The feasibility of using electronic health records to inform clinical decision making for community-onset urinary tract infection in England.

Patrick Rockenschaub

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Institute of Health Informatics

University College London

15th July 2021

I, Patrick Rockenschaub, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Urinary tract infections (UTIs) are a major source of morbidity, yet differentiating UTI from other conditions and choosing the right treatment remains challenging. Using case studies from English primary and secondary care, this thesis investigates the potential use of electronic health records (EHR) — i.e., data recorded as part of routine care — to aid the diagnosis and management of community-onset UTI.

I start by introducing sources of uncertainty in diagnosing UTI (Chapter 1) and review how EHRs have previously been used to study UTIs (Chapter 2). In Chapter 3, I discuss EHR sources available to study UTIs in England. In Chapter 4, I explore how EHRs from primary care can be used to guide antibiotic prescribing for UTI, by evaluating harms of delaying treatment in key patient groups. In Chapters 5 and 6, I explore the use of EHR data as a diagnostic tool to guide antibiotic de-escalation in patients with suspected UTI in the emergency department (ED).

Cases of community-onset UTI could be identified in both primary and secondary care data but case definitions relied heavily on coarse diagnostic codes. A lack of information on patients' acute health status, clinical observations (e.g., urine dipstick tests), and reasons for antibiotic prescribing resulted in heterogeneous study cohorts, which likely confounded estimated effects of antibiotic treatment in primary care. In secondary care, early prediction of bacteriuria to guide antibiotic prescribing decisions in the ED proved promising, but model performance varied greatly by patient mix and variable definitions.

Better recording of clinical information and a combination of retrospective EHR analysis with prospective cohorts and qualitative approaches will be required to derive actionable insights on UTI. Results based solely on currently available EHR data need to be interpreted carefully.

Impact Statement

Antibiotics are often used for patients that look like they might have a urinary tract infection (UTI). Often antibiotics are right for these patients and help them get better, but they only work if the patient really has a bacterial infection. If they don't, antibiotics are ineffective but may still cause side effects and promote the emergence of antimicrobial resistance. Deciding who does and does not need antibiotics is difficult, and current clinical rules or diagnostic tests frequently get it wrong. Doctors therefore often opt to use antibiotics when in doubt to ensure that their patients are as safe as possible. New strategies are needed to reduce the frequent over-prescribing for UTI in both general practice and hospitals.

When patients make use of the healthcare system, an abundance of information is generated. More and more of this information is captured within local and national research databases in the form of structured electronic health records (EHRs). Over the last two decades, EHRs have already been increasingly used to describe the epidemiology of UTIs. As the breadth and depth of recorded information increases, the same data may further be used directly to support doctors in diagnosing and managing UTIs — e.g., by estimating the probability of a patient having UTI and/or estimating the likely patient benefit of taking antibiotics. This thesis critically assesses the feasibility of using data captured in EHRs to answer these questions.

In doing so, this thesis provides context on how results of previous studies that used EHRs to investigate UTIs may be interpreted. Findings of this thesis have already been or will be disseminated at international conferences and in peer-reviewed scientific journals. All publications arising from this thesis try to present an honest view of the quality of evidence that may be derived from EHR

data for a given research question, clearly discussing any biases with regards to the way that data on UTIs is currently captured within EHRs. It contributes towards a better understanding of the strengths and limitations of EHR data, establishing a foundation for future EHR research that aims to improve the diagnosis and management of community-onset UTIs in England.

Where clinical results of this thesis were judged promising, work is underway to translate the findings into real-world clinical practice and apply them to improve patient care in the English National Health Service. I have obtained seed funding from the UCL-led Precision AMR funding initiative to externally validate the clinical risk prediction model for bacteriuria developed in this thesis at University College London Hospital. Embedded in a wider team of researchers, I am actively engaging with clinicians and patients to use the findings of this thesis as a basis for discussing how statistical models may be introduced into the healthcare system to tackle the current over-use of antibiotics for suspected UTI.

Acknowledgements

I would like to express my sincere thanks to all those who made this thesis possible. First, I am deeply thankful to Dr Laura Shallcross, who guided me since I arrived at UCL, kindled my passion for research, and supported me in finding my way back whenever that same passion led me down another rabbit hole. I am extremely grateful for the trust you put in me and it was a genuine pleasure to work alongside and learn from you. I further would like to thank Prof Nick Freemantle and Prof Andrew Hayward, who helped me to always keep track of the bigger picture that sometimes gets lost in the fray of a PhD.

During my time at the UCL Institute of Health Informatics, I had the opportunity to work with and learn from many amazing colleagues. Chief among those, I would like to thank: Nonie Alexander for countless pep talks over cups of tea, Selina Patel for always providing a new perspective, Dr Anna Aryee and Dr Arnoupe Jhass for incredible help (and patience) with interpreting my clinical findings, and Dr Peter Dutey-Magni for always being there to spitball my most random research ideas. I would also like to thank Dr Martin Gill and Dr David McNulty at University Hospitals Birmingham NHS Foundation Trust, who provided me with clinical and technical support to navigate the rich data at their institution and without whom I'd been left with an uninterpretable mess.

Nothing I accomplished over the last years would have been possible without the love and support of Anna Lena. Thank you for sticking with me through all the ups and downs that make up a PhD. In this, she was joined by my parents, brother, and friends, who bore with me whenever I wouldn't shut up about my research and who kept me sane despite my best efforts to the contrary.

Contents

List of Figures

List of Tables

Chapter 1

Uncertainty in the diagnosis and management of UTIs

Abstract

Urinary tract infections (UTIs) affect an estimated 150 million individuals worldwide every year, causing substantial morbidity and healthcare costs. Ideally, doctors would be able to diagnose UTIs instantly and without ambiguity. Antibiotics would be prescribed only when they are needed and when they are effective. Unfortunately, it is difficult to differentiate patients with genuine UTI from those with other conditions. Over-treatment of UTI in primary and secondary care is common. This has a negative impact on patients by delaying their access to effective treatment and unnecessarily exposing them to side-effects such as antibiotic resistance, which result from inappropriate antibiotic therapy. In this first chapter of my thesis, I describe the epidemiology of UTI and discuss the uncertainties associated with correctly diagnosing and managing UTIs. I end the chapter by introducing the aims and objectives of my thesis, which will explore how routinely collected electronic health record (EHR) data in England may be used across healthcare settings to reduce these uncertainties and improve the diagnosis and management of community-onset UTI.

1.1 Epidemiology

While rough estimates suggest that 150 million individuals world-wide are affected by urinary tract infections (UTIs) each year [1], precise numbers for individual countries are hard to obtain. Differences in health systems, health seeking

behaviour, data recording, and study designs make it difficult to estimate the burden of disease [2, 3, 4]. Despite the uncertainties in the estimates, findings agree on their relative importance. UTIs are generally considered one of the most common infectious reasons for consultation in primary care in the United States (US) [2], France [4], the Netherlands [5], and the United Kingdom (UK) [6]. UTIs feature among the top ten reasons for Accidents & Emergency visits in the UK [7], and account for an estimated 40% of healthcare-associated infections worldwide [1]. The incidence of UTI-related healthcare contacts has increased since the turn of the century in both the US and the UK, especially among women and the elderly [8, 9].

Infections of the urinary tract are particularly frequent in women, with almost 40% of women reporting at least one UTI in their lifetime [3]. More than 10% of female respondents in both the US and the UK indicated that they had experienced symptoms of UTI in the previous year [3, 10]. Many of those women experienced multiple recurrent UTIs[1] during that time [3]. Infections in men, on the other hand, are rare and usually linked to other risk factors — e.g., history of prostatitis or catheter use [1]. The risk of experiencing UTI increases in both sexes above the age of 65 years [12]. Other factors predisposing to (complicated) UTI include pregnancy, comorbidities like diabetes, and underlying urological abnormalities [2].

Due to their frequency, UTIs not only represent a major source of morbidity but also incur substantial healthcare costs. Like estimates of incidence, estimated healthcare costs attributable to UTI are imprecise and vary from country to country but UTIs were estimated to account for 6 billion dollars in direct healthcare costs worldwide in 1994 [1].

1.2 Pathophysiology

Most UTIs are infections of the lower urinary tract — i.e., the urethra and bladder — and are managed in primary care. Lower UTI may progress to more serious infections by ascending to the kidneys (pyelonephritis) or causing a life-threatening systemic inflammatory response (urosepsis). The latter almost always requires

[1] The European Association of Urology defines recurrent UTIs as the occurrence of ≥ 2 UTIs within 6 months or ≥ 3 UTIs within 12 months [11].

urgent hospitalisation [13]. Most UTIs are caused by *Escherichia coli* bacteria, followed by *Klebsiella pneumoniae* and *Proteus Mirabilis* [14]. The distribution of causative pathogens is relatively stable across countries, although antibiotic sensitivities may vary substantially [15]. Uncomplicated lower UTI in the UK are in first instance treated with the oral antibiotics nitrofurantoin and trimethoprim [13]. Pyelonephritis and urosepsis usually warrant more broad spectrum antibiotics and may even require intravenous antibiotics (see Section 1.4.1 for a more detailed discussion of recommended treatments) [16], although exact guidelines commonly vary between hospital trusts.

Remark (Uncomplicated community-onset UTI).

UTIs may be classified as community-onset if the signs and symptoms of infection first appear while the patient is at home — i.e., not currently admitted to a healthcare facility such as a hospital or care home. A closely related concept are community-acquired and hospital-acquired (nosocomial) UTIs [2], although these terms make a stronger assumption about the setting in which the infection was acquired (rather than the setting in which symptoms first showed). While community-onset UTIs will therefore often also be community-acquired, they may be hospital-acquired if the patient was recently discharged from hospital or may be more generally considered as healthcare associated if they are for example the likely consequence of a recent surgery. UTIs may be considered catheter-associated if they occur in the presence of an indwelling catheter.

UTIs may further be categorised as complicated. Complicated UTIs describe infection in the presence of factors that increase the risk of bad outcomes or infectious complication [2, 17]. Complicating factors may include pregnancy, male gender, anatomical abnormalities of the urinary tract, suppression of the immune system, or recent discharge from hospital. Complicated UTIs often warrant their own treatment guidelines in order to mitigate the increased risks of negative outcomes [17].

1.3 Diagnosis

The gold standard definition of bacterial UTI is the presence of two or more urinary symptoms (e.g., increased urinary frequency or dysuria; Table 1.1) as well as microbiological confirmation of bacteria in the urine (called bacteriuria) [17]. In practice, obtaining both can be difficult due to time pressures imposed by limited resources [18] or the need to start treatment quickly [19].

Microbiological urine culture results, however, may take several days to return [20]. No reliable rapid diagnostic test is currently available in routine practice. In order to minimise patients' need for re-consultation or risk of hospitalisation, General Practitioners (GPs) therefore often rely on faster but less reliable evidence from near-patient (dipstick) tests, allowing them to make decision quicker at the cost of increased uncertainty [21, 22]. The difficulties of obtaining microbiological confirmation in primary care are reflected in national guidelines, which do not recommend routine urine cultures for lower UTI in the absence of complicating factors such as male sex or pregnancy [13, 23]. Routine urine culture is more commonly recommended in secondary care [24] but the delay until culture results are available remains. In the meantime, the need to initiate treatment early is exacerbated in hospital due to the usually sicker patient population. Hospital doctors therefore also frequently rely on dipstick tests to guide initial prescribing decisions [25]. On top of this, the increased severity of illness in patients presenting to hospital may add further uncertainty to the diagnostic process by preventing clinicians from asking for a reliable history of symptoms from the patient [26].

Much of the uncertainty described above can be traced back to an inability of current diagnostic tests and criteria to provide a timely, accurate, and cost-effective estimate of a patient's likelihood of UTI. The following sections therefore review the most common diagnostic tools used to diagnose UTI, and discuss their strengths and limitations in quickly establishing a reliable diagnosis.

1.3.1 Clinical symptoms

A central step in diagnosing the presence of UTI is the examination of the patient and the patient's medical history. Clinical presentation of UTI often

Table 1.1: Commonly described symptoms of UTI.

Category	Symptom
Urinary symptoms	dysuria, haematuria, pyuria, urinary urgency, urinary frequency, cloudy urine, malodorous urine
Pain	abdominal pain, suprapubic pain, flank pain, pelvic pain
Unspecific symptoms	fever, confusion, altered mental status, functional decline

UTI, urinary tract infection.

Sources: Little *et al.* (2006), Schmiemann *et al.* (2010).

includes urinary symptoms such as painful or difficult urination (dysuria), waking up repeatedly at night to urinate (nocturia), frequent or urgent need to urinate during the day (frequency), blood in the urine (haematuria), cloudy urine, and offensive-smelling (malodorous) urine (Table 1.1) [17, 23, 27]. Abdominal pain or flank pain may also be signs of UTI, with the latter suggesting that UTI may have ascended from the bladder into the kidneys [17, 28].

The presence of any single one of the above symptoms isn't necessarily specific to UTI. Symptoms such as abdominal pain can overlap with a range of causes including other infections (e.g., gastroenteritis or appendicitis) as well as non-infectious disorders such as urinary calculi [29]. More than one symptom usually needs to be present to indicate a UTI. In a population of primary care patients, Little *et al.* (2006) [27] reported that a simple clinical decision rule based entirely on the presence of two or more of dysuria, nocturia, malodorous urine, or cloudy urine can achieve a sensitivity of 65% and a specificity of 69% in predicting presence of bacteriuria. Multiple urinary symptoms alone therefore already provide some evidence of the presence or absence of UTI but large uncertainty remains.

In elderly patients in particular, non-specific signs and symptoms such as confusion, altered mental status, or functional decline are widely considered indicative of UTI [28], despite a lack of definitive evidence for association [32]. The sole presence of unspecific symptoms is problematic due to a wide range of alternative causes and a high prevalence of asymptomatic bacteriuria in certain patient populations [32]. The prevalence of asymptomatic bacteriuria in the general population increases with age. While less than 5% of women below the age of

Remark (Asymptomatic bacteriuria). Asymptomatic bacteriuria describes the growth of uropathogens during culture of mid-stream urine in the absence of urinary symptoms [30]. It is not yet fully understood why some patients might experience symptoms while others do not, but antibiotic treatment of asymptomatic bacteriuria is not usually recommended outside of pregnancy [31]. Symptoms of UTI are therefore indispensable for the diagnosis and management of UTIs.

50 years are routinely found to have asymptomatic bacteriuria, this proportion increases to ∼20% among women aged 65 years or more [33]. Prevalence of asymptomatic bacteriuria is generally lower in men but nevertheless affects ∼10% of men above the age of 65 years [34, 35]. Finally, in patients aged 80 years or more, up to 50% of women and 20% of men have been found to routinely exhibit bacteriuria, although estimates vary by study [35]. Importantly, diagnostic tests for UTI — including microbiological culture and urine dipstick tests discussed next — are unable to distinguish between asymptomatic and symptomatic bacteriuria, and do not necessarily indicate UTI in the absence of symptoms [36].

1.3.2 Microbiological culture of urine

The gold standard for diagnosing bacteriuria includes microbiological culture of mid-stream urine [17]. Once incubated, urine cultures usually take between 24–48 hours to produce a result [20]. Classifications of culture results may vary between laboratories. Most laboratories and clinical guidelines require $\geq 10^5$ colony-forming units per millilitre (cfu/mL) to rule in a diagnosis of UTI [17]. However, others have suggested increasing thresholds to 10^6 cfu/mL [37], lowering thresholds to 10^2 cfu/mL in the presence of symptoms or catheter use [11, 38], or do not specify definitive thresholds at all [39]. Laboratory classifications and guidelines may also differ for men and women, requiring for example two consecutive urine samples to confirm asymptomatic bacteriuria in female patients without urinary symptoms but only one culture-positive sample in male patients

Remark (Mid-stream urine). During mid-stream urine sampling the first part of the urine is excluded and sampling only starts after some urine has been passed into the toilet. Mid-stream urine samples are meant to avoid sample contamination by minimizing the risk of bacteria present on the skin, genitals, or lower urethra entering the urine during sample collection. While the risk of contamination is indeed reduced in mid-stream urine, a substantial proportion of samples still show signs of contamination in clinical practice [41].

[11]. Prior antibiotic treatment might influence a samples propensity to grow bacteria, either requiring longer incubation of the urine sample or altogether preventing it from growing [20]. Even variations in healthcare processes — such as a delay of a few hours in processing the urine — may influence the probability of bacteria growing [40].

Despite the common use of mid-stream urine, sample contamination is common, further complicating classification and interpretation of urine culture results. According to a widely accepted definition by the College of American Pathologists, a urine sample is considered contaminated if two or more pathogens are isolated at $\geq 10^4$ cfu/mL [42], commonly referred to as mixed growth. The study authors note that it is difficult to define a single decision rule to identify contaminated samples: "There are circumstances in which two organisms in quantities $\geq 10^4$ cfu/mL both contribute to a urinary tract infection, just as there are instances in which a single organism found in quantities $\geq 10^5$ cfu/mL is a contaminant" [42]. The UK Standards for Microbiology Investigations (SMI) therefore define contamination based on a combination of cfu/mL, number of isolates, specimen type, urinary white blood cell (WBC) counts, presence of risk factors like an indwelling catheter, and urinary symptoms [43]. Contaminated samples add uncertainty because they mask the presence or absence of relevant pathogens in the urinary tract. If a mixed culture was obtained despite a clear indication, UK guidelines therefore recommend repeating the culture [43]. However, many patients will have been started on antibiotics and/or sent home at

this point.

In addition to determining the presence or absence of bacteria in the urine sample, microbiological cultures are the only commonly used diagnostic test that also determines the antimicrobial sensitivities of the isolated pathogens and allows targeted treatment based on the pathogen's resistance pattern, although results will usually only be available after empirical antibiotics have started.

1.3.3 Urinalysis

Urinalysis is the "analysis of urine, using physical, chemical and microscopical tests, to determine the proportions of its normal constituents and to detect alcohol, drugs, sugar, or other abnormal constituents" [45]. Urinalysis is recommended for the confirmation of a range of urinary disorders, including infectious and non-infectious urinary tract and renal disorders [46].

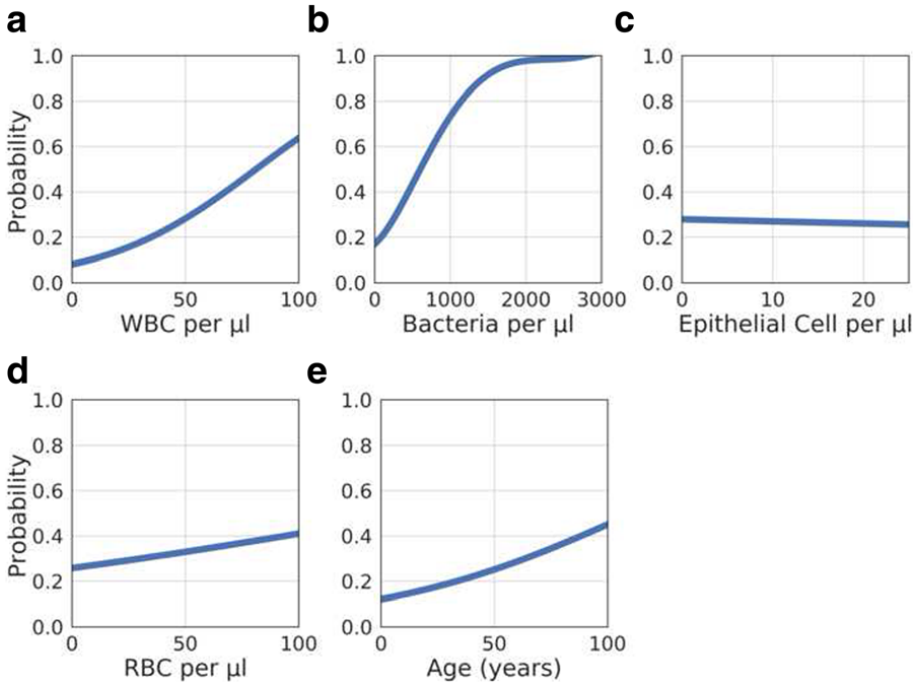

Figure 1.1: Probability of a positive urine culture result in relation to **(a)** urinary WBC counts, **(b)** bacteria counts, **(c)** urinary epithelial cell counts, **(d)** urinary RBC counts, and **(e)** age.

RBC, Red blood cell; WBC, white blood cell.

Figure source: Burton *et al.* (2019) [44] reused under Creative Commons Attribution 4.0 International License; image caption was adapted from the original image caption.

1.3.3.1 Visual examination and flow cytometry

A central part of urinalysis is the visual inspection of the urine sample, optionally with the help of a microscope. Visual inspection of the urine can determine its colour and cloudiness. Microscopy further allows for the detection and approximate quantification of WBCs, red blood cells (RBCs), epithelial cells, and bacteria [44, 47]. Presence of these particles in the urine has been found to be associated with bacterial growth during urine culture (Figure 1.1). Other particles detectable via urine microscopy include parasites, yeast cells, casts[2], lipids, and crystals [46].

More recently, automated optical particle analysis via flow cytometry has been used to estimate particle quantities in the laboratory [46]. During flow cytometry, the sample is guided past one or more light sources, and the light scatter created when the light bounces of particles is measured and analysed [48]. Automated urinalysis systems have been found to perform as well or better than their manual counterparts [47]. Due to their fast turn-around time, automated particle analyses are increasingly used as screening tools in hospitals. Decision rules based on low urinary WBC and bacteria counts may be used to filter out samples that are highly unlikely to grow a priori. This process is sometimes termed reflex culture [49].

Urinalysis based on visual examination and flow cytometry is much quicker than urine culture, but cannot definitively confirm or foreclose the presence of colony-forming units. While a complete absence of bacteria determined via urinalysis is a good indicator of the absence of bacteriuria, the presence of bacteria or other particles does not necessarily imply the presence of colony-forming units or actual culture growth [44, 50, 51]. Despite these limitations, urinalysis can be a highly effective screening tool that can reduce the number unnecessarily cultured urine samples by an estimated 40-50% [44] and — if appropriately fed back to clinicians — might reduce unnecessary initiation of antibiotic therapy for patients who are highly unlikely to have a positive culture.

[2] Cylindrical structures produced in the upper renal tract.

Remark (Viscosity and sampled quantity). Urine samples may not always be sufficiently analysed visually due to a high viscosity, preventing the samples from being passed through the flow cytometry machine [24]. Urine samples may further be too small to allow for visual analysis. Other diagnostic tests like urine dipsticks or urine cultures are less affected by these issues.

1.3.3.2 Dipstick test

Some substances present in the urine cannot be detected via microscopy but may be detected using chemical reactions. In practice, these chemical reactions tend to be measured using a urine dipstick test. A dipstick test consists of dipping a thin paper strip with several allotted reagents into a urine sample. If the substance corresponding to each reagent is present in sufficient quantities, that reagent will change colour. Dipstick tests offer a quick and inexpensive alternative to both microbiological culture of a urine sample and urine flow cytometry. Due to their speed and the minimal requirements for equipment and labour, dipstick tests are ubiquitously used as a first screening tool in both primary and secondary care [27].

Dipsticks commonly test for the presence of nitrites, leukocyte esterase, blood, proteins, glucose, ketones, and bilirubin [52], among other urinary parameters like acidity (pH) and concentration. Particularly nitrites and leukocyte esterase are prominent indicators of UTI. Nitrites often signal a presence of gram negative organisms in the urine — most notably E. *coli* — which produce nitrite as a byproduct of anaerobic respiration [53]. The presence of nitrites has a high estimated specificity (>90%) for subsequent growth of bacteria during urine culture but low sensitivity (28%; Table 1.2) [27], although exact estimates vary by study [52]. Leukocyte esterase is an enzyme produced by WBCs and indirectly indicates the presence of WBCs in the urine. Since WBCs are cells produced by the immune system to fight infection, a presence of WBCs in the urine provides some evidence for UTI, although it might also signal other disorders that cause non-infective inflammation of the urinary tract [52]. Contrary to nitrites, the presence of leukocyte esterase has a relatively low estimated specificity (53%) but high sensitivity (85%;

Table 1.2: Performance of dipstick results in predicting urine culture growth in a single primary care study.

	Bacterial growth during urine culture		
	Yes	No	p-value
Total (col-%)	254 (100)	154 (100)	
Dipstick results (col-%)			
Nitrites	72 (28)	7 (9)	<0.001
Leukocyte esterase	217 (85)	72 (47)	<0.001
Blood	186 (73)	71 (46)	<0.001
Proteins	119 (47)	47 (31)	0.643

Note: Column percentages do not add up to 100% since a urine sample might contain more than one substance.

UTI, urinary tract infection.

Source: Little *et al.* (2006).

Table 1.2) [27]. A third commonly used marker of UTI is blood in the urine, which has a specificity and sensitivity comparable to that of leukocyte esterase [27]. Combined into a clinical decision rule, these three markers have previously been found to achieve a specificity of 70% and a sensitivity of 77% when predicting urine culture growth in non-pregnant adult women who present with suspected UTI in primary care and do not have a likely alternative diagnosis — e.g., vaginal symptoms [27].

While helpful as an initial screening tool, dipstick results — just like results obtained from urine microscopy or flow cytometry — cannot be used to reliably diagnose UTI. In the above-cited primary care study, strictly applying the decision rule would have misclassified a quarter of the study population, missing 23% of true bacteriuria cases and wrongly labelling 30% of patients without bacteriuria as suffering from UTI [27]. Despite these limitations, urine dipstick tests are widely used in primary and secondary care — often as the primary diagnostic tool — due to a lack of alternatives [21]. As a result, previous studies have repeatedly reported an over- or misinterpretation of dipstick results [21, 22] and consequently high levels of antibiotic treatment in the absence of (symptomatic) bacteriuria. This is increasingly acknowledged in national guidelines [23] and quality schemes [54], which particularly discourage dipstick-based diagnoses in the absence of symptoms

and in the elderly (due to high prevalence of asymptomatic bacteriuria in these patients).

1.3.4 Polymerase chain reaction assays

Polymerase chain reaction (PCR) assays are a relatively new addition to the diagnostic arsenal for UTI. They are the most commonly used nucleic acid detection method and are used to detect bacterial DNA in the urine. Due to their short time-to-result, previous studies have investigated the use of PCR as a rapid diagnostic test for *E. coli* bacteriuria in both children [55] and adults [56]. While PCR tests are generally viewed as fast and reliable, the high cost per test currently prohibits their use as a routine screening test [57]. As a result, PCR tests for urinary pathogens are not in routine use in the NHS at the time of writing. Notably, PCR results are solely based on the presence of bacterial DNA and are thus unable to distinguish between viable bacteria and dead cells [58]. It is possible that a PCR test indicates the presence of bacteria that have already been eliminated by the patient's immune system or prior antibiotic treatment, and which would not grow in a standard microbiological culture. Like culture and unlike urinalysis, however, PCR can determine species and guide the choice of antibiotic treatment [56].

Overall, the lack of affordable rapid and reliable diagnostic tests for bacterial UTI undermines clinicians' ability to distinguish patients who will benefit from antibiotic treatment from those who will not, and risks delaying effective therapy.

1.4 Disease management

Diagnostic uncertainty directly impacts the management of UTI, forcing clinicians to make prescribing decisions in the absence of a definitive diagnosis. Even if we were to find good predictors of bacterial UTI, however, further uncertainty remains as to which patient groups require immediate antibiotic treatment, which patients may perhaps get better on their own, which drug to use, and how long to treat for.

Remark (Empirical antibiotic prescribing). In view of the delay in obtaining microbiological culture of urine, antibiotics are usually prescribed empirically for suspected UTI, based on knowledge of the most likely uropathogens in a specific patient based on factors such as their age, gender, and underlying health conditions [13, 16]. However, an unusual pathogen or unknown antimicrobial resistance may render the initial antibiotic ineffective [59]. In order to minimize the risk of treatment failure, prescribing guidelines are therefore often optimised towards local prevalence of antimicrobial resistance, and hospitals usually provide hospital-specific antibiotic prescribing guidelines.

1.4.1 Choice and duration of antibiotic therapy

National guidelines suggest short three-day courses of empirical oral narrow-spectrum antibiotics — nitrofurantoin or trimethoprim — as first line treatment for lower UTI [13]. Narrow-spectrum antibiotics are antibiotics that only act against a limited number of species or bacteria, allowing them to directly target the suspected pathogen without introducing unnecessary selection pressure on the host microbiome [60]. For example, the active spectrum of nitrofurantoin is limited to most uropathogens, with few known resistances [61]. Broad-spectrum antibiotics, on the other hand, act against a range of causative agents, making them less likely to fail during empirical treatment at the cost of increased cross-resistance in other, non-targeted bacteria [60]. Broad-spectrum antibiotics therefore provide clinicians with reassurances even if they misdiagnose the pathogen or source of infection [62]. Broad-spectrum antibiotics are not generally indicated for lower UTI, and should only be used if initial treatment failed, if indicated by susceptibility profiles, or if symptoms indicate more severe infection [13]. Oral broad-spectrum antibiotics (cefalexin, co-amoxiclav, and ciprofloxacin) for seven to 14 days are recommended as first line treatment for pyelonephritis [16]. Intravenous antibiotics are indicated if oral antibiotics can't be administered — e.g., due to vomiting — or if very severe

UTI or urosepsis are suspected [16]. Specific guidelines may vary by region or hospital based on local resistance profiles.

Remark (Delayed antibiotic prescribing). Instead of making a binary decision to immediately prescribe antibiotics or to not prescribe antibiotics at all, clinicians might instead opt to delay antibiotic prescribing [63]. When delaying antibiotics, patients are handed a prescription but are asked to only take the medication if symptoms do not abate or grow worse. Instead of a direct prescription, patients might alternatively be provided with a postdated prescription or an option to simply collect a prescription later from the reception [64]. In any case, delayed antibiotic prescribing strategies aim to reduce the unnecessary initiation of antibiotic therapy while simultaneously making it easy to get an antibiotic quickly if symptoms persist.

In primary care, retrospective studies of English general practices indicated reasonable — but not perfect — adherence to guidelines, with 70–80% of patients treated according to the recommended first line treatment [6, 65]. However, they also reported considerable use of broad-spectrum antibiotics. It is difficult to ascertain what proportion of those patients had symptoms that were indeed indicative of pyelonephritis [16], or whether doctors might have discounted the long-term consequences posed by broad-spectrum antibiotics — i.e., antibiotic resistance and increased risk of future treatment failure — and prioritised the avoidance of infection-related adverse consequences [62]. Length of treatment was also frequently found to exceed the recommended treatment duration of three days, with $>$50% of courses being prescribed for five or seven days [66].

Data on antibiotic prescribing in relation to clinical indication are lacking in secondary care in the UK, due to an ongoing lack of electronic prescribing in this setting [67]. This makes it difficult to investigate the congruence of prescribing for lower UTI in secondary care against local guidelines. In a single-centre study of antibiotic prescribing in patients presenting to the emergency department (ED)

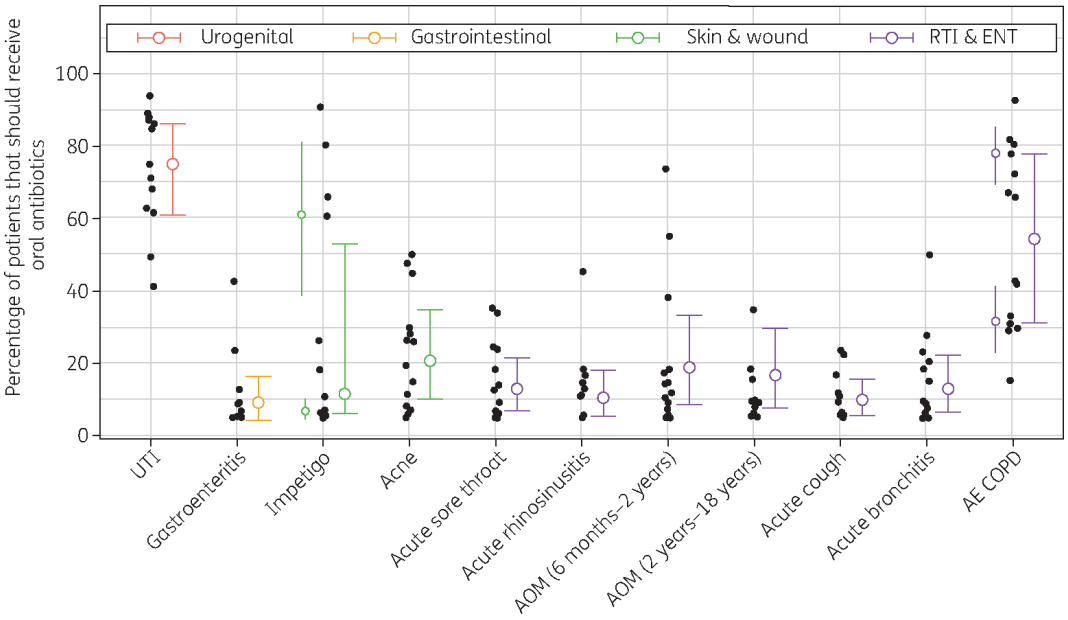

Figure 1.2: Ideal antibiotic prescribing proportions (median and interquartile range) in primary care for key infections, as elicited via expert consensus in a previous primary care study.

Figure source: Smith *et al.* (2018) [68]. Copyright © 2018, Oxford University Press; reused under Crown Copyright; image caption was adapted from the original image caption.

at Queen Elizabeth Hospital Birmingham, we previously found that 33.3% of patients admitted for lower UTI received piperacillin / tazobactam, although local guidelines recommend nitrofurantoin as first line treatment [24]. This relationship was also seen in patients with suspected pyelonephritis or urosepsis, where 60–70% of patients received empirical treatment with piperacillin / tazobactam. The recommended first line treatment co-amoxiclav was prescribed in only ∼5% of cases, indicating substantial deviation from guideline recommendations [24].

1.4.2 Need for immediate antibiotic treatment

While oral antibiotics are the recommended first-line treatment for UTI [13], the reliance on antibiotics as the exclusive treatment for (lower) UTI is increasingly challenged. A recent series of interviews with infectious disease experts revealed a notable disagreement on recommended treatment strategies for a range of common infectious conditions in primary care (Figure 1.2) [68]. When asked to estimate the proportion of uncomplicated lower UTI episodes that truly require antibiotic treatment, opinions ranged from 40% to 95% of cases. Indeed, increasing evidence

from randomized controlled trials in young and otherwise healthy women suggests that uncomplicated lower UTIs might resolve spontaneously in some of these patients[3] [69, 70, 71].

While many patients in this patient group clearly require antibiotic treatment for UTI, these results suggest that it might be safe to delay or withhold antibiotic treatment in some patients with suspected UTI. A prospective cohort study among Danish women with uncomplicated UTI found that a third of patients would consider delaying antibiotic treatment when asked by their primary care physician [63]. After one week, 55% of the women who were willing to delay antibiotics reported not having used antibiotics to treat their UTI, and none of the patients in the study developed serious complications. The risk of minor complications was comparable between patients that were willing to delay antibiotic treatment and those that were treated according to standard care — i.e., who received immediate antibiotics [63].

