Egeghy. 2003. “Methods of Dealing with Values Below the

Limit of Detection Using SAS." *Presented at Southeastern SAS User Group, St. Petersburg, FL, September 22-24, 2003.*

Davies, S., J. Watson, and L. Shallcross. 2015. "A Global Overview of Antimicrobial Resistance." http://www.globalhealthdynamics.co.uk/wp-content/uploads/2015/06/02_-Davies-Watson-Shallcross.pdf.

Davies, Timothy J, Nicole Stoesser, Anna E Sheppard, Manal Abuoun, Philip Fowler, T Phuong Quan, David Griffiths, et al. 2019. "Reconciling the Potentially Irreconcilable? Genotypic and Phenotypic Amoxicillin-Clavulanate Resistance in Escherichia Coli." *BiorXiv*. https://doi.org/10.1101/511402.

Decraene, V, H T T Phan, R George, D H Wyllie, O Akinremi, Z Aiken, P Cleary, and A Dodgson. 2018. "A Large, Refractory Nosocomial Outbreak of Klebsiella Pneumoniae Carbapenemase-Producing Escherichia Coli Demonstrates Carbapenemase Gene Outbreaks Involving Sink Sites Require Novel Approaches to Infection Control." *Antimicrobial Agents and Chemotherapy* 62 (12): 1–12.

DePestel, Daryl D, and David M Aronoff. 2013. "Epidemiology of Clostridium Difficile Infection." *J Pharm Pract* 26(5) (October): 464–75. https://doi.org/10.1038/jid.2014.371.

Devices, Susceptibility, and Human Services. 2003. "Class II Special Controls Guidance Document : Antimicrobial Susceptibility Test (AST) Systems ; Guidance for Industry and FDA," 1–42.

Didelot, Xavier, A Sarah Walker, Tim E Peto, Derrick W Crook, and Daniel J Wilson. 2016. "Within-Host Evolution of Bacterial Pathogens." *Nature Reviews Microbiology* 14 (January): 150. https://doi.org/10.1038/nrmicro.2015.13.

Didelot, Xavier, and Daniel J. Wilson. 2015. "ClonalFrameML: Efficient Inference of

Recombination in Whole Bacterial Genomes." *PLoS Computational Biology* 11 (2): 1–18. https://doi.org/10.1371/journal.pcbi.1004041.

Díez-Aguilar, María, María-Isabel Morosini, Lorena López-Cerero, Álvaro Pascual, Jorge Calvo, Luis Martínez-Martínez, Francesc Marco, et al. 2015. "Performance of EUCAST and CLSI Approaches for Co-Amoxiclav Susceptibility Testing Conditions for Clinical Categorization of a Collection of *Escherichia Coli* Isolates with Characterized Resistance Phenotypes." *Journal of Antimicrobial Chemotherapy* 70 (April): 2306–10. https://doi.org/10.1093/jac/dkv088.

Dingle, Kate E., Xavier Didelot, T. Phuong Quan, David W. Eyre, Nicole Stoesser, Tanya Golubchik, Rosalind M. Harding, et al. 2017. "Effects of Control Interventions on Clostridium Difficile Infection in England: An Observational Study." *The Lancet Infectious Diseases* 17 (4): 411–21. https://doi.org/10.1016/S1473-3099(16)30514-X.

Drouin, Alexandre, Sébastien Giguère, Maxime Déraspe, Mario Marchand, Michael Tyers, Vivian G. Loo, Anne Marie Bourgault, François Laviolette, and Jacques Corbeil. 2016. "Predictive Computational Phenotyping and Biomarker Discovery Using Reference-Free Genome Comparisons." *BMC Genomics* 17 (1): 1–15. https://doi.org/10.1186/s12864-016-2889-6.

Drouin, Alexandre, and Mario Marchand. 2018. "Interpretable Genotype-to-Phenotype Classifiers with Performance Guarantees," 1–16. https://doi.org/10.1101/388348.

Duerden, Brian, Carole Fry, Alan P Johnson, and Mark H Wilcox. 2015. "The Control of Methicillin-Resistant Staphylococcus Aureus Blood Stream Infections in England." *Open Forum Infect Dis*. https://doi.org/10.1093/ofid/ofv035.

Earle, Sarah G., Chieh Hsi Wu, Jane Charlesworth, Nicole Stoesser, N. Claire Gordon, Timothy M. Walker, Chris C.A. Spencer, et al. 2016. "Identifying Lineage Effects When

Controlling for Population Structure Improves Power in Bacterial Association Studies." *Nature Microbiology* 1 (5): 1–8. https://doi.org/10.1038/nmicrobiol.2016.41.

Edelsberg, John, Derek Weycker, Rich Barron, Xiaoyan Li, Hongsheng Wu, Gerry Oster, Sejal Badre, Wendy J. Langeberg, and David J. Weber. 2014. "Prevalence of Antibiotic Resistance in US Hospitals." *Diagnostic Microbiology and Infectious Disease* 78 (3): 255–62. https://doi.org/10.1016/j.diagmicrobio.2013.11.011.

European Centre for Disease Prevention and Control. 2016. "Surveillance of Antimicrobial Resistance in Europe: Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net)."

European Committee on Antimicrobial Susceptability Testing. 2016. "EUCAST: MIC Distributions and ECOFFs." 2016. http://www.eucast.org/mic_distributions_and_ecoffs/.

European Committee on Antimicrobial Susceptibility Testing. 2019. "Clinical Breakpoints for Bacteria v 9.0." 2019. http://www.eucast.org/clinical_breakpoints/.

Eyre, David W, Madeleine L Cule, Daniel J Wilson, David Griffiths, Alison Vaughan, Lily O'Connor, Camilla L C Ip, et al. 2013. "Diverse Sources of C. Difficile Infection Identified on Whole-Genome Sequencing." *New England Journal of Medicine* 369 (13): 1195–1205. https://doi.org/10.1056/NEJMoa1216064.

"Farr Institute." 2019. 2019. https://farrinstitute.org.

Farr, W. 1875. "Supplement to the 35th Annual Report of the Registrar-General (for 1861-1870)." London.

Fedson, David S, and J A Baldwin. 1982. "Previous Hospital Care as a Risk Factor for Pneumonia: Implications for Immunization With Pneumococcal Vaccine." *JAMA* 248 (16): 1989–95. https://doi.org/10.1001/jama.1982.03330160037022.