However, it remains currently unclear for which patient groups it is safe to delay antibiotics. All studies discussed in this section suffered from relatively small sample sizes. Serious complications were generally rare ($\leq 2\%$) and most studies were therefore under-powered to ascertain differences in risks between treatment groups. The risk of developing a complication most likely differs based on patient characteristics such as age and previous medical history [72]. In order to ascertain the safety of different prescribing strategies definitively, rigorous studies based on large numbers of high quality data are required. Unless they use prospective randomization, these studies will need to carefully consider the risk for confounding by indication [72], since the characteristics of patients who are prescribed antibiotics immediately are likely to differ from those who receive a delayed prescription.

[3] These studies and their limitations are discussed in detail in Chapter 4.

1.5 Clinical impact of reducing uncertainty in UTI

The inherent uncertainty in diagnosing and managing suspected UTI outlined above has important implications for clinical care and patient outcomes. Reducing this uncertainty has the potential to improve patient management by reducing the risk of misdiagnosis, mitigating the emergence of antimicrobial resistance, enabling conservative, prompt, and effective treatment of (drug-resistant) infection, and by reducing side effects of antibiotics. Improved management of UTI could decrease the number and length of hospital stays, reduce associated healthcare costs, and decrease the number of complications of UTI.

Reducing misdiagnosis The difficulty to quickly and accurately diagnose UTI means that many patients are over-diagnosed for UTI [24, 73]. Patients have repeatedly been found to be unnecessarily treated with antibiotics for suspected UTI that later could not be confirmed by microbiological culture, meaning that the antibiotics most likely did not benefit the patient. An incorrect assumption of UTI may further delay treatment for other diseases with similar presentation — like sexually transmitted diseases [73] — as well as other infectious and non-infectious diagnoses [24]. Reducing the uncertainty in diagnosing UTI can aid clinicians in identifying the underlying cause of clinical symptoms and initiated appropriate treatment early.

Mitigating antimicrobial resistance The overuse of antibiotics is an established driver of antimicrobial resistance [74, 75]. Nine out of ten episodes of suspected UTI are treated with antibiotics, making them a key driver of antibiotic prescribing in both primary and secondary care [24, 76]. A substantial proportion of this prescribing is thought to be unnecessary [24, 68]. Improved certainty of UTI diagnosis could give doctors confidence to withhold antibiotics without fear of an increase in adverse outcomes. Earlier confirmation of susceptibility profiles could further optimise the choice of empirical antibiotics [59, 77], allowing clinicians to prescribe the right antibiotic to the right patient and limiting the risk of treatment failure in the presence of uncommon pathogens or drug resistance.

Reducing antibiotic side effects Aside from an increased risk of antimicrobial resistance, the frequent use of antibiotics has other unintended consequences for patients. Common side effects of antibiotics include nausea, vomiting, diarrhoea, bloating, indigestion, abdominal pain, or loss of appetite [78]. Allergies to antibiotics are common with 5-15% of patients reporting a penicillin allergy [79]. While allergies are thought to be over-reported and tend to be mild if indeed present, antibiotic use can lead to severe adverse outcomes like anaphylaxis or Stevens-Johnson syndrome[4] [79]. Furthermore, frequent antibiotic use puts patients at risk of opportunistic infections like *Clostridium difficile*, particularly in the elderly [80, 81]. Delaying or withholding antibiotics in those that are highly unlikely to have bacterial UTI avoids putting patients at risk of experiencing these side effects.

Shorter hospital stays and reduced healthcare costs Appropriate antibiotic use has further been linked to a moderate reduction in the length of hospital stays [82]. If used frequently, on the other hand, antibiotics have been associated with an increased risk of future hospitalisations [83]. While this evidence is from observational studies only — and thus at high risk of confounding by indication — it is plausible that long-term use of antibiotics increases the risk of more severe, drug-resistant infection that require hospitalisation. Diagnostic uncertainty has been identified as a key reason for divergence from appropriate prescribing behaviours recommended in national and local guidelines [62]. Clinicians might feel more confident in following recommendations if diagnostic uncertainty and risk of complications are reduced. A reduction in the number and length of hospital stays accompanying increased guideline adherence may significantly reduce healthcare costs associated with UTI [2, 4].

1.6 Aims of this thesis

In this thesis, I investigate the use of routinely collected medical data in form of electronic health records (EHRs) to improve the diagnosis and management of

[4] A rare skin disorder characterised by a severe blistering rash often requiring treatment in an intensive care or burn unit [79].

UTI in England. Using case studies from primary care and the ED, I examine the availability of information on urinary symptoms, diagnosis of UTI, urine dipstick tests, and microbiological culture results recorded in EHR data from each setting. I then explore whether the available information allows for the identification of episodes of community-onset UTI, which present the majority of UTI cases in these settings. Finally, I assess whether these and other information routinely captured in EHR databases allow to obtain valid findings on the diagnosis and management of community-onset UTI from EHR data and to account for confounding factors that may bias these results.

This thesis is structured as follows. In Chapter 2, I perform a scoping review of studies that used EHR data to assess the diagnosis and management of UTIs. In Chapter 3, I then describe several major English EHR databases and the information captured within them. In Chapter 4, I present a retrospective observational database study assessing the association between antibiotic prescribing for community-onset lower UTI in English primary care and subsequent risk of complications. In Chapter 5, I describe the creation of a risk prediction model for the presence of bacterial growth in urine samples collected in the ED at Queen Elizabeth Hospital Birmingham. In Chapter 6, I expand on these results and evaluate variations in the estimated performance of this model when looking at clinically relevant patient subgroups or when changing key definitions. In Chapter 7, I summarise my findings, discuss the strengths and limitations of this thesis, comment on the translation of EHR studies into clinical practice, and provide recommendations for future EHR studies and data collection.

Chapter summary

- Microbiological culture of urine samples may take up to 48 hours, introducing a bottleneck for evidence-based diagnosis and management of UTIs.
- Rapid diagnostic tests like urine dipsticks are commonly used to

guide empirical prescribing decisions until culture results are available. These tests have relatively poor predictive performance.

- As a result, previous studies have repeatedly reported over-diagnosis and over-treatment of UTIs in clinical practice.
- Reducing the uncertainty inherent in the diagnosis and management of UTIs may reduce unnecessary prescribing for UTI, and avoid side effects of antibiotic treatment and emergence of antimicrobial resistance.

Chapter 2

Use of EHR data to guide diagnosis and management of suspected community-onset UTI in adults: a scoping review

Abstract

Introduction: Chapter 1 described common sources of uncertainty that complicate the diagnosis and management of urinary tract infections (UTIs). Analysis of patient data routinely collected during each healthcare visit in the form of electronic health records (EHRs) may provide an opportunity to support the diagnosis of UTI and aid clinicians in choosing the optimal treatment in light of this uncertainty. In this chapter, I undertake a scoping review to identify previous research studies that have used EHR data to guide the diagnosis and management of suspected community-onset UTI in primary or secondary care.

Background: Over the course of the last two decades, routine data have increasingly been used to study the onset and course of disease. With the increasing breadth and depth of EHR data sources, they may provide an opportunity to aid doctors directly in confirming suspicion of UTI and choosing adequate treatments, either by informing the development of comprehensive clinical guidelines or in form of clinical decision support systems.

Methods: I performed a scoping review of peer-reviewed studies describing the use of EHR data to estimate the risk of — or risk factors for — microbiologically confirmed bacterial UTI and infectious complications in adult patients consulting with suspected community-onset UTI in primary or secondary care. I identified

retrospective observational studies that used data from routinely collected EHRs through a systematic search of two bibliographic databases (Embase and MEDLINE), including all studies published after 1974 in English language. I summarised the characteristics of included studies and provided a narrative synthesis. Risks of bias (ROBs) and applicability of studies to the review question were appraised using an adapted Newcastle - Ottawa Quality Assessment Scale and the Prediction model Risk of Bias Assessment Tool (PROBAST).

Results: I screened 4,334 records identified during the database search. After adding one study identified via manual reference screening, eight studies were included in the review (UK: 5 studies; US: 3 studies). Included studies primarily investigated the use of EHR data in primary (5 studies) or primary/secondary care (2 studies), with only one study being exclusively performed in secondary care. Almost all studies (6/8) assessed the association between antibiotic treatment choices and subsequent complications of UTI. The remaining two studies attempted to develop risk prediction models to predict patients' risk of developing complications after consulting in primary care or the probability of microbiological growth in urine samples for patients consulting in the emergency department. Studies were generally judged at high ROB, most prominently due to non-random treatment assignment and difficulties in identifying the intended target population from EHR data.

Discussion: EHR data provide a valuable resource for research on suspected community-onset UTI but few studies have been conducted to date, particularly in secondary care. Increasingly detailed data collection may provide further opportunities to use EHR data to investigate the diagnosis and management of community-onset UTI, and help to overcome the ROBs identified in this review. Particularly data on patients' health status at consultation — paired with careful definition of study cohorts and appropriate statistical methodology — may be able to account for common biases and ensure more reliable results.

2.1 Introduction

In Chapter 1, I described how uncertainty in the diagnosis of community-onset urinary tract infections (UTIs) influences clinicians' management of this infection in primary and secondary care. I discussed the wide range of signs and symptoms potentially attributable to UTI, how these overlap with other infectious and non-infectious conditions, the potential consequences of delaying antibiotic treatment, and how a lack of reliable rapid diagnostic tests for UTI forces doctors to rely heavily on their clinical experience when treating patients with suspicion of UTI.

A large number of research studies have investigated strategies to improve the diagnosis and management of UTI. Clinical trials have compared different choices and durations of antibiotic treatment for community-onset UTI [84, 85] and a number of trials have investigated the safety and efficacy of withholding antibiotics in low risk women with suspected UTI [69, 86]. However, these studies are expensive, they often exclude important population sub-groups, and treatment effects identified through trials do not always translate effectively into clinical practice [87]. Alongside clinical trials, observational studies have evaluated the use of novel diagnostic tests [88] or clinical scoring systems [27, 89] to aid the diagnosis and management of UTI, yet few of those are used in routine practice. Novel diagnostic tests like PCR are expensive [57] and need to show substantial clinical benefit in order to justify those excess costs. Simple clinical scoring systems on the other hand have shown limited performance in identifying UTIs [90], still require clinicians to consider the risk of false negatives carefully [90], and may quickly become complex and hard to apply without automated support.

Analysis of detailed electronic health record (EHR) data collected as part of routine care may provide a novel opportunity to aid clinicians in reducing uncertainty in the diagnosis and management of community-onset UTIs. The large quantities of real-world data captured in structured EHRs may allow for better (and potentially automated) risk stratification of patients for whom community-onset UTI is suspected. For example, EHR data can be used to develop risk prediction

models which stratify patients by risk of complications and thus guide antibiotic treatment decisions.

In this chapter, I use a scoping review approach to examine how previous studies have used EHR data to support the diagnosis and management of patients presenting with suspected community-onset UTI in primary or secondary care. The aim of this review is to provide context for the work that is presented in subsequent chapters of my thesis. In contrast to a systematic review, this chapter does not specifically assess the impact of a treatment or clinical practice, but instead attempts to synthesis the basic characteristics of studies that have used EHR to guide UTI care and discuss similarities or differences between them [91].

2.2 Background

Clinicians record data during each patient consultation as part of standard clinical practice. With the rise of EHR systems and electronic prescribing in primary and secondary care, this information has become increasingly accessible for research and offers a valuable resource for medical research (I give an overview over the different EHR systems available in English primary and secondary care in Chapter 3) [92]. Advantages of using routinely collected EHR data include the often large numbers of patients covered within them, relatively fast and cheap access to the data, and their depiction of real-world clinical practice.

An early use of EHR data in England has been the surveillance of infectious diseases. As early as 1999, data from EHR sources were used to estimate the incidence of common infections and patterns of antibiotic use in English primary care [93, 94]. Over the past two decades, EHRs have thus been used to investigate the epidemiology and management of UTI. Examples include their use to investigate the effects of patient characteristics — e.g., diabetes mellitus — on the incidence of UTI [95], to assess concordance of antibiotic prescribing for UTI with local or national guidelines [9, 24, 65, 66, 96], and to establish local or national patterns of (urological) pathogens and antibiotic susceptibilities [24, 97]. By depicting routine medical care, EHRs and EHR studies therefore allow

policy makers and doctors to adapt guidelines to real-world clinical practice — e.g., through adjusting recommended first and second line treatments to observed resistance patterns [97].

As the digital maturity of healthcare providers grows, and the breadth and depth of data available for each individual healthcare contact increases — e.g., through increased linkage of data sources [98, 99] — tools and evidence derived from EHR sources may be used to directly aid clinicians in diagnosing and managing suspected UTI at the point of care. Information on clinical presentation and outcomes from thousands of patients with suspected UTI may be used to enable a targeted, "personalised" approach to the management of UTIs. Unlike conventional clinical guidelines, which may be limited in their achievable complexity, tools based on EHR data may simultaneously account for a large number of patient factors and communicate their combined impact on an individual patient [100]. Ready access to large cohorts of patients consulting for UTI further allows researchers to assess even small effect sizes or investigate rare outcomes linked to empirical treatment choices. Evidence derived this way may again inform the development or adaption of clinical guidelines for UTI, or may provide the basis for clinical decision support systems — either by simplifying them into easy to handle clinical scores (if possible) or by directly integrating them into the electronic patient management systems of healthcare providers.

The value of these models for clinical care, however, heavily depends on the quality of data recorded in EHRs [92]. EHRs are primarily designed for clinical care and are seldom collected with research in mind. Analysts are limited to those data items that were judged relevant for delivery of care, and information not routinely measured or recorded in standard care tends to be missing. As a result, careful analysis and evaluation of EHR data are paramount to ensure that reliable findings are derived from them. The following pages present a scoping review of studies that used routinely collected EHR data to guide the diagnosis and management of UTI and describe the approaches and findings of those studies. Being mindful of the limitations of EHR research, this chapter also includes a detailed analysis of the

various risks of bias encountered in those studies, and how they were or weren't accounted for in each study.

2.3 Aims and Objectives

To identify, summarise, and appraise published studies that used routinely collected EHR data to estimate the risk of UTI-related outcomes in order to aid the diagnosis and management of community-onset UTI at the point of care.

Objectives:

2.1 To identify peer-reviewed studies describing the use of EHR data to estimate the risk of — or risk factors for — microbiologically confirmed bacterial UTI or infectious complications of UTI in adult patients consulting primary or secondary care with suspected community-onset UTI.

2.2 To assess the risk of bias (ROB) in identified studies, and identify potential concerns about their direct applicability to the target population of patients consulting with suspected community-onset UTI.

No review protocol was prospectively registered for this review.

2.4 Methods

2.4.1 Eligibility criteria

I followed the Population – Concept – Context framework recommended by the Joanna Briggs Institute (JBI) to define the scope of this review [101].

2.4.1.1 Population

I included all studies relating to adult patients consulting for suspected community-onset UTI in primary or secondary care. Studies exclusively investigating UTIs in patients that warrant special care such as pregnant women, immunosuppressed patients (e.g., due to neoplasms or autoimmune disease), patients pre- or post-surgery, or in patients with trauma (e.g., spinal cord injury)

were excluded from the review. Studies relating to healthcare-acquired infections, catheter-associated infections, and recurrent UTIs were also excluded.

2.4.1.2 Concept

I reviewed all studies that used routinely collected EHR data (see Remark box below for a definition of EHR for the purposes of this review) to generate evidence that supports clinicians in diagnosing or managing suspected community-onset UTIs during initial patient presentation. Studies were included if they used EHR data to estimate the risk of microbiologically confirmed bacterial UTI or infectious complications in patients consulting primary or secondary care for suspected UTI. I excluded UTI focused studies that estimated only population incidence rates of UTI over time, or studies that described antibiotic prescribing patterns and adherence to local or national antibiotic prescribing guidelines. I also excluded studies that utilised EHR data to evaluate local antimicrobial stewardship interventions for UTI outside of standard clinical care. In order to be eligible, studies could only use

Remark (Electronic health records). EHRs were defined as a "longitudinal collection of electronic health information about individual patients" [102], including information on patient demographics, medical diagnoses, clinical observations, laboratory tests, and prescriptions. For the purposes of this thesis, eligible sources of EHR data were those that at a minimum contained information on healthcare contacts (e.g., consultation or admission dates) and clinical diagnoses. Notably, this requirement excluded stand-alone computer-aided laboratory or pharmacy systems, unless they were linked to aforementioned administrative and diagnostic information. In order to be considered in this review, sources of EHR data were further required to allow for the automated extraction of all relevant clinical information. This excluded electronic patient management systems that electronically stored documents like discharge letters but which needed to be reviewed manually (e.g., by a clinician or a study nurse) in order to extract the necessary information.

data contained within the EHR, and this could not be supplemented by manual note review.

2.4.1.3 Context

I included studies that described consultations in primary care or new hospital visits for community-onset UTI. Studies that were limited to care home residents or patients in whom suspicion of UTI only arose after they had been hospitalised for non-UTI reasons were not considered to investigate community-onset UTI and were therefore excluded.

2.4.2 Types of evidence sources

I included original quantitative retrospective observational studies published as peer-reviewed papers, and systematic reviews of such studies. Qualitative studies and studies that involved active collection of data outside of routine care (e.g., prospective observational studies or randomized controlled trials) were excluded from the analysis, even if some of their data were collected through EHR systems. Studies published only as conference abstracts or preprints were also excluded.

2.4.3 Search strategy

Two bibliographic databases (Embase and MEDLINE) were systematically searched for articles published in English language between January 1st 1974 and December 10th 2020[1]. No geographical restrictions were applied. Following standard recommendations for scoping reviews issued by JBI, a three part literature search was performed [101]. First, an initial search was performed in both Embase and MEDLINE using a preliminary list of search terms. Titles, abstracts, and index terms of a random sample of 50 studies identified during the initial search as well as known relevant articles were reviewed, and the search strategy was adapted accordingly. A second search using the updated search terms was performed (see Table 2.1 for detailed results from Embase), and all retrieved studies were exported into the web-based systematic review software DistillerSR (Evidence Partners,

[1] Articles were limited to English language for feasibility reasons. The year 1974 was the earliest year available in the bibliographic database Embase and was judged the absolute lower limit for any possible use of EHR data.

Table 2.1: Final search strategy applied to the Embase bibliographic database.

#	Search term	Results
1	human/	21,658,278
2	adult/	7,487,346
3	aged/	3,063,902
4	2 or 3	8,516,579
5	urinary tract infection/	105,014
6	electronic health record/	19,788
7	medical record/	180,757
8	"medical record review"/	133,353
9	record.ti,ab,kw.	199,693
10	records.ti,ab,kw.	423,827
11	6 or 7 or 8 or 9 or 10	745,617
12	1 and 4 and 5 and 11	3,753

Ottawa, Canada). Finally, the bibliography of all studies included in the review were manually screened for any missed articles. The search was last updated using this strategy on January 17th 2021.

2.4.4 Selection of studies

All records identified via the above described search strategy were automatically screened for duplicates using the DistillerSR software. Suggested duplicates were manually reviewed and removed if a duplicate entry was confirmed. The selection of records for inclusion into the review was performed in multiple iterations. I first screened the titles of all de-duplicated records. Next, I reviewed the abstracts of all remaining records. Finally, I retrieved the full-text articles for all records left after title and abstract review and compared them to the eligibility criteria. If an eligibility criterion could not be definitively confirmed at any stage of the review, the article was put through to the next review stage. Before each review step, I piloted the format and layout of the screening forms using 25 randomly selected studies.

2.4.5 Data extraction

For each study included in the final review, I recorded the following information: author(s), year of publication, countries, setting (primary care and/or secondary care), number of patients, exposures, outcomes, covariates, statistical methodology, and key findings (see Appendix A for the data extraction template).

2.4.6 Assessment of risk of bias and applicability

Each study included in the final review was assessed for its ROB and applicability to the intended target population of this review based on widely used quality assessment tools (see Appendix B for templates of each used tool). Cohort and case control studies of exposure effects were assessed using a recent adaption of the well-known Newcastle - Ottawa Quality Assessment Scale (NOS) proposed by Guyatt and Busse [103]. The adapted NOS was used in place of the original version since it includes an assessment of the quality of all predictors (and not only the exposure) and explicitly distinguishes between ROB — i.e., risk to the internal validity of the study — and applicability of the study results to the target population [103]. Since risk prediction models pursue a very different aim than the more traditionally epidemiological cohort and case control studies, the recently developed Prediction model Risk of Bias Assessment Tool (PROBAST) [104] was used to assess ROB and concerns about applicability in studies that developed or evaluated clinical risk prediction models. The items of both the adapted NOS and PROBAST were grouped according to the following categories: participants, predictors, outcomes, and analysis. A separate rating was made for ROB and applicability in each of those categories (analysis was appraised for ROB only).

2.4.7 Analysis of the evidence and presentation of results

A narrative description of all included studies was produced, summarising the study characteristics, main design choices, and key findings. A numerical analysis of the geographical distribution, healthcare settings, exposures, outcomes, and statistical methodologies was performed. Study characteristics and ROB assessment for each study were presented in tabular form. Year of publication and number of included

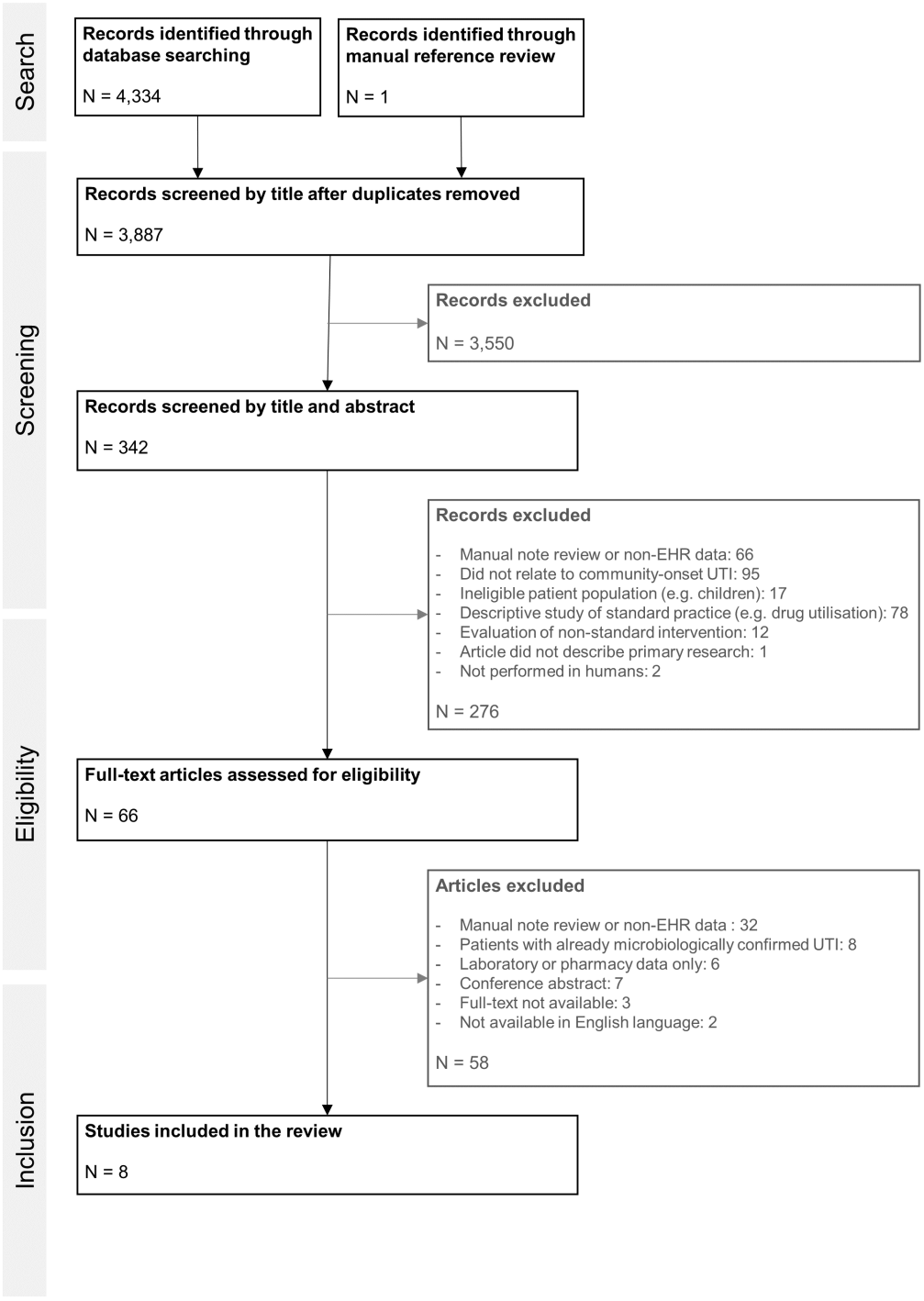

Figure 2.1: Flow chart of reviewed studies and reasons for exclusion.

patients were presented graphically for all studies included in the full-text review. All results were reported following the PRISMA Extension for Scoping Reviews reporting guideline (see Appendix G) [105].

2.5 Results

The search in Embase and MEDLINE resulted in 4,334 identified records (Embase: 3,753; MEDLINE: 581; Figure 2.1)[2]. One additional record was identified through manual reference screening. After removing duplicates, I screened 3,887 titles which resulted in 342 records for detailed screening based on title and abstract. Due to the broad search criteria, a large proportion of the remaining records described studies that used manual note review or non-EHR data (66 records), described other kinds of infection or other types of UTI (95 records), or provided a descriptive analysis of guideline adherence and patterns of care (78 records). A total of 66 articles were retrieved for full-text review, of which eight were included in the final data extraction step and appraised for ROB and applicability to the intended target population (Figure 2.1).

2.5.1 Identification of EHR studies

A primary challenge at each stage of the review was determining whether or not a study used automatically collected, structured EHR data (included) or required manual patient note review (excluded). Most reviewed full-text articles did not include a clear description of the data source and frequently reported a not further defined "retrospective review of medical records". Articles whose methodology section suggested a probable manual note review were excluded during full-text review. A total of 28 studies were excluded this way. These studies tended to be published in or before 2015 (20 out of 28 excluded studies) and to include less than 1,000 patients (24 out of 28 excluded studies; Figure 2.2 purple circles), strengthening the conclusion that they did not describe the analysis of automatically extracted EHR data. Studies that were judged to describe EHR research (Figure 2.2

[2]The much larger number of records identified within Embase compared to MEDLINE was unexpected. In order to avoid any unintended exclusion of records in MEDLINE, I performed additional searches in MEDLINE to assess the sensitivity to individual search terms. The largest difference was observed for keywords relating to (electronic) medical records, which may be more narrowly defined in MEDLINE. Of the final seven articles identified through database search, five were found in both Embase and MEDLINE, one did not contain any reference to medial records in MEDLINE and was thus only found in Embase, and one was not indexed in MEDLINE at all. This suggests that while there was some difference between the records returned by Embase and MEDLINE, most of these differences were due to studies unrelated to EHR research.

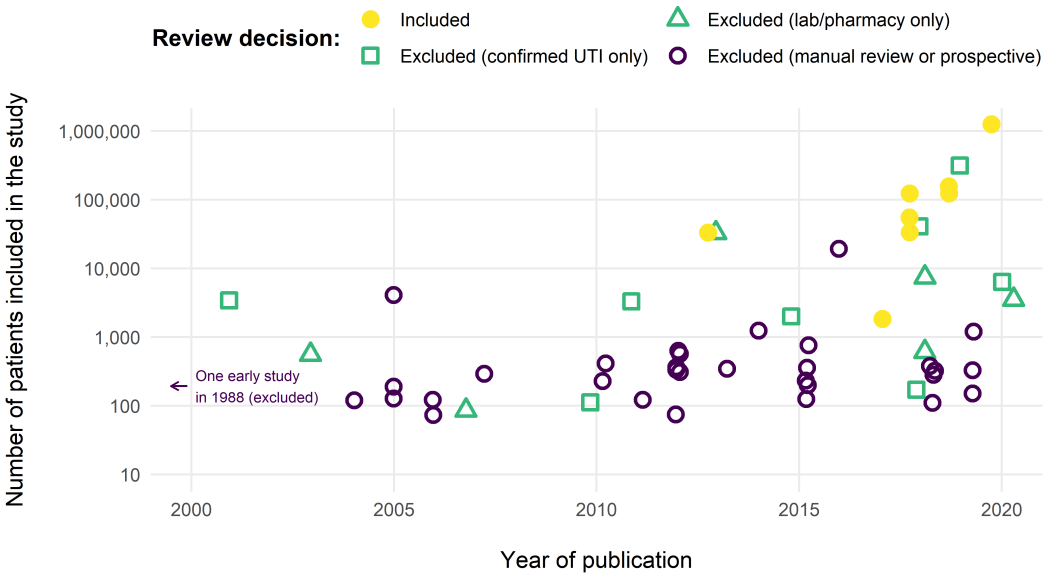

Figure 2.2: Year of publication and number of included patients among 54/66 full-text peer-reviewed articles assessed for eligibility. 12 studies which were eligible for full-text review were excluded (and were omitted from this figure) because they were conference abstracts (7 studies), could not be retrieved (3 studies), or were not in English language (2 studies).

yellow dots) were all published in or after 2013 and — with the exception of one study — all included more than 10,000 patients. Several studies were similar in size and scope to included studies but either included only microbiologically confirmed (rather than suspected) cases of UTI (Figure 2.2 green squares) or were performed in stand-alone laboratory or pharmacy datasets (Figure 2.2 green triangles).

2.5.2 Study characteristics and key findings

Of the eight studies included in the review, five were conducted in the UK and three were conducted in the US (Table 2.2). All UK studies used the same primary care database (the Clinical Practice Research Datalink [CPRD], see Chapter 3 for a detailed description). US-based studies used national data from outpatient clinics of the Veterans Affairs healthcare system (1 study; both primary and secondary care), three local outpatient clinics (1 study; both primary and secondary care) or four local emergency departments (ED) (1 study; secondary care only). Four studies were limited to patients aged 65 years or more, and two studies included only men. Six studies conducted a retrospective cohort study, estimating the effect of a single

Table 2.2: Characteristics of studies included in the scoping review.

Authors	Year	Ctry.	Population	Setting	Exposure	Outcome	Statistical methodology
[106] Drekonja *et al.*	2013	US	Adult men consulting with recorded suspicion/diagnosis of UTI and who were prescribed antibiotic treatment	Primary / secondary	Duration of antibiotics	Re-consultation; recurrence of infection	Multivariable logistic regression
[107] Grigoryan *et al.*	2017	US	Adult women consulting with recorded suspicion/diagnosis of UTI and who were prescribed antibiotic treatment	Primary / secondary	Duration of antibiotics	Re-consultation; recurrence of infection	Multivariable logistic regression
[108] Ahmed *et al.*	2018	UK	Patients aged $\geq$ 65 years consulting with recorded suspicion/diagnosis of UTI and who were prescribed antibiotic treatment	Primary	Type of antibiotic	Re-consultation; hospitalisation; death	Multivariable mixed effects logistic regression; propensity score matching
[109] Taylor *et al.*	2018	US	Adult patients consulting with recorded suspicion/diagnosis of UTI — or symptoms attributable to UTI — and who had a urine culture ordered	Secondary	†	Culture growth $> 10^4$ cfu/mL	Multivariable (penalised) logistic regression; machine learning models
[110] Ahmed *et al.*	2019	UK	Patients aged $\geq$ 65 years consulting with recorded suspicion/diagnosis of UTI and who were prescribed nitrofurantoin, cefalexin, ciprofloxacin, or co-amoxiclav	Primary	Type of antibiotic	Re-consultation; hospitalisation; death	Multivariable mixed effects logistic regression; propensity score matching
[111] Ahmed *et al.*	2019	UK	Men aged $\geq$ 65 years consulting with recorded suspicion/diagnosis of UTI and who were prescribed antibiotic treatment	Primary	Duration of antibiotics	Re-consultation; hospitalisation; death	Multivariable mixed effects logistic regression; propensity score matching
[112] Gharbi *et al.*	2019	UK	Patients aged $\geq$ 65 years consulting with recorded suspicion/diagnosis of UTI	Primary	Delayed antibiotics	Bloodstream infection; death	Multivariable logistic and Cox regression
[113] Mistry *et al.*	2020	UK	Patients of all ages consulting with recorded suspicion/diagnosis of UTI	Primary	†	Hospitalisation	Multivariable Cox regression

† These studies developed predictive models and therefore did not have a dedicated exposure variable. UK, United Kingdom; US, United States of America; UTI, urinary tract infection.

exposure after adjusting for confounding factors. Two studies attempted to predict individual patient risks of bacterial growth in urine samples (1 study) or risk of infectious complication (1 study).

2.5.2.1 Cohort studies of exposures

Most (6/8) included studies performed an inferential analysis of the effects of different kinds of empirical antibiotic treatment on clinical outcomes. These studies assessed the association between adverse outcomes of infection and treatment choice including: type of antibiotic (2 studies) [108, 110]; duration of prescribed antibiotics (3 studies) [106, 107, 111]; and delaying or withholding antibiotics (1 study) [112]. The considered outcomes were one or more of: re-consultation for UTI[3]; recurrence of UTI; hospitalisation for infection-related complications; or death. Outcomes were measured within 14 to 60 days after initial consultation except recurrence of UTI, which was measured up to one year after consultation. With the exception of one study [112], studies that assessed treatment choices were limited to patients who received antibiotics during or soon after their index consultation.

All studies that estimated effects of empirical antibiotic treatment used multivariable logistic regression to estimate effect sizes of interest. Multivariable Cox regression was used in addition to logistic regression by one study, although without justification as to why logistic regression was used for one outcome (bloodstream infection within 60 days) and Cox regression for the other (death within 60 days) [112]. Three studies by the same group of authors further used a multilevel modelling framework to account for correlation by primary care practice and employed propensity score matching to adjust for differential treatment assignment [108, 110, 111].

Studies assessing the duration of treatment for community-onset UTI in primary care provided conflicting evidence, reporting that shorter courses of treatment increase [111], decrease [107] or do not effect odds of re-consultation for UTI [106]. The three studies included similar antibiotics, but studies that found

[3]Interpreted as an indicator of treatment failure.

an increase or no change in the odds of re-consultation for UTI were limited to male patients only [106, 111]. In the only study that also investigated the more severe outcomes of hospitalisation and death, shorter courses of antibiotics in male patients above the age of 65 years did not increase the risk of those outcomes [111]. In a further cohort study investigating the use of nitrofurantoin in male patients aged $\geq$ 65 years who had renal impairment (defined via an estimated glomerular filtration rate [eGFR] $<$ 60 mL/minute/1.73m^2) and consulted for community-onset UTI in primary care, Ahmed *et al.* concluded that use of nitrofurantoin did not lead to an increased risk of hospitalisation and was instead associated with lower odds of re-consultation or death in men with eGFR of 45-59 and 30-44 mL/minute/1.73m^2 respectively [108]. In a second cohort study, the same authors compared nitrofurantoin to cefalexin, ciprofloxacin, or co-amoxiclav in both men and women aged $\geq$ 65 years consulting primary care for community-onset UTI [110]. They reported a slightly higher risk of re-consultation when being prescribed nitrofurantoin instead of cefalexin, ciprofloxacin, or co-amoxiclav. However, they also found increased odds of hospitalisation for sepsis when being prescribed cefalexin or ciprofloxacin, and increased odds of death when being prescribed cefalexin [110]. Finally, Gharbi *et al.* (2019) [112] suggested that delaying or withholding antibiotics in older patients aged $\geq$ 65 years who consulted primary care for community-onset UTI was associated with a 7–8-fold increase in the odds of bloodstream infection by comparison with patients of similar age that received immediate antibiotic treatment.

2.5.2.2 Risk prediction studies

Two studies did not assess the effects of a single exposure variable and instead attempted to develop clinical risk prediction models to aid the diagnosis and management of UTI. Mistry *et al.* (2020) [113] used routine English primary care data from CPRD to predict the 30-day risk of infection-related hospitalisation for all patients with an incident episode of (suspected) UTI and who did not receive antibiotic treatment during their initial consultation. The authors included almost 1,000,000 episodes of UTI. They modelled the risk of hospitalisation

using multivariable Cox regression and adjusting for demography, comorbidities, previous hospitalisations, previous prescriptions other than antibiotics, flu vaccination status, and date of consultation. The model was externally validated in data from Welsh primary care practices. The authors concluded that their model achieved good, generalisable performance (area under the receiver operating characteristic curve [AUROC]: 0.821), suggesting it might serve as a basis to evaluate a patient's risk of developing complications.

In the only included study that had access to urine culture results and used data from four EDs in the US, Taylor *et al.* (2018) [109] used clinical observations and medical history to predict bacterial growth of $\geq 10^4$ colony forming units per millilitre (cfu/mL) in urine samples from patients visiting the ED with suspected UTI. Suspected UTI was defined as a recorded symptom or diagnosis comparable with UTI, including both urinary symptoms as well as less specific symptoms such as altered mental status, abdominal pain, or fever. They used multivariable logistic regression and machine learning algorithms — including support vector machines, random forests, extreme gradient boosting, and neural networks — to model factors that influence a patient's risk of bacterial growth [109]. Reporting favourable model performance (AUROC: 0.904), they also concluded that their model might be used by clinicians to guide early diagnosis of bacterial UTI.

2.5.3 Risk of bias and concerns about applicability

Most studies were judged at high or unclear ROB for at least one of the assessed categories (Table 2.3). Included predictors and outcomes were generally considered at low ROB. Predictors were primarily limited to variables that could be directly derived from EHR records and which tend to be recorded well, including demography, presence of chronic diseases, previous healthcare visits, and past prescriptions. Two studies were judged at high ROB for the assessment of their outcomes — re-consultation for and recurrence of UTI — since they were limited to outpatient data only and failed to cover all possible healthcare settings in which these outcome might occur [106, 107]. Two study were considered at unclear ROB for how they defined their primary outcome. First, there were questions about the

reliability of identifying the outcome of bloodstream infection in Gharbi *et al.* from hospital discharge codes without microbiological culture results [112]. Second, it remained unclear whether included predictors like laboratory flow cytometry results in Taylor *et al.* may have unduly influenced which urine samples underwent microbiological culture — and may thus have introduced selection bias in the outcome [109].

Although predictors that were included in the analyses were usually well measured, studies that aimed to estimate exposure effects frequently failed to sufficiently account for differences between exposed and unexposed patients [106, 107, 112], introducing confounding by indication. For example, in Gharbi *et al.* patients who did not receive immediate antibiotics during their initial consultation in primary care tended to be notably older and were more likely to have recently been to hospital or received antibiotic treatment, raising questions about the comparability of the treatment groups [112]. Difficulties in ascertaining the clinical rationale behind delaying or withholding antibiotics from EHR data might have contributed to these difficulties (see Chapter 4 for a more detailed discussion of this issue). Ahmed *et al.* used propensity score matching to account for similar imbalances between treatment groups in their studies [108, 110, 111]. While this might have plausibly accounted for the imbalances between patients who received long versus short courses of treatment [108, 111], it remained less clear whether this approach could also account for differences between patients who received the standard first line therapy of nitrofurantoin as opposed to other, more broad-spectrum antibiotics like cefalexin [110].

The analyses of included prediction modelling studies were also judged at high risk of bias. Mistry *et al.* only assessed model performance via AUROC and ignored the substantial class imbalance in their outcome [113]. This means that their model would have an immense false positive or false negative rate when used in clinical practice. Taylor *et al.* on the other hand did not perform adequate external validation [109]. Instead, they evaluated their model on a simple hold-out set covering the exact same patient population as their training set, which is not generally considered

Table 2.3: ROB and concerns about applicability of included studies.

	ROB				Applicability		
	Participants	Predictors	Outcomes	Analysis	Participants	Predictors	Outcomes
[106]	+	+	–	–	?	+	+
[107]	+	+	–	–	?	+	+
[108]	+	+	+	+	?	+	+
[109]	–	+	?	–	–	+	+
[110]	+	+	+	?	?	+	+
[111]	+	+	+	+	?	+	+
[112]	+	–	?	–	+	+	+
[113]	–	+	+	–	–	+	+

+ indicates low ROB/low concern regarding applicability; – indicates high ROB/high concern regarding applicability; and ? indicates unclear ROB/unclear concern regarding applicability. Studies [109] and [113] were assessed using the Prediction model Risk of Bias Assessment Tool (PROBAST) [104]. All other studies were assessed using an adaption of the frequently used Newcastle - Ottawa Quality Assessment Scale (NOS) proposed by Guyatt and Busse [103].

ROB, risk of bias.

appropriate external validation (see Chapter 5 for an in-depth discussion of this issue) [114].

ROB in the selection of participants was generally judged low for all studies of exposures (Table 2.3). However, there were unclear concerns about the applicability of those study populations to the review question. Concerns related to the fact that study population included in some of the studies might considerably differ from the target population which is of greatest clinical relevance and in whom the research would likely be applied in clinical practice. For example, several studies included in the review limited the analysis to patients who received antibiotic treatment [106, 107, 108, 110, 111] but failed to compare patient characteristics among those who did (included) and did not (excluded) receive treatment. While this might be less of a concern for settings in which most patients have been shown to receive antibiotics [108, 110, 111, 112], Drekonja *et al.* (2013) [106] reported that only 37.5% of eligible patients with suspected UTI in their study were prescribed an antibiotic, introducing a high risk of selection bias.

ROB in the selection of participants was also judged high for both studies that developed risk prediction models. Mistry *et al.* trained their model — which

aimed to predict risk of complication in all primary care patients consulting with UTI — only in patients who did not receive immediate antibiotic treatment, thereby implicitly assuming that they were also representative of patients who did receive antibiotics for suspected UTI [113]. Taylor *et al.* trained their model in patients with a requested urine culture and symptoms that were broadly comparable with UTI but which are also seen in a range of other common conditions seen in the ED. As they did not assess differences in model performance in patient subgroups [109], it is possible that the strong predictive performance in this study was therefore driven by a high proportion of individuals with no evidence — and very low a priori probability — of UTI. While this may still be a valuable prediction target, these patients are not representative of a population with clear suspicion of UTI (see Chapters 5 and 6 for a more detailed discussion of this issue).

2.6 Discussion

In this chapter, I used a scoping review to identify, summarise, and evaluate studies that used routinely collected EHR data to investigate the diagnosis and management of suspected community-onset UTI. In the eight studies which met the inclusion criteria, EHR data were primarily used to investigate the effects of antibiotic treatment choices on preventing complications of infection. Two studies further attempted to use EHR data to predict individual patient's probability of microbiologically confirmed UTI or subsequent risks of hospitalisation for complications of UTI. Both areas of research were relatively recent, with the earliest included study being published in 2013 and all remaining studies in or after 2017. Almost all studies were performed in primary care.

The ROBs identified in this review related to common challenges in retrospective observational studies. A lack of randomisation made it difficult for included studies to ensure comparability between treatment groups. If patient factors that aren't reliably measured in EHR data strongly affect both treatment assignment as well as the risk of experiencing the outcome, these analyses will be at a high risk of confounding by indication [115]. It is likely that such confounding

played a role in some of the studies identified by this scoping review. Clinicians base their treatment decisions on a careful examination of the patient and the choice of treatment usually carries information about the state of that patient — e.g., about the (perceived) disease severity [115]. The use of auxiliary information and adequate methodology may allow to reduce such confounding. For example, in all three studies by Ahmed *et al.* the authors used propensity score matching to approximate randomisation. It remains unclear, however, whether the covariates available from EHR data were sufficient to account for the most severe sources of confounding — such as disease severity — which is not well recorded in structured EHR data. Other included studies either did not have sufficient information about patient characteristics to plausibly account for non-random treatment assignment [106, 107] or did not sufficiently account for the stark difference between treatment groups [112], and were consequently considered at high ROB.

Aside from non-random treatment assignment, this review revealed difficulties in identifying patients with suspected UTI reliably from routine EHR data. Included studies usually used diagnostic codes with records of antibiotic prescribing or urinalysis to ascertain a suspected UTI. None of the included studies formally evaluated these approaches, and it remains unclear how accurately they identify the target population with suspected UTI in primary or secondary care. For example, if inclusion criteria were too strict, they may have selectively excluded an important subgroup of the target population. Studies that required patients to receive a certain treatment such as antibiotic prescribing may bias their cohort to more severe UTI that requires immediate treatment and exclude less severe cases. The extent of this bias depends on how representative patients who receive treatment are of all patients who consult with suspected community-onset UTI. While this issue may play a lesser role in UK primary care — where the majority of patients receives antibiotics [76] — it may have biased estimates in Drekonja *et al.* where the requirement to receive immediate antibiotics excluded almost two-thirds of all eligible patients [106]. Too broad a definition, on the other hand, may lead to a very heterogeneous patient population and similarly cast doubt on the applicability of findings. This

was primarily seen in Taylor *et al.* who included patients with symptoms that are compatible with UTI, but also with a range of other infectious and non-infectious conditions. The authors did not investigate how well their model performed in patients with clear symptoms of UTI and whether individual patient groups may have driven the reported model performance.

2.6.1 Strengths and limitations

This chapter provides a systematic search of published literature using standard scoping review methodology in two major bibliographic databases. Closely following guidelines on conducting [101] and reporting [105] scoping reviews, I was able to identify previous studies in a reproducible manner and highlight approaches and pitfalls common to studies using EHR data in order to investigate community-onset UTIs.

Since the main aim of this review was to inform the later chapters of this thesis, its scope was defined in a very targeted manner. Included studies were limited to research that used *routinely collected EHR data* to estimate the risk of UTI-related outcomes to aid the diagnosis and management of community-onset UTI *at the point of care*. For one, this excluded studies that used prospectively collected data — including randomised controlled trials as well as prospective cohort studies — or manual case note reviews, even when retrospective. Findings of this review therefore cannot be interpreted as a comprehensive summary of studies assessing the diagnosis and treatment of suspected community-onset UTI but instead represent a summary of the ways that routinely collected EHR data have been used as one (novel) approach to study UTI. Included studies were further limited to those that had access to administrative information (e.g., admission times) and clinical diagnoses. These are the types of systems that would usually be available to doctors at the point of care, and will be used in the remainder of this thesis. As a result, studies that aimed to assess suspected UTI using pharmacy or laboratory data only were excluded. Only peer-reviewed articles were included, which may have missed very recent literature presented at conferences or published on preprint servers. Finally, due to time constraints I undertook this review by

myself. Ideally, a second reviewer would have independently reviewed titles and abstracts to mitigate any subjectivity in the application of inclusion criteria [101].

2.6.2 Conclusion

While the analysis of routinely collected EHR data holds great potential to guide the diagnosis and management of suspected community-onset UTI, this scoping review has shown that such analyses need to be conducted carefully and interpreted with caution. I was able to identify several sources of bias in previously published EHR studies investigating community-onset UTIs, owing to the observational, retrospective, and routine nature of the data. These biases potentially limit the applicability of some of the study findings to real-world clinical practice, although the precise extent of bias remains unclear. After describing the breadth and depth of data available in English EHR data sources in Chapter 3, the remainder of this thesis therefore develops two case studies that investigate if and how prudent study design and careful statistical methodology may account for the biases described in this review.

The majority of included studies were conducted in primary care. In the first case study in Chapter 4, I thus assess the association between delayed or withheld antibiotic prescribing in primary care and complications of infection in adult women consulting with suspected community-onset UTI. I use data from the same English primary care database (CPRD; see Chapter 3) that was used in five out of the eight included studies. Mindful of the difficulties of identifying community-onset UTI from EHR data described in this review, I utilise linkage to secondary care data (HES; see Chapter 3) to ensure that I only include actual community-onset infections. I further investigate the extent of any imbalance between treatment groups, as reported by some of the reviewed studies. I apply statistical balancing methods (matching and inverse probability of treatment weighting) to balance the characteristics of patients who did and did not receive immediate antibiotic prescriptions. Finally, I use these balanced cohorts to (re-)estimate the association between delayed or withheld antibiotic prescribing for community-onset UTI and complications of infection *as well as* additional adverse outcomes not directly

related to UTI. The latter was used to assess the success of the balancing methods in accounting for confounding in the data. If balancing were indeed successful in approximating a randomised patient population, we would not expect to find any association with adverse outcomes that are not usually a consequence of UTI.

There was only one study in this review that used routinely collected data to study community-onset UTI at presentation in secondary care. This study used data from four EDs operated by a single US healthcare provider to predict bacteriuria in patients with suspected UTI. In my second case study described in Chapters 5 and 6, I use a bespoke EHR dataset from Queen Elisabeth Hospital Birmingham to develop a similar risk prediction model for bacteriuria in English patients that presented with suspected community-onset UTI in the ED. Addressing the concerns identified in this review, I carefully assess the presence of patient heterogeneity and spurious associations in the study population, and evaluate their impact on the reliability and performance of my model. I further evaluate the performance of the model over time and compared to clinicians' judgement, obtaining an estimate of the expected stability and reliability of model predictions if the model were to be used in English clinical practice.

Chapter summary

- Detailed EHR data collected as part of routine care may provide an opportunity to identify patients at low risk of UTI or UTI-related complications, allowing clinicians to avoid antibiotic treatment in this group altogether or stop antibiotics early.
- While many studies have used EHR data to establish the incidence of infection or assess guideline adherence, few studies have used EHR data to estimate risk factors to guide the diagnosis and management of suspected community-onset UTI at the point of care.
- Among those studies that did estimate risks, almost all were judged to be at high risk of bias, owing to inadequate methods and/or a difficulty

to ascertain important confounding factors from EHR data.

- Data captured in EHR databases are not collected for research. Careful study design and interpretation is therefore necessary to guarantee results that are meaningful for clinical practice.

Chapter 3

Recording of UTI in English EHR databases

Abstract

In the previous chapter, I reviewed literature using routinely collected electronic health record (EHR) data to create evidence that may guide the treatment of community-onset urinary tract infection (UTI) at the point of care. I discussed how EHR data may provide a novel approach to reduce the substantial uncertainty inherent to diagnosing and managing UTI.

In this chapter, I give an overview of key EHR databases that can be used to investigate UTIs in English primary and secondary care. I describe the content and limitations of these datasets in order to lay the foundation for the research studies presented in Chapters 4-6. While providing an overview of the wider data availability in England, this chapter pays particular attention to the Clinical Practice Research Datalink (CPRD) — a UK-wide primary care dataset used in five out of the eight studies reviewed in Chapter 2 — and bespoke emergency department data collected from the EHR system at Queen Elizabeth Hospital Birmingham. Both datasets are later analysed in Chapters 4-6. I end the chapter by summarising the depth, breadth, and reliability of UTI data in the covered data sources and briefly discuss what an ideal EHR dataset for research of UTI may look like.

3.1 Healthcare and medical data in England

The National Health Service (NHS) has been providing free healthcare for English residents at the point of care since 1948 [116]. Services are mainly funded through a tax-based system. Fees might be levied for prescriptions, dental services, and opticians. Patients can request an exemption from prescription charges, and

approximately 90% of drugs are dispensed to patients whose fees have been waived [116].

Provision of services is divided into primary and secondary/tertiary care. The former is mostly performed by General Practitioners (GPs) in local practices, with an average of four to six GPs per practice [116]. Each patient is registered with a single primary practice of their choosing. GPs act as the first point of contact for patients, and as gatekeepers for specialist care. If judged necessary, GPs refer patients to hospital for specialised treatment that cannot be provided in primary care. Notable exceptions include emergency care or sexual health, which can be attended immediately without referral [116].

Health care policy for England is the responsibility of the central government [117], with the Department of Health and Social Care defining a national health strategy. NHS England, an independent governing body, is charged with implementing this strategy. Where sensible, it does so itself on a national level [118]. In the early 2000's a substantial part of the decision making was devolved to local and regional authorities. As of 2014, there were 211 Clinical Commissioning Groups (CCGs) who managed two thirds of the NHS' annual budget [119]. Services commissioned by CCGs include most secondary care services and parts of local primary care. Besides NHS England and local CCGs, the Department of Health funds a number of additional arm's length bodies. These focus on specialised aims such as promoting public health (Public Health England [PHE]), compiling evidence-based guidance (National Institute for Health and Care Excellence [NICE]), and regulating health service quality (Care Quality Commission [CQC]) [118].

Data generated in the English health system is governed by NHS Digital, a non-governmental public organisation [120]. As part of this role, it collects medical datasets itself and acts as a trusted third party to link patient data across externally or internally curated healthcare datasets.

3.2 Primary care

Primary care is the first point of contact with the healthcare system for most patients in England [116]. As a result, most uncomplicated UTIs are first seen in primary care [3]. The following pages describe several major databases used to conduct large-scale population health research in English primary care, and discuss their strengths and limitations for research on the diagnosis and management of UTI. This list does not aim to be exhaustive, but rather introduces data sources that have been used previously (see particularly Chapter 2) as well as data sources that provide a similar breadth and depth and may thus be used as an alternative.

3.2.1 Clinical Practice Research Datalink (CPRD)

The Clinical Practice Research Datalink (CPRD) is a database of primary care health records spanning the entire United Kingdom (UK)[1]. Data are collected directly from the GP patient management software Vision and more recently EMIS [121, 122]. Data collected via Vision is available as CPRD GOLD [121], whereas data collected via EMIS is called CPRD Aurum [122]. Aurum was designed to be similar to GOLD, but some differences exist due to underlying differences in Vision and EMIS. Many conclusions about GOLD nevertheless also hold for Aurum, and the two databases may be pooled to increase sample sizes. As Aurum has only recently been made available, all primary care analysis presented in this thesis is based on GOLD data only and the following description of CPRD will consequently focus exclusively on GOLD. In this chapter as well as the rest of this thesis, the term CPRD will be used to refer to CPRD GOLD [121] and all analyses presented in this thesis used CPRD GOLD linkage set 15.

Active data collection in CPRD dates back as far as the 1990s. In 2013, CPRD included data from more than eleven million patients from 674 practices, 4.4 million of which were active ($\sim$7% of the UK population) [121]. Patients in CPRD have been shown to be broadly representative of the UK population in terms of age, sex, ethnicity, and body mass index (Figure 3.1 and [121, 123, 124]). Patients enter the database when they register with a participating GP practice and stop providing data

[1] England, Wales, Scotland, and Northern Ireland.

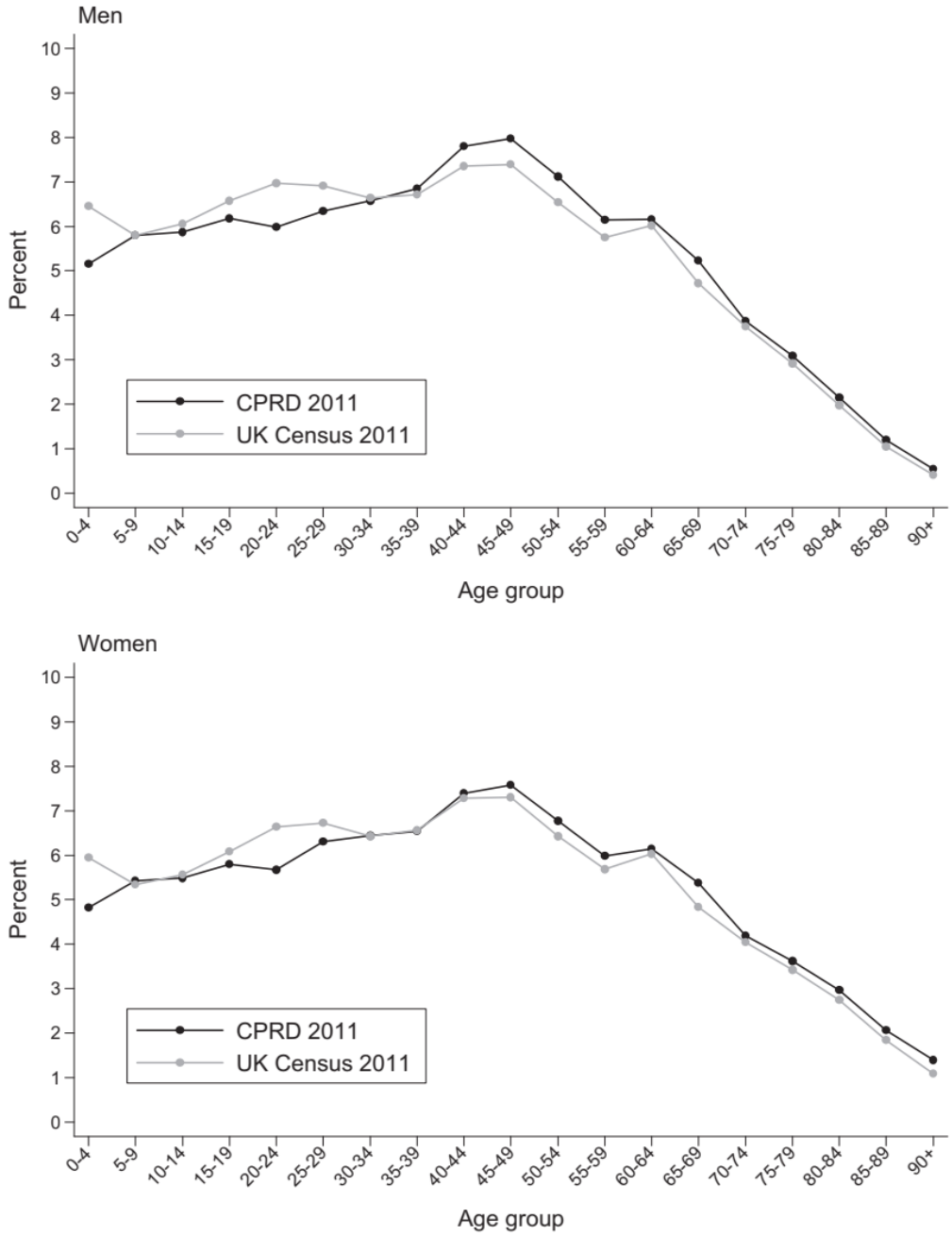

Figure 3.1: Age distribution in a random sample of one million patients from CPRD primary care data compared with UK census data in 2011. Top panel: men. Lower panel: women. All patients were of acceptable research quality as ascertained by CPRD.

CPRD, Clinical Practice Research Datalink.

Figure source: Herrett *et al.* (2015). Copyright © 2015, Oxford University Press; reused under Creative Commons Attribution 4.0 International License; image caption was adapted from the original image caption.

when they transfer out, die, or if the practice withdraws from CPRD. The average active follow-up time is 5.1 years [121], although information on diagnoses may be available from historic records and can date back before electronic data collection.

Data in CPRD are stored using pseudonymised patient and practice identifiers. To prevent re-identification, no directly identifiable information such as name or address are available. However, CPRD includes a patient's gender, year of birth, and ethnicity (as recorded by the clinician). The month of birth is added for patients less than 16 years old to provide additional granularity [125]. A postcode-based social deprivation score (Index of Multiple Deprivation 2015 or Townsend score) allows to account for a patient's approximate socio-economic status [126].

Patients who for any reason do not want to share their information can opt out of CPRD through their GP. No reasons need to be given to opt out, and patients continue to receive the same level of care as before. Any data collected on patients who opted out is removed from CPRD, even if some of it was collected before the opt-out. Since 2018, opt-outs can be performed online (https://your-data-matters.service.nhs.uk).

3.2.1.1 Available clinical data items

Diagnoses The medical diagnoses and symptoms available within CPRD were recorded by GPs as part of routine care and are stored together with a corresponding date and consultation identifier. While a more detailed summary of the consultation is captured in free-text format — which is not generally available for research [127] — clinicians are asked to enter the most important clinical information in form of Read codes, a medical terminology used in UK primary care. Read codes entered by the GP are a primary source of information on clinical activity in CPRD. The terminology used in CPRD is Read v2 [128]. In total, there are about 110,000 Read codes available, but more than 50% of recorded diagnoses are covered by the most common 1,000 codes (estimated via the CALIBER read code browser[2]). Medical

[2] The Cardiovascular Disease Research using Linked Bespoke Studies and Electronic Health Records (CALIBER) is a data resource hosted at the UCL Institute of Health Informatics that links CPRD with several other routine data sources or patient cohorts to enable translational research on cardiovascular disease [129].

concepts can often be recorded via multiple broadly interchangeable Read codes, allowing for different levels of granularity. This flexibility comes at the cost of a high number of redundant codes.

Remark (SNOMED-CT in English primary care). In April 2018, the NHS started to retire Read codes in favour of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [130], which solved several legacy issues of Read [131] and enabled greater compatibility of coding with secondary care and international databases. Given the lag in data collection, however, it will take time until these codes enter the data, and no SNOMED-CT codes were used in this thesis. Judging from experiences with previous changes of clinical terminology in other countries, changes of coding behaviour might be expected as a result and will need to be considered in future analyses [132].

Prescriptions Unlike diagnoses, almost all primary care prescribing is automatically captured in CPRD and does not rely on explicit coding by the GP [121]. Prescriptions given to a patient are directly generated by the practice management software. The primary exception are prescriptions made out-of-hours or during home visits, which must be entered manually and may therefore be missing [133]. Data on prescriptions include the date of prescribing, the consultation during which the prescription was made, the prescribed substance, formulation and strength. Additional information on the dosing schedule might be available if the prescribing GP filled in the relevant field during the prescribing process. Drugs are classified via the British National Formulary. For antibiotics, roughly 80% of all prescriptions contain dosage information (as estimated from the data used in this thesis). Although prescriptions are captured automatically, CPRD does not explicitly record the reasons for prescribing. If required, the prescribing indication may be derived from related diagnosis codes based on custom rules and algorithms. Options to infer indications include utilizing a temporal relationship

between medication and diagnoses — e.g., on the same day —, linking codes with the same consultation identifier, or carrying previous indications forward [6]. Which method is used depends on the drug under investigation, the medical condition studied, and the specific research question.

Clinical observations and laboratory tests Data on vital signs and clinical measurements like height, weight, or blood pressure are potentially available in CPRD. Requests for tests are recorded and results can be obtained if they have been added to the patient record via electronic links to the laboratory [121]. Like medical diagnoses, both measurements and lab tests are identified via Read codes. Depending on the test, different result values are stored. Most commonly, results include the date, value and measurement unit. Additional information may include reference ranges and test-specific qualifiers. Whether a measurement is available in CPRD depends on how often it is performed in primary care and whether it tends to be recorded in structured fields rather than free text.

3.2.1.2 Linkage to other EHR databases

On its own, CPRD mainly covers treatment in primary care. Information on hospital admissions and mortality — although potentially available — is limited. When patients are treated in hospital, information on the hospital stay is derived from discharge letters and only recorded if entered into the system by practice staff. Thus, information on hospital stays may be entered late, may be incomplete, or may not be entered at all [134]. GPs also record if and when a patient died. However, when compared to external data, recording of a patient's death is usually delayed and exact time of death can't be established from CPRD alone [135].

To obtain more accurate information, CPRD offers the possibility to link patient records to external databases. The most important among these are Hospital Episode Statistics (HES) for information on hospital stays and the Office for National Statistics (ONS) for mortality and census information. A detailed description of secondary care datasets can be found later in this chapter (see Section 3.3). Further available linkages include among others the Hospital Treatment Insights database [136], the Myocardial Ischaemia National Audit Project [129],

Table 3.1: Frequency of commonly recorded Read codes indicative of possible UTI

Absolute and relative frequency of commonly recorded Read codes in CPRD indicative of a diagnosis of possible UTI.

Rank	Read code	Description	Frequency	%	cum-%
1	K15..00	Cystitis	1,344,447	23.7	23.7
2	K190z00	Urinary tract infection, site not specified NOS	1,339,843	23.7	47.4
3	K190.00	Urinary tract infection, site not specified	1,176,556	20.8	68.2
4	1J4..00	Suspected UTI	1,087,845	19.2	87.4
5	1AG..00	Recurrent urinary tract infections	157,952	2.8	90.2
6	K190311	Recurrent UTI	68,239	1.2	91.4
7	K150.00	Acute cystitis	66,900	1.2	92.6
8	14D4.00	H/O: recurrent cystitis	61,918	1.1	93.7
9	K101.00	Acute pyelonephritis	48,276	0.9	94.5
10	K190.11	Recurrent urinary tract infection	43,875	0.8	95.3

All numbers were obtained via the CALIBER Read code browser [129].

CALIBER, Cardiovascular Disease Research using Linked Bespoke Studies and Electronic Health Records; CPRD, Clinical Practice Research Datalink; cum-%, cumulative percentage; H/O, history of; NOS, not otherwise specified; UTI, urinary tract infection.

the Cancer Registration Data [137], and the Mental Health Dataset [122]. Separate approval must be sought to access these sources. CPRD also allows linkage to local datasets where necessary and appropriate [121].

The use of linked datasets is tightly regulated and only approved if it is essential to the research question at hand, since increasing the number of linked datasets raises the possibility for deductive disclosure of patients' identity. When linkage is approved, only immediately relevant data fields are made available. A justification is required for each field used. Linkage between different datasets is limited to patients and practices that have agreed to linkage. This is the case for ∼75% of practices that contributed data to CPRD in England and for ∼58% across the UK [121].

3.2.1.3 Recording of UTI

Which information is routinely captured in English primary care and the way clinical information is recorded within CPRD has important implications for retrospective research of UTI. Urine cultures are not routinely indicated in primary care for patients with suspected UTI, unless they have a history of recurrent or drug resistant infections [13]. As a result, microbiological confirmation of UTI isn't

usually available and identification of UTI in CPRD relies primarily on clinical coding. This introduces several issues. Clinical coding of UTI is usually coarse and may include patients with a wide range of clinical symptoms and disease severity. For example, the Read codes *Urinary tract infection, site not specified NOS* and *Urinary tract infection, site not specified* together account for almost half of all recorded UTI codes in CPRD (2,516,399; 44.4%; Table 3.1). Part of this ambiguity certainly stems from the uncertainty regarding diagnosis in primary care in the absence of reliable rapid tests (see Chapter 1 for a more detailed discussion of this issue). However, sparsity of coded clinical information also reflects the fact that CPRD data are not recorded with research in mind. Little is known about the patients current health status or detailed presenting complaint. Clinical symptoms have been reported to be rarely recorded [112], as are results from diagnostic tools such as urine dipstick tests [65]. While this information may be found in free-text in the GP's notes, these notes are not usually available to researchers [127].

In addition to difficulties defining and identifying patients with UTI, prescribing records only capture intended prescribing in primary care. The absence of a prescribing record in CPRD does not guarantee that the patient did not receive a drug, as they may have received treatment from an out of hours GP. If a patient was given drugs in secondary care, those will not have been recorded in CPRD either, unless patients were required to continue the medication following discharge from hospital [121]. Vice versa, an existing record in CPRD does not necessarily indicate that an antibiotic was subsequently dispensed to the patient, and much less that it was taken as intended [138]. Importantly for analysing treatment decisions for UTI, CPRD also does not reliably identify delayed prescribing (see Section 1.4.1 for a definition of delayed prescribing). Although a code for delayed prescribing exists within the Read terminology, it has been reported to be under-used and thus unreliable when trying to identify delayed prescribing strategies within primary care databases [139].

Finally, previous studies noted a discrepancy between diagnosis codes for UTI and antibiotic prescribing records in CPRD. While prescribing records are for the

most part captured automatically and are therefore likely to be reliable, the use of diagnosis codes describing the reason for prescribing is at the GP's discretion. As a result, almost half of all prescriptions of nitrofurantoin — an antibiotic which is only indicated for the treatment of lower UTI — does not have an accompanying diagnosis of UTI [6], suggesting that analyses based on diagnostic codes alone miss a considerable proportion of UTI cases in primary care. A possible reason for this observation could be that due to its unambiguous indication, GPs do not feel the need to redundantly record an indication [6], confining all documentation of the reason for nitrofurantoin prescribing to free-text notes.

Issues related to the recording of UTI in English primary care — and in CPRD in particular — are discussed in more detail in Chapter 4 in relation to a study of the association between delayed or withheld antibiotics for community-onset lower UTI in primary care and subsequent risk of infectious complication and/or hospitalisation.

3.2.2 Other sources of primary care data in England

3.2.2.1 THIN, QResearch, and ResearchOne

The Health Improvement Network (THIN), QResearch, and ResearchOne are alternative English primary care databases that provide individual-level records similar in scope and size to CPRD. These databases mainly differ in the exact number of included practices and the practice management software they are based on [140]. THIN was built on Vision — the same software underlying CPRD GOLD — and resembles it in structure and coverage [141]. A proportion of the practices contributing to CPRD GOLD are also available within THIN and data for these practices should be equivalent in the two databases. QResearch collects data from practices that run EMIS software [140], which more recently has also been added to CPRD via its Aurum database (see Section 3.2.1 earlier) [122]. Finally, the ResearchOne database contains data from practices using SystmOne software [142]. Due to their inherent similarities and embedding in the same primary care environment, theses databases share all major strengths and limitations with CPRD with regards to research on UTI, or primary healthcare more generally. Data from

more than one vendor/database may be pooled to increase the sample sizes or obtain independent samples for external validation [138, 143]. Most results and conclusions discussed in this thesis for CPRD therefore directly apply to THIN, QResearch, and ResearchOne.

3.2.2.2 Discover and KID

In addition to nation-wide databases created in collaboration with specific software vendors, local and regional medical data warehouses have been built by local CCGs over the last couple of years. Two examples for such initiatives are Discover in North West London [140] and the Kent Integrated Dataset (KID) in Kent and Medway [144]. These initiatives were originally created with the aim of commissioning and service evaluation but have recently been extended to enable research [140]. Although smaller in size than their nation-wide counterparts — and thus likely less generalisable — these local datasets are more comprehensively linked to additional local data including secondary care from local acute trusts, community health services, social services, and mental health [140, 144]. Since many of the limitations for UTI research discussed in the context of CPRD are due to the way clinical information is recorded in structured EHR data and free text isn't available in these local databases either [144], limitations of these data resources for the research on UTI are expected to be similar to CPRD.