Ferguson, J. A., M. J. Goldacre, J. Henderson, and A. J. Bron. 1991. “Ophthalmology in the Oxford Region: Analysis of Time Trends From Linked Statistics.” *Eye*, no. 5: 379–84.

Finney, John M., A. Sarah Walker, Tim E A Peto, and David H. Wyllie. 2011. “An Efficient Record Linkage Scheme Using Graphical Analysis for Identifier Error Detection.” *BMC Medical Informatics and Decision Making* 11 (1). https://doi.org/10.1186/1472-6947-11-7.

Fitzpatrick, J M, J S Biswas, J D Edgeworth, J Islam, N Jenkins, R Judge, A J Lavery, and M Melzer. 2016. “Gram-Negative Bacteraemia ; a Multi-Centre Prospective Evaluation of Empiric Antibiotic Therapy and Outcome in English Acute Hospitals.” *Clinical Microbiology and Infection* 22: 244–51. https://doi.org/10.1016/j.cmi.2015.10.034.

Gágyor, Ildikó, Jutta Bleidorn, Michael M Kochen, Guido Schmiemann, Karl Wegscheider, and Eva Hummers-pradier. 2015. “Ibuprofen versus Fosfomycin for Uncomplicated Urinary Tract Infection in Women : Randomised Controlled Trial.” *BMJ* 351. https://doi.org/10.1136/bmj.h6544.

Ganguly, Nirmal K, NK Arora, Sujith J Chandy, Mohameed Nadeem Fairoze, J.P S Gill, Usha Gupta, Shah Hossain, et al. 2011. “Rationalizing Antibiotic Use to Limit Antibiotic Resistance in India+.” *Indian Journal of Medical Research* 134 (3): 281–94. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed14&NEWS=N&AN=369625618.

Garner, Julia S, William R Jarvis, T Grace Emori, Teresa C Horan, and James M Hughes. 1988. “CDC Definitions for Nosocomial Infections.” *American Journal Infect Control*, no. 16: 128–40.

GBD 2017 Causes of Death Collaborators. 2018. “Global, Regional, and National Age-Sex-Specific Mortality for 282 Causes of Death in 195 Countries and Territories, 1980 –

2017: A Systematic Analysis for the Global Burden of Disease Study 2017." *Lancet* 392: 1736–88.

Genomics England. 2017. "Whole Genome Sequencing to Diagnose TB." 2017. https://www.genomicsengland.co.uk/whole-genome-sequencing-to-diagnose-tb/.

Gerver, Sarah, Mary Sinnathamby, Sabine Bou-Antoun, Sophina Kauser, Malcolm Canvin, Julia Abernethy, John Davies, and Russel Hope. 2014. "Annual Epidemiological Commentary: Mandatory MRSA, MSSA and E. Coli Bacteraemia and C. Difficile Infection Data, 2013/14." *Public Health England*, no. July: 53.

Geyter, Deborah De, Lieve Blommaert, Nicole Verbraeken, Mark Sevenois, Luc Huyghens, Helena Martini, Lieve Covens, Denis Piérard, and Ingrid Wybo. 2017. "The Sink as a Potential Source of Transmission of Carbapenemase-Producing Enterobacteriaceae in the Intensive Care Unit." *Antimicrobial Resistance and Infection Control* 6 (1): 1–6. https://doi.org/10.1186/s13756-017-0182-3.

Goldacre, Alexander, Julia Pakpoor, and Michael Goldacre. 2017. "Perinatal Characteristics and Obstetric Complications in Mothers with Multiple Sclerosis: Record-Linkage Study." *Multiple Sclerosis and Related Disorders* 12 (February): 4–8. https://doi.org/10.1016/j.msard.2016.11.011.

Goto, Michihiko, Amy M. J. O'Shea, Daniel J. Livorsi, Jennifer S. McDanel, Makoto M. Jones, Kelly K. Richardson, Brice F. Beck, et al. 2016. "The Effect of a Nationwide Infection Control Program Expansion on Hospital-Onset Gram-Negative Rod Bacteremia in 130 Veterans Health Administration Medical Centers: An Interrupted Time-Series Analysis." *Clinical Infectious Diseases* 63 (5): 642–50. https://doi.org/10.1093/cid/ciw423.

Hand, David J. 2009. "Measuring Classifier Performance: A Coherent Alternative to the Area under the ROC Curve." *Machine Learning* 77 (1): 103–23.

https://doi.org/10.1007/s10994-009-5119-5.

Hand, David J. 2013. “When Is the Area under the ROC Curve an Appropriate Measure of Classifier Performance ?” 34 (xxx): 1–7.

Hart, C A, and S Kariuki. 1998. “Antimicrobial Resistance in Developing Countries.” *BMJ*, no. 317: 647–50.

Hasman, H., F. M. Aarestrup, O. Lund, A. M. Seyfarth, E. Zankari, M. V. Larsen, Y. Agerso, and R. S. Kaas. 2012. “Genotyping Using Whole-Genome Sequencing Is a Realistic Alternative to Surveillance Based on Phenotypic Antimicrobial Susceptibility Testing.” *Journal of Antimicrobial Chemotherapy* 68 (4): 771–77. https://doi.org/10.1093/jac/dks496.

Hassoun, Ali, Peter K. Linden, and Bruce Friedman. 2017. “Incidence, Prevalence, and Management of MRSA Bacteremia across Patient Populations-a Review of Recent Developments in MRSA Management and Treatment.” *Critical Care (London, England)* 21 (1): 211. https://doi.org/10.1186/s13054-017-1801-3.

Hastie, T, R Tibshirani, and J Friedman. 2009. *Elements of Statistical Learning*. Springer.

Hay, Alastair D, Michael Thomas, Alan Montgomery, Mark Wetherell, Andrew Lovering, Cliodna Mcnulty, Deirdre Lewis, Becky Carron, Emma Henderson, and Alasdair Macgowan. 2005. “The Relationship between Primary Care Antibiotic Prescribing and Bacterial Resistance in Adults in the Community : A Controlled Observational Study Using Individual Patient Data.” *Journal of Antimicrobial Chemotherapy*, no. May: 146–53. https://doi.org/10.1093/jac/dki181.

“Health Data Research UK.” 2019. 2019. https://www.hdruk.ac.uk.

Health Protection Agency. 2008a. “Escherichia Coli Bacteraemia in England, Wales and Northern Ireland, 2003 to 2007.”

———. 2008b. “Report for the National Audit Office: Trends in Rates of Healthcare Associated Infection in England 2004 to 2008.”

———. 2008c. “Voluntary Reporting of Staphylococcus Aureus Bacteraemia in England, Wales and Northern Ireland, January-December 2007.”

Health Protection Scotland and Information Services Division. 2016. “Scottish Antimicrobial Use and Resistance in Humans in 2015.”

Health Protection Scotland and National Services Scotland. 2011. “Protocol for Surveillance of Antimicrobial Resistance in Urinary Isolates in Scotland.” http://www.documents.hps.scot.nhs.uk/hai/amr/protocol-for-surveillance-of-amr-in-urinary-isolates-in-scotland-v1.0.pdf.

Hebert, Courtney, Jessica Ridgway, Benjamin Vekhter, Eric C Brown, Stephen G Weber, and Ari Robicsek. 2012. “Demonstration of the Weighted-Incidence Syndromic Combination Antibiogram: An Empiric Prescribing Decision Aid.” *Infection Control and Hospital Epidemiology* 33 (4).

Herrman, H. E., J. A. Baldwin, and David Christie. 1983. “A Record-Linkage Study of Mortality and General Hospital Discharge in Patients Diagnosed as Schizophrenic.” *Psychological Medicine* 13 (3): 581–93. https://doi.org/10.1017/S0033291700048005.

Hesp, Ayla, Kees Veldman, Jeanet van der Goot, Dik Mevius, and Gerdien van Schaik. 2019. “Monitoring Antimicrobial Resistance Trends in Commensal Escherichia Coli from Livestock, the Netherlands, 1998 to 2016.” *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 24 (25): 1–13. https://doi.org/10.2807/1560-7917.ES.2019.24.25.1800438.

Higgins, Julian P T, Simon G Thompson, Jonathan J Deeks, and Douglas G Altman. 2003. “Measuring Inconsistency in Meta-Analyses.” *BMJ (Clinical Research Ed.)* 327 (7414):

557–60. https://doi.org/10.1136/bmj.327.7414.557.

Hillier, Sharon, Zoe Roberts, Frank Dunstan, Chris Butler, Anthony Howard, and Stephen Palmer. 2007. “Prior Antibiotics and Risk of Antibiotic-Resistant Community-Acquired Urinary Tract Infection : A Case – Control Study.” *Journal of Antimicrobial Chemotherapy* 60 (1): 92–99. https://doi.org/10.1093/jac/dkm141.

Ho, Pak-Leung, Kin-Hung Chow, Eileen L Lai, Eric H Y Lau, and Vincent C C Cheng. 2012. “Extended-Spectrum-β-Lactamase-Positive Escherichia Coli Mainly Adds to, Rather than Replaces, Extended-Spectrum-β-Lactamase-Negative E. Coli in Causing Bacteraemia in Hong Kong, 2000–10.” *Journal of Antimicrobial Chemotherapy* 67 (3): 778–80. http://dx.doi.org/10.1093/jac/dkr502.

Hunt, Martin, Alison E Mather, Leonor Sánchez-Busó, Andrew J Page, Julian Parkhill, Jacqueline A Keane, and Simon R Harris. 2017. “ARIBA: Rapid Antimicrobial Resistance Genotyping Directly from Sequencing Reads.” *Microbial Genomics* 3 (10). https://doi.org/10.1099/mgen.0.000131.

Inouye, Michael, Harriet Dashnow, Lesley Ann Raven, Mark B. Schultz, Bernard J. Pope, Takehiro Tomita, Justin Zobel, and Kathryn E. Holt. 2014. “SRST2: Rapid Genomic Surveillance for Public Health and Hospital Microbiology Labs.” *Genome Medicine* 6 (11): 1–16. https://doi.org/10.1186/s13073-014-0090-6.

Ironmonger, Dean, Obaghe Edeghere, Neville Q Verlander, Savita Gossain, Susan Hopkins, Bridget Hilton, and Peter M Hawkey. 2018. “Effect of General Practice Characteristics and Antibiotic Prescribing on Escherichia Coli Antibiotic Non-Susceptibility in the West Midlands Region of England : A 4 Year Ecological Study.” *Journal of Antimicrobial Chemotherapy*, no. 73: 787–94. https://doi.org/10.1093/jac/dkx465.

Johnson, Alan P. 2015. “Surveillance of Antibiotic Resistance.” *Phil. Trans. R. Soc. B* 370

(20140080). https://doi.org/10.1097/00001432-199908000-00002.

Jorgensen, J. H. 1993. “Selection Criteria for an Antimicrobial Susceptibility Testing System.” *Journal of Clinical Microbiology* 31 (11): 2841–44.

Jorgensen, James H, and Mary Jane Ferraro. 2009. “Antimicrobial Susceptibility Testing : A Review of General Principles and Contemporary Practices AND THE RATIONALE FOR PERFORMING.” *Clinical Infectious Diseases* 49 (1 December): 1749–55. https://doi.org/10.1086/647952.

Kahlmeter, G, P Menday, and O Cars. 2003. “Non-Hospital Antimicrobial Usage and Resistance in Community-Acquired Escherichia Coli Urinary Tract Infection.” *Journal of Antimicrobial Chemotherapy* 52 (6): 1005–10. https://doi.org/10.1093/jac/dkg488.

Kallonen, Teemu, Hayley J Brodrick, Simon R Harris, Jukka Corander, Nicholas M Brown, Veronique Martin, Sharon J Peacock, and Julian Parkhill. 2017. “Systematic Longitudinal Survey of Invasive Escherichia Coli in England Demonstrates a Stable Population Structure Only Transiently Disturbed by the Emergence of ST131.” *Genome Research*, 1437–49. https://doi.org/10.1101/gr.216606.116.Freely.

Kennedy, James. 2010. “Particle Swarm Optimization.” In *Encyclopedia of Machine Learning*, 760–66. Springer US.

Khaledi, Ariane, Aaron Weimann, Monika Schniederjans, Tzu-hao Kuo, Antonio Oliver, Gabriel Cabot, Axel Kola, et al. 2019. “Fighting Antimicrobial Resistance in Pseudomonas with Machine Learning-Enabled Molecular Diagnostics.”