3.2.2.3 PHE Fingertips and OpenPrescribing

All data sources discussed so far contain sensitive, individual-level data. Access to these datasets is therefore restricted, and associated with substantial costs and administrative burdens, including ethical approval and secure data storage. For simple descriptive studies that do not require linkage of prescriptions and diagnoses at an individual patient level, alternative data sources exist to avoid added costs. When analysing antibiotic usage in England, PHE's Fingertips database [145] and the OpenPrescribing project run by University of Oxford [146] warrant special mention. Both databases are publicly available and are accessible free of charge.

Fingertips was developed by PHE and provides summary information on all primary care practices with a list size of at least 800 patients participating in

the Quality and Outcomes Framework [145]. Data are available by year and for predefined indicators, including antimicrobial resistance. While Fingertips represents a valuable resource for the comparison of practice performance on predefined indicators, its value for research on UTI is limited. The only available indicator related to UTI is the ratio of trimethoprim to nitrofurantoin prescribing, which is an indicator for the adherence to national prescribing guidance [147].

OpenPrescribing is an interface to raw prescribing data published monthly by NHS Digital [146]. As with PHE Fingertips, data are aggregated at the practice level. OpenPrescribing contains more detailed information on primary care prescribing, as it is not limited to indicators but covers all issued prescriptions. It does not contain any other demographic or clinical information on patients. Name, size, location, and related CCG are available for all practices. Since prescribing cannot be linked back to clinical diagnoses, however, the usefulness of OpenPrescribing for research on UTI is limited to the comparison of prescribing volumes of antibiotics (almost) exclusively used for the treatment of UTI — i.e., trimethoprim and nitrofurantoin [147].

3.3 Secondary care

Secondary care in the UK provides specialised care (e.g., elective operations) that can't be offered by a GP, and is usually based in hospital. Secondary care further includes urgent or emergency services for acute conditions such as severe infection or sepsis. For the purposes of this thesis, the term secondary care also included highly specialised tertiary care.

3.3.1 Hospital Episode Statistics (HES)

Hospital Episode Statistics (HES) is the largest secondary care database in England. Hospital discharge information are submitted monthly by each trust to the central Secondary User Service (SUS) data warehouse, cleaned, and annually extracted as HES after a final update and approval by each trust [148]. Since 2004, SUS data has also been linked to reimbursement as part of the Payment by Results

programme [148]. HES covers all NHS CCGs in England[3] and provides data on hospital admissions dating back to 1989, outpatient appointments from 2003/04, and emergency department (ED) visits since 2007/08 [148]. These different types of hospital activity are separated into distinct but linkable databases. HES Admitted Patient Care (HES APC) data covers all inpatient activity and is structured into finished consultant episodes (FCEs), which are defined as the time spent under the care of one consultant at a particular hospital [148]. Moving wards might or might not start a new episode, depending on whether the patient's main clinician changes or not. All episodes that make up a single hospital stay are combined within so-called hospital spells [148]. A spell combines one or more consecutive FCEs. HES Outpatient records cover patients that were referred to hospital but who were not assigned a bed and do not stay overnight. HES Accident & Emergency (A&E) records, on the other hand, relate to a single visit to the ED, which may or may not be followed by an admission to hospital[4]. Patients in each of these databases are identified by a pseudonomysed patient identifier called a HESID, which is uniquely generated for each data extract provided to researchers in order to minimise any risk of cross-referencing between unrelated extracts. Additional information on adult critical care and maternity episodes is available on reasonable request [148].

3.3.1.1 Available clinical data items

Diagnoses HES contains up to 20 diagnoses per FCE [148] or outpatient visit [149]. Recorded diagnoses might change between FCEs that are part of a single hospital spell, reflecting changes in the patient's status, test results and investigations. Diagnoses are encoded using the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10). Per FCE or outpatient visit, there is one primary diagnosis which represents the main reason for treatment [148]. All remaining codes are secondary health conditions relevant to the FCE — e.g., diabetes if diabetes wasn't the main reason for admission. The order of secondary codes has no inherent meaning. Unlike for

[3] Excluding private hospitals.
[4] An admission following an ED visit would not be part of HES A&E but is instead covered by HES APC.

example in CPRD (see Section 3.2.1), where diagnoses are usually recorded by a GP or nurse [121], diagnoses in HES are retrospectively recorded by accredited clinical coding personnel based on written discharge summaries provided by the treating physician [148]. In addition to diagnoses codes, HES also contains information on procedures performed while in hospital. Procedures are encoded using the Office of Population Censuses and Surveys' Classification of Surgical Operations 4th revision (OPCS-4) and up to 24 procedures can be entered into the system [148, 149]. HES A&E data uses its own, small set of 39 diagnosis codes together with codes on investigations and treatments performed while the patient was in A&E [150].

Remark (SNOMED-CT in English secondary care). Like primary care, hospitals in England have been required to use SNOMED-CT by April 2020 [130]. Their use is limited to document direct patient care, whereas ICD-10 and OPCS-4 codes will continue to be used for national reporting including HES.

Prescriptions, clinical observations and laboratory tests Unlike the primary care databases covered earlier, HES is not directly extracted from patient management systems but is instead based on simplified data submitted for high-level service evaluation and reimbursement [148]. It therefore neither contains electronic prescribing information[5] nor information on laboratory tests or vital signs.

3.3.1.2 Linkage to other EHR databases

HES data from APC, Outpatients, and A&E provided as part of a single extract may be linked using the HESID provided alongside the extract [148]. HES data is further routinely linked to the primary care databases described earlier as well as to census data from ONS. Additional linkage to other databases like the National Joint Registry, the UK Renal Registry and cohort studies such as Whitehall II has been previously performed and may be available to researchers upon reasonable request [64]. Besides existing linkage, HES can be custom linked to further data sources

[5] Electronic prescribing is largely unavailable in NHS trusts, with a report by the Nuffield Trust from 2016 stating that it had only been commissioned in 12% of trusts [67].

using NHS Digital's trusted third-party bespoke linkage service.

3.3.1.3 Recording of UTI

Due to the lack of prescribing and laboratory data in HES, the identification of UTI cases in HES exclusively rests upon coded diagnoses. While secondary diagnoses and procedures allow for the identification of underlying comorbidities and conditions predisposing to UTI, no information is available about the acute health state of the patient. Microbiological confirmation of UTI is not available in HES, nor is any information on treatment choice. Due to the coarse nature of HES A&E diagnosis codes, ED visits for UTI cannot be distinguished from other urological conditions [150].

Issues arising from this paucity of information on the reason for admission and treatment in secondary care are discussed in more detail in Chapter 4 in relation to a study of the association between delayed or withheld antibiotics for community-onset lower UTI in primary care and subsequent risk of infectious complication and/or hospitalisation.

3.3.2 Queen Elizabeth Hospital Birmingham (QEHB)

There is currently no national secondary care EHR resource that includes detailed information on diagnoses, prescriptions and laboratory data. A continuing paucity of electronic prescribing systems and lack of IT system integration prevents the collection of this data on a national scale [67]. However, more detailed data for research on secondary care may be obtained from individual, digitally-mature hospital trusts like Queen Elizabeth Hospital Birmingham (QEHB) — although at the loss of national generalisability.

QEHB is a large tertiary teaching hospital located in central Birmingham in the English Midlands. It was re-opened in 2010 to replace a previous hospital of the same name. Healthcare activities were moved to the new hospital complex starting in June 2010 and the transfer was finished in November 2011. The complete restructuring of hospital infrastructure allowed the hospital to also invest in new technologies. As a result, University Hospitals Birmingham NHS Foundation

Trust — which includes QEHB — was later chosen as one of twelve centres of digital excellence within NHS England [151]. Comprehensive electronic patient management systems at QEHB capture high-quality clinical data including patient demographics, diagnoses, prescribing, clinical observations, examinations, and laboratory results starting from 2010/11 [152].

Unlike the data available in HES, which is collected for reimbursement and reporting reasons [148], data stored in QEHB's patient management systems is stored primarily to aid patient management. Clinical data are recorded digitally by healthcare personnel to inform clinical decision making. As far as this information is entered in a structured format — and not as free-text information — it may allow to approximate the information available to clinicians and other healthcare professionals at the time of decision making and create models that can better support the diagnosis and management of UTI in secondary care.

3.3.2.1 Available clinical data items

Diagnoses QEHB separately records (suspected) diagnoses made in the ED and discharge diagnoses for admitted patients. Up to October 2011, ED diagnoses were captured in QEHB's ED patient management system using UK national A&E diagnosis codes — the same set of 39 codes used in HES A&E [150]. These codes were replaced with a more detailed set of 462 local, bespoke codes in November 2012, and recording was changed again to conform with the recording standard used in the UK Emergency Care Data Set (ECDS) in December 2017. ECDS codes include 762 possible diagnoses that are mapped to a condensed set of SNOMED-CT codes [153]. Discharge diagnoses and procedures for patients admitted to QEHB are recorded using ICD-10 and OPCS-4. The data structure for discharge diagnoses and procedures resembles that employed by HES, with diagnoses and procedures grouped by finished consultant episodes. Each episode has one recorded primary diagnosis and up to 23 secondary diagnoses. Episodes could have up to 24 recorded procedures. The first 20 diagnoses and 24 procedures are reported for reimbursement to HES, and should be identical for the subset of

HES records pertaining to QEHB[6].

Prescriptions Data at QEHB includes electronic prescribing records and medication administration records stored within the hospital's Prescribing Information and Communications System (PICS). Available information includes the medications' name and substance, date and time of prescribing, formula (e.g. tablet or injection), route of administration (e.g. oral or intravenous), frequency of administration (e.g. once a day or three times a day), and dosage. For medications administered in hospital, the system further records the date and time of each administration.

Investigations and laboratory tests PICS also records the results of clinical observations and investigations, including vital signs like heart rate, respiratory rate, systolic blood pressure, body temperature, oxygen saturation, AVPU score[7] as well as blood measurements such as white blood cell counts, red blood cell counts, C-reactive protein, creatinine, bilirubin, and alkaline phosphatase. While generally available within the dataset, availability of these variables may differ by ward. For example, the ED at QEHB has not always been fully electronically integrated with PICS and data on some measurements therefore might not be available for patients seen in the ED.

Microbiology Data at QEHB further includes patient-linked laboratory data on all flow cytometry and microbiological cultures performed at the in-house laboratory. Available information includes sample type, time of collection, time of receipt in the laboratory, time the results were available to the clinician, performed tests, cultured isolates, and sensitivity to individual antibiotics.

3.3.2.2 Linkage to other EHR databases

Patient data from QEHB may be linked to national HES records and ONS death records provided appropriate ethical approval is sought. These additional data may allow to track a patient's medical care outside of QEHB, providing a more reliable and comprehensive picture of his or her medical history.

[6] This is difficult to ascertain, however, since individual hospitals are not identifiable within HES.

[7] The AVPU score is a measure of consciousness in which patients are rate as either alert (A), responsive to voice (V), responsive to pain (P), or unconscious (U).

3.3.2.3 Recording of UTI

The higher depth and breadth of data recorded at QEHB allows for a potentially more reliable identification of UTI diagnoses and management decisions taken as a result of such diagnoses. Patients with (suspected) UTI consulting the ED at QEHB may be identified using suspected ED diagnosis, requests for and results of urine culture, as well as prescription and administration of antibiotics. Although neither free-text nor a full list of presenting complaints in the ED is accessible, ED diagnoses and urine culture results may be used to define clinically-confirmed bacterial UTI (see Section 1.3 for a detailed discussion on the diagnosis of UTI, and Chapter 5 for definition of UTI at QEHB). Supplementary information on vital signs and blood parameters measured on arrival in the ED — if widely available — may further allow to infer the patient's approximate health status on arrival. Prescription records may be used to infer empirical treatment regimes in the ED. If the patient was admitted to hospital as a result of their visit to the ED, discharge diagnoses may be used to confirm UTI as the primary reason for admission to hospital, or investigate discrepancies between suspected diagnoses in the ED, microbiological evidence, and final diagnoses listed on the discharge summaries [24]. Prescribing and administration records may further be used to define ongoing antibiotic treatment and assess changes in treatment in response to culture results — e.g., a step down to narrow spectrum antibiotics.

3.3.3 Other sources of secondary care data in England

3.3.3.1 Hospital Treatment Insights

The HES database described earlier does not include information on drug use in hospital. This prevents it from being used for research on antibiotic use in secondary care, or indeed most pharmacoepidemiological studies. However, since HES may be linked to other databases, data can be added that is missing within HES itself. Hospital Treatment Insights (HTI) is one such linkage, combining HES data for a subset of 43 out of 153 English hospital trusts with dispensing information from their hospital pharmacies [136]. Due to the way that drug dispensing is captured

in pharmacy records, though, care needs to be taken when using HTI. Pharmacy records only contain a linkable patient identifier if a drug was specifically requested for a named patient. If a drug was dispensed to a ward in bulk to be used on demand, on the other hand, pharmacies do not retain information on the patient and these records are therefore notably missing in HTI. I previously showed that this severely affects the recording of antibiotics in HTI [136], limiting its usability for research on UTI.

3.3.3.2 Discover and KID

Both the Discover dataset in North West London [140] as well as the KID in Kent and Medway [144] discussed earlier in Section 3.2 also capture information on secondary care visits. Both datasets obtain secondary care information from SUS, and available data items are therefore highly comparable to those derived from HES.

3.3.3.3 PHE Fingertips

In addition to data on GP practices, PHE Fingertips also captures summary indicators of antibiotic prescribing and antimicrobial resistance for English acute hospital trusts. Available indicators for UTI include the percentage of antibiotic prescriptions for lower UTI in older people meeting NICE and PHE guidance and the percentage of *E. coli* blood specimens susceptible to gentamicin, ciprofloxacin, piperacillin / tazobactam, cephalosporins, and carbapenems [145].

3.4 Conclusion

Several EHR data sources are available to investigate the diagnosis and management of UTIs in English primary and secondary care. Most allow for the identification of individual patients presenting with (suspected) UTI and — depending on the database — may further allow to establish the presence of diagnostic criteria (e.g., recorded symptoms and/or urine culture), ascertain the use of antibiotic treatment, or monitor subsequent complications of infection. Linkage between datasets — e.g., by linking primary and secondary care data from CPRD and HES — may further enhance the information captured in any single data source, providing a more comprehensive and perhaps less biased picture of UTI. However, none

Table 3.2: Summary of information on the diagnosis and management of UTI available in major English EHR datasets.

Dataset	Setting	Scope	Diagnoses				Treatment	
			SYM	DIAG	UA	UC	PRSC	RSN
CPRD	Primary	National	∼	+	–	–	+	–
THIN	Primary	National	∼	+	–	–	+	–
QResearch	Primary	National	∼	+	–	–	+	–
ResearchOne	Primary	National	∼	+	–	–	+	–
OpenPrescribing	Primary	National	–	–	–	–	+	–
Discover	Prim./Sec.	Local	∼	+	–	–	+	–
KID	Prim./Sec.	Local	∼	+	–	–	+	–
PHE Fingertips	Prim./Sec.	National	–	–	–	∼	∼	–
HES	Secondary	National	–	+	–	–	–	–
QEHB	Secondary	Local	∼	+	∼	+	+	–
HTI	Secondary	National	–	+	–	–	∼	–

+ indicates that information is generally available within the dataset, ∼ indicates that information is partially available in the dataset, – indicates that information is not available in the dataset. Please note that this table represents a simplified summary of the available data. For example, CPRD sometimes contains information on urinary symptoms or urine cultures. However, these data have been shown to be seldom recorded and were thus considered not available. Please refer to the detailed descriptions of each dataset for a detailed account of available data items.

CPRD, Clinical Practice Research Datalink; DIAG, Diagnoses; EHR, Electronic health records; HES, Hospital Episode Statistics; HTI, Hospital Treatment Insights; KID, Kent Integrated Dataset; PHE, Public Health England; PRSC, Prescribing; QEHB, Queen Elizabeth Hospital Birmingham; RSN, Reason for prescribing (or non-prescribing); SYM, Symptoms; THIN, The Health Improvement Network; UA, Urinalysis; UC, Urine culture; UTI, urinary tract infection.

of the sources discussed in this chapter are perfect. They often miss detailed information on the patients' health status and likelihood of true bacterial UTI at initial consultation, and make it difficult to reconstruct the rationale behind treatment choices (Table 3.2). As discussed in Chapter 2, capturing this information is crucial in order to be able to reliably account for diagnostic uncertainty and differential treatment assignment.

Following this reasoning, an ideal UTI dataset would enable to capture all information available to the treating physician(s). This involves a comprehensive documentation of all observed clinical information and diagnostic reasoning, including presence or absence of urinary symptoms, the results of any performed diagnostic tests, and an indicator for the (perceived) severity of disease. In addition, an ideal dataset would capture the treatment decisions in detail. It would not only record the type and quantity of all received antibiotics but also document any reasons for deviation from standard practice — e.g., prescribing broad-spectrum

instead of narrow-spectrum antibiotics, or opting to treat with delayed antibiotics. It would provide the same detail of information for all past instances of UTI as well as across both primary and secondary care. Finally, an ideal dataset would allow to track the progression of disease from the onset of symptoms (as reported by the patient) until the infection resolves, including adverse outcomes such as treatment failure, recurrence of infection, progression to pyelonephritis or sepsis, and death. It would clearly allow to link these outcomes to the initial infection — e.g., by distinguishing urosepsis from sepsis with other infectious origin [133].

Having reviewed previous studies using EHR data to investigate UTIs (Chapter 2) and the breadth and depth of available EHR data sources in England (this chapter), the following Chapters 4-6 examine how real-world data from CPRD-HES-ONS and QEHB can be used to answer concrete research questions on the diagnosis and management of UTIs.

Chapter summary

- Several nation-wide pseudonomised primary care databases exist in England and the rest of the UK, which collect medical information directly from the patient management software of participating practices.
- Identification of UTI in English primary care databases relies primarily on diagnosis codes entered manually by the physician, whereas other key information such as urinary symptoms, diagnostic tests, or urine cultures are either not performed or not well recorded.
- Prescriptions are automatically captured in most of primary care but the decision process leading to the prescriptions is not.
- Nation-wide recording of secondary care data is currently limited to diagnosis and procedures, and no prescribing information is captured due to a continuing lack of wide-spread electronic prescribing.

- Information collected at local, digitally mature hospital sites offers a more detailed picture of care processes, and includes data on prescriptions, observations, laboratory tests, and microbiology that may allow for a better classification of cases.

Chapter 4

Can EHR data guide the management of UTI in primary care: a case study using linked data from CPRD to evaluate the relationship between prescribing and risk of adverse outcomes

Abstract

Introduction: Previously, I discussed how uncertainty in the diagnosis of community-onset urinary tract infection (UTI) influences how clinicians manage this infection, and how routinely-collected electronic health records (EHRs) have been used to aid clinicians in their decision making. In this chapter, I explore the use of primary care data from the Clinical Practice Research Datalink (CPRD) to support the management of uncomplicated, community-onset UTIs in women.

Background: A key question in the management of community-onset UTI is whether it is safe to delay issuing an antibiotic prescription and see if symptoms resolve without treatment. This approach is widely used for patients with respiratory tract infection, but evidence to support the use of this approach for UTI based on randomised controlled trials is conflicting. The aim of this study was to explore whether analysis of primary care EHRs can support this evidence base and inform decisions around the safety of delaying or withholding antibiotics for community-onset lower UTI in adult women of different age-groups.

Methods: I undertook a retrospective cohort study of the association between delayed or withheld antibiotics and adverse outcomes within 30 days — including progression to severe UTI (pyelonephritis, sepsis, hospitalisation for UTI), death, and hospitalisation for reasons unrelated to UTI — in adult women who consulted for community-onset lower UTI in primary care between 2007 and 2015. I used English data from CPRD linked to Hospital Episode Statistics (HES) and census data. Delayed or withheld prescribing was defined as the absence of antibiotic prescribing on the day of initial consultation for UTI. I estimated the associations between delayed or withheld prescribing and adverse outcomes for different age groups — accounting for patient characteristics including demography, comorbidities, and medical history. I assessed the impact of residual confounding using propensity score analysis (PSA) and coarsened exact matching (CEM).

Results: I observed 650,416 episodes of community-onset UTI among 1.9 million women providing 8.4 million patient-years at risk, of which 589,063 (90.6%) were treated with antibiotics immediately. Progression to severe UTI and death within 30 days was seen in 7,947 (1.2%) and 2,320 (0.4%) episodes respectively. Delaying or withholding antibiotics was associated with increased odds of progressing to severe UTI (adjusted odds ratio [aOR] 1.61, 95% confidence interval [95% CI] 1.51–1.72, p-value < 0.001) and dying (aOR 1.45, 95% CI 1.29–1.62, p-value < 0.001). However, it was also associated with higher risk of hospitalisation for reasons unrelated to UTI, and patients with delayed or withheld antibiotics were more likely to have had recorded risk factors pre-disposing to adverse outcomes. Results remained unchanged when applying PSA or CEM.

Discussion: Patients with delayed or withheld antibiotics in this study were at an estimated higher risk of infectious complications and death. Difficulties in ascertaining severity of infection and reasons for delaying or withholding antibiotics cast doubt on the validity of the obtained estimates. Better recording of (urinary) symptoms at initial presentation, vital signs and other markers of severity of disease, and a decision to delay prescribing are required to fully understand and analyse delayed prescribing strategies in EHR data.

4.1 Introduction

General Practitioners (GP) in England on average have eleven minutes to spend on a patient during a routine consultation [154]. In those eleven minutes, they need to enquire about the presenting complaint, make clinical observations and measurements, derive a likely diagnosis, and make a treatment decision. This time pressure in primary care — combined with limited resources — may impact patient management, affording GPs less time to perform thorough anamnesis and clinical investigations [18]. In the context of urinary tract infection (UTI), lack of resources and time are exacerbated by the absence of fast, cheap, and unambiguous diagnostic tests (see Section 1.3 for a detailed discussion). Faced with time pressures and limited information, GPs need to make decisions about whether the patient requires antibiotic treatment, and whether this should be initiated immediately or whether it is worth delaying to see if the patient improves without antibiotics. The decision must balance the risk of adverse outcomes associated with delaying antibiotic treatment for genuine bacterial infection (e.g., progression to pyelonephritis or sepsis) versus the adverse consequences of unnecessary antibiotic treatment for the individual (e.g., antibiotic side effects or future drug treatment failure) and the wider population (e.g., increased antimicrobial resistance).

Chapter 2 discussed how large-scale electronic health records (EHRs) coupled with appropriate statistical analysis may support GPs in making these decisions. By pooling information across a wide range of patients, EHR data may provide robust and reliable evidence on which GPs may base their empirical decision making. How well evidence derived from EHRs can support GPs, however, depends on the depth and breadth of information captured within primary care. In Chapter 3, I concluded that some of the information central to the investigation of UTIs — such as urinary symptoms, urinalysis results, urine culture results, and reasons for or against a decision to prescribe antibiotics — are mostly missing from English primary care datasets. It remains unclear to what extent this sparsity has affected previous studies reviewed in Chapter 2 and whether the information that *is* captured routinely is sufficient to avoid biased results. In order to investigate this question,

in this chapter I use linked data from the Clinical Practice Research Datalink to quantify the risk of delaying or withholding antibiotic treatment in adult women consulting with suspected community-onset lower UTI. I critically assess whether the available data allows for the reliable estimation of those risks, and how my results inform the way currently available EHR data can and cannot be used to guide GP decision making for the treatment of community-onset lower UTI.

4.2 Background

Frequent antibiotic use is a major driver of emerging antimicrobial resistance [75], and the reduction of unnecessary antibiotic prescribing has become a national priority [97]. More than 90% of patients consulting with uncomplicated lower UTI in English primary care are prescribed antibiotic treatment [76]. Given the potential impact on antimicrobial resistance, reliance on antibiotics as a universal treatment for lower UTI has therefore recently been questioned [68]. Several randomized controlled trials suggested that it might be safe and feasible to delay antibiotic prescribing for young, non-pregnant women (see Section 1.4.1 for a definition of delayed antibiotic prescribing) [63, 69, 70, 71, 155]. However, this conclusion was contradicted by two recent trials which reported an increased incidence of pyelonephritis among women whose lower UTI was treated with painkillers instead of antibiotics [156, 157]. The true increase in the risk of infectious complications due to delayed prescribing remains uncertain. Sample sizes in previous studies were relatively small — the maximum number of patients per study was 1,000 and most had less than 400 — and pyelonephritis continued to be a rare outcome even in patients who were not treated with antibiotics, with less than 7 (<4%) cases of pyelonephritis in any single study.

Evidence for the safety of delaying antibiotic prescribing for uncomplicated lower UTI in elderly patients ($\geq$ 65 years of age) is even sparser. Although some of the above mentioned trials included women up to the age of 90 years [63, 69, 70], the majority of included patients was younger and no age-stratified risks of infectious complications were reported. This lack of evidence was recently

addressed by Gharbi *et al.* (2019) [112] who published a retrospective observational study based on large-scale EHR data that assessed the relationship between delaying antibiotic prescribing for suspected UTI and risk of bloodstream infection or death within 60 days in elderly patients (both men and women). They used data on more than 300,000 UTI episodes from the Clinical Practice Research Datalink (CPRD) between 2008 and 2015. The large sample size provided sufficient statistical power to investigate even rare outcomes. They observed 1,539 (0.5%) cases of sepsis and 6,193 (2.0%) deaths within their study population and estimated that delaying or withholding antibiotics increased patients' odds of sepsis 7–8-fold [112]. If true, these results would raise serious doubts about the safety of delaying antibiotics in elderly patient populations. However, baseline characteristics of patients with and without antibiotic prescribing differed considerably in this study, and several clinicians and researchers have expressed concerns that the results may be affected by confounded by indication [72].

Against the backdrop of emerging antimicrobial resistance, strategies like delayed antibiotic prescribing have become more popular in managing respiratory infections. Importantly, the safety of delayed antibiotic prescribing in respiratory infections was confirmed in a large systematic review [158]. Similarly clear evidence for the use of delayed prescribing in UTIs is lacking and remains controversial. The conflicting evidence on the safety of delayed antibiotic prescribing for lower UTI may deter GPs from considering it as a treatment option in primary care, potentially missing an opportunity to reduce a driver of antimicrobial resistance. Conducting a sufficiently large randomised controlled trial would be costly and time-consuming, however, and is thus unlikely to happen soon. Gharbi *et al.* demonstrated that EHR databases — with their large, nationally representative patient populations — may provide a fast and cheap alternative to generate evidence in the absence of randomised controlled trials, although at the risk of bias and misleading results due to an inherent lack of randomisation.

The analysis presented in this chapter therefore utilises a nationally representative database of primary care records to revisit the evidence for a link

between delaying or withholding antibiotic prescribing in women consulting for community-onset lower UTI in primary care and an increased risk of infectious complications. Using a careful definition of community-onset lower UTI, I estimated the protective effect of immediate antibiotic prescribing and assessed whether conclusions change when considering different age groups. I applied inverse probability weighting and matching strategies to account for previously observed differences in patients who do and do not receive immediate antibiotic prescribing [112], attempting to minimise the effect of confounding by indication on the obtained results. I end the chapter by discussing whether the applied methodology is likely to have mitigated important sources of bias in the analysis and allowed me to obtain reliable estimates of the safety of delaying antibiotic treatment for community-onset lower UTI in primary care.

4.3 Aims and Objectives

To investigate the utility of EHR data in generating reliable estimates of the risk of infectious complications associated with delaying or withholding antibiotics in adult women presenting with community-onset lower UTI in English primary care.

Objectives:

4.1 To identify episodes of community-onset lower UTI in a large routinely collected primary care database.

4.2 To estimate the relative odds of progressing to severe UTI or dying in adult women treated with systemic antibiotics during an initial consultation for community-onset lower UTI in primary care, compared to those that did not receive antibiotic prescribing during their initial consultation.

4.3 To investigate variations in the estimated protective effect of antibiotic prescribing by patient age at initial consultation.

4.4 To perform secondary analyses of the relationship between immediate antibiotic treatment and hospitalisation for lower respiratory tract infection

(LRTI) or hospitalisation due to other causes.

4.5 To use inverse probability weighting and matching procedures to evaluate the presence of — and account for — observed confounding factors associated with a decision to prescribe, delay or withhold antibiotics during a patient's initial consultation.

4.4 Methods

Study design: Retrospective observational cohort study.

Study setting: 674 English primary care practices.

4.4.1 Data source and management

Data for this chapter was extracted from retrospective electronic primary care records available in the CPRD database [121]. For more information on CPRD and the information collected within CPRD, please refer to Chapter 3 where CPRD is introduced in detail. Briefly, CPRD is a large, routinely collected database with information on patient consultations from participating primary care practices. Data for this chapter was taken from the English subset of CPRD practices (75% of English practices, 58% of all UK practices) which had a valid link to secondary care (Hospital Episode Statistics [HES]) and census data (Office for National Statistics [ONS]). Data from HES included both data from admitted patients as well as information from emergency department (ED) visits.

Raw pseudonomysed patient data from CPRD was extracted by Dr Kenan Direk and transferred into a dedicated environment within the UCL Data Safe Haven. All further data processing and analyses were performed by me within this secure environment. Due to the large number of patients included in the extract, I imported the raw data into a relational SQL database set up specifically for this purpose. Patient cohorts within the raw data were identified using codelists published previously [112] and refined in close collaboration with Dr Laura Shallcross and Dr Anna Aryee (see Appendix H for a list of included codes). Additional raw secondary care data from HES and census data from ONS were

extracted by NHS Digital after I provided them with a list of pseudonomysed patient identifiers for all included patients. Episodes in the subset of HES linked to CPRD were extracted by NHS Digital and transferred to the UCL Data Safe Haven, where they were stored in the same database as the previously described CPRD records. I included all ICD-10 diagnoses and OPCS-4 procedures in the extract that related to infection — including UTI and alternative diagnoses such as lower respiratory tract infection — and/or adverse outcomes of infection (see Appendix H), as well as any records available for these patients within HES A&E.

Remark (UCL Data Safe Haven). The UCL Data Safe Haven is an ISO 27001 — an international standard for information security management — certified research environment that provides secure storage of, access to, and analysis of patient data. All analyses in this thesis were performed within this Data Safe Haven.

I linked data from primary care, secondary care and census using a common patient identifier. I identified the patient cohort for the analysis described in this Chapter from the SQL database using the R programming language [159] and included information on patient demographics, consultation dates, and medical diagnoses. All data was converted into a single matrix containing one row per UTI episode.

4.4.2 Ethical approval

CPRD requires researchers to obtain ethical approval from the Independent Scientific Advisory Council (ISAC; https://www.cprd.com/ISAC/), an independent review panel appointed by the Medicines and Healthcare products Regulatory Agency (MHRA). A separate ISAC protocol must be completed for each intended study, specifying the study background, cohort definition, sample size calculations, analysis plan and code lists. Changes made to the protocol after approval must be submitted as amendments to the protocol.

I obtained access to the CPRD linked to HES and census data as part of the

Preserving Antibiotics through Safe Stewardship (PASS) programme grant. I wrote the first draft of the ISAC protocol (protocol Nr. 17 048), which was revised and edited by my supervisors (Dr Laura Shallcross and Prof Andrew Hayward). The protocol with the title *The use and protective effect of antibiotics against complications of infection in patients in primary care: a cohort study using linked data from CPRD, HES, and ONS* was approved on May 19th 2017.

4.4.3 Patient population

All female patients in CPRD were included in the analysis if they had been registered with a participating practice between April 1st 2007 and March 31st 2015. Patients were only included if their records were up-to-standard as defined by CPRD [121]. Patients entered the study population on the latest of the following dates: one year of continuous registration, the practice's up-to-standard date, the date the patient turned 18 years, or April 1st 2007. Patients exited the cohort at their date of death or 30 days before the earliest of: date the patient left the practice, or 31st March 2015. Start and end dates chosen for this analysis correspond to the NHS financial year, which starts in April and ends in March of each year.

The analyses reported in this chapter formed part of a wider range of analyses undertaken as part of the *Preserving Antibiotics through Safe Stewardship (PASS)* project, which included the investigation of patterns of antibiotic prescribing to inform the "co-design [of] antimicrobial stewardship interventions across healthcare settings, by integrating data-science, evidence-synthesis, behavioural-science and user-centred design" [160].

4.4.4 Episodes of community-onset lower UTI

In order to mimic the conditions analysed in earlier randomised controlled trials as closely as possible, the primary unit of observation for this analysis were episodes of new community-onset lower UTI in adult women [63, 69, 70, 71, 155, 156, 157].

4.4.4.1 Defining UTI episodes

In order to define the start date of each UTI episode, I identified all primary care consultations from CPRD or hospital admissions from HES that were associated

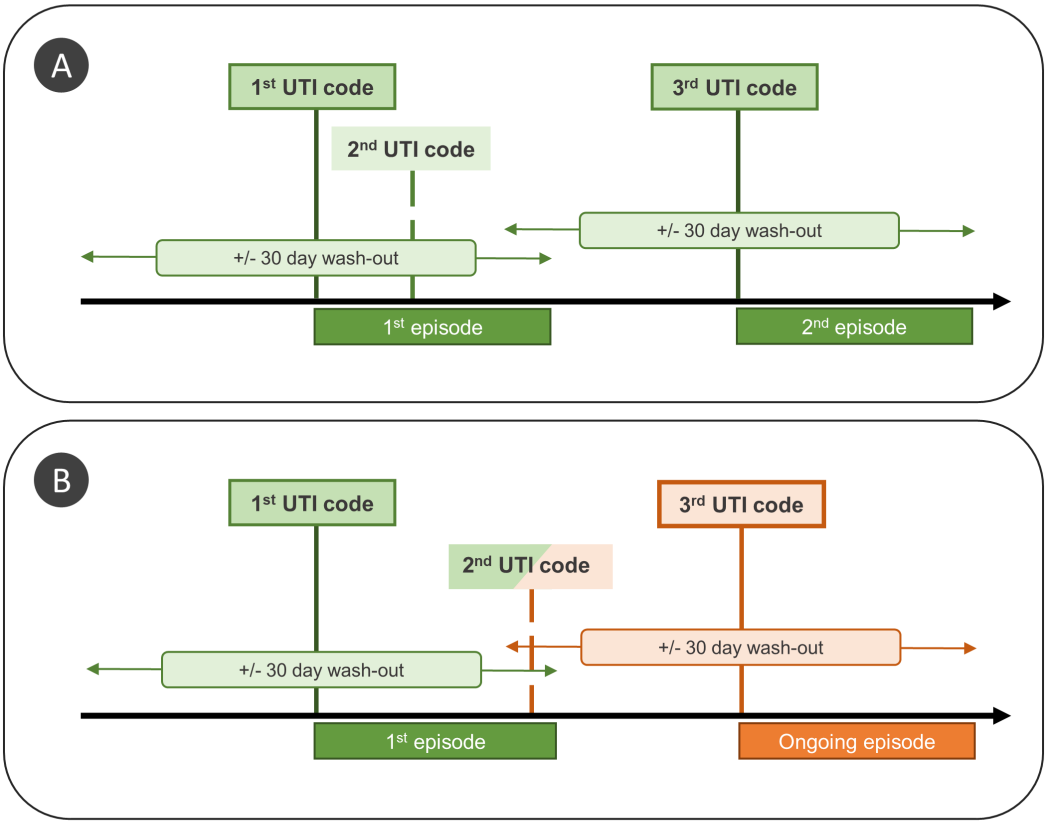

Figure 4.1: Classification of UTI episodes for two scenarios of a patient with three records of UTI, which are identical except for the timing of the second UTI code. In both panels the first UTI code marks the start of a new UTI episode (first episode). The second UTI code occurs within 30 days and is therefore considered to be part of the first episode. The third UTI code occurs more than 30 days after the start of the first episode and is classified as: **A)** a new episode because the last evidence of UTI was recorded more than 30 days earlier; **B)** an ongoing episode that is excluded from the analysis because the last evidence of UTI — i.e., the second UTI code — was recorded less than 30 days before and may therefore represent an ongoing episode of infection.