Kim, Eric. 2015. “Everything You Wanted to Know about the Kernel Trick (But Were Too Afraid to Ask).” 2015. http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html.

Kim, Jiwoong, David E Greenberg, Reed Pfifer, Shuang Jiang, Guanghua Xiao, Yang Xie,

Samuel A Shelburne, Andrew Koh, and Xiaowei Zhan. 2019. “VAMPr: VAriant Mapping and Prediction of Antibiotic Resistance via Explainable Features and Machine Learning.” *BioRxiv*, 537381. https://doi.org/10.1101/537381.

Kong, Ling Yuan, David W Eyre, Jacques Corbeil, Frederic Raymond, A Sarah Walker, Mark H Wilcox, Derrick W Crook, et al. 2018. “Clostridium Difficile: Investigating Transmission Patterns Between Infected and Colonized Patients Using Whole Genome Sequencing.” *Clinical Infectious Diseases* 68 (2): 204–9. https://doi.org/10.1093/cid/ciy457.

Kotwani, Anita, Chand Wattal, Shashi Katewa, P. C. Joshic, and Kathleen Holloway. 2010. “Factors Influencing Primary Care Physicians to Prescribe Antibiotics in Delhi India.” *Family Practice* 27 (6): 684–90. https://doi.org/10.1093/fampra/cmq059.

Kouchaki, Samaneh, Yang Yang, Timothy M Walker, A Sarah Walker, Daniel J Wilson, Timothy E A Peto, Derrick W Crook, et al. 2018. “Application of Machine Learning Techniques to Tuberculosis Drug Resistance Analysis.” *Bioinformatics*, no. November: 1–7. https://doi.org/10.1093/bioinformatics/bty949.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” *Advances in Neural Information Processing Systems* 2: 1097–1105.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Laupland, Kevin B., D. B. Gregson, D. L. Church, T. Ross, and J. D D Pitout. 2008. “Incidence, Risk Factors and Outcomes of Escherichia Coli Bloodstream Infections in a Large Canadian Region.” *Clinical Microbiology and Infection* 14 (11): 1041–47. https://doi.org/10.1111/j.1469-0691.2008.02089.x.

Laxminarayan, Ramanan, Zulfiqar a Bhutta, Adriano Duse, Thomas O Brien, Iruka N Okeke,

Ariel Pablo-mendez, and Keith P Klugman. 2003. “Drug Resistance.” In *Disease Control Priorities in Developing Countries. 2nd Edition. Washington (DC): The International Bank for Reconstruction and Development / The World Bank*, edited by Dean T Jamison, Joel G Breman, Anthony R Measham, George Alleyne, Mariam Claeson, David B Evans, Prabhat Jha, Anne Mills, and Philip Musgrove, 1031–51. https://doi.org/http://files.dcp2.org/pdf/DCP/DCP55.pdf.

Leitner, Eva, Gernot Zarfel, Josefa Luxner, Kathrin Herzog, Shiva Pekard-Amenitsch, Martin Hoenigl, Thomas Valentin, et al. 2015. “Contaminated Handwashing Sinks as the Source of a Clonal Outbreak of KPC-2-Producing Klebsiella Oxytoca on a Hematology Ward.” *Antimicrobial Agents and Chemotherapy* 59 (1): 714–16. https://doi.org/10.1128/AAC.04306-14.

Leverstein-van Hall, Maurine A., Karola Waar, Jan Muilwijk, James Cohen Stuart, L. J M Sabbeon, H. M E Frenay, B. Maraha, et al. 2013. “Consequences of Switching from a Fixed 2:1 Ratio of Amoxicillin/Clavulanate (CLSI) to a Fixed Concentration of Clavulanate (EUCAST) for Susceptibility Testing of Escherichia Coli.” *Journal of Antimicrobial Chemotherapy* 68 (11): 2636–40. https://doi.org/10.1093/jac/dkt218.

Lin, Lawrence I-kuei. 1989. “A Concordance Correlation Coefficient to Evaluate Reproducibility.” *Biometrics* 45 (1): 255–68.

Lishman, H, C Costelloe, S Hopkins, A Johnson, R Hope, R Guy, A Holmes, and P Aylin. 2016. “Effect of Trimethoprim / Nitrofurantoin Prescribing on the Incidence and Antibiotic Susceptibility Patterns of E . Coli Bacteraemia Nationally at the GP Practice Level (2012-2014).” In *26th European Congress of Clinical Microbiology and Infectious Diseases*. Amsterdam, Netherlands.

Liss, Blassius, and Oliver A Cornely. 2016. “Burden and Benefit of Antibiotic Prophylaxis in

Cancer Chemotherapy." *Lancet Infectious Diseases* 16 (June): 2016.

Livermore, David M, Russell Hope, Rosy Reynolds, Ruth Blackburn, Alan P Johnson, and Neil Woodford. 2013. "Declining Cephalosporin and Fluoroquinolone Non-Susceptibility among Bloodstream Enterobacteriaceae from the UK : Links to Prescribing Change?" *Journal of Antimicrobial Chemotherapy* 68 (June): 2667–74. https://doi.org/10.1093/jac/dkt212.

Livermore, David M, Peter Stephens, Julius Weinberg, Alan P Johnson, Tiffany Gifford, Dorabella Northcott, Dorothy James, Robert C George, and David C E Speller. 2000. "Regional Variation in Ampicillin and Trimethoprim Resistance in Escherichia Coli in England from 1990 to 1997, in Relation to Antibacterial Prescribing." *Journal of Antimicrobial Chemotherapy* 46: 411–22.

Llewelyn, Martin J, Kieran Hand, Susan Hopkins, and A Sarah Walker. 2015. "Antibiotic Policies in Acute English NHS Trusts: Implementation of 'Start Smart—Then Focus' and Relationship with Clostridium Difficile Infection Rates." *Journal of Antimicrobial Chemotherapy* 70: 1230–35. https://doi.org/10.1093/jac/dku515.

Logan, Latania K., and Robert A. Weinstein. 2017. "The Epidemiology of Carbapenem-Resistant Enterobacteriaceae: The Impact and Evolution of a Global Menace." *Journal of Infectious Diseases* 215 (Suppl 1): S28–36. https://doi.org/10.1093/infdis/jiw282.