UTI, urinary tract infection.

with a relevant diagnosis code. UTI was defined using a previously published codelist which included codes for both lower UTI and pyelonephritis (see Appendix H for a list of all included codes). Events in CPRD were identified using Read codes, the standard medical terminology used in English primary care until April 2020. Events in HES were identified using 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes. The resulting data were a list of recorded UTI events for each patient — including both primary care consultations and hospitalisations — each of which considered the possible start of a UTI episode. All UTI *events* in a patient's history were sorted chronologically and the earliest observed event was set as the start date of that patient's first UTI *episode*[1], irrespective of whether it originated in primary or secondary care. I then looked at each subsequent UTI event in order. If the second event was recorded within 30 days of the current episode's start date, the event was considered to be part of the same episode of care (Figure 4.1 A). For example, we could imagine a patient whose symptoms did not resolve after an initial consultation (episode start) and who attended for a follow-up consultation two weeks after the initial visit. This follow-up visit should not be considered the start of a new episode, as it is a result of the original infection and thus part of the same episode. On the other hand, if the second event in a patient's history was recorded more than 30 days after the most recent episode start, I assumed the code was the beginning of a new episode and compared all subsequent codes to the new episode start. After considering all codes of a single patient, each code had been mapped to exactly one episode.

A wash-out window of 30 days was chosen for the definition of UTI episodes to force enough time between consecutive episodes without unduly excluding truly distinct episodes. Even when using a 30-day wash-out window, situations may arise in which the initial episode had resolved but another infection was acquired within the wash-out period (Figure 4.1 B). For example, a patient might have had

[1] Note the difference between events and episodes. A UTI event is each Read code or ICD-10 code for UTI linked to a consultation in CPRD or hospital admission in HES. A UTI episode is the period of time for which a patient was considered to suffer from the same, continuous infection.

a recorded UTI event on day one (episode start) and two more UTI events on days 28 and 40. Applying the above classification rule, we would assign the event on day 28 to the first episode and treat the event on day 40 as the start of a second episode. However, it is unclear whether the event on day 28 truly belonged to the first episode, or whether in fact it was already the start of the second episode[2]. To guard against such scenarios, I further required truly new episodes to have no other UTI code within 30 days prior to the episode start date. All episodes that did not fulfil this requirement were considered ongoing episodes and were excluded from the analysis (Figure 4.1 B).

4.4.4.2 Defining community-onset lower UTI

Each identified episode was classified as describing a lower UTI or pyelonephritis, depending on the diagnostic code recorded at episode start. If both lower UTI and pyelonephritis codes were entered at the initial encounter, the episode was classified as pyelonephritis. Episodes were also classified into UTIs initially treated in the community — i.e., the first record of the episode was found in primary care records — and UTIs originating in hospital — i.e., the first record of the episode was found in hospital records. Episodes for which there was a record of UTI in both primary and secondary care on the day the episode started were considered to have been treated in hospital. To ensure that I was only including uncomplicated lower UTIs originating in the community, I then discarded all episodes that were classified as upper UTI or that were treated in hospital. I further excluded episodes in which the patient attended the ED after seeing his or her GP, was referred to specialist care, or died on the same day as the episode start. Episodes were also excluded if the linked HES record showed that the patient was an inpatient in hospital on the date that their UTI episode was recorded as having supposedly started in primary care.[3]

[2] Note that the second event is closer in time to the third event than to first.

[3] The exact reason for why this might occur in the dataset is unclear. A possible explanation is that GPs retrospectively entered information from hospital discharge letters into their practice management systems and backdated them to the day when the patient was seen in hospital.

4.4.4.3 Urinary symptoms and dipstick results

For each episode of community-onset lower UTI, all recorded information on urinary symptoms and urine dipstick tests on the day of initial consultation and within 30 days prior to initial consultation were extracted from the database. Urinary symptoms were categorised into dysuria, haematuria, incontinence, urinary urgency, abnormal appearance of urine, malodorous urine, and difficulty urinating (see Appendix H for a list of Read codes associated with each condition). For dipstick tests, recorded use of the test as well as recorded positive results for leukocyte esterase (+ to +++), nitrites (positive, negative), and blood (+ to +++) were similarly identified using Read codes.

4.4.5 Exposure

The main exposure of interest was delayed or withheld treatment with systemic antibiotics, defined as the *absence* of a prescription recorded on the same day as the episode start date — i.e., the date of primary care consultation for lower UTI. Systemic antibiotics were defined as a prescribing record for any oral or intravenous drug included in the British National Formulary chapter 5.1, excluding anti-tuberculosis (5.1.9.) and anti-leprotic drugs (5.1.10). This definition was chosen in concordance with recent literature to capture all systemic antibiotic prescribing while excluding other forms of antibiotics such as inhalers, eye drops, and creams unrelated to bacterial UTIs [6, 161]. Based on the presence or absence of a prescribing record for systemic antibiotics on the day of episode start, I classified patients into two groups: patients who had a record of antibiotic prescribing on the same day as their episode start date (immediate prescribing) and patients who did not have a record of antibiotic prescribing when their episode started (delayed or withheld prescribing). Since delayed prescribing is not well recorded in CPRD [139], no additional attempt was made to identify delayed prescribing using Read codes, or to distinguish between delayed and withheld prescriptions.

4.4.6 Outcome

The primary outcome of interest was progression to severe UTI defined as either a recorded diagnosis of pyelonephritis or bloodstream infection in primary care or a hospitalisation related to UTI — including lower UTI, pyelonephritis, and bloodstream infection — within 30 days of the episode start. Secondary outcomes included all-cause mortality defined as any death recorded by ONS, hospitalisation for LRTI, and hospitalisation due to reasons other than UTI, LRTI or sepsis. The latter two outcomes were included as control conditions to assess association with outcomes not directly related to UTI. All secondary outcomes were also ascertained within 30 days of episode start. Sepsis was identified using ICD-10 diagnosis codes only, as no microbiological outcomes were available. For the purposes of this analysis, the terms bloodstream infection, bacteriuria, sepsis, severe sepsis, and septic shock were used interchangeably. Coding granularity did not allow me to distinguish between these more nuanced medical terminologies. Sepsis was also not limited to urosepsis but instead included sepsis of any origin, again due to a lack of precision in diagnostic coding. All diagnosis codes used to identify the above outcomes can be found in Appendix H.

4.4.7 Covariates

The risk of progressing to more severe disease after consulting for community-onset lower UTI in primary care is potentially confounded by patient characteristics that are associated with both the likelihood of being prescribed an antibiotic and an a priori risk of severe infection or infectious complication. I adjusted for the following risk factors in my analysis: age at study entry, quintile of socio-economic status (Index of Multiple Deprivation [IMD] 2015), practice region (South of England, London, East of England and Midlands, North of England and Yorkshire), Charlson Comorbidity Index (CCI) [162], and smoking status (non-smoker, ex-smoker, current smoker). To account for short-term effects due to general ill-health, recent illness, lack of social support, or exposure to hospital environments, I additionally adjusted for recent hospital stays (discharge from hospital 30 days prior to index visit, number of admissions within one year prior to index visit, and total number

of nights spent in hospital within one year prior to index visit), recent ED visits (visit 30 days prior to index visit, number of visits within one year prior to index visit), and prescription of systemic antibiotics in primary care within 30 days prior to index visit. In an attempt to capture factors that influenced the choice of antibiotic or the probability of a prescription not being recorded, I further considered whether the index consultation was a home visit — which do not allow for automatic capture of antibiotic prescribing and might therefore be wrongly classified as a consultation without immediate antibiotic prescribing [138] — and if the episode represented an episode of recurrent UTI. Recurrent UTIs were defined as an explicit code for recurrent UTI, a prescription of nitrofurantoin or trimethoprim for 28 days or more spanning the episode start date (prophylactic treatment), or two or more consultations for UTI in the 12 months prior to episode start [11]. All variables were ascertained on the date of episode start. CCI and smoking status were calculated using the patients' entire medical history in primary care up to episode start date. Patients without a recorded smoking code were considered non-smokers. Patients whose latest record indicated a non-smoker but who had a previous record of smoking were reclassified as ex-smokers. All numerical variables were included linearly into the analysis, unless explicitly specified otherwise. Covariates were chosen based on previous literature [72, 112], clinical plausibility, and data availability. Notably, no explicitly recorded information was available on the severity of infection or the patient's general health status at the time of consultation. Due to the way variables were derived from the data, no (explicitly) missing data could arise in the analysis presented in this chapter (see Section 5.4.6.2 for a more detailed discussion of missing data in EHRs and ways to account for it).

4.4.8 Statistical analysis

I analysed the association between delayed or withheld antibiotic treatment and subsequent adverse outcomes within 30 days of consultation in primary care using univariable and multivariable regression models. Crude odds ratios (OR) and adjusted odds ratios (aOR) were calculated using generalised estimating equations (GEEs) with a logit link. An exchangeable correlation structure was

chosen to account for correlations between multiple episodes of the same patient. Huber-White robust sandwich estimators were used to derive consistent — although conservative — 95% confidence intervals (95% CI) [163]. The multivariable analysis adjusted for all covariates irrespective of observed strength of univariable association. To provide an alternative, more interpretable estimate of the clinical impact of any observed association between delayed or withheld antibiotics and all included outcomes, I calculated the number needed to be exposed to harm (NNEH) [164]. 95% CIs for the NNEH were obtained by running the same analysis in 200 bootstrapped samples.

Remark (Number needed to be exposed to harm). In this study, the NNEH represents the expected number of people in whom treatment needs to be delayed or withheld in order to observe one additional outcome. The NNEH is calculated by taking the inverse of the average risk difference in the *unexposed* patient population, that is those *with* immediate antibiotic treatment[a]. The NNEH is closely related to the more commonly known number needed to treat (NNT). The NNEH and the NNT are in fact identical if the distribution of covariates in the exposed and unexposed groups are equal, but will differ depending on the covariate patterns if they are not [164].

[a] Note that in this analysis people *with* immediate antibiotic are considered unexposed, since they are not exposed to delayed or withheld antibiotic prescribing strategies. This differs from most other studies where exposure is defined by receiving medication.

I also investigated the impact of age on the estimated effect of antibiotic treatment. I did so in two ways. First, I stratified the analysis by age, repeating the analysis in patients aged <65 years and $\geq$65 years. Second, I included an interaction term between age and treatment status in the full cohort. I considered multiple functional forms of age including linear, quadratic and cubic effects as well as natural cubic splines with one knot (at 65 years) or two knots (at 50 and 85 years). The final form was chosen via the Quasi-likelihood under the Independence Model Criterion (QIC), a quasi-likelihood adaption of the Akaike Information Criterion (AIC) [165].

Remark ($\mathscr{L}_1$ measure). Imbalance in covariates is commonly assessed by calculating the univariate mean differences. However, mean balance on single variables does not guarantee balanced multivariate distributions between treatment groups. Iacus, King & Porro (2011) [166] therefore proposed a multivariate measure of imbalance $\mathscr{L}_1$, which measures the distance between multivariate histograms using the L_1 norm. Matching and weighting methods can thereby be compared in how well they reduced not only univariate imbalance but multivariate differences between treatment groups. On the downside, the value of $\mathscr{L}_1$ depends on an arbitrary choice of bins, which may influence the exact value of $\mathscr{L}_1$ obtained for different balancing methods [166]. Multiple choices of bins should therefore be compared when $\mathscr{L}_1$ is used to compare balancing methods.

Sensitivity to residual confounding was investigated using propensity score (PS) analysis in combination with either matching or inverse probability of treatment weighting (IPTW) [167]. Each patient's probability of being prescribed an antibiotic was calculated using a logistic regression analysis on all covariates included in the main analysis. The common support between exposed and unexposed patients was assessed graphically. Covariate imbalance between treatment groups was jointly analysed before and after each matching/IPTW strategy via univariate mean standard differences and the multivariate $\mathscr{L}_1$ imbalance measure [166]. Matching on propensity scores, however, has recently been criticised by King et al. (2019) [168]. They reported that balance in the joint distribution of covariates (as opposed to univariate mean differences in each single variable) may actually deteriorate after PS matching. They traced the reasons for this paradox to the fact that matching is performed on a scalar (the PS π) rather than the full covariate matrix X. While equality in X implies equality in π, the reverse is not true and two patients with the same PS π may have very different covariates X. Acknowledging this criticism, I also reran the matched analysis using coarsened exact matching (CEM) and compared the results to PS matching [168]. For CEM,

Table 4.1: Number and proportion of urinary symptoms and urine dipstick tests recorded on the day of or within 30 days prior to UTI episode start in primary care.

	On the day of episode start	Within 30 days prior to episode start
Total number of episodes (%)	650,416 (100)	650,416 (100)
Number of urinary symptoms (%)		
No symptom	637,168 (98.0)	621,022 (95.5)
One symptom	10,282 (1.6)	25,748 (4.0)
Two symptoms	2,501 (0.4)	3,097 (0.5)
Three or more symptoms	465 (<0.1)	549 (<0.1)
Urinary symptoms (%)		
Dysuria	7,882 (1.2)	15,938 (2.5)
Urinary frequency	5,560 (0.9)	10,393 (1.6)
Haematuria	935 (0.1)	2,278 (0.4)
Abnormal appearance of urine	804 (0.1)	985 (0.2)
Urinary urgency	797 (0.1)	1,692 (0.3)
Incontinence	509 (<0.1)	1,742 (0.3)
Malodorous urine	200 (<0.1)	369 (<0.1)
Difficulty urinating	51 (<0.1)	265 (<0.1)
Urine dipstick test (%)		
Dipstick recorded	80,294 (12.3)	106,328 (16.3)
Positive leukocyte esterase	2,621 (0.4)	3,507 (0.5)
Positive nitrites	2,411 (0.4)	2,896 (0.4)
Positive blood	1,931 (0.3)	2,495 (0.4)

UTI, urinary tract infection.

age was coarsened into decades and all other variables were grouped into equal sized bins on a square root scale (i.e. 0, 1, 2-4, 5-8, ...).

All results were reported following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [169] and the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) [170] — which extends STROBE for electronic health records research (see Appendix G).

4.5 Results

Between 2007 and 2015, 1.9 million adult women were registered with a GP practice and met the study inclusion criteria, contributing 8.4 million patient-years at risk (Figure 4.2). Included women had an average duration of follow-up of 4.5 years. Out of these, 331,901 (17.7%) patients had one or more recorded episodes

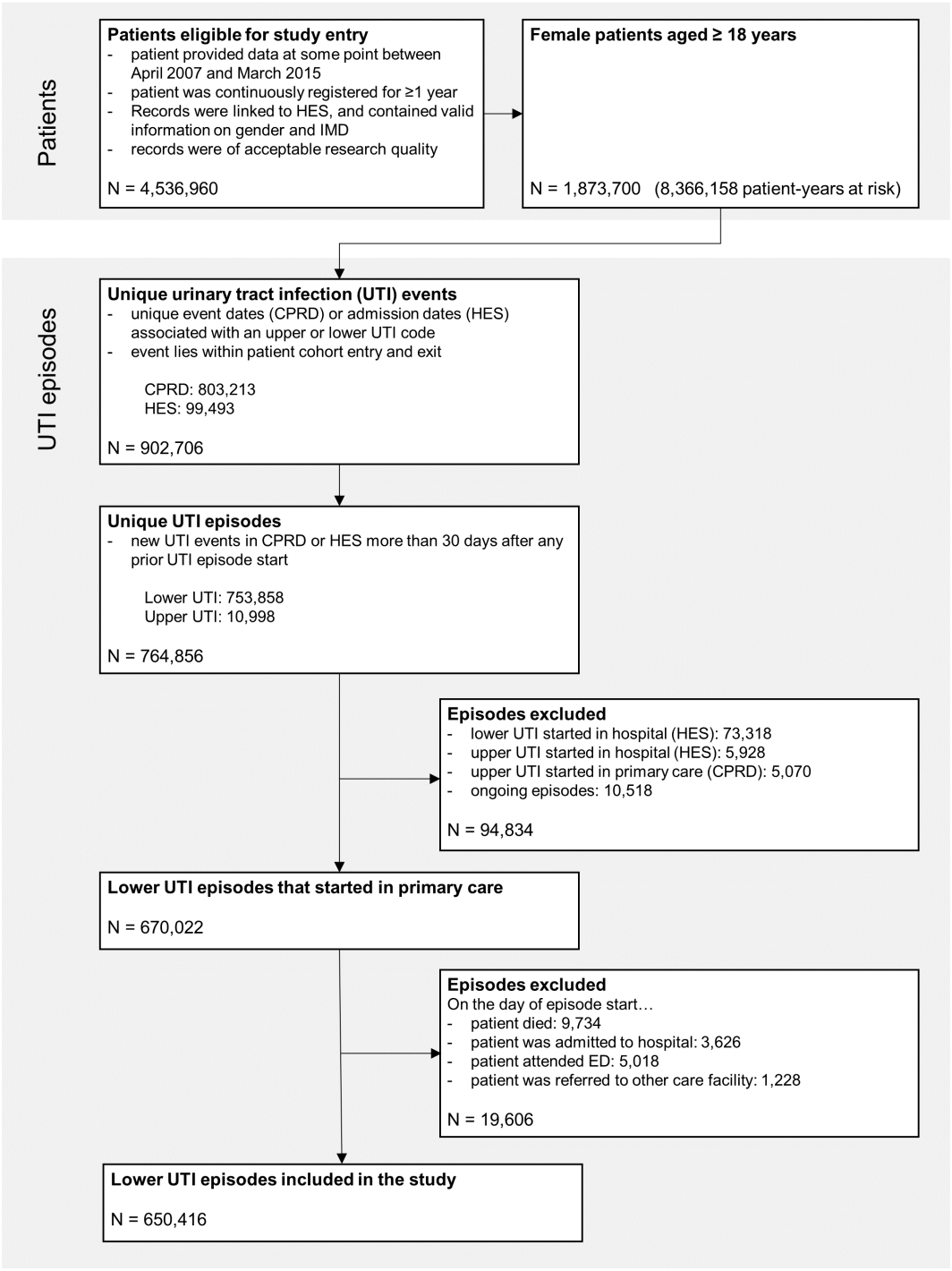

Figure 4.2: Flow chart of cohort selection for community-onset lower UTI in primary care.

Table 4.2: Characteristics of women consulting for community-onset lower UTI in primary care.

	Overall	Immediate antibiotic prescribing		p-value[1]
		Yes	No	
Total (row-%)	650,416 (100)	589,063 (90.6)	61,353 (9.4)	
Age (SD)	55.6 (21.2)	55.5 (21.1)	55.9 (22.5)	<0.001
IMD quintiles (%)				
Q1 (least deprived)	151,666 (23.3)	137,528 (23.3)	14138 (23.0)	<0.001
Q2	148,242 (22.8)	134,208 (22.8)	14,034 (22.9)	
Q3	136,657 (21.0)	124,059 (21.1)	12,598 (20.5)	
Q4	115,919 (17.8)	105,221 (17.9)	10,698 (17.4)	
Q5 (most deprived)	97,932 (15.1)	88,047 (14.9)	9,885 (16.1)	
Region (%)				
South	255,515 (39.3)	231,709 (39.3)	23,806 (38.8)	0.003
London	72,719 (11.2)	65,706 (11.2)	7,013 (11.4)	
East and Midlands	181,915 (28.0)	164,478 (27.9)	17,437 (28.4)	
North and Yorkshire	140,267 (21.6)	127,170 (21.6)	13,097 (21.3)	
CCI (SD)	1.01 (1.55)	1.00 (1.54)	1.10 (1.63)	<0.001
Smoking status (%)				
Non-smoker	398,436 (61.3)	360,284 (61.2)	38,152 (62.2)	<0.001
Ex-smoker	146,659 (22.5)	132,768 (22.5)	13,891 (22.6)	
Smoker	105,321 (16.2)	96,011 (16.3)	9,310 (15.2)	
Financial year (%)				
2007	86,906 (13.4)	78,834 (13.4)	8,072 (13.2)	<0.001
2008	82,479 (12.7)	74,573 (12.7)	7,906 (12.9)	
2009	86,980 (13.4)	78,540 (13.3)	8,440 (13.8)	
2010	88,113 (13.5)	79,555 (13.5)	8,558 (13.9)	
2011	85,397 (13.1)	77,261 (13.1)	8,136 (13.3)	
2012	84,951 (13.1)	77,192 (13.1)	7,759 (12.6)	
2013	77,221 (11.9)	70,198 (11.9)	7,023 (11.4)	
2014	58,369 (9.0)	52,910 (9.0)	5,459 (8.9)	
Recurrent UTI (%)	110,216 (16.9)	95,632 (16.2)	14,584 (23.8)	<0.001
Recent antibiotic[2] (%)	98,317 (15.1)	82,494 (14.0)	15,823 (25.8)	<0.001
Index event was home visit (%)	11,354 (1.7)	8,396 (1.4)	2,958 (4.8)	<0.001
Hospital stays				
Recent hospitalisation[3] (%)	31,378 (4.8)	27,646 (4.7)	3,732 (6.1)	<0.001
Number of stays[2] (SD)	0.19 (0.66)	0.19 (0.65)	0.26 (0.78)	<0.001
Number of nights[2] (SD)	1.8 (9.2)	1.7 (8.9)	2.5 (11.2)	<0.001
ED visits				
Recent visit[3] (%)	21,345 (3.3)	17,723 (3.0)	3,622 (5.9)	<0.001
Number of visits (SD)	0.38 (1.12)	0.37 (1.09)	0.50 (1.34)	<0.001
Outcomes[2] (%)				
Progression to severe UTI	7,947 (1.2)	6,586 (1.1)	1,361 (2.2)	<0.001
Death	2,320 (0.4)	1,895 (0.3)	425 (0.7)	<0.001
Hospitalisation (LRTI)	773 (0.1)	647 (0.1)	126 (0.2)	<0.001
Hospitalisation (Other)	12,146 (1.9)	10,303 (1.7)	1,843 (3.0)	<0.001

[1] Calculated via Kruskal-Wallis Rank Sum Test (continuous variables) and χ^2 test (categorical variables).
[2] Within 12 months prior to episode start. [3] Within 30 days prior to episode start.

CCI, Charlson Comorbidity Index; ED, emergency department; IMD, Index of Multiple Deprivation; LRTI, lower respiratory tract infection; SD, standard deviation; UTI, urinary tract infection.

of lower UTI in primary care during their follow-up. The total number of observed episodes was 650,416, amounting to 78.0 (95% CI 77.7–78.3) episodes per 1,000 patient-years at risk. Women who had at least one episode of lower UTI contracted on average two UTIs during follow-up. Only 2% of episodes had a urinary symptom recorded in addition to a diagnosis on the day of episode start (Table 4.1), and 4.5% had a urinary symptom recorded on that day or within 30 days prior to episode start. Urine dipstick tests were more frequently recorded, with 12.3% of episodes having a recorded dipstick test on the day of episode start. However, a positive result for leukocyte esterase, nitrites, or blood was recorded for less than one percent of episodes (Table 4.1).

The average age at the start of a lower UTI episode was 56 years (Table 4.2). The study population had lower levels of social deprivation than the general population. Nine out of ten patients were prescribed an antibiotic immediately. Age and socioeconomic status were broadly comparable between women who were immediately treated with antibiotics and those with delayed or withheld antibiotics, although statistical tests nevertheless indicated a significant difference due to the large sample size. Treatment groups also varied only marginally on geographical regions, comorbidity scores and smoking status, suggesting relatively little influence of demography and long-term patient factors on short-term prescribing decisions. Patients who did not receive immediate treatment did, however, differ on their recent medical history. They were more likely to have recently attended the ED (5.9% versus 3.0% in patients treated immediately with antibiotics, p-value < 0.001) or to have been admitted to hospital (6.1% versus 4.7%, p-value < 0.001) in the month before their UTI episode. They were 1.5-times as likely to be classified as presenting with recurrent UTI (23.8% versus 16.2%, p-value < 0.001), twice as likely to have been prescribed systemic antibiotics in the 30 days prior to episode start (25.8% versus 14.0%, p-value < 0.001) and more then three times as likely to have their UTI diagnosed during a home visit (4.8% versus 1.4%, p-value < 0.001).

4.5.1 Associations with progression to severe UTI

Patients in whom antibiotics were delayed or withheld were more likely to progress to severe UTI — i.e., pyelonephritis, bloodstream infection, or hospitalisation for UTI — within 30 days of episode start (2.2% versus 1.1%; OR 1.97, 95% CI 1.85–2.09, p-value <0.001; Tables 4.2 & 4.3). After adjusting for all other covariates, delaying or withholding antibiotics continued to be associated with an increased likelihood of progressing to severe UTI (aOR 1.61, 95% CI 1.51–1.71, p-value < 0.001). The corresponding NNEH was estimated at 150 (95% CI 132–179), meaning that antibiotics would have needed to be withheld or delayed in 150 additional patients in order to observe one more progression to severe UTI.

Several covariates were associated with an increased risk of progressing to severe UTI (Table 4.3). Increased odds were estimated for older patients (aOR 1.11 per each 5 years, 95% CI 1.10–1.11, p-value <0.001), patients with higher scores of CCI (aOR 1.12, 95% CI 1.11–1.13, p-value <0.001), patients who smoked (aOR 1.18, 95% CI 1.09–1.26, p-value <0.001), episodes that were labelled as a recurrent episode of UTI (aOR 1.20, 95% CI 1.14–1.27, p-value <0.001), episodes following antibiotic prescribing (aOR 1.32, 95% CI 1.25–1.39, p-value <0.001), episodes that were a home visit (aOR 2.35, 95% CI 2.15–2.57, p-value <0.001), episodes following a recent hospital stay (aOR 1.46, 95% CI 1.34–1.59, p-value <0.001) and episodes following a recent ED visit (aOR 1.36, 95% CI 1.22–1.51). The odds of developing severe UTI were further associated with the number and length of hospital stays in the previous 12 months and marginally increased over the study period. Reduced risks of progressing to severe UTI were observed for smokers in the univariate analyses, likely due to unadjusted confounding.

4.5.2 Associations with other outcomes

Delaying or withholding antibiotics was associated with all other adverse outcome considered in this study (Figure 4.3). Delayed or withheld antibiotics increased the estimated odds of death within 30 days by aOR 1.45 (95% 1.29–1.62, p-value < 0.001). It was also associated with an observed increase in outcomes not directly related to UTI. Patients whose antibiotic treatment was delayed or withheld had

Table 4.3: Univariable and multivariable associations between delayed or withheld antibiotic prescribing for community-onset lower UTI and progression to severe UTI within 30 days, adjusting for covariates using generalised estimating equations and Huber-White sandwich estimators.

	Univariable analysis		Multivariable analysis	
	OR (95%-CI)	p-value	aOR (95%-CI)	p-value
Delayed or withheld antibiotics	1.97 (1.85–2.09)	<0.001	1.61 (1.51–1.72)	<0.001
Age (per 5 years)	1.17 (1.16–1.18)	<0.001	1.11 (1.10–1.11)	<0.001
IMD quintiles				
Q1	1		1	
Q2	1.21 (1.13–1.30)	<0.001	1.16 (1.08–1.25)	<0.001
Q3	1.26 (1.17–1.35)	<0.001	1.18 (1.10–1.27)	<0.001
Q4	1.18 (1.10–1.28)	<0.001	1.16 (1.07–1.25)	<0.001
Q5	1.27 (1.18–1.37)	<0.001	1.22 (1.13–1.33)	<0.001
Region				
South	1		1	
London	1.09 (1.01–1.18)	0.029	1.13 (1.04–1.22)	0.003
Midlands and East	1.14 (1.08–1.21)	<0.001	1.12 (1.06–1.18)	<0.001
North and Yorkshire	1.10 (1.04–1.17)	0.002	1.04 (0.97–1.11)	0.246
Financial year				
2007	0.75 (0.69–0.83)	<0.001	0.79 (0.72–0.87)	<0.001
2008	0.88 (0.80–0.96)	0.004	0.91 (0.83–0.99)	0.034
2009	0.95 (0.87–1.03)	0.226	0.96 (0.88–1.05)	0.385
2010	1		1	
2011	0.96 (0.88–1.05)	0.361	0.95 (0.87–1.04)	0.243
2012	1.07 (0.98–1.16)	0.117	1.06 (0.97–1.15)	0.194
2013	1.05 (0.96–1.14)	0.317	1.03 (0.94–1.12)	0.552
2014	1.14 (1.04–1.25)	0.004	1.12 (1.02–1.22)	0.021
CCI	1.31 (1.29–1.32)	<0.001	1.12 (1.11–1.13)	<0.001
Smoking status				
Non-smoker	1		1	
Ex-smoker	1.11 (1.05–1.18)	<0.001	0.98 (0.92–1.03)	0.385
Smoker	0.82 (0.77–0.88)	<0.001	1.18 (1.09–1.26)	<0.001
Recurrent UTI	1.69 (1.61–1.79)	<0.001	1.20 (1.14–1.27)	<0.001
Recent antibiotic[1]	1.87 (1.78–1.97)	<0.001	1.32 (1.25–1.39)	<0.001
Index event was home visit	5.69 (5.23–6.19)	<0.001	2.35 (2.15–2.57)	<0.001
Hospital stays				
Recent hospitalisation[1]	2.65 (2.46–2.85)	<0.001	1.46 (1.34–1.59)	<0.001
Number of stays (per 5)[2]	8.03 (6.98–9.25)	<0.001	2.59 (2.13–3.15)	<0.001
Number of nights (per 7)[2]	1.15 (1.14–1.16)	<0.001	1.05 (1.04–1.06)	<0.001
ED visits				
Recent visit[1]	2.50 (2.29–2.73)	<0.001	1.36 (1.22–1.51)	<0.001
Number of visits (per 5)[2]	2.14 (2.03–2.26)[3]	<0.001	1.06 (0.92–1.20)	0.427

[1] Within 30 days prior to episode start. [2] Within 12 months prior to episode start. [3] Due to extremely skewed nature of this variable, GEE analysis did not converge for the number of ED visits during univariable analysis. Results presented here were obtained via standard GLM, which did converge, as did multivariable analysis.

95% CI, 95% confidence interval; aOR, adjusted odds ratio; CCI, Charlson Comorbidity Index; ED, emergency department; GEE, generalised estimating equations; GLM, generalised linear model; IMD, Index of Multiple Deprivation; OR, odds ratio; UTI, urinary tract infection.

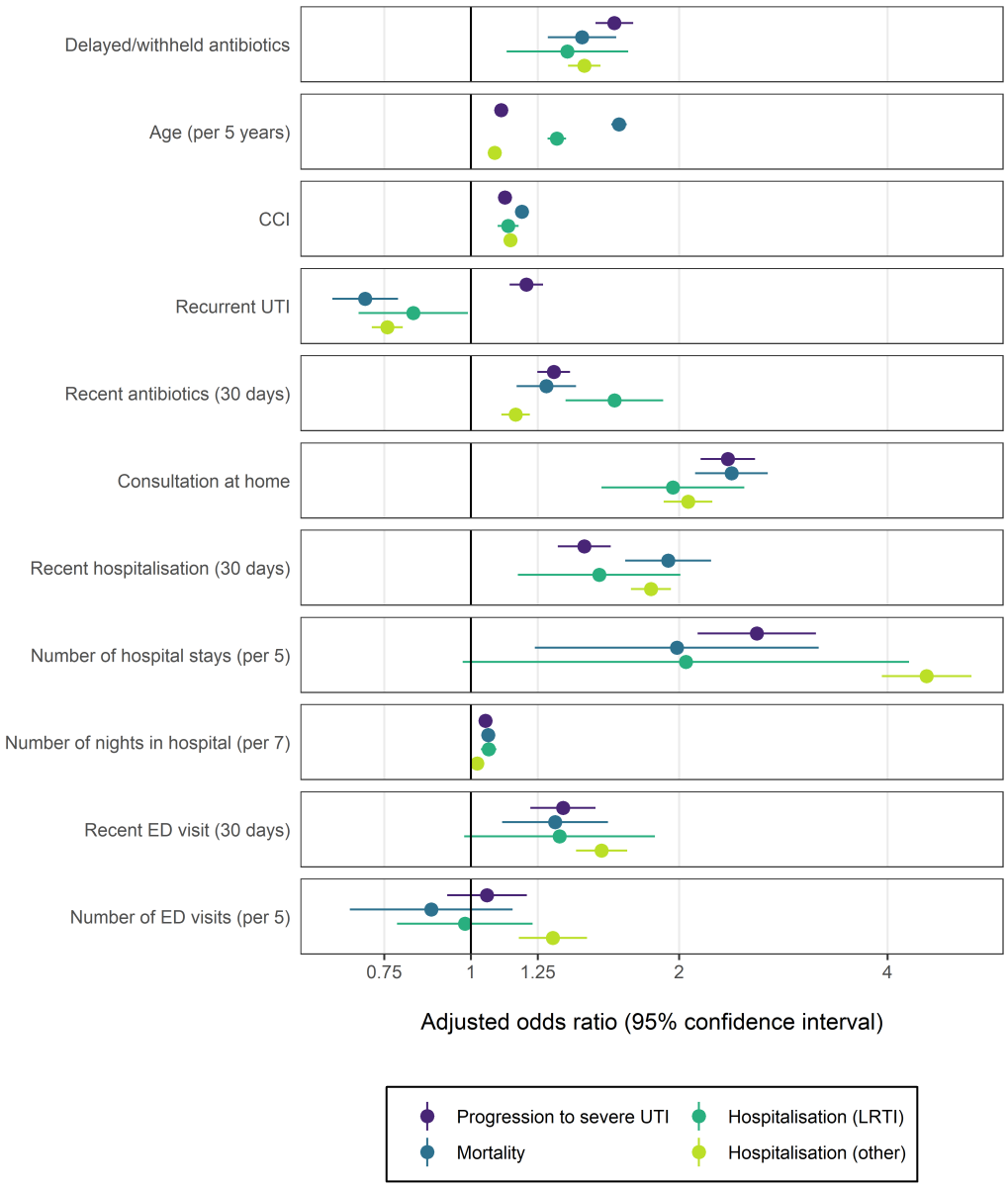

Figure 4.3: Association of covariates with included complications, estimated via GEE.