London School of Hygiene & Tropical Medicine. 2016. "Supporting Surveillance Capacity for Antimicrobial Resistance LABORATORY CAPACITY STRENGTHENING FOR DRUG RESISTANT INFECTIONS IN LOW AND MIDDLE INCOME COUNTRIES."

Ludden, Catherine, Sandra Reuter, Kim Judge, Theodore Gouliouris, Beth Blane, Francesc Coll, Plamena Naydenova, et al. 2017. "Sharing of Carbapenemase-Encoding Plasmids between Enterobacteriaceae in UK Sewage Uncovered by MinION Sequencing."

Microbial Genomics 3 (7): e000114. https://doi.org/10.1099/mgen.0.000114.

Lunn, M., and D. McNeil. 1995. "Applying Cox Regression to Competing Risks." *Biometrics*, 524–32. https://doi.org/10.2307/2532940.

Mathers, Amy J., Kasi Vegesana, Ian German Mesner, Katie E. Barry, Aaron Pannone, Josh Baumann, Derrick W. Crook, et al. 2018. "Intensive Care Unit Wastewater Interventions to Prevent Transmission of Multispecies Klebsiella Pneumoniae Carbapenemase-Producing Organisms." *Clinical Infectious Diseases* 67 (2): 171–78. https://doi.org/10.1093/cid/ciy052.

Mathers, Amy J, Heather L Cox, Brandon Kitchel, Hugo Bonatti, Karen C Brassinga, Joanne Carroll, and W Michael Scheld. 2011. "Molecular Dissection of an Outbreak of Carbapenem-Resistant Enterobacteriaceae Reveals Intergenus KPC Carbapenemase Transmission through a Promiscuous Plasmid." *MBio* 2 (6): e00204. https://doi.org/10.1128/mBio.00204-11.Editor.

McVean, Gil. 2009. "A Genealogical Interpretation of Principal Components Analysis." *PLoS Genetics* 5 (10). https://doi.org/10.1371/journal.pgen.1000686.

Mee-Marquet, Nathalie L van der, Dominique S Blanc, Houssein Gbaguidi-Haore, Sandra Dos Santos Borges, Quentin Viboud, Xavier Bertrand, Roland Quentin, and (Infection Control Regional Working Group of the Reseau des Hygienistes du Centre). 2015. "Marked Increase in Incidence for Bloodstream Infections Due to Escherichia Coli, a Side Effect of Previous Antibiotic Therapy in the Elderly." *Frontiers in Microbiology* 6 (June). https://doi.org/10.3389/fmicb.2015.00646.

Meyer, Keith C. 2005. "Aging." *Proceedings of the American Thoracic Society* 2 (5): 433–39. https://doi.org/10.1513/pats.200508-081JS.

Miller, Jason R, Sergey Koren, and Granger Sutton. 2010. "Assembly Algorithm for Next-

Ganeration Sequencing Data (Manuscript).” *Genomics* 95 (6): 315–27. https://doi.org/10.1016/j.ygeno.2010.03.001.Assembly.

Miller, R., A. S. Walker, K. Knox, D. Wyllie, J. Paul, E. Haworth, D. Mant, T. Peto, and D. W. Crook. 2010. “Feral and Wild-Type Methicillin-Resistant Staphylococcus Aureus in the United Kingdom.” *Epidemiology and Infection* 138 (5): 655–65. https://doi.org/10.1017/S0950268809991294.

Murphy, K. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Musicha, Patrick, Jennifer E. Cornick, Naor Bar-Zeev, Neil French, Clemens Masesa, Brigitte Denis, Neil Kennedy, et al. 2017. “Trends in Antimicrobial Resistance in Bloodstream Infection Isolates at a Large Urban Hospital in Malawi (1998–2016): A Surveillance Study.” *The Lancet Infectious Diseases* 17 (10): 1042–52. https://doi.org/10.1016/S1473-3099(17)30394-8.

National Institute of Dental and Craniofacial Research. 2020. “Periodontal Disease in Seniors (Age 65 and Over).” 2020. https://www.nidcr.nih.gov/research/data-statistics/periodontal-disease/seniors.

Newcombe, H. B, A. P. James, and S. J. Axford. 1957. “Family Linkage of Vital and Health Records.”

Nguyen, Marcus, Thomas Brettin, S. Wesley Long, James M. Musser, Randall J. Olsen, Robert Olson, Maulik Shukla, et al. 2018. “Developing an in Silico Minimum Inhibitory Concentration Panel Test for Klebsiella Pneumonia.” *Scientific Reports* 8 (1): 1–11. https://doi.org/10.1038/s41598-017-18972-w.

Nguyen, Marcus, S. Wesley Long, Patrick F McDermott, Randall J Olsen, Robert Olson, Rick L Stevens, Gregory H Tyson, Shaohua Zhao, and James J Davis. 2018. “Using Machine Learning to Predict Antimicrobial Minimum Inhibitory Concentrations and Associated

Genomic Features for Nontyphoidal Salmonella." *Journal of Clinical Microbiology* 57 (2): 380782. https://doi.org/10.1128/JCM.01260-18.

NHS Digital. 2019. "Hospital Episode Statistics." 2019. https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics.

Nys, Sita, Peter H. Terporten, Jacomina A.A. Hoogkamp-Korstanje, and Ellen E. Stobberingh. 2008. "Trends in Antimicrobial Susceptibility of Escherichia Coli Isolates from Urology Services in The Netherlands (1998-2005)." *Journal of Antimicrobial Chemotherapy* 62 (1): 126–32. https://doi.org/10.1093/jac/dkn151.

O'Neil, Jim. 2014. "Antimicrobials in Agriculture and the Environment: Reducing Unnecessary Use and Waste," no. December.

O'Neill, J. 2014. "Antimicrobial Resistance : Tackling a Crisis for the Health and Wealth of Nations." *Review on Antimicrobial Resistance*, no. December: 1–16.

Olsson, O., S. Bergstrom, F. P. Lindberg, and S. Normark. 1983. "AmpC Beta-Lactamase Hyperproduction in Escherichia Coli: Natural Ampicillin Resistance Generated by Horizontal Chromosomal DNA Transfer from Shigella." *Proceedings of the National Academy of Sciences* 80 (24): 7556–60. https://doi.org/10.1073/pnas.80.24.7556.

ONS. 2018. "Clinical Commissioning Group Mid Year Population Estimates." 2018. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/clinicalcommissioninggroupmidyearpopulationestimates.

Orlek, A., H. Phan, A. E. Sheppard, M. Doumith, M. Ellington, T. Peto, and N. Stoesser. 2017. "Ordering the Mob: Insights into Replicon and MOB Typing Schemes from Analysis of a Curated Dataset of Publicly Available Plasmids." *Plasmid* 91: 42–52. https://doi.org/10.1016/j.plasmid.2017.03.002.

Paisitkriangkrai, Paul. 2012. "Linear Regression and Support Vector Regression." 2012.

https://cs.adelaide.edu.au/~chhshen/teaching/ML_SVR.pdf.

Pedregosa-Izquierdo, Fabian. 2015. “Feature Extraction and Supervised Learning on FMRI: From Practice to Theory.” Université Pierre et Marie Curie - Paris VI. https://doi.org/<NNT: 2015PA066015> <tel-01100921v2>.

Pedregosa, Fabian, Francis Bach, and Alexandre Gramfort. 2014. “On the Consistency of Ordinal Regression Methods” 18: 1–35. http://arxiv.org/abs/1408.2327.

Pedregosa, Fabian, Ron Weiss, and Matthieu Brucher. 2011. “Scikit-Learn : Machine Learning in Python” 12: 2825–30.

Pesesky, Mitchell W., Tahir Hussain, Meghan Wallace, Sanket Patel, Saadia Andleeb, Carey Ann D. Burnham, and Gautam Dantas. 2016. “Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-Negative Bacilli from Whole Genome Sequence Data.” *Frontiers in Microbiology* 7 (NOV): 1–17. https://doi.org/10.3389/fmicb.2016.01887.

Poolman, Jan T., and Annaliesa S. Anderson. 2018. “Escherichia Coli and Staphylococcus Aureus: Leading Bacterial Pathogens of Healthcare Associated Infections and Bacteremia in Older-Age Populations.” *Expert Review of Vaccines* 17 (7): 607–18. https://doi.org/10.1080/14760584.2018.1488590.

Prager, Elliot. 1978. “Distribution of Cancers of the Large Bowel.” *JAMA: The Journal of the American Medical Association* 239 (17): 1744. https://doi.org/10.1001/jama.1978.03280440028009.

Public Health England. 2015a. “English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR) 2010 to 2014.”

———. 2015b. “Polymicrobial Bacteraemia and Fungaemia in England, Wales and Northern Ireland, 2014.” Vol. 9.

———. 2015c. “Thirty-Day All-Cause Fatality Subsequent to MRSA, MSSA and E . Coli Bacteraemia and C . Difficile Infection.”

———. 2016a. “Annual Epidemiological Commentary Mandatory MRSA, MSSA and E. Coli Bacteraemia and C. Difficile Infection Data 2015/16.”

———. 2016b. “Mandatory Enhanced MRSA, MSSA and Escherichia Coli Bacteraemia, and Clostridium Difficile Infection Surveillance.”

———. 2017. “English Surveillance Programme for Antimicrobial Utilisation and Resistance (ESPAUR).”

———. 2019. “Annual Epidemiological Commentary: Bacteraemia, MSSA Bacteraemia and C . Difficile Infections, up to and Including Financial Year April 2018 to March 2019,” 1–88.

Public Health England and NHS Improvement. 2017. “Guidance on the Definition of Healthcare Associated Gram-Negative Bloodstream Infections,” no. July.

Purohit, Manju Raj, Lars Falkdalen Lindahl, Vishal Diwan, Gaetano Marrone, and Cecilia Stålsby Lundborg. 2019. “High Levels of Drug Resistance in Commensal E. Coli in a Cohort of Children from Rural Central India.” *Scientific Reports* 9 (1): 1–11. https://doi.org/10.1038/s41598-019-43227-1.

Quainoo, Scott, Jordy P. M. Coolen, Sacha A. F. T. van Hijum, Martijn A. Huynen, Willem J. G. Melchers, Willem van Schaik, and Heiman F. L. Wertheim. 2017. “Whole-Genome Sequencing of Bacterial Pathogens : The Future of Nosocomial Outbreak Analysis.” *American Society for Microbiology* 30 (4): 1015–64.

Quan, T. Phuong, Zharain Bawa, Dona Foster, Tim Walker, Carlos Del Ojo Elias, Priti Rathod, Zamin Iqbal, et al. 2018. “Evaluation of Whole-Genome Sequencing for Mycobacterial Species Identification and Drug Susceptibility Testing in a Clinical Setting: A Large-Scale

Prospective Assessment of Performance against Line Probe Assays and Phenotyping." *Journal of Clinical Microbiology* 56 (2): 1–14. https://doi.org/10.1128/JCM.01480-17.

Quan, T Phuong, Nicola J Fawcett, John M Wrightson, John Finney, David Wyllie, Katie Jeffery, Nicola Jones, et al. 2016. "Increasing Burden of Community-Acquired Pneumonia Leading to Hospitalisation, 1998 – 2014." *BMJ Thorax*, 1–8. https://doi.org/10.1136/thoraxjnl-2015-207688.

R Core Team. 2017. "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing. 2017.

Randhawa, Varinder, Syed Sarwar, Sandra Walker, Marion Elligsen, Lesley Palmay, and Nick Daneman. 2014. "Weighted-Incidence Syndromic Combination Antibiograms to Guide Empiric Treatment of Critical Care Infections: A Retrospective Cohort Study." *Critical Care* 18 (3). https://doi.org/10.1186/cc13901.

Rennie, Jdm, and Nathan Srebro. 2005. "Loss Functions for Preference Levels: Regression with Discrete Ordered Labels." *IJCAI Workshop on Advances in Preference Handling*, 6. http://ttic.uchicago.edu/~nati/Publications/RennieSrebroIJCAI05.pdf.

Richet, H, A Reynaud, N Y Caro, E Espaze, and I Be. 1999. "Mutations in the AmpC Promoter of Escherichia Coli Isolates Resistant to Oxyiminocephalosporins without Extended Spectrum Beta-Lactamase Production." *FEMS Microbiology Letters* 173: 459–65.

Riley, Richard D, Julian P T Higging, and Jonathan J Deeks. 2011. "Interpretation of Random Effects Meta-Analyses." *BMJ* 342 (February): 1–7. https://doi.org/https://doi.org/10.1136/bmj.d549 (Published.