CCI, Charlson Comorbidity Index; ED, emergency department; GEE, generalised estimating equations; UTI, urinary tract infection.

Table 4.4: Multivariable associations between delayed or withheld antibiotic prescribing for community-onset lower UTI and any included complication within 30 days, stratifying by or interacting with age and adjusted for all other covariates.

Patient subgroup	Progression to severe UTI	Death	Hospitalisation (LRTI)	Hospitalisation (Other)
Stratification				
<65 years	1.50 (1.35–1.67)	2.16 (1.44–3.22)	1.72 (1.08–2.73)	1.38 (1.27–1.51)
≥65 years	1.66 (1.54–1.79)	1.50 (1.34–1.68)	1.35 (1.09–1.67)	1.49 (1.40–1.60)
Interactions[1]				
25 years	1.35 (1.15–1.59)	1.97 (0.71–5.43)	2.99 (1.53–5.85)	1.35 (1.19–1.54)
45 years	1.72 (1.56–1.89)	2.35 (1.49–3.71)	2.24 (1.44–3.49)	1.51 (1.39–1.63)
65 years	1.85 (1.68–2.04)	2.23 (1.79–2.78)	1.68 (1.31–2.15)	1.55 (1.42–1.68)
85 years	1.56 (1.45–1.68)	1.50 (1.33–1.69)	1.26 (1.02–1.56)	1.40 (1.30–1.50)

[1] In the interaction models a continuous interaction term between age (covariate) and antibiotic prescribing (exposure) was included in the model. Effect sizes are reported for four ages (25 years, 45 years, 65 years, 85 years) that roughly span the range of observed patient ages.

LRTI, lower respiratory tract infection; UTI, urinary tract infection.

both increased odds of hospitalisation for LRTI (aOR 1.38, 95% CI 1.13–1.69, p-value < 0.001) and hospitalisation for other causes unrelated to UTI (aOR 1.46, 95% CI 1.38–1.54, p-value < 0.001). The estimated effects of other covariates was broadly comparable between outcomes, with the notable exception of age and recurrent UTI (Figure 4.3). Age was more strongly associated with the odds of death or hospitalisation for LRTI. Recurrent UTI, on the other hand, was positively associated with the risk of progressing to severe UTI but appeared "protective" for all other outcomes. Variables relating to recent hospitalisations or ED visits were particular important to the odds of being hospitalised for any reason other than UTI, LRTI or bloodstream infection.

4.5.3 Interactions with age

The association between delayed or withheld antibiotics and progression to severe UTI depended on age (Figure 4.4). An interaction between age and delayed or withheld antibiotics[4] suggested that the effect of delaying or withholding antibiotics peaked between the age of 35 years and 75 years. The relative odds of progressing to severe UTI when delaying or withholding antibiotics was estimated to be aOR

[4] Corresponding to an analysis in which only prescribing status is interacted with age on a continuous scale, resulting in a coefficients for prescribing that gradually changes with age. Coefficients of other covariates are shared among all patients.

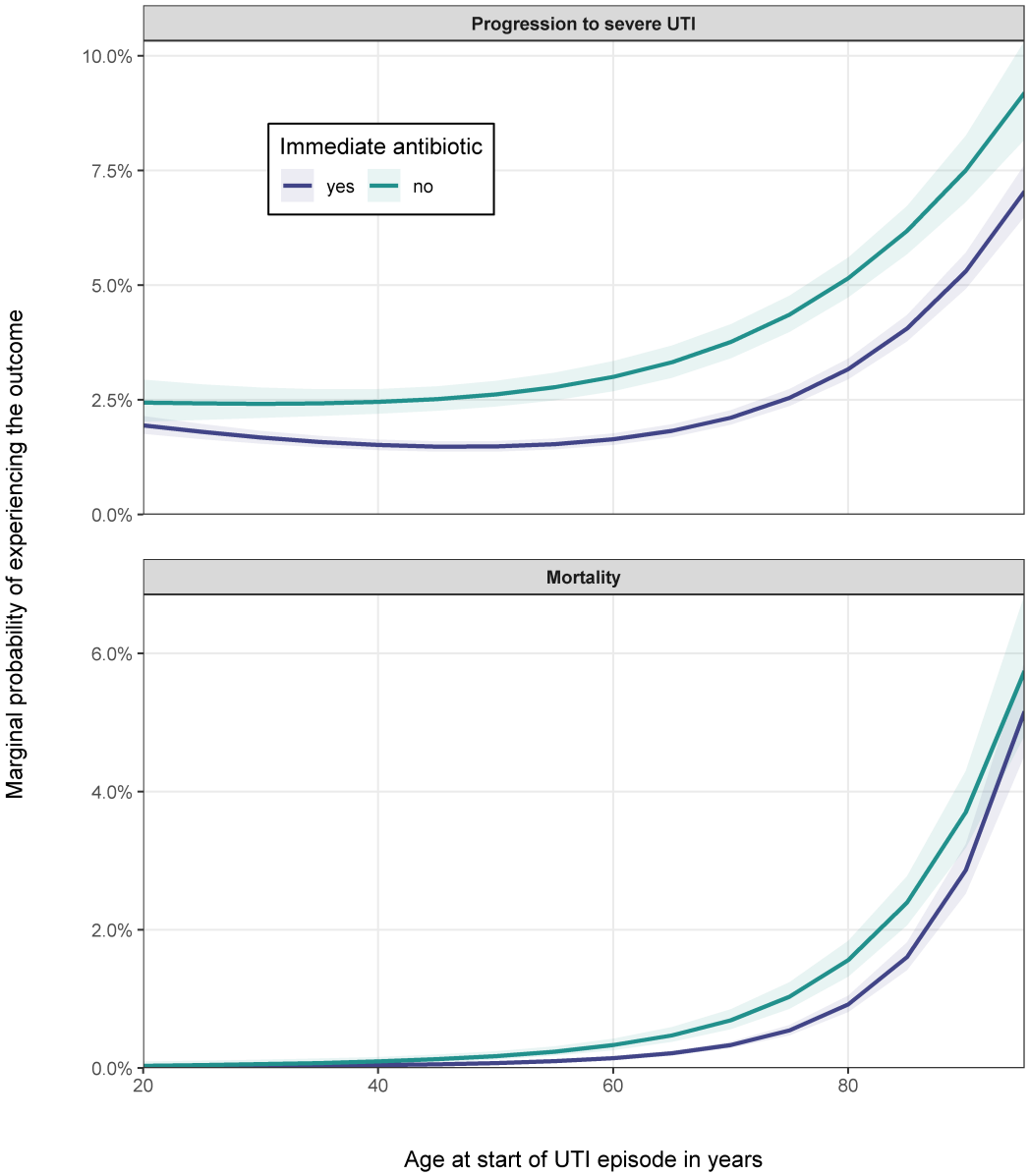

Figure 4.4: Relationship between age and marginal probability of progression to severe UTI or all-cause mortality, by treatment status.

UTI, urinary tract infection.

1.85 (95% CI 1.68–2.04) at the age of 65 years compared to aOR 1.35 (95% CI 1.15–1.59) at the age of 25 years and aOR 1.56 (95% CI 1.45–1.68) at the age of 85 years. The QIC favoured a model with prescribing-age interaction over a simpler model without interactions (ΔQIC -10.6). Due to the peak around 65 years, stratified analysis[5] in patients aged <65 years and ≥65 years suggested little difference in the estimated effect of delaying or withholding antibiotic prescribing (aOR 1.50, 95% CI 1.35–1.67 in patients <65 years versus aOR 1.66, 95% CI 1.54–1.79 in patients ≥65 years; Table 4.4). Owing to the increased incidence of progression to severe UTI with increased age, the NNEH was estimated to be 92 (95% CI 81–109) in patients aged ≥65 compared to 243 (213–290) in patients aged <65 years.

The estimated odds of death showed a similar relationship with age (ΔQIC -14.2). The relative odds of death was largest in patients aged 35 to 75 years (aOR 2.35, 95% CI 1.49–3.71 at the age of 45 years and aOR 2.23, 95% CI 1.79–2.78 at the age of 65 years; Table 4.4). Unlike progression to severe UTI or death, odds of hospitalisation for LRTI appeared to linearly decrease with age. The odds of hospitalisation for reasons other than UTI, LRTI and bloodstream infection showed little trend with age. In either case a simple model without age interactions was preferred by the QIC (ΔQIC -1.4 and ΔQIC -0.8 respectively).

Remark: (Model choice). Included patients could contribute more than one UTI episode to the analysis. This may introduce correlation between observations, as observations coming from the same patient will tend to be more similar to each other than to episodes from another patient [171]. This violates the independence assumption of generalised linear models (GLMs), potentially leading to too narrow confidence intervals and incorrect inference. I used GEEs with an exchangeable correlation structure to account for violations of the independence assumption. GEEs are a semi-parametric modelling technique that relies only on the first two moments and —

[5] Corresponding to an analysis in which each covariate is interacted with an indicator variable $I(age \geq 65)$, resulting in two sets of coefficients, one for patients below the age threshold and one for patients above.

under mild regularity conditions — provide consistent estimates of the average effects in the patient population [172]. Generalised linear mixed models (GLMMs) are a popular alternative model class that can similarly account for correlated clusters in the data. GLMMs are a fully parametric method that estimates one or more random effects for each cluster that capture cluster-specific information not accounted for by the covariates [171]. Whereas GEEs marginalise over the entire patient population and estimate the average change in the mean population response with one unit increase, GLMMs estimate the average change in an individual patient's response given the patient's characteristics. However, interpretation converges with decreasing correlation [172].

GEEs were chosen over GLMMs mainly for practical reasons. The large number of small clusters (60% of patients had only one episode, 90% of patients had 4 or fewer episodes) made it hard to fit GLMMs and accurately estimate random effects for each patient. This is reflected in the results from the primary GEE model, which estimated a correlation coefficient α of 0.019 (95% CI: -0.045–0.082) and confidence bounds on the coefficients that were virtually indistinguishable from a standard GLM fit, suggesting that there is little correlation between the multiple observations from a single patient or at the very least that the data is insufficient to detect the correlation accurately.

4.5.4 Differences between treatment groups

The estimated probability of *not* receiving antibiotics immediately ranged from 5% to 80%, with most patients belonging to three groups / modes centered around 7%, 13% and 17% (Figure 4.5). The common support — i.e. the overlap and shape — of the propensity score distribution was comparable between the treatment groups across most of the observed values but scores tended to be slightly higher among the not immediately prescribed. The original sample had an estimated median $\mathscr{L}_1$ covariate imbalance of 0.391. Although 1-nearest neighbour matching reduced

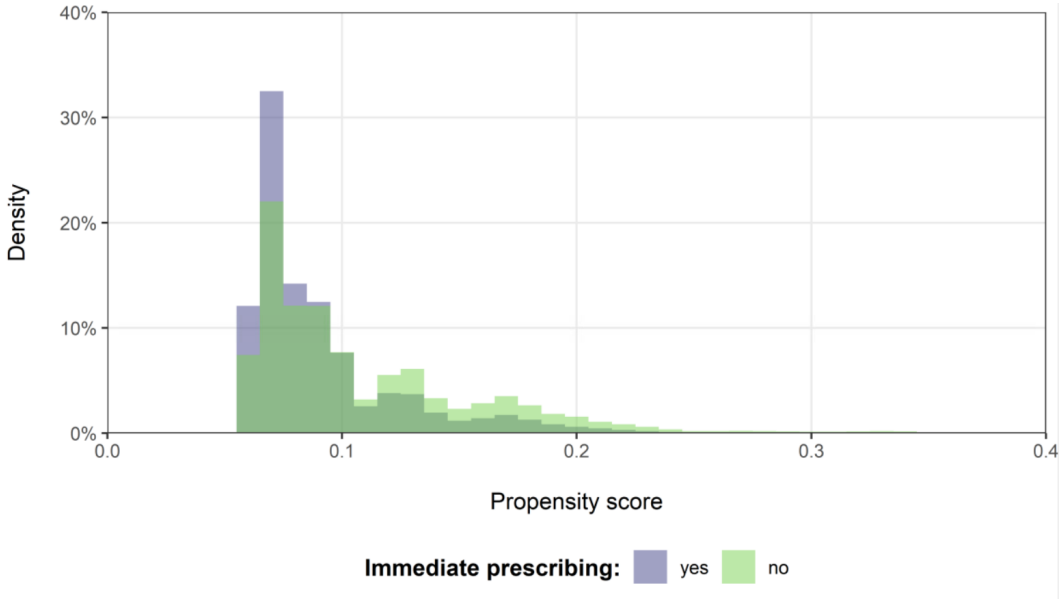

Figure 4.5: Distribution of PS for delayed or withheld antibiotics estimated via logistic regression on all covariates, by observed treatment status.

PS, propensity scores.

univariate mean standard differences in the cohort, it did not reduce imbalance in the joint distribution of covariates as measured by $\mathscr{L}_1$. IPTW on PS improved multivariate imbalance to $\mathscr{L}_1$ 0.318 (Table 4.5 and Appendix D). CEM similarly reduced imbalance with an $\mathscr{L}_1$ of 0.313. However, although CEM retained 70.4% of all samples, it only retained 43.8% (3,481 / 7,947) of patients who progressed to severe UTI. PS matching, on the other hand, retained only 18.9% of patients but 28.2% (2,244 / 7,947) of those progressing to severe UTI, and IPTW on PS — by design — retained all patients.

Estimated effect sizes after re-balancing remained comparable to those obtained in the main analysis (Table 4.5 and Appendix D). The largest estimates were obtained using CEM, which resulted in an estimated association between withheld or delayed antibiotics prescribing and progression to severe UTI of aOR 1.87 (95% 1.72–2.04). Yet, estimated associations with all other outcomes — including hospitalisation for reasons unrelated to UTI, LRTI or bloodstream infection — remained similarly high after re-balancing, suggesting that underlying unobserved confounding remained.

4.6 Discussion

4.6.1 Clinical findings

In this analysis of a large English primary care database, I found that one out of every five adult women had at least one community-onset lower UTI recorded during follow-up, demonstrating the substantial burden of disease caused by lower UTIs in women. The vast majority of those patients (>90%) were immediately prescribed systemic antibiotics to treat their UTI episode. The incidence of infectious complications in the 30 days following initial consultation was rare in both patients with immediate prescribing and in those with delayed or withheld antibiotics. However, delaying or withholding antibiotics was associated with a 1.6-fold increase in the odds of progressing to severe UTI — defined as pyelonephritis, sepsis, or hospitalisation for UTI. Delaying or withholding antibiotics was further found to be associated with a 1.5-fold increase in the odds of death over the same time period. The estimated adverse effects varied by age, but delayed or withheld prescribing was associated with increased risks of complications across all age groups.

While these results suggest that it might not be safe to withhold or delay antibiotics in an unselected group of women consulting for community-onset lower UTI in primary care, analysis of secondary outcomes also found an association between delayed or withheld prescribing and the risk of being hospitalised for LRTI or for other reasons unrelated to UTI. We would not generally expect these risks to be influenced by prescribing decisions for uncomplicated UTI. These findings might therefore point towards the role of other underlying differences between treatment groups that may have caused the observed associations between prescribing decision and adverse outcomes. Patients for whom antibiotics were delayed or withheld were more likely to have presented with recurrent UTI, to have a history of recent hospital visits and/or antibiotic exposure, and to have required a home visit by the GP. With the exception of recurrent UTI, these same factors were also associated with an increased risk of all outcomes considered — both UTI related and others. Rather than representing a purposeful decision to delay

or withhold antibiotic treatment, the lack of a prescribing record during the initial consultation for UTI may instead have signalled underlying differences in patient characteristics and disease severity. For example, patients who recently attended the hospital might already be taking antibiotics, which may be the reason why they did not receive a prescription at presentation in primary care. Similarly, a home visit by the GP — e.g., because the patient is too sick to leave the house — does not allow for automatic capture of prescribing and makes it less likely that all prescribing decisions are recorded faithfully [72]. As a result, the observed increased risk of adverse outcomes may have been driven by circumstances that influenced treatment rather than the presence or absence of treatment itself.

4.6.2 Methodological findings

Evidence from randomised controlled trials discussed earlier has been conflicting regarding the safety of delaying or withholding antibiotics for women with community-onset lower UTI, with some authors reporting no (significant) increase in the risk of subsequent pyelonephritis [63, 69, 70, 71, 155] whereas others do [156, 157]. The small sample sizes were generally insufficient to capture the rare progression to pyelonephritis with good certainty, and patient populations across trials were mostly limited to women younger than 65 years old. In this analysis of real-world clinical practice, I was able to identify a large cohort of adult women of all ages presenting with community-onset lower UTI, which may allow to provide further evidence for or against the safety of delaying antibiotic prescribing for community-onset lower UTI. However, case definitions of UTI in this study exclusively relied on recorded diagnosis codes. Little information was available on urinary symptoms, urine dipstick results, and overall perceived severity of disease (see Section 4.6.3 for a detailed discussion of this issue). Consequently, there was a mismatch between the information on which the doctor based his or her prescribing decision, and the information that was retrospectively available to reconstruct and analyse this decision. In particular, a failure to account for underlying disease severity tends to result in underestimated treatment effects in observational studies through confounding by indication [115], which may in turn have overestimated

the safety of delaying antibiotics of UTI.

Results remained largely unchanged after balancing the treatment groups on observed covariates using PS analysis or CEM. While this might be seen as adding additional credibility to the results, balancing the treatment groups also did not impact the estimated associations between delayed or withheld antibiotic prescribing for community-onset lower UTI and adverse outcomes *not* directly related to UTI. We may therefore conclude that the information available within the EHR database is insufficient to account and adjust for the non-random treatment assignment (or recording of treatment). A reason for this apparent failure to adjust for important confounders may be due the exclusive availability of long- or mid-term risk factors such as comorbidities and recent hospitalisations. If prescribing or recording of prescribing were mostly driven by short-term factors pertaining to the index consultation itself, confounding factors would have been largely unobserved and neither balancing technique would have been able to account for those factors. This makes it difficult to appraise the reliability and importance of the obtained results, and suggests that improved data collection and/or an integration of qualitative and quantitative research methods will be needed to answer questions about the safety of delaying or withholding antibiotics for suspected UTI.

4.6.3 Strengths and limitations

A major strength of this study is its large sample size of more than 600,000 episodes of community-onset lower UTI in a representative patient population of $\sim$2 million adult women consulting in English primary care. This sample size allowed me to precisely estimate associations with relatively rare outcomes such as progression to severe UTI (1.2%), death (0.4%) and hospitalisation for LRTI (0.1%) within 30 days after consulting for lower UTI in primary care. The routine nature of the data guaranteed that the patients were representative of the wider patient population. Additional linked data from hospital and national census records allowed to exclude healthcare-acquired UTIs and to comprehensively identify complications following initial consultation in primary care. The latter mitigated likely under-reporting of

Table 4.5: Multivariate imbalance ($\mathscr{L}_1$) and multivariable associations (adjusted odds ratios and 95% confidence intervals) between delayed or withheld antibiotic prescribing for community-onset lower UTI and any included complication within 30 days, after adjusting for confounding using PS or CEM.

	Propensity score analysis		**Coarsened exact matching**[1]
	Matching	IPTW	
Retained patients (%)	122,706 (18.9)	650,416 (100.0)	458,049 (70.4)
Covariate imbalance ($\mathscr{L}_1$)	0.423	0.318	0.312
Outcomes (aOR, 95% CI)			
Progression to severe UTI	1.57 (1.44–1.71)	1.69 (1.58–1.80)	1.87 (1.72–2.04)
Death	1.31 (1.13–1.53)	1.57 (1.39–1.76)	1.44 (1.20–1.73)
Hospitalisation (LRTI)	1.34 (1.02–1.76)	1.55 (1.26–1.90)	1.59 (1.18–2.14)
Hospitalisation (Other)	1.39 (1.29–1.50)	1.49 (1.41–1.58)	1.58 (1.46–1.70)

[1] ED visit in the prior 30 days and number of hospitalisations in the prior year where excluded as covariates in the CEM analysis due to small remaining numbers after matching.

95% CI, 95% confidence interval; aOR, adjusted odds ratio; CEM, coarsened exact matching; IPTW, inverse probability of treatment weighting; LRTI, lower respiratory tract infection; PS, propensity score; UTI, urinary tract infection.

hospital activity in primary care records. The analysis further contains a large range of sensitivity analyses that helped to judge the robustness of the results and identify limitations stemming from the nature of EHR data.

While a large enough sample was certainly necessary to detect the treatment effects of interest, the validity of the estimates obtained in this analysis depends on the successful identification of the date of initial consultation for UTI, the presence of antibiotic prescribing on the day of initial consultation in primary care, and the occurrence of infectious complications as a result of the initial consultation for lower UTI. Valid derivation of these quantities from routine patient records demands complete and correct recording of relevant information in the database [173]. Data obtained from CPRD is observational and retrospective, and collected with routine patient care rather than academic research in mind. The recording of information is governed by national guidelines, clinical need, and time pressures [121]. The data used in this study would be at substantial risk of confounding if clinical (coding) practices resulted in systematic differences in cohort selection (e.g. differences in the date of lower UTI), treatment assignment (e.g. differences in the propensity to prescribing systemic antibiotic), or outcomes (e.g. a priori differences in the odds

of pyelonephritis within 30 days of initial consultation).

Diagnosis of UTI in this study was based on the presence of one or more relevant diagnosis codes in the patient's medical records. Diagnoses are entered by the treating GP and generally represented suspected rather than confirmed UTI. As discussed in Chapter 1, a full clinical confirmation of UTI is based on a combination of microbiological culture of urine and the presence of urinary symptoms [17]. Since urine cultures are not routinely indicated in English primary care in the absence of additional risk factor [13], culture results were not normally available — neither in CPRD nor to the treating GP during or after the initial consultation. Treatment of suspected UTI is therefore for the most part empirical and unconfirmed. Recorded diagnoses of UTI in primary care were relatively broad, with information on urinary symptoms or urine dipsticks at presentation available only in a small subset of patients. UTI episodes included in this study might have included a wide variety of disease severities, ranging from mild UTIs and asymptomatic bacteriuria to more serious infections. Each of these might have been indiscriminately recorded as "urinary tract infection". In some extreme cases, this might have even include pyelonephritis, blurring the line between index event and outcomes. Besides potential heterogeneity in UTI severity, it is also likely that I missed a notable portion of new or recurrent UTIs. A recent study of antibiotic usage in primary care found that up to 40% of nitrofurantoin prescriptions — which are solely indicated for the treatment of lower UTIs — had no recorded reason for prescribing [6]. The extent and impact of either of these issues on the current analysis remains uncertain. The availability of admitted patients and ED data from hospitals allowed for the identification of episodes that originated in hospital, that required immediate hospitalisation, or that represented retrospective entries of discharge information. While this certainly helped to reduce some of the heterogeneity in UTI severity, it is plausible or even likely that important heterogeneity remained.

Treatment status for each episode of UTI was based on the presence or absence of a prescription for systemic antibiotics on the day of episode start.

Prescriptions in English primary care are issued electronically and are automatically captured within CPRD [121]. While antibiotic prescriptions are thus generally well recorded, delayed prescribing has been found to be recorded poorly [139]. A low rate of observed delayed prescribing might be due to a lack of recording, it might reflect an infrequent use of delayed antibiotics in general practice [3], or a combination of both. Even in cases where antibiotics were prescribed for immediate use, the patient might have decided not to file the prescription, not take the antibiotic as instructed, or not take the antibiotic at all. The analysis presented in this chapter assumed that each patient who was prescribed an antibiotic used it as prescribed. If a large proportion of our treatment group failed to take the antibiotics immediately, this might have reduced the association between delayed or withheld antibiotics and infectious complications of UTI. A second — potentially opposing — source of bias in the assignment of treatment status stems from selective under-recording of antibiotic prescribing. Adding home visits to the model for example single-handedly reduced the estimated crude effect of not prescribing antibiotics on progression to severe UTI from OR 1.97 (95% CI 1.85–2.09) to a partially adjusted OR of 1.76 (95% CI 1.65–1.87) — more than half-way to its fully adjusted aOR of 1.61 (95% CI 1.51–1.72). Although confidence intervals were overlapping for all of these estimates, the substantial drop in point estimates likely reflected the fact that prescriptions were less likely to be recorded electronically during or after a home visit, as has been noted elsewhere [72]. Since they are home bound, patients who require home visits are almost certainly less healthy than their mobile counterparts. This puts them at an a priori higher risk of adverse outcomes irrespective of whether or not they were prescribed an antibiotic. While I was able to adjust for episodes that represented home visits, I likely missed other circumstances that affect both the prescribing decision and baseline risk of complication in a similar way.

4.6.4 Comparison with existing literature

Delaying or withholding antibiotics in this study was found to be associated with an increased observed risk of progressing to severe UTI, which is in line with

findings from recent randomized controlled trials [156, 157]. The findings were also compatible with evidence from earlier trials that generally found increased but statistically non-significant risks of pyelonephritis in women who did not receive immediate antibiotics [63, 69, 70, 71]. While the direction of the estimated effects generally agreed, effect sizes estimated here from routine data were much smaller than those estimated in trials. In trials that observed at least one case of pyelonephritis during follow-up, estimated effects ranged from OR 2.98 to OR 15.34. It is unclear why the data reported here would lead to smaller effect sizes but unrecorded use of antibiotics in the delayed/withheld group and a failure to account for disease severity may be possible explanations. Notably, the largest estimated increases in risk were all observed in trials that compared the use of antibiotics to non-steroidal anti-inflammatory drugs (NSAIDs) [71, 156, 157], whereas trials that compared antibiotics to placebo generally reported smaller increases [69]. The NSAID groups in all of those trials also had a much higher incidence of pyelonephritis (2.0%-4.5%) than the placebo arm of the only placebo-controlled study of comparable size (0.4%) [174]. While these differences may be partially explained by better follow-up in the NSAID trials, a recent systematic review suggested that the observed difference might also be due to NSAIDs themselves, which have been associated with worse outcomes in other infections [86]. I was unable to estimate any possible impact of NSAIDs on the results presented here, since prescriptions for NSAIDs were not available in the originally requested data.

In the only directly comparable study performed using EHR data (see Chapter 2), Gharbi *et al.* (2019) recently report a 7–8-fold increase in the odds of bloodstream infection when delaying or withholding antibiotics in patients aged 65 years or more. Their results were based on very similar observational data from CPRD. My results suggested a much lower increase in the odds progressing to severe UTI of aOR 1.66 (95% CI 1.54–1.79) in patients aged $\geq$65. It remains unclear what caused the above described differences in findings but the approach presented here differs from that employed by Gharbi *et al.* in several important ways. First, in line with the above described trials, I limited my analysis to

women — who represent the majority of uncomplicated lower UTI in primary care [65] — and chose a broader outcome of progression to severe UTI. Second, I used more stringent inclusion criteria that utilised hospital information to exclude episodes originating in secondary care. Third, I adjusted for a number of additional variables, of which home visits was the most important. Fourth, I used a shorter follow-up period of 30 days instead of the 60 days used by Gharbi *et al.*, which more accurately reflects the usual length of UTI episodes [71, 157] and the period of increased risk of adverse outcome thereafter. Finally, I treated delayed prescribing and withheld prescribing as a single group, acknowledging the fact that delayed prescribing is poorly recorded in CPRD [139].

A PS analysis similar to that presented in this chapter was also employed by Ahmed *et al.* to account for non-random treatment choice when investigating the association between prescribe duration of antibiotic prescribing [111] or type of antibiotic [108, 110] and infectious complications (see Chapter 2). While the authors reported balance on all measured covariates after PS matching, my findings in this chapter may suggest that this matching may not have been able to balance the treatment groups on important factors that may have influenced treatment choice.

4.7 Conclusion

In the analysis presented in this chapter, I assessed if routinely collected EHR data can be used to guide management of community-onset lower UTI in primary care by estimating the safety of delaying or withholding antibiotic treatment during the index consultation. I was able to identify a large cohort of women presenting with lower UTI. Careful linkage to hospital records enabled me to create a stricter definition of community-onset UTI than those previously used. However, detailed clinical information — such as urinary symptoms or vital signs — to ascertain the presence and severity of UTI were not well recorded, and neither were the reasons for delaying or withholding prescribing. Instead, delayed or withheld prescribing needed to be inferred indirectly from the presence or absence of prescribing records, which resulted in considerable differences in

patient characteristics between treatment groups. The evidence found in this chapter against the safety of delaying or withholding antibiotic treatment in a subpopulation of women presenting with lower UTI in primary care therefore needs to be interpreted cautiously. Associations between a (presumed) decision to delay or withhold antibiotics and adverse outcomes were not only found for outcomes linked to UTI (progression to severe UTI or death) but also for other outcomes not usually related to UTI (hospitalisation for LRTI or other non-UTI reasons), suggesting a more complex relationship with the exposure. Applying popular methods to balance treatment groups led to similar estimates for all outcomes, suggesting that they were unable to account sufficiently for likely confounding due to factors which are difficult to capture in and infer from routinely collected primary care EHR data (see Chapter 3). These results also cast doubt on the success of similar balancing methods applied in some of the earlier studies reviewed in Chapter 2, although confounding by indication may have been less severe for exposures such as duration of treatment. Improved data recording for UTI in primary care records and complementary research using prospective data collection or qualitative methods will be necessary to better understand the various reasons for not prescribing antibiotics for community-onset lower UTI in primary care, and any selective biases arising from it. Some of these data items are already recorded in secondary care datasets (see Chapter 3). Chapters 5 and 6 will therefore use one such dataset from Queen Elizabeth Hospital Birmingham to assess the utility of better data recording to identify and predict microbiologically confirmed UTI in the ED.

Chapter summary

- Careful use of data from primary and secondary care can identify cases of community-onset lower UTI from diagnosis codes but additional information on urinary symptoms and dipstick test results will need to be collected to confirm and validate those cases.

- Delayed or withheld antibiotic prescribing could only be inferred indirectly, and there were considerable differences between treatment groups with high-risk patients counter-intuitively being more likely to *not* be prescribed antibiotics.
- Attempting to account for confounding using PS or CEM led to similar results, suggesting that important confounding factors could not be identified from EHR data.
- More information on the patient's acute health status and on the reasons for prescribing or non-prescribing are necessary to reliably estimate the risks and benefits associated with delaying antibiotic prescribing for lower UTI in primary care.

My role in the work presented in this chapter

I wrote the first draft of the study protocol, which was revised and edited by two of my supervisors (Dr Laura Shallcross and Prof Andrew Hayward). I designed the study described in this chapter with input from my supervisors and undertook all the statistical analysis with guidance from Prof Nick Freemantle. Raw primary care data for this study was extracted from the CPRD research database by Dr Kenan Direk and raw data from HES and ONS was provided by NHS Digital (see Chapter 3). I performed all further data linkage, pre-processing, and analyses within the UCL Data Safe Haven. I performed all analyses presented in this chapter. Computer code for the data extraction, cohort definition, and main analysis was reviewed and validated by Dr Ruth Blackburn. I interpreted all findings with help from Dr Laura Shallcross, Prof Irwin Nazareth, Prof Nick Freemantle, and Prof Andrew Hayward. I wrote the chapter with feedback from Dr Laura Shallcross, Prof Nick Freemantle, and Prof Andrew Hayward.

Software and code used in this chapter

All analysis in this chapter was performed using R (v3.6.2) and RStudio (v1.2.5033) on Windows 10. Data processing was performed using the *tidyverse* (1.3.0) and *data.table* (1.12.8) packages. All model building was performed using base R and the *geepack* (0.1.0) package. All code is publicly available at https://github.com/prockenschaub/phd_code.

Publications resulting from this chapter

Shallcross L, Rockenschaub P, Blackburn R, Nazareth I, Freemantle N & Hayward A. Antibiotic prescribing for lower UTI in elderly patients in primary care and risk of bloodstream infection: A cohort study using electronic health records in England. *PLoS Med.* 2020;17:e1003336

Chapter 5

Using EHR data to predict bacteriuria in the ED: a case study using data from QEHB

Abstract

Introduction: In the previous chapter, I attempted to estimate the association between antibiotic prescribing for urinary tract infections (UTIs) in primary care and risk of infectious complications. My findings were limited by a large risk of unobserved confounding due to a lack information on key variables. Some of the information that was crucially missing in Chapter 4, however, is more commonly recorded in secondary care. In this chapter, I therefore investigate the use of secondary care data from a large English teaching hospital to develop a risk prediction model for microbiological growth in urine samples collected in the emergency department (ED).

Background: UTIs are a leading cause of emergency visits, yet correctly diagnosing UTI in the ED remains difficult. The lack of reliable rapid diagnostic tests is exacerbated by a sicker patient population and pressure on ED clinicians to make rapid treatment decisions. Risk prediction models based on a patient's medical history and data routinely stored in the electronic health record (EHR) might provide an opportunity to predict the presence of bacteria in the patient's urine hours before urine culture results are available, providing diagnostic information to inform targeted antibiotic prescribing.

Methods: Adult patients visiting the ED at Queen Elizabeth Hospital Birmingham between 2011 and 2019 and who had a urine sample sent for microbiological culture within 24 hours of arrival were identified from local EHR systems. I extracted

information on patient demographics, comorbidities, clinical observations, blood tests, urine flow cytometry, and previous healthcare activity. The probability of bacterial growth $\geq 10^4$ colony-forming units per millilitre was predicted using a range of linear and tree-based prediction algorithms, additionally comparing pre-processing and missing data imputation methods. Models were trained and internally validated using data from 2011 to 2017 with 3-times repeated 10-fold cross-validation. Model predictions were re-calibrated and externally validated on temporally independent test data from 2018/19.

Results: Overall, 12,680 visits were included in the analysis, of which 4,677 (36.9%) showed positive culture growth. An extreme gradient boosting tree model achieved the highest area under the receiver operating characteristic of .814 (95% confidence interval .793—.835). Most predictive power was based on urine flow cytometry measurements, particularly bacteria and urinary white blood cell counts. More flexible pre-processing or imputation methods did not meaningfully improve model performance. Raw predicted probabilities exhibited sub-optimal calibration and underestimated risk of bacterial growth in 2018/19, whereas re-calibration of probabilities led to better but slightly too extreme predicted probabilities.

Discussion: The observed predictive performance implied scope to use risk prediction models to aid diagnosis of UTI in the ED. The results agreed with previous studies on the importance of urine flow cytometry tests to predict bacterial growth in urine cultures, almost to the exclusion of any other included clinical information. However, the performance observed in this chapter was considerably lower than that reported in previously published studies, suggesting possible differences in variable definitions or patient case mix. These differences are further explored in Chapter 6.