Rizk, Guillaume, Dominique Lavenier, and Rayan Chikhi. 2013. "DSK: K-Mer Counting with Very Low Memory Usage." *Bioinformatics* 29 (5): 652–53. https://doi.org/10.1093/bioinformatics/btt020.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. 2007. “A Review of Feature Selection Techniques in Bioinformatics.” *Bioinformatics* 23 (19): 2507–17. https://doi.org/10.1093/bioinformatics/btm344.

Samuel, A. L. 1959. “Some Studies in Machine Learning Using the Game of Checkers.” *IBM Journal of Research and Development*, 71–105.

Schlackow, Iryna, Nicole Stoesser, A Sarah Walker, Derrick W Crook, Tim E A Peto, and David H Wyllie. 2012. “Increasing Incidence of Escherichia Coli Bacteraemia Is Driven by an Increase in Antibiotic-Resistant Isolates: Electronic Database Study in Oxfordshire 1999-2011.” *The Journal of Antimicrobial Chemotherapy* 67 (6): 1514–24. https://doi.org/10.1093/jac/dks082.

Schlackow, Iryna, a Sarah Walker, Kate Dingle, David Griffiths, Sarah Oakley, John Finney, Ali Vaughan, et al. 2012. “Surveillance of Infection Severity: A Registry Study of Laboratory Diagnosed Clostridium Difficile.” *PLoS Medicine* 9 (7): e1001279. https://doi.org/10.1371/journal.pmed.1001279.

Seale, Anna C., N. Claire Gordon, Jasmin Islam, Sharon J. Peacock, and J. Anthony G. Scott. 2017. “AMR Surveillance in Low and Middle-Income Settings - A Roadmap for Participation in the Global Antimicrobial Surveillance System (GLASS).” *Wellcome Open Research* 2. https://doi.org/10.12688/wellcomeopenres.12527.1.

Seemann, Torsten. 2019. “ABRICATE.” 2019. https://github.com/tseemann/abricate.

Sheppard, Anna E., Nicole Stoesser, Daniel J. Wilson, Robert Sebra, Andrew Kasarskis, A Luke W. Anson, Adam Giess, et al. 2016. “Nested Russian Doll-Like Genetic Mobility Drives Rapid Dissemination of the Carbapenem Resistance Gene BlaKPC.” *Antimicrobial Agents and Chemotherapy* 60 (6): 3767–78. https://doi.org/10.1128/AAC.00464-16.Address.

Sheppard, S. K., X. Didelot, G. Meric, A. Torralbo, K. A. Jolley, D. J. Kelly, S. D. Bentley, M. C. J. Maiden, J. Parkhill, and D. Falush. 2013. “Genome-Wide Association Study Identifies Vitamin B5 Biosynthesis as a Host Specificity Factor in Campylobacter.” *Proceedings of the National Academy of Sciences* 110 (29): 11923–27. https://doi.org/10.1073/pnas.1305559110.

Shields, Ryan K, Liang Chen, Shaoji Cheng, Kalyan D Chavda, and Ellen G Press. 2017. “Emergence of Ceftazidime-Avibactam Resistance Due to Plasmid-Borne BlaKPC-3 Mutations during Treatment of Carbapenem-Resistant Klebsiella Pneumoniae Infections.” *Antimicrob Agents Chemother.* 61 (3): 1–11. https://doi.org/10.1128/AAC.02097-16.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. “Mastering the Game of Go with Deep Neural Networks and Tree Search.” *Nature* 529 (7587): 484–89. https://doi.org/10.1038/nature16961.

Snitkin, Evan S., Adrian M. Zelazny, Pamela J. Thomas, Frida Stock, David K. Henderson, Tara N. Palmore, and Julia A. Segre. 2012. “Tracking a Hospital Outbreak of Carbapenem-Resistant Klebsiella Pneumoniae with Whole-Genome Sequencing.” *Sci Transl Med* 4(148). https://doi.org/10.1126/scitranslmed.3004129.

Stoesser, N., A. E. Sheppard, G. Peirano, L. W. Anson, L. Pankhurst, R. Sebra, H. T.T. Phan, et al. 2017. “Genomic Epidemiology of Global Klebsiella Pneumoniae Carbapenemase (KPC)-Producing Escherichia Coli.” *Scientific Reports* 7 (1): 1–11. https://doi.org/10.1038/s41598-017-06256-2.

Stoesser, N, E M Batty, D W Eyre, M Morgan, D H Wyllie, C Del Ojo Elias, J R Johnson, A S Walker, T E A Peto, and D W Crook. 2013. “Predicting Antimicrobial Susceptibilities for

Escherichia Coli and Klebsiella Pneumoniae Isolates Using Whole Genomic Sequence Data," 1–11. https://doi.org/10.1093/jac/dkt180.

The World Bank. 2017. "Waterlife: Improving Access to Safe Drinking Water in India," 1–32. https://www.innovationpolicyplatform.org/system/files/Waterlife Case_April 20 2017_0.pdf.

Vanwinckelen, Gitte, and Hendrik Blockeel. 2012. "On Estimating Model Accuracy with Repeated Cross-Validation." *21st Belgian-Dutch Conference on Machine Learning*, 39–44. https://doi.org/10.2307/1868898.

Vernet, G, C Mary, D.M. Altmann, O. Doumbo, S. Morpeth, and Z.A. Bhutta. 2014. "Surveillance for Antimicrobial Drug Resistance in Under-Resourced Countries." *Emerging Infectious Diseases* 20 (3): 434–41. https://doi.org/http://dx.doi.org/10.3201/eid2003.121157.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the Metafor Package." *Journal of Statistical Software* 36 (3): 1–48.

Vihta, Karina-Doris, Nicole Stoesser, Martin J Llewelyn, T Phuong Quan, Tim Davies, Nicola J Fawcett, Laura Dunn, et al. 2018. "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." *The Lancet Infectious Diseases* 18 (10): 1138–49. https://doi.org/http://dx.doi.org/10.1016/S1473-3099%2818%2930353-0.

Vlieghe, E., A.M. Bal, and I.M. Gould. 2010. "Chapter 26 Surveillance of Antibiotic Resistance in Developing Countries: Needs, Constraints and Realities." In *Antimicrobial Resistance in Developing Countries*, 463–75. https://doi.org/10.1007/978-0-387-89370-9.