5.1 Introduction

In Chapter 4, I investigated the association between antibiotic prescribing decisions for lower urinary tract infection (UTI) in primary care and subsequent risk of adverse outcomes. I found that the interpretability of the results was limited by

a high risk of residual confounding, even after accounting for several common risk factors. In particular, information on disease severity was not recorded reliably in primary care records. This was complicated by the fact that urine samples were not usually submitted for microbiological culture in primary care, forcing me to rely on clinical coding to define the study cohort. These limitations are therefore a direct consequence of how data are routinely measured and recorded in primary care.

Electronic health record (EHR) systems in secondary care potentially capture some of the information that was crucially missing from the primary care records. Urine cultures and cell counts are much more frequently performed due to the easier access to laboratories, and physiological parameters such as vital signs and blood biomarkers are often measured in the emergency department (ED), potentially providing a means to assess disease severity at and during admission. The availability of these measurements may allow us to go beyond the analyses presented in Chapter 4 and make it feasible to use EHR data to support diagnostic decisions for UTI. Yet, the scoping review in Chapter 2 revealed that there have been very few studies to date that investigated the diagnosis and management of community-onset UTI in EHR data from secondary care, whether in England or elsewhere. One possible explanation for this is the fragmentation of hospital patient management systems and continuing lack of wide-spread electronic prescribing in hospital [67]. Over the last decade some English hospitals like Queen Elizabeth Hospital Birmingham (QEHB), however, have become increasingly digitally mature and now provide access to rich longitudinal hospital data [152].

In this chapter, I assess whether data routinely collected at QEHB contains sufficient information to support better diagnostic decisions for suspected UTI by developing statistical models that predict the presence of bacteria in the urine (called bacteriuria) at or shortly after arrival in the ED. Earlier access to this information could help clinicians to refine empirical prescribing decisions[1]. There are two ways in which this could work: 1) in low risk patients, doctors may wait for the model result before initiating antibiotics and 2) in all other patients, doctors may

[1]Culture based diagnosis of bacteriuria takes 24-48 hours, see Chapter 1.

use the model to revise their empirical prescribing decisions once model results are available. As there is no consensus on the best predictive modelling strategy, I evaluated the impact of a range of algorithmic choices and pre-processing steps and compared their performance to previously developed models [51, 109] and clinical decision rules [27, 89].

5.2 Background

UTIs are a leading cause of preventable emergency hospital admissions, accounting for 14.2% of all avoidable emergencies [175]. Despite their frequency, it can be difficult to reliably diagnose UTIs in the ED. Sources of diagnostic uncertainty include variations in clinical presentation, lack of reliable rapid diagnostic tests, and a need to initiate treatment early. These issues were already discussed in detail in Chapter 1 but are further exacerbated by the nature of ED medicine. First, patients tend to be sicker on arrival and may often require rapid decision making. Second, when traditional symptoms of UTI (e.g., dysuria, urinary frequency and urgency) are present, previous studies found them to be less reliable predictors of UTI in ED populations [73] than they are in primary care populations [27]. The predictive power of rapid diagnostic tests like urine dipsticks was also found to be lower in ED patient populations [73]. Possible reasons for this reduced performance in ED patients include spectrum bias — i.e., differences in test performance due to a differences in the clinical manifestation of UTI between primary or secondary care [176] — or an increased use of dipstick tests to screen ED patients with ambiguous symptoms [177]. Finally, on top of less reliable diagnostic information, ED clinicians usually do not have access to a patient's medical history outside of hospital [178] and are therefore more reliant on patient anamnesis, which can prove challenging if patients are confused or otherwise unresponsive. As a result, previous studies have repeatedly reported over-diagnosis and over-treatment of UTI in the ED [24, 73]. Furthermore, once treatment has been initiated patients regularly continue on antibiotics even after culture results indicated little evidence of UTI [24]. Possible reasons for these conservative prescribing decisions include the above

described large diagnostic uncertainty and an ensuing fear of delaying antibiotic treatment for severe infections [179]. Clinicians need to balance these risks against the adverse consequences associated with unnecessary antibiotic prescribing, which include antibiotic resistance and treatment side-effects (see Section 1.5).

With increasing availability of EHRs, attention has focused on whether data contained within these records may be used to support earlier, more informed prescribing decisions through the use of risk prediction models. Early and reliable prediction of UTI may affirm clinicians in their diagnosis, while providing them with additional certainty to withhold immediate treatment in low-risk patients with little evidence of infection. This strategy could potentially reduce the number of patients who are treated with antibiotics unnecessarily. An ideal model would predict a patient's probability of clinically confirmed UTI — i.e., bacteriuria in the presence of relevant urinary symptoms [17]. While bacteriuria can be measured objectively using microbiological culture results, urinary symptoms are unfortunately not well recorded in structured ED data [24]. This makes it challenging to derive a reliable ground-truth of clinically confirmed UTI. Barring improvements in the recording of urinary symptoms, bacteriuria is therefore arguably the most sensible target for such a prediction model, although at the risk of overly emphasising the importance of culture growth[2].

Chapter 2 demonstrated that few studies have looked into using routinely collected EHR data to guide diagnosis of UTI in the ED. In the only study that was eligible for inclusion in the review, Taylor *et al.* (2018) [109] used machine learning models to predict bacteriuria in 80,000 patients with UTI symptoms consulting at the ED of several US hospitals. They compared model predictions to clinicians empirical prescribing decisions and diagnosis at discharge based on ICD-10 codes. The authors reported that their final model significantly outperformed clinical judgement as measured by discharge diagnoses and/or antibiotic treatment initiation, increasing specificity of diagnosis by up to 33 percentage points. However, they did not externally validate the likely future performance of their

[2]Since asymptomatic bacteriuria — i.e., bacteriuria in the absence of symptoms — does not usually require treatment [13].

model, and there is no evidence that it was translated into real-world clinical practice. It is therefore difficult to judge whether the model could be used in English secondary care, and whether its use as a tool to inform antibiotic prescribing decisions in clinical practice would indeed lead to improved patient care.

To assess whether a similar model could be applied in English hospitals, the analysis presented in this chapter develops models to predict bacteriuria in English ED patients who had a urine sample sent for microbiological culture, and evaluates their predictive performance in a temporally independent test set. I used high-quality hospital records from QEHB — one of the NHS's centres of digital excellence [151] — to train the models. Since there is no single agreed-upon strategy to risk prediction modelling in health care, various approaches to pre-processing, missing data imputation, and modelling were used. Each approach was internally validated using repeated cross-validation to obtain unbiased estimates of expected future performance. The best performing models were then externally validated in temporally independent data from the same hospital, and any differences between expected and observed performance are discussed.

5.3 Aims and Objectives

To use EHR data from a large English tertiary teaching hospital to predict probability of bacterial growth in urine samples among individuals with suspected community-onset UTI in the ED.

Objectives:

5.1 To develop models that predict bacterial growth in urine samples collected during ED visits based on clinical information recorded in the patient's medical history and during their stay in the ED.

5.2 To compare changes in the estimated model performances due to the use of different pre-processing steps, imputation methods, and model architectures.

5.3 To evaluate the likely future performance of these models in a temporally independent test set.

A protocol for the analysis presented in this chapter was prospectively published in Rockenschaub *et al.* (2020) [180].

5.4 Methods

Study design: Retrospective observational cohort study.

Study setting: One English tertiary teaching hospital.

5.4.1 Data source and management

This chapter used a custom dataset extracted from QEHB, one of the largest teaching hospitals in England and an early-adopter of electronic record keeping with detailed records dating to 2010 [180]. Data available at QEHB is described in detail in Section 3.3.2.

I identified data items relevant to the diagnosis and management of UTI at QEHB in close collaboration with Dr Laura Shallcross and Dr David McNulty. Data for all patients with a ED diagnosis of UTI, a blood or urine sample submitted for microbiological culture within 48 hours of arrival in hospital, or a discharge diagnosis of UTI were extracted from QEHB's patient management systems into text files by Dr David McNulty (see Appendix H for a list of diagnosis codes). Local patient identifiers were replaced with pseudonomysed patient identifiers at QEHB. After pseudonomysation, the data was securely transferred to the UCL Data Safe Haven, where I imported the text files into the R programming language [159], linked all provided information via pseudonomysed patient identifiers, and transformed the data into a tabular format suitable for analysis. A detailed definition of the resulting analysis cohort and inclusion criteria as well as the methods employed to deal with missing data is provided in Chapter 5.

5.4.2 Ethical approval

Access to data related to the diagnosis and management of UTI at QEHB was approved by the Health Research Authority (HRA, reference number 17/HRA/3427). No further ethical approvals were required due to the retrospective and pseudonomysed nature of the data, which was collected as part of routine care.

The request for approval was drafted and submitted by Dr Laura Shallcross.

5.4.3 Patient population

I included pseudonomysed patient information on all adult patients who attended the ED at QEHB between November 1st 2011 and March 31st 2019[3] and who had a urine sample sent for microbiological testing within 24 hours of arrival (Figure 5.1). I excluded patients without a valid record of age or sex, patients aged <18 years, pregnant women (defined as the presence of a 10th revision of the International Statistical Classification of Diseases and Related Health Problems [ICD-10] code related to pregnancy or a pregnancy test recorded within ±9 months of arrival; Appendix H), and patients whose earliest urine sample was taken more than 24 hours after their recorded arrival in the ED. A window of 24 hours was chosen to account for delays in delivering samples to the microbiological laboratory, particularly overnight. Following a similar rationale as that presented in Chapter 4, samples from patients with a UTI diagnosis at QEHB within 30 days prior to arrival in the ED were excluded in order to limit samples to likely community-onset UTI.

Samples recorded at 12am midnight of the same day as an ED visit were treated as if they were recorded during that visit. All other samples taken before arrival in the ED were excluded from the analysis to restrict the analysis to only those samples that were likely taken in the ED. I further excluded samples for which the recorded collection date was after the date that the sample was received in the laboratory — indicating a likely data entry error — and samples for which the recorded collection date was more than 72 hours before the date that the sample was received in the laboratory — potentially relating to the re-analysis of a historic sample. Finally, samples in-between two immediately consecutive ED visits — where it was unclear during which visit the sample was collected — were also excluded from the analysis. Fewer than one percent of all urine samples were excluded this way (Figure 5.1).

[3] In November 2011, electronic recording of ED diagnosis at QEHB was switched to a more granular set of diagnosis codes that allowed for the identification of suspected UTI; see Section 3.3.2.

Remark (Recorded sample times at midnight). Upon inspection, a disproportional number of samples had a recorded time of 12am. Since laboratories at QEHB are closed overnight, it is likely that records supposedly recorded at 12am represent samples with missing information on the sample time rather than a peak of sampling frequency at midnight. To account for these inaccuracies in sampling time, samples recorded at 12am of the same day as an ED visit were treated as if they were recorded during that visit.

Patients were allowed to contribute more than one visit to the study population. For readability, I will continue to refer to visits as "patients" as if they were unique, although some visits may relate to the same person. The impact of any potential correlation between visits of the same patients was assessed (see Section 5.5.4).

5.4.4 Outcome

The primary outcome was predominant bacterial growth in the ED urine sample. Samples were considered positive if they grew a predominant pathogen $\geq 10^4$ colony-forming units per millilitre (cfu/mL) during microbiological culture. This label was used as the prediction target for all algorithms discussed in this chapter. Instead of one or two predominant organisms, urine culture may also show growth of several different organisms at once (so-called mixed growth; see Section 1.3.2). Rather than evidence for infection, mixed growth is widely regarded as contamination of the urine sample [42]. Following standard procedure at QEHB, mixed growth without a predominant organism was therefore considered negative, except where *Escherichia Coli* was explicitly reported.

Urine samples at QEHB were analysed in accordance with UK standard laboratory procedures (UK Standards for Microbiology Investigations: SMI B41, Investigation of Urine [43]). Following additional local standard operating procedures meant to reduce the number of cultured urines that have a very low prior probability of growing organisms, not all urine samples sent for microbiological analysis were eventually tested. Thresholds based on urinary white blood cell

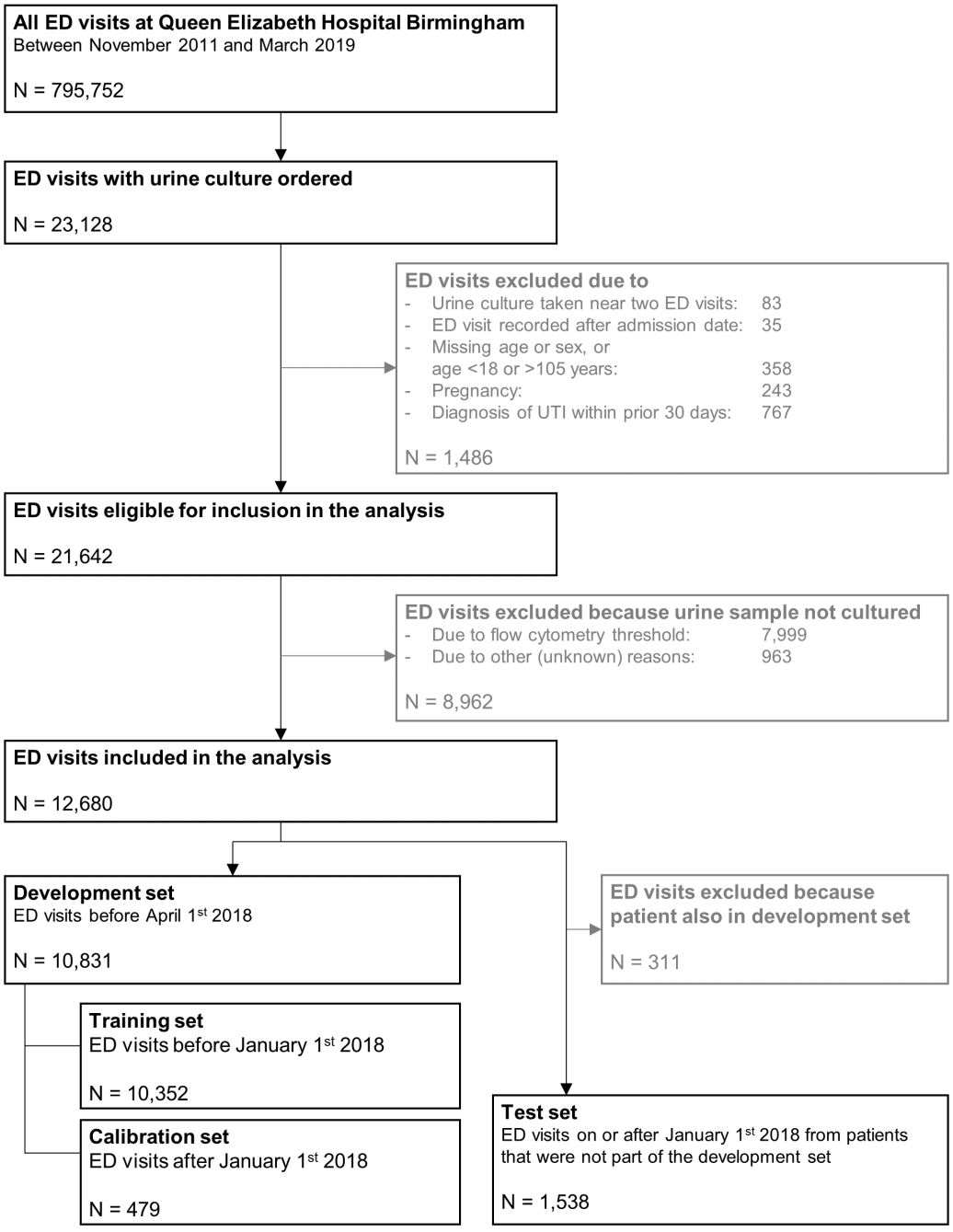

Figure 5.1: Flow chart of cohort selection for community-onset UTI in the ED at QEHB.

ED, emergency department; QEHB, Queen Elizabeth Hospital Birmingham; UTI, urinary tract infection.

counts (WBC) and bacteria counts obtained via urine flow cytometry[4] were used at QEHB to decide which samples were ultimately cultured and which weren't. At the start of the study period, this meant that only samples with a urinary WBC count >40/μL or bacterial counts >4000/μL were cultured. From October 2015 on-wards, these limits were raised to urinary WBC count >80/μL or bacterial counts >8000/μL. No urine counts could be obtained for samples of less than 4mL or which were too viscous to pass through the instrument. These samples were always cultured [180]. Samples whose urine flow cytometry results that did not exceed the above thresholds — and thus were not cultured — were excluded from the analysis. Approximately 40.8% of all eligible urine samples collected within the study period were excluded this way (38.5% before and 43.5% after October 2015).

In order to ensure that I did not miss bloodstream infections originating from a urinary source, urine samples were also considered positive if they showed bacterial growth $< 10^4$ cfu/mL but the same pathogen was grown from a blood sample taken from the same patient within 24 hours of arrival. This additional criteria affected <5 (<0.1%) samples.

Remark (Defining positive culture growth). Chapter 1 briefly touched on the choice of thresholds to define significant bacterial growth in urine cultures. Suggested ideal cut-offs in the literature range from 10^2–10^6 cfu/mL [37, 38]. Data from QEHB distinguished 10^3–10^4, 10^4–10^5, and $\geq 10^5$ cfu/mL. A threshold of $\geq 10^4$ cfu/mL was chosen in this analysis because it covered the majority of observed samples and allowed for a direct comparison with results reported previously [51, 89, 109].

[4] See Chapter 1 for a detailed description of urine flow cytometry and its role in detecting bacteriuria in hospital.

Table 5.1: List of candidate predictors for the prediction of bacterial growth in ED urine cultures.

Category	Variables
Demography	Age, Sex, Ethnicity
Clinical presentation	Suspected diagnosis (as recorded by the ED clinician)
Vital signs	Body temperature, Respiratory rate, Heart rate, O^2 saturation (via pulse oximetry), Systolic blood pressure, AVPU[a], SEWS
Urine flow cytometry	Bacteria, WBC, RBC, Epithelial cells, Small round cells, Casts, Crystals, Conductivity
Blood tests	WBC, Platelets, CRP, Creatinine, ALP, Bilirubin
Comorbidity	CCI, Cancer[b], Immunosuppression[a,b], Renal disease[c], Urological disease[c], Renal or urological surgery[c]
Previous urine culture[b]	Culture submitted, Culture positive, Culture resistant[a]
Previous antibiotic exposure[b]	Prescription of systemic antibiotics
Previous healthcare activity[b]	Number of hospitalisations for UTI, Number of hospitalisations for any cause, Number of ED visits for UTI, Number of ED visits for any cause
Time of consultation	Month, Day of year, Day of week, Time of day

[a] Dropped prior to analysis due to small numbers (< 5% of patients). [b] Within 1 year prior to arrival.
[c] Within 5 years prior to arrival. Codelists used to define variables in this table can be found in Appendix H.

ALP, Alkaline Phosphatase; CCI, Charlson Comorbidity Index; CRP, C-reactive protein; ED, emergency department; RBC, Red blood cells; SEWS, Standardised early warning score; UTI, urinary tract infection; WBC, White blood cells.

5.4.5 Predictors

Potential candidate predictors were selected based on clinical expertise, previous literature [17, 51, 89, 109], and availability of data within the EHR system (Table 5.1). Relevant information on the patient's health status at arrival in the ED included vital signs (heart rate, respiratory rate, temperature, oxygen saturation, systolic blood pressure, AVPU scale, standardised early warning score [SEWS]), blood tests (white blood cells [WBC], platelets, C-reactive protein [CRP], creatinine, alkaline phosphatase [ALP], bilirubin), and urine flow cytometry (WBC, red blood cells [RBC], epithelial cells, small round cells[5], bacteria, casts, crystals, conductivity). For each patient, all measurements recorded while the patient was in the ED were included in the analysis. If more than one value was recorded for a variable, the mean value was included. ED clinicians also recorded one or more suspected

[5] Renal tubular cells and transitional epithelial cells [181].

diagnoses based on clinical judgement[6].

General patient characteristics and information on patients' medical histories were further derived from the local hospital records. Available demographic information included age at arrival (in decades), gender, and ethnicity (Asian, Black, White, Other). Medical histories were derived using varying look-back windows. Due to the long-term nature of the included conditions, the Charlson Comorbidity Index (CCI) was calculated using all ICD-10 codes in the patient's medical history at QEHB (see Appendix H for a detailed list of all included codes). Presence of underlying renal/urological conditions or surgery was ascertained using ICD-10 diagnostic codes and OPCS-4 procedure codes within five years before the ED visit. Finally, the following medical histories were obtained using data for the one-year period immediately prior to the visit: recorded diagnosis of cancer or immunosuppression (ICD-10 codes), number of ED visits and/or inpatient admissions overall and for UTI, previous urine cultures and urine culture results performed at QEHB, and prior oral or intravenous prescription of systemic antibiotics at QEHB (see Section 4.4.5 for a detailed definition of systemic antibiotics). In addition to patient characteristics, predictors related to the date and time of arrival (month, day of the year, day of the week, time of day) were included as candidate predictors.

Throughout the study period, results from urine dipstick tests that were performed in the ED were only recorded as free-text and were thus unavailable. Furthermore, data on socio-economic status (Index of Multiple Deprivation 2015) was only available for inpatients, and was therefore excluded since all models included both patients admitted to hospital and patients that were discharged directly from the ED. Although the data used in this study could in theory be linked to national HES data, the strict confines of the ethical approval did not allow for such a linkage. Consequently, the medical history available in this analysis was entirely based on visits to QEHB. Data relevant to UTI were gathered across multiple systems covering the ED, inpatient wards, e-prescribing, and laboratories (see

[6] Suspected diagnoses in the ED were recorded using proprietary codes until November 2017 and Emergency Care Data Set (ECDS) codes thereafter; see Section 3.3.2.

Section 3.3.2 for a detailed description of data management and ethical approval).

5.4.6 Statistical analysis

5.4.6.1 Feature engineering

All candidate predictors were included in the primary model building. In order to avoid undue influence of very large or small predictor values, all continuous predictors were capped at the 1^{st} and 99^{th} percentile. For each model, three different transformations were considered for continuous predictors: identity/untransformed, log transformed (after adding an offset of one to make all values strictly larger than zero), and Yeo-Johnson transformation (a family of power transformations that transforms the variable to more closely resemble a normal distribution). Transformations were applied to improve linearity in the relationship between outcome and each predictor in the considered linear models. Categorical predictors were encoded either in full-rank or via one-hot encoding, depending on requirements of each model class. Where predictors were highly correlated — i.e., had correlation coefficient > 0.9 using Spearman's rank correlation [180] —

Remark (Variable encoding). Categorical variables may be encoded as full rank — also called dummy encoding — or via one hot encoding. In full rank encoding a variable with k categories is represented by an intercept (the reference category) and $k-1$ binary indicator variables for the remaining categories. Effect sizes are estimated relative to the reference category, ensuring identifiability of the model when using standard regression techniques. This contrasts to one-hot encoding popular in the machine learning community. In one-hot encoding, each categorical variable is represented by k indicator variables, one for each category. No intercept or reference class is used. In the absence of inherently meaningful reference classes, one-hot encoding has been observed to result in slightly better performance for models that can deal with collinearity of predictors [182] and was thus preferred when possible.

one of the correlated variables was removed prior to the analysis based on clinical judgement. Similarly, variables that were measured in less than 5% of patients were excluded prior to the analysis (Table 5.1). In addition to a model using all candidate predictors, a parsimonious set of predictors was specified a priori using only the following variables previously identified in the literature as central to the prediction of bacterial growth in urine cultures: age, sex, history of positive urine culture, and all available urine flow cytometry results [109].

5.4.6.2 Missing data

EHR data is primarily recorded with patient management in mind. Information is only recorded if measuring *and* recording the information in the EHR system was deemed necessary by the healthcare personnel in charge of patient care. Missing data is therefore ubiquitous when using EHR data for research. Negative information — i.e., information on the absence of an event — is seldom recorded explicitly, making it difficult to distinguish failure to record a disease from a genuine absence of disease. For example, a record of renal disease in a patient's medical history explicitly implies that a diagnosis has previously been made. Absence of renal disease, on the other hand, is usually implicitly signaled by an *absence* of diagnosis codes. As a consequence, I was unable to distinguish between cases in which a patient truly did not have a disease, and cases where the patient did suffer from the disease but the disease was not recorded, either because it hasn't been diagnosed yet or because the diagnoses was not captured in EHR records. For variables that describe the presence or absence of an event — i.e., those related to comorbidity diagnoses and past healthcare activity — I therefore assumed that an absence of a record meant that the event did not take place.

Missingness could be ascertained, however, in clinical information relating to general patient characteristics and clinical observations — i.e., demographic information, urine flow cytometry results, vital signs, and blood tests. For example, a missing record of heart rate can be unambiguously interpreted[7]. I graphed

[7]The reason for missingness, however, still remains uncertain. Data may be missing if measurements aren't routinely indicated for certain patient groups, if measurements are at the discretion of the clinician, or if measurements aren't always recorded after they are measured.

the patterns of missingness for variables in which missingness could be clearly identified and imputed likely values using several common imputation methods. These methods differed in their computational complexity but also statistical capability of faithfully reflecting the uncertainty caused by missing data. They may therefore differ in their impact on model performance. The following four imputation methods were applied to assess their impact on model performance:

Mean imputation and missing indicators For mean imputation each missing numerical value was replaced by the mean observed value in the training set. For categorical variables, missing values were assigned to a "missing" category. To allow the models to distinguish between patients with observed values and patients for whom the value was imputed with the mean, the models were additionally supplied with dummy variables that indicated the presence (0) or absence (1) of the value in the original, unimputed data.

k-nearest neighbours For k-nearest neighbours (KNN) imputation, each patient was matched with k patients most similar to him or her. Similarity between observations was estimated using Gower's distance, which is able to handle both continuous and categorical predictors [183]. Once the k neighbours were identified for a patient, his or her missing numerical values were imputed with the mean value among the k neighbours and categorical values were imputed with the mode. A value of $k = 5$ was chosen for this analysis.

k-means clustering Following the approach adopted by Taylor et al. (2018) [109], imputation via k-means clustering was performed by fitting a univariate unsupervised k-means clustering algorithm to each numeric variable. Each numerical variable was then categorised by replacing its value with the group it was assigned to during clustering, converting all variables into categorical variables. Missing values were assigned to a "missing" category. A value of $k = 5$ was chosen for this analysis.

Multiple imputation All previous imputation methods produce only a single imputation for each missing value. The prediction model would treat those values

as if they were actually observed[8], ignoring the inherent uncertainty involved in the imputation process [184]. Multiple imputation addresses this issue and imputes each missing value with M random draws from an imputation model — e.g. linear regression or predictive mean matching — resulting in M imputed datasets. A separate prediction model was then fit to each imputed dataset, and the M predictions averaged across datasets. By considering multiple plausible imputations, multiple imputation actively accounts for the uncertainty surrounding the imputation process. Imputation was performed using the multivariate imputation by chained equations (MICE) algorithm with predictive mean matching (continuous variables), logistic regression (binary variables), and multinomial regression (categorical variables) [184, 185]. Five datasets were imputed, and the algorithm was run for 10 iterations to achieve approximate convergence. All variables were assumed to be missing at random, meaning that the probability of a value being missing depended only on the values of observed covariates. An example of this missingness mechanism would be if creatinine were less frequently measured in younger, otherwise healthy patients, and the probability of it being measured only depended on age and comorbidities (which are covariates in the data)[9]. Following standard recommendations, the outcome was included as a predictor in the imputation models [186]. Since the outcome won't be available during model deployment — i.e., in real-time on the hospital ward — a second set of datasets was imputed *without* the outcome and used during model evaluation. This ensured that the evaluation faithfully reflected the eventual intended use of the model [180, 187].

5.4.6.3 Modelling

In order to compare the performance of different algorithms for the prediction of bacterial growth in urine samples, I fitted the following linear and tree-based

[8] This is not strictly true for imputation with missing indicators. However, although this might allow models to fit a different mean to observations with and without observed values, it still does not adequately reflect the additional uncertainty introduced by imputation.

[9] Whether this assumption holds for all included predictors is uncertain. Given the large number of covariates included in the model, I am able to account for many factors that may govern the measurement of variables but it remains likely that some unmeasured factors remain.

Remark (Hyperparameters). The term hyperparameter describes model parameters which are not directly optimised by the training algorithm. For example, while an elastic net algorithm finds the optimal coefficients β (parameters) given a mixture proportion α and a penalty weight C (hyperparameters), it cannot directly detect the best values for α and C using training data alone. If all we know is the loss on the training dataset, no regularisation would always appear better and we would be compelled to choose $C = 0$ each time. Hyperparameters are therefore commonly optimised using training-test splits, which allow to compare the performance of different hyperparameters on a test set that was left out during training.

model classes: standard logistic regression (LR), logistic regression with fractional polynomials (LR-FP), elastic net (E-NET), random forest (RF), and extreme gradient boosting trees (XGB). Support Vector Machines (SVM) with linear and radial basis function kernels were also considered but abandoned due to computational restrictions within the UCL Data Safe Haven. For LR-FP, up to four degrees of freedom (equivalent to two polynomial terms) were considered [188]. The best fitting fractional polynomials were chosen using the Akaike Information Criterion (AIC). For E-NET, RF, and XGB, 30 hyperparameter combinations were fit and the best performing combination was chosen for each model class. All hyperparameters were randomly sampled from each model's parameter space [189].

Remark (Estimating model performance under class imbalance). The area under the receiver operating characteristic curve (AUROC) or c-statistic measures how well a model is able to differentiate between patients with and without the outcome of interest. The AUROC quantifies this separation in a very intuitive way: when randomly choosing one patient with and one patient without the outcome, the AUROC is the probability that the patient with the outcome will be assigned a higher risk score [190]. A model that assigns risk scores at random will therefore have a score close to 0.5, whereas a model that

is able to perfectly classify each patient will have a score of one. Since the AUROC compares sensitivity and specificity across all possible classification thresholds, it is insensitive to changes in the prevalence of the outcome. In the presence of class imbalance — for example when the negative outcome is much more common in the study population than the positive outcome — this can lead to situations where the achieved positive predictive value is extremely low despite seemingly good results in terms of AUROC. The area under the precision-recall curve (AUPRC) has been proposed as a more reliable performance measure in the presence of imbalance. Instead of specificity, the AUPRC compares sensitivity (or recall) at each threshold to the positive predictive value (also called precision). The AUPRC does not have a fixed lower baseline at 0.5 but instead has a lower bound equal to the prevalence of the outcome [190]. For example, if 30 out of 100 of patients experienced the outcome, the expected AUPRC is 0.3 for a model that assigns risk scores at random. If 80 out of 100 patients in the sample had the outcome, the expected AUPRC would be 0.8 for the same random model, thus naturally accounting for the prevalence of the outcome.

5.4.6.4 Model validation

The ability of each model to distinguish positive and negative urine samples (model discrimination) was evaluated using AUROC, AUPRC, specificity, and negative predictive value (NPV). Specificity and NPV were evaluated at an a priori sensitivity of 95%, which I considered the minimal sensitivity acceptable to rule-out bacterial growth in urine samples in clinical practice [180]. In addition to discriminatory power, each model was evaluated with regards to its calibration. Model calibration was assessed via calibration plots as well as by calculating the calibration intercept, calibration slope, and the Hosmer-Lemeshow H-statistic.

In order to obtain realistic and reliable estimates of future model performance, internal validation using repeated k-fold cross-validation was performed. Additional

external validation was performed based on temporally independent data. Data collected before January 1st 2018 was used as training data, and the remaining data was set aside as a temporally independent test set. The test set was further split into a calibration set (all visits between January 1st and March 31st 2018) and an evaluation set (after March 31st 2018). Resampling was performed entirely on the training set and final models were re-calibrated using the calibration set prior to evaluation on the evaluation set.

Remark (Model calibration). Although less commonly reported than measures of model discrimination, model calibration is an important measure of model performance that describes how reliably a model can predict absolute risks. That is, if a model predicts a risk of 20% for given patient, the long-term probability of experiencing the outcome for patients *like* this patient should be 20%. If predicted risk and observed probabilities agree, a model can be said to be well calibrated. Model calibration is usually assessed by comparing predicted risks to the observed proportion of positive outcomes among patients with similar predicted risks. Researchers have further proposed a hierarchy of model calibration, ranging from weak calibration (predictions are correct on average) to strong calibration (predictions are correct for every covariate pattern) [191].

Internal validation I performed 3-times repeated 10-fold cross-validation to internally validate each model. During each repetition, the training set was randomly split into ten mutually exclusive folds. Models were trained on the combined data from nine folds (including pre-processing steps and missing data imputation) and the fitted model was evaluated on the tenth fold. Since the tenth fold was withheld during training, the model hadn't previously seen any of the data it was evaluated on, providing an unbiased estimate of model performance in future data[10]. This process was done ten times — leaving out each fold once

[10] Assuming that the future data originates from the same data generating process as the data used for training. Changes in the data generating process may occur for example if new clinical guidelines are introduced that change the relationships between predictors and the outcome in future data.

Remark (Resampling). Numerous studies have shown that evaluating a predictive model on the same data that it was trained on results in overly optimistic performance estimates [193]. By naively choosing the model that performs best on the training data, one likely ends up with an overfit model. Splitting the data once into a training set and a test set (split-sample analysis) accounts for some of these issues but the results heavily depend on which samples were included in the training and test sets. More reliable results can be obtained by resampling the data several times and averaging the model results across resamples. Two types of resampling are commonly recommended in practice: cross-validation and bootstrap. For cross-validation, the data is split into k folds and the model is alternatingly fit on the combined data from $k-1$ folds and evaluated on the kth left-out fold [194]. In bootstrap, the data is repeatedly and randomly resampled with replacement, and model performance is calculated as the estimated performance in those samples not included in the current resample or as a weighted average of the performance in the original data, the resampled data, and those samples not included in the resample [194].

— resulting in ten models and ten estimates of model performance for each model class and hyperparameter combination. To ensure stable estimates, I repeated this process three times, choosing ten new folds for each repetition and resulting in a total of 30 estimates for each model class and hyperparameter combination. Model performance was then summarised by the mean and standard deviation of the estimated performances across folds and repeats. Averaged observed differences in model performances were tested for statistical significance using Bayesian generalised linear mixed models estimated via Markov Chain Monte Carlo sampling with four chains of 2,000 warm-up iterations and 2,000 sampling iterations [192].

Remark (Optimism-adjusted bootstrap). Which resampling method is most appropriate for internal validation is an ongoing debate, with some authors advocating for the use of optimism-adjusted bootstrap over cross-validation [195, 196]. The bootstrap has the advantage of building models on resamples of the same size as the training data, and has been shown to be the preferred method for logistic regression models [193]. However, optimism-adjusted bootstrap does not fully hold out data and partially re-uses the training data to estimate performance. When applied to the problem presented in this chapter, both gradient boosting trees and random forests showed extremely inflated results of AUROCs >0.990 and inappropriately tight confidence limits when estimating performance via optimism-adjusted bootstrap. A similar issue has actually been hinted at in the seminal paper on bootstrap validation [197], and more recently in a prominent machine learning textbook [194] and in unpublished sources [198]. Bootstrap can lead to inflated performance estimates in situations where model flexibility or a large number of predictors relative to sample size ($n \approx p$) allow the model to overfit the data, allowing the model to exploit the missing separation of training and test sets. For this reason, all validation presented here was performed using cross-validation.