Wald, Abraham. 1943. "Tests of Statistical Hypotheses Concerning Several Parameters

When the Number of Observations Is Large." *Transactions of the American Mathematical Society* 54 (3): 426. https://doi.org/10.2307/1990256.

Walker, Sarah, Tim E A Peto, Lily O'Connor, Derrick W Crook, and David Wyllie. 2008. "Are There Better Methods of Monitoring MRSA Control than Bacteraemia Surveillance? An Observational Database Study." *PloS One* 3 (6): e2378. https://doi.org/10.1371/journal.pone.0002378.

Walker, Timothy M., Thomas A. Kohl, Shaheed V. Omar, Jessica Hedge, Carlos Del Ojo Elias, Phelim Bradley, Zamin Iqbal, et al. 2015. "Whole-Genome Sequencing for Prediction of Mycobacterium Tuberculosis Drug Susceptibility and Resistance: A Retrospective Cohort Study." *The Lancet Infectious Diseases* 15 (10): 1193–1202. https://doi.org/10.1016/S1473-3099(15)00062-6.

Wilson, Daniel J. 2019. "The Harmonic Mean P-Value for Combining Dependent Tests." *Proceedings of the National Academy of Sciences of the United States of America* 116 (4): 1195–1200. https://doi.org/10.1073/pnas.1814092116.

Wirth, Thierry, Daniel Falush, Ruiting Lan, Frances Colles, Patience Mensa, Lothar H. Wieler, Helge Karch, et al. 2006. "Sex and Virulence in Escherichia Coli: An Evolutionary Perspective." *Molecular Microbiology* 60 (5): 1136–51. https://doi.org/10.1111/j.1365-2958.2006.05172.x.

Wood, Jonathan. 2019. "Penicillin: The Oxford Story." 2019. http://www.ox.ac.uk/news/science-blog/penicillin-oxford-story.

World Health Organisation. 2015. "Global Action Plan on Antimicrobial Resistance." *WHO Press*, 1–28. https://doi.org/ISBN 978 92 4 150976 3.

———. 2018. "The Top 10 Causes of Death." 2018. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

———. 2019. “Ten Threats to Global Health in 2019.” 2019. https://www.who.int/emergencies/ten-threats-to-global-health-in-2019.

World Health Organization. 2012. “Antimicrobial Resistance Surveillance.” 2012. https://www.who.int/medicines/areas/rational_use/AMR_Surveillance/en/.

———. 2015. “Global Antimicrobial Resistance Surveillance System.” http://apps.who.int/iris/bitstream/10665/188783/1/9789241549400_eng.pdf?ua=1.

———. 2019a. “Defined Daily Dose.” 2019. https://www.whocc.no/ddd/definition_and_general_considera/.

———. 2019b. “GLASS Enrolled Countries.” 2019. https://docs.google.com/spreadsheets/d/14QJ4tUfqmS5YF60BOpXZzZffwr6cRlu_vEZ9_oYfpRA/edit#gid=0.

Wyllie, D H, A S Walker, T E A Peto, and D W Crook. 2007. “Hospital Exposure in a UK Population, and Its Association with Bacteraemia.” *Journal of Hospital Infection*, no. 67: 301–7. https://doi.org/10.1016/j.jhin.2007.08.018.

Wyllie, David H., Derrick W. Crook, and Tim E.A. Peto. 2006. “Mortality after Staphylococcus Aureus Bacteraemia in Two Acute Hospitals in Oxfordshire, 1997-2003: Cohort Study.” *British Medical Journal* 333 (7562): 281–84. https://doi.org/10.1136/bmj.38834.421713.2F.

Wyllie, David H., Tim E.A. Peto, and Derrick Crook. 2005. “MRSA Bacteraemia in Patients on Arrival in Hospital: A Cohort Study in Oxfordshire 1997-2003.” *British Medical Journal* 331 (7523): 992–95. https://doi.org/10.1136/bmj.38558.453310.8F.

Wyllie, David H, a Sarah Walker, Ruth Miller, Catrin Moore, Susan R Williamson, Iryna Schlackow, John M Finney, Lily O’Connor, Tim E a Peto, and Derrick W Crook. 2011. “Decline of Meticillin-Resistant Staphylococcus Aureus in Oxfordshire Hospitals Is

Strain-Specific and Preceded Infection-Control Intensification." *BMJ Open* 1 (1): e000160. https://doi.org/10.1136/bmjopen-2011-000160.

Xie, C., D. Mcd Taylor, B. P. Howden, and P. G P Charles. 2012. "Comparison of the Bacterial Isolates and Antibiotic Resistance Patterns of Elderly Nursing Home and General Community Patients." *Internal Medicine Journal* 42 (7): 21–24. https://doi.org/10.1111/j.1445-5994.2011.02436.x.

Yang, Yang, Timothy M Walker, A Sarah Walker, Daniel J Wilson, Timothy E A Peto, Derrick W Crook, Farah Shamout, Tingting Zhu, and David A Clifton. 2019. " DeepAMR for Predicting Co-Occurrent Resistance of Mycobacterium Tuberculosis ." *Bioinformatics*, no. January: 1–10. https://doi.org/10.1093/bioinformatics/btz067.

Yelin, Idan, Olga Snitser, Gal Novich, Rachel Katz, Ofir Tal, Miriam Parizade, Gabriel Chodick, Gideon Koren, Varda Shalev, and Roy Kishony. 2019. "Resistance of Urinary Tract Infections." *Nature Medicine* 25 (July).

Zankari, Ea, Henrik Hasman, Salvatore Cosentino, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M. Aarestrup, and Mette Voldby Larsen. 2012. "Identification of Acquired Antimicrobial Resistance Genes." *Journal of Antimicrobial Chemotherapy* 67 (11): 2640–44. https://doi.org/10.1093/jac/dks261.

Zerbino, Daniel R., and Ewan Birney. 2008. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." *Genome Research* 18 (5): 821–29. https://doi.org/10.1101/gr.074492.107.

Zhou, Xiang, and Matthew Stephens. 2012. "Genome-Wide Efficient Mixed Model Analysis for Association Studies." *Nature Genet.* 44 (7): 821–24. https://doi.org/10.1038/ng.2310.