External validation Model discrimination and calibration of the best performing models were evaluated on a single temporally independent test set[11] — i.e., all ED visits after March 31st 2018. Since internal validation results in as many fitted model instances as there are resamples, a final model was fitted for each model class by applying the best hyperparameter combination identified during internal validation to the entire training data. Approximate 95% confidence intervals were obtained by bootstrapping the test data 1,000 times and using the 2.5% and 97.5% percentiles. Model calibration was assessed twice, once using the raw predictions

[11] Nested-cross validation is sometimes considered preferable to single-split external validation when there is no qualitative difference between the training and test set [114]. The reason for this is closely linked to the argument against singe-split validation described earlier, since estimated external performance again heavily depends on the (often random) choice of test set. Nested-cross validation was not appropriate as a replacement for the external validation performed here, which attempted to evaluate the models resilience to potential changes in the patient case mix over time.

of each model and once after re-calibrating the raw predictions using Platt scaling on the calibration set — i.e., all ED visits between January 1st and March 31st 2018. During Platt scaling, a univariable logistic regression model is trained using only the raw model predictions as inputs [199], thus shifting and scaling the raw predictions to better correspond to the probabilities observed in the calibration set. While Platt scaling changes calibration, it does not change the ranks of the prediction, leaving AUROC and AUPRC unchanged [199].

All results were reported following the strengthening the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (see Appendix G) [200].

5.5 Results

Between November 2011 and March 2019, 795,752 ED visits were recorded at QEHB (Figure 5.1). Out of those, 23,128 (2.9%) visits from 18,353 unique patients had a urine sample submitted for microbiological analysis. After excluding visits with contradictory admission dates, missing data on age or gender, visits from patients aged <18 or >105 years, visits by pregnant women, visits from patients who were diagnosed with UTI in hospital in the previous 30 days, and visits for which samples were sent for microbiological analysis but not cultured (predominantly due to low flow cytometry counts; Figure 5.1), 12,680 ED visits with urine culture results were included in the analysis. Of these visits, 10,352 (81.6%) visits were recorded before January 1st 2018 (training set), 479 (3.8%) visits were recorded between January 1st and March 31st 2018 (calibration set), and 1,538 (12.1%) visits were recorded on or after March 31st 2018 (test set)[12].

The absolute number of included ED visits with a urine sample showed a clear pattern, with numbers increasing until 2014 before slightly declining in 2015 and levelling out afterwards (Figure 5.2 A). The drop in included visits was both due a decline in the number visits for which a urine sample was sent for

[12] An additional 311 (2.5%) visits after March 31st 2018 were excluded from the test set because an earlier visit by the same patient was already included in the training or calibration set.

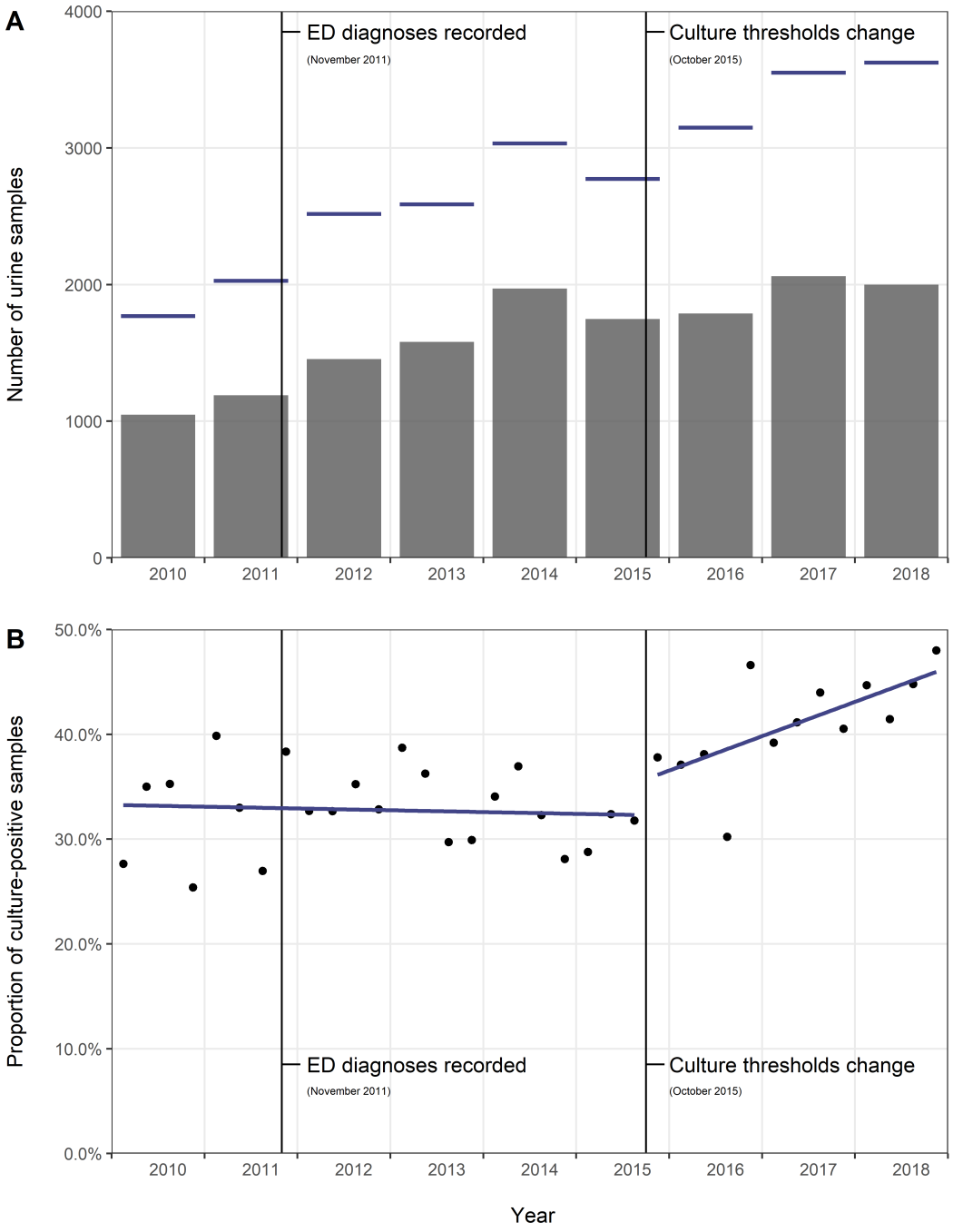

Figure 5.2: A) Yearly distribution of ED visits with a urine sample sent for microbiological culture (blue lines), and number of ED visits for which the urine sample was ultimately cultured (grey bars). Although visits before November 2011 are presented here to show an overall trend, they were not included in the main analysis since ED diagnosis were not yet recorded for these visits. **B)** Quarterly proportion of cultured urine samples that showed predominant bacterial growth (black dots) and linear trend (blue lines) before and after the change in urinalysis thresholds in October 2015.

ED, emergency department.

microbiological analysis as well as a drop in the proportion of requested samples that were ultimately cultured, coinciding with changes in the laboratory procedures that raised the urine flow cytometry threshold necessary to culture a sample (see Section 5.4.4). The effects of this change could also be seen in the quarterly proportion of culture-positive urine samples, which fluctuated around 33% before the procedure change and — with the exception of one outlying quarter in 2016 — increased thereafter up to an average of approximately 44% in 2018 (Figure 5.2 B).

Roughly half (51.9%) of all included visits were patients above the age of 65 years and two-thirds (66.0%) were women (Table 5.2). Three out of four (73.0%) patients were white, with a minority of Asian (13.2%) and Black (4.4%) patients. 23.8% and 17.9% of patients had a CCI of 1–2 and $\geq$3 respectively, and a history of renal (21.6%) or urological (28.5%) disease was common. Strikingly, almost half of all included patients had a previous hospital visit (47.8%) and/or urine sample (48.9%) recorded at QEHB in the year prior to their ED visit. Slightly more than one third of urine samples came from patients who had been diagnosed with a UTI syndrome in the ED (lower UTI, pyelonephritis, and urosepsis; Table 5.3). Samples from patients without an explicit ED diagnosis of UTI syndrome but a record of symptoms that may be compatible with this diagnosis — including urinary symptoms (dysuria, haematuria, urinary retention, and problems related to urine catheters), abdominal pain, and a record of "altered mental status" — accounted for another 13.7%. All remaining samples either came from patients with a primary diagnosis of non-urinary infection (13.4%) or non-infectious conditions (26.9%). Out of all cultured ED samples, 4,677 (36.9%) grew bacteria. The most commonly grown organism was *E. Coli* (3,504; 73.3%), followed by *Klebsiella pneumoniae* (378; 7.9%) and *Proteus mirabilis* (209; 4.4%).

5.5.1 Univariable associations

Bacterial growth was more commonly found in urine samples from older patients (40.2% of samples from patients aged $\geq$ 65 years versus 33.3% of samples from patients age <65) and women (39.0% of samples from women versus 32.9% of samples from male patients). Other characteristics that were associated with

Table 5.2: Characteristics and medical histories of patients with ED urine cultures.

	Overall	**Bacterial growth**		**p-value**[1]
		Yes	No	
Total number of visits (%)	12,680 (100.0)	4,677 (36.9)	8,003 (64.1)	
≥65 years[2]	6,584 (51.9)	2,645 (40.2)	3,939 (59.8)	<0.001
Female[2] (%)	8,368 (66.0)	3,260 (39.0)	5,108 (61.0)	<0.001
Ethnicity (%)				
White	9,256 (73.0)	3,496 (37.8)	5,760 (62.2)	<0.001
Asian	1,671 (13.2)	555 (33.2)	1,116 (66.8)	
Black	5,52 (4.4)	175 (31.7)	377 (68.3)	
Mixed or other	526 (4.1)	183 (34.8)	343 (65.2)	
Not recorded	675(5.3)	268 (39.7)	407 (60.3)	
CCI (%)				
0	7,386 (58.2)	2,686 (36.4)	4,700 (63.6)	0.295
1-2	3,023 (23.8)	1,126 (37.2)	1,897 (62.8)	
≥3	2,271 (17.9)	865 (38.1)	1,406 (61.9)	
Individual comorbidities (%)[2]				
Cancer	915 (7.2)	322 (35.2)	593 (64.8)	0.286
Underlying renal condition	2,733 (21.6)	1,011 (37.0)	1,722 (63.0)	0.913
Underlying urological condition	3,614 (28.5)	1,333 (36.9)	2,281 (63.1)	1.000
Renal or urological surgery	2,484 (19.6)	838 (33.7)	1,646 (66.3)	<0.001
Hospital activity in prior year (%)[2]				
Any hospitalisation	6,067 (47.8)	2,259 (37.2)	3,808 (62.8)	0.446
Urine sample taken	6,195 (48.9)	2,251 (36.3)	3,944 (63.7)	0.217
Urine sample positive	3,062 (24.1)	1,355 (44.3)	1707 (55.7)	<0.001
Antibiotics in hospital	3,194 (25.2)	1,149 (36.0)	2,045 (64.0)	0.225

[1] Obtained via χ^2 tests.
[2] These were binary yes/no variables. For legibility, only positive/yes categories are shown. p-values represent comparisons with the negative/no category.

Note: Percentages in the overall column represent column-%, whereas percentages in the bacterial growth columns represent row-%.

CCI, Charlson Comorbidity Index.

an increased probability of bacterial growth were white ethnicity and a positive urine culture within one year prior to consultation (Table 5.2). Previous renal or urological surgery was associated with a slightly lower probability of bacterial growth.

Urine samples from patients who had been diagnosed with a UTI syndrome in the ED had the highest probability of bacterial growth, ranging from 40-50%. Samples from patients with recorded symptoms that may be compatible with this diagnosis had variable probability of bacterial growth. The proportion of samples

Table 5.3: Recorded ED diagnoses of patients with ED urine cultures.

	Overall	Bacterial growth		p-value[1]
		Yes	No	
UTI diagnosis (%)				
Lower UTI	3,720 (29.3)	1,721 (46.3)	1,999 (53.7)	<0.001
Pyelonephritis	686 (5.4)	340 (49.6)	346 (50.4)	<0.001
Urosepsis	592 (4.7)	229 (38.7)	363 (61.3)	0.376
Symptoms attributable to UTI (%)				
Urinary symptoms	645 (5.1)	179 (27.8)	466 (72.2)	<0.001
Altered mental status	570 (4.5)	206 (36.1)	364 (63.9)	0.739
Abdominal pain	522 (4.1)	91 (17.4)	431 (82.6)	<0.001
Other infections (%)				
Sepsis (other)	539 (4.3)	166 (30.8)	373 (69.2)	0.003
LRTI	568 (4.5)	175 (30.8)	393 (69.2)	0.002
Other infection	577 (4.6)	154 (26.7)	423 (73.3)	<0.001
Other diagnoses (%)				
Genitourinary problem	344 (2.7)	94 (27.3)	250 (72.7)	<0.001
Other reason	3,070 (24.2)	1,055 (34.4)	2,015 (65.6)	0.001
Not recorded	847 (6.7)	267 (31.5)	580 (68.5)	0.001

[1] Obtained via χ^2 tests, testing each diagnosis against a joint category made up of all other diagnoses. Note: Percentages in the overall column represent column-%, whereas percentages in the bacterial growth columns represent row-%. A visit may further be associated with more than one ED diagnosis. If more than one diagnosis was recorded, the visit was assigned in a hierarchical fashion, preferring codes that were more likely to be the reason for ordering a urine culture. For example, if a visit had both a recorded diagnosis of urinary symptoms and substance abuse, the visit was labelled as relating to urinary symptoms. Only 406 (2.6%) of visits had diagnoses falling into more than 1 category, and for 267 (65.8%) of those the discarded diagnosis fell into "other reasons".

LRTI, lower respiratory tract infection; UTI, urinary tract infection.

with bacterial growth was notably lower for patients with only urinary symptoms (27.8%) or abdominal pain (17.4%), whereas it was comparable to the overall average in patients with altered mental status (36.1%). All remaining samples had probabilities of bacterial growth ranging from 26-35%.

In terms of physiological measurements, urine samples that later grew bacteria mainly differed in their flow cytometry results and CRP counts (Table 5.4). Unsurprisingly, they had much higher urinary counts of bacteria (median 13.5, IQR 5.2–33.3 versus median 5.5, IQR 1.0–14.3; $10^3/\mu$L) and WBC (median 540, IQR 163–1,880 versus median 249, IQR 96–820; /μL) at urine flow cytometry. The number of RBC and epithelial cells in the urine, on the other hand, was generally lower in samples that grew bacteria. Among blood markers, CRP showed an association with bacterial growth (median 32, IQR 7–98 versus median 24, IQR 6–89) while WBC counts in the blood displayed only little association (median 10.9,

Table 5.4: Urine flow cytometry results, vital signs, and blood biomarkers of patients with ED urine cultures. Values were summarised using the median and IQR.

	Overall	Bacterial growth		p-value[1]
		Yes	No	
Urine flow cytometry (median/IQR)				
Bacteria	8.1 (2.3–21.3)	13.5 (5.2–33.3)	5.5 (1.0–14.3)	<0.001
Epithelial cells	21 (6–54)	14 (5–40)	25 (9–62)	<0.001
RBC	35 (13–132)	32 (12–104)	37 (14–158)	<0.001
WBC	328 (111–1191)	540 (163–1881)	249 (96–821)	<0.001
Vital signs (median/IQR)				
Heart rate	85 (74–97)	85 (73–96)	85 (74–98)	0.563
Respiratory rate	18 (17–19)	18 (17–19)	18 (16–19)	0.627
Temperature (C)	36.6 (36.1–37.1)	36.6 (36.2–37.2)	36.6 (36.1–37.1)	0.248
Systolic BP	125 (111–144)	125 (111–146)	124 (111–143)	0.291
O^2 saturation	96.8 (95.1–98.0)	96.6 (95.1–98.0)	96.9 (95.1–98.0)	0.147
Blood biomarkers (median/IQR)				
WBC	10.8 (8.0–14.5)	10.9 (8.2–14.6)	10.7 (7.9–14.5)	0.088
CRP	27 (6–93)	32 (7–98)	24 (6–89)	<0.001
Platelets	231 (182–292)	227 (180–284)	233 (183–296)	0.001
Creatinine	84 (66. 118)	83 (65–115)	84 (66–120)	0.045
Bilirubin	9 (6–14)	9 (6–14)	9 (6–14)	0.140
ALP	88 (69–117)	88 (69–117)	88 (69–119)	0.897

[1] Obtained via non-parametric Kruskal-Wallis rank sum tests.

ALP, Alkaline phosphatase; BP, blood pressure; CRP, C-reactive protein; IQR, interquartile range; RBC, red blood cells; WBC, white blood cells.

IQR 8.2–14.6 versus median 10.7, IQR 7.9–14.5). Platelet and creatinine counts showed a slightly negative relationship with bacterial growth. Bilirubin and ALP did not show any association with positive urine cultures. Vital signs recorded in the ED also showed little association with bacterial growth, although these results have to be interpreted with caution due to the large proportion of missing information in these variables (Figure 5.3).

5.5.2 Missing data

The majority of patients (11,318 out of 12,680, 89%) exhibited one of ten patterns of missingness in key predictor variables from urine flow cytometry, vital signs, and blood biomarkers (Figure 5.3). Only 526 (4.1%) patients had fully observed data across all those domains. Predictor values tended to be missing in groups. For example, urine flow cytometry results (blue) were either always present together or missing together, as were vital signs (green). The overall missingness of urine

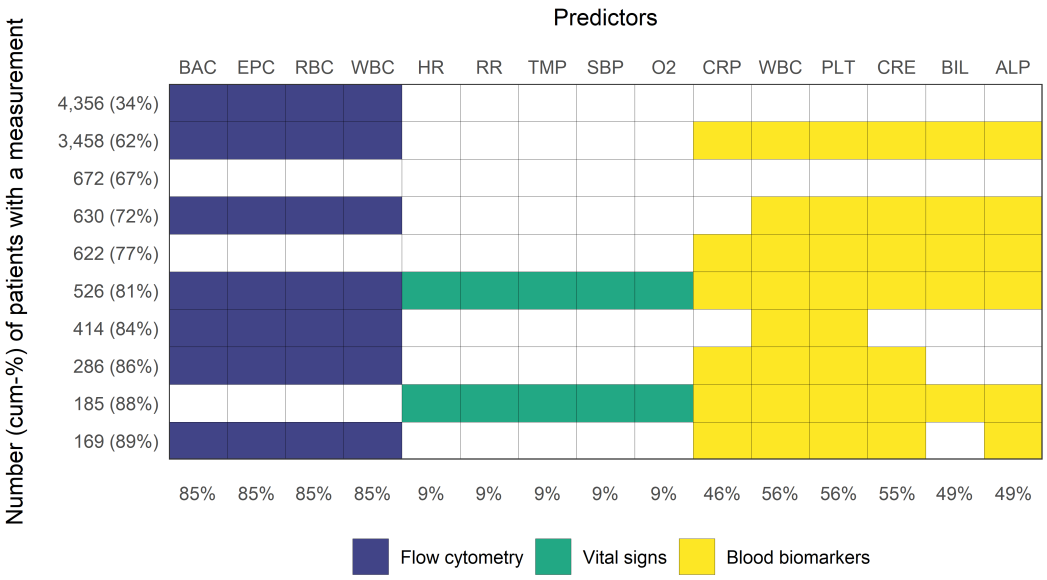

Figure 5.3: Top ten patterns of missingness in key variables for the prediction of bacterial growth in ED urine cultures. Rows represent patterns of missingness, with coloured tiles indicating predictors that were measured and white tiles indicating predictors that were missing. The number and cumulative proportion of visits with each distinct missingness pattern is shown in the left row margin. The univariable — i.e., column-wise — proportion of missingness in each predictor is displayed at the bottom.

BAC, bacteria; EPC, epithelial cells; RBC, red blood cells; WBC, white blood cells; HR, heart rate; RR, respiratory rate; TMP, temperature; SBP, systolic blood pressure; O2, O^2 saturation; CRP, C-reactive protein; PLT, platelets; CRE, creatinine; BIL, bilirubin; ALP, alkaline phosphatase.

flow cytometry was relative low with 85% of patients having an observed record of urinary bacterial, epithelial cell, RBC and WBC count. If urine flow cytometry results were missing, this was mainly due to viscosity that prevented the sample to be run through the machine or due to a particularly small urine sample size. In contrast, only 9% of patients had vital signs recorded during their ED visit. Rather than the absence of measurement, the low proportion of vital signs likely reflects an underlying difference in the IT systems used during ED visit and inpatient stays[13]. Blood markers obtained in the laboratory (yellow) were available in about half of the cohort but were not always recorded together. Blood WBC and platelet counts were the most common measurements requested (56% of patients), and were usually measured together. Bilirubin and ALP results were similarly correlated but only

[13] Patients who had vital signs recorded were more likely to be admitted to hospital then those without vital signs (83.6% versus 60.9%), suggesting that vital signs are entered or omitted from the system non-randomly and confounded by disease severity.

measured in 49% of patients. Creatinine was measured in 55%, and CRP was the least commonly requested test at 46%.

Due to the way categorical variables were defined in this study, only two could by definition contain missing data: ethnicity and ED diagnosis. Both showed low levels of missingness (5.3% and 6.7% respectively). It is likely that other categorical variables such as urinary comorbidities also contained missing information but due to the nature of how diagnoses were recorded, I was unable to distinguish them from truly absent information (see Section 5.4.6.2).

5.5.3 Internal validation

5.5.3.1 Discriminative performance

Seven out of the ten most predictive variables (as judged by their univariable discriminative performance) were derived from urine flow cytometry results. Bacteria counts in the urine alone achieved an AUROC of .631 (95% CI .625–.636) using LR with mean imputation and .662 (95% CI .657–.667) using XGB without imputation (Table 5.5). XGB was also able to make better use of urinary WBC, achieving an AUROC of .608 (95% CI .602–.614). The only urine flow cytometry parameter that did not feature among the top 20 variables was presence of urine crystals, which showed no predictive power (XGB rank 37; AUROC .503, 95% CI .497–.510). Among non-urinary predictors, suspected ED diagnosis was ranked highest at positions two (LR) and three (XGB) with the an AUROC of .603 in either case. Casts in the urine, age, small round cells in the urine, and missing urine flow cytometry achieved AUROCs around .550 each. All other variables provided univariate AUROCs of <.550. No blood markers including CRP (XGB rank 15; AUROC .518, 95% CI .512–0.524) and WBC (XGB rank 28; AUROC .508, 95% CI .502–.513), were among the ten most informative variables in either model. The only missing indicators that had some predictive power were those related to urine flow cytometry, suggesting little informative missingness in the remaining variables.

The best performing multivariable model was an XGB model including all predictors, which achieved an AUROC of .808 (95% CI .802–.814) and an AUPRC of .678 (95% CI .670–.687; Table 5.6 and Figure 5.4). At a pre-defined sensitivity

Table 5.5: Univariable discriminative performance of the top ten candidate predictors when predicting bacterial growth during internal validation.

LR		XGB	
Variable	AUROC (95% CI)	Variable	AUROC (95% CI)
UA bacteria	.631 (.625–.636)	UA bacteria	.662 (.657–.667)
ED diagnosis	.603 (.598–.608)	UA WBC	.608 (.602–.614)
UA epithelial cells	.585 (.577–.593)	ED diagnosis	.603 (.598–.609)
UA WBC	.581 (.575–.587)	UA epithelial cells	.580 (.573–.588)
UA casts	.561 (.556–.566)	UA casts	.560 (.555–.565)
UA small round cells	.555 (.549–.562)	UA small round cells	.559 (.552–.565)
UA RBC	.549 (.543–.556)	Age	.548 (.542–.553)
Age	.548 (.542–.553)	Missing flow cytometry	.546 (.539–.554)
Missing flow cytometry	.546 (.539–.554)	UA RBC	.543 (.535–.550)
Previous bacteriuria	.539 (.534–.545)	Previous bacteriuria	.539 (.534–.545)

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; CI, confidence interval; ED, emergency department; LR, logistic regression; RBC, red blood cells; WBC, white blood cells; UA, urinary (flow cytometry); XGB, extreme gradient boosting trees.

of 95%, the model achieved a specificity of 34.4% (95% CI 32.6–36.2) and had an NPV of 92.4% (95% CI 92.0–92.8). This model performed statistically significantly better than all other models when compared by AUROC (all p-values < 0.005) except a random forest using all predictors (p-value = 0.082), which was the second-best performing model with an AUROC of .805 (95% CI: .799–.811). Log transformation of laboratory and urine flow cytometry predictors in combination with mean imputation were preferred during pre-processing and imputation (see Appendix E). There was only a moderate differences in discriminative performance between models including all predictors and models that relied on a reduced set of predictors — i.e., age, sex, record of a culture-positive urine sample in the previous 12 months, and all urinalysis parameters. An XGB model using only these eleven variables achieved an estimated AUROC of .795 (95% CI .789–.801), surpassing the performance of all linear models irrespective of whether they used a full or reduced set of predictors (Table 5.6).

5.5.3.2 Calibration

Calibration intercepts suggested a moderately but systematically overestimated risk of bacterial growth when using RF with all predictors and an underestimated risk when using RF with the reduced set of predictors (Table 5.7). XGB showed good

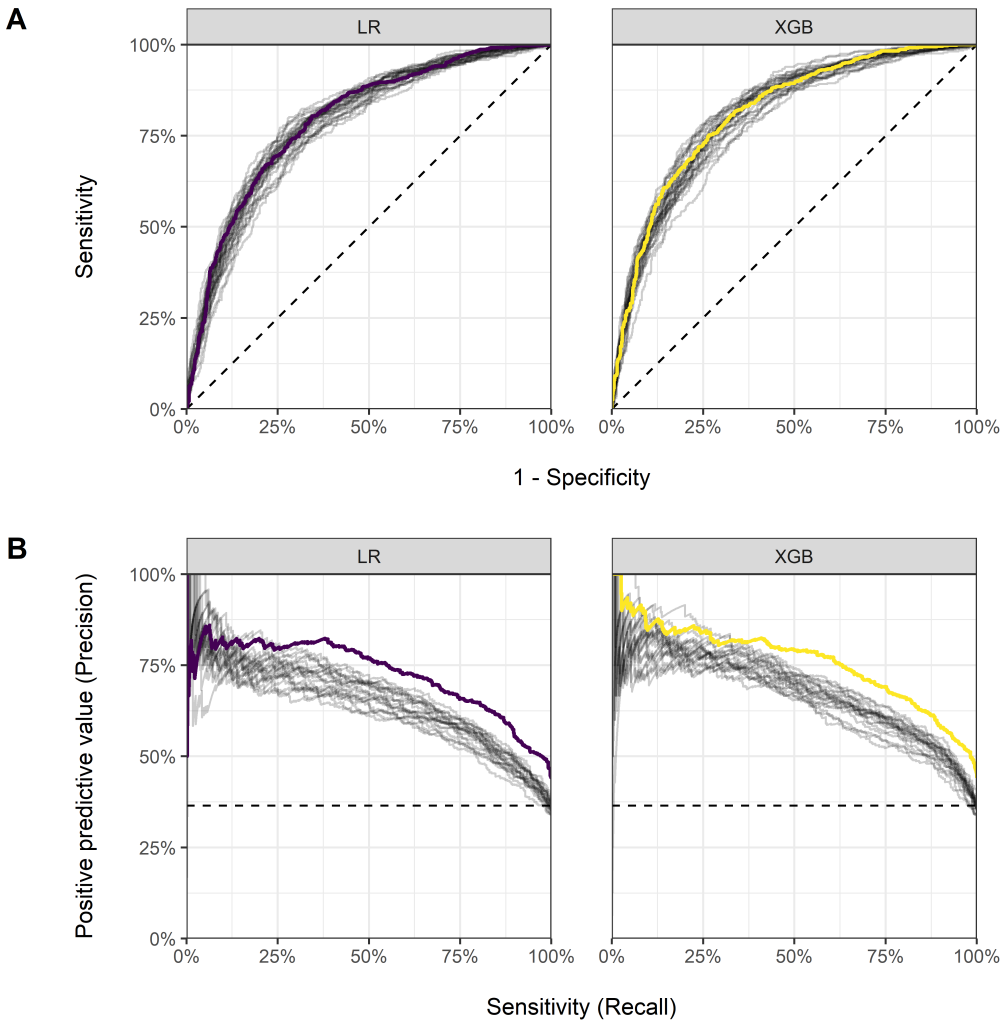

Figure 5.4: (A) Receiver operating characteristic and (B) precision-recall curves of LR and XGB models when predicting bacterial growth using all available predictors. Grey lines represent curves of re-samples during internal validation. Coloured lines represent curves from external validation on the test set. Dashed lines represent the expected performance of a random classifier.

LR, logistic regression; XGB, extreme gradient boosting trees.

Table 5.6: Multivariable discriminative performance when using full and reduced predictor sets to predict bacterial growth during internal validation.

	AUROC (95% CI)	**AUPRC** (95% CI)	**Specificity**[1] (95% CI)	**NPV**[1] (95% CI)
All candidate predictors				
XGB	.808 (.802–.814)	.678 (.670–.687)	34.4 (32.6–36.2)	92.4 (92.0–92.8)
RF	.805 (.799–.811)	.677 (.667–.687)	35.1 (33.3–36.8)	92.5 (92.1–92.9)
E-NET	.791 (.784–.797)	.651 (.642–.661)	30.5 (28.8–32.2)	91.5 (91.0–92.0)
LR	.788 (.782–.793)	.651 (.642–.661)	28.9 (27.7–30.2)	91.1 (90.7–91.6)
LR-FP	.782 (.776–.788)	.646 (.637–.656)	29.1 (27.5–30.7)	91.1 (90.5–91.6)
Reduced set of predictors				
XGB	.795 (.789–.801)	.666 (.658–.674)	34.6 (32.9–36.3)	92.4 (92.0–92.8)
E-NET	.777 (.770–.783)	.635 (.625–.645)	29.8 (28.3–31.4)	91.2 (90.7–91.7)
LR	.776 (.769–.783)	.635 (.625–.646)	29.3 (27.7–30.9)	91.1 (90.5–91.6)
LR-FP	.770 (.763–.777)	.630 (.620–.640)	27.8 (26.2–29.4)	90.5 (89.9–91.1)
RF	.767 (.761–.774)	.650 (.641–.660)	12.4 (11.5–13.3)	77.6 (76.0–79.2)

[1] At a preset sensitivity of 95%. Note: All models presented in this table used log-transformation and mean imputation. Performance metrics represent average performance from 3-times 10-fold cross-validation.

95% CI, 95% confidence interval; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic; E-NET, elastic net; LR, logistic regression; LR-FP, logistic regression with fractional polynomials; NPV, negative predictive value; RF, random forest; XGB, extreme gradient boosting trees.

calibration according to the calibration intercept. No calibration intercepts were calculated for regression-based models (LR, LR-FP, and E-NET) during internal validation, since they are always expected to result in a value close to zero[14]. Estimated calibration slopes suggested underfitting for both XGB models, both E-NET models, and an RF using all predictors. LR using all predictors, LR-FP using all predictors, and RF using the reduced set of predictors on the other hand showed evidence of overfitting. The calculated H-L H-statistics suggested moderate miscalibration except for RF using the reduced set of predictors, which showed severe miscalibration (p-value <0.001).

5.5.4 External validation

Model performance on independent test data from April 1st 2018 to March 31st 2019 was slightly higher than estimated during internal validation (Table 5.8 and

[14] The calibration intercept quantifies the difference between the mean predicted risk and the mean estimated risk. If evaluated in the same sample that the model was trained in, these will be numerically zero for regression models. During internal validation, the values will randomly vary around zero but the average risk prediction is still expected to be equal to the average risk [201, 202].

Table 5.7: Calibration when using full and reduced predictor sets to predict bacterial growth during internal validation.

	Calibration		H-L H-statistic	
	intercept	slope	statistic	p-value
All candidate predictors				
XGB	0.00 (-0.02–0.03)	1.26 (1.21–1.31)	18.4 (15.2–21.6)	0.018
RF	-0.04 (-0.06– -0.01)	1.28 (1.22–1.34)	20.6 (17.0–24.3)	0.008
E-NET	-	1.07 (1.03–1.12)	20.8 (17.7–24.0)	0.008
LR	-	0.96 (0.92–1.00)	18.0 (15.4–20.6)	0.021
LR-FP	-	0.95 (0.90–1.00)	14.9 (11.7–18.1)	0.061
Reduced set of predictors				
XGB	0.00 (-0.02–0.03)	1.11 (1.07–1.15)	12.0 (9.2–14.9)	0.151
E-NET	-	1.09 (1.05–1.14)	23.2 (19.5–26.8)	0.003
LR	-	0.99 (0.95–1.04)	21.4 (17.3–25.4)	0.006
LR-FP	-	0.99 (0.94–1.03)	15.2 (12.4–18.0)	0.055
RF	0.04 (0.02–0.07)	0.94 (0.89–0.99)	105.4 (86.3–124.5)	<0.001

Note: All models presented in this table used log-transformation and mean imputation. Performance metrics represent average performance from 3-times 10-fold cross-validation.

95% CI, 95% confidence interval; E-NET, elastic net; H-L, Hosmer-Lemeshow; LR, logistic regression; LR-FP, logistic regression with fractional polynomials; RF, random forest; XGB, extreme gradient boosting trees.

Table 5.8: Discriminatory performance when using full and reduced predictor sets to predict bacterial growth during external validation.

	AUROC (95% CI)	**AUPRC** (95% CI)	**Specificity**[1] (95% CI)	**NPV**[1] (95% CI)
All candidate predictors				
XGB	.815 (.794–.836)	.760 (.725–.795)	35.2 (30.5–41.8)	89.9 (87.8–91.5)
RF	.807 (.785–.828)	.740 (.702–.780)	35.2 (28.2–40.7)	89.9 (87.1–91.3)
E-NET	.799 (.775–.821)	.729 (.689–.764)	28.8 (24.7–36.5)	87.9 (85.4–90.2)
LR	.795 (.772–.817)	.727 (.689–.767)	28.8 (24.4–34.4)	87.9 (85.3–89.7)
LR-FP	.788 (.766–.809)	.711 (.673–.751)	32.7 (26.7–38.2)	89.2 (86.5–90.8)
Reduced set of predictors				
XGB	.807 (.782–.828)	.738 (.699–.775)	34.0 (30.7–39.4)	89.6 (87.9–90.9)
E-NET	.789 (.766–.811)	.712 (.671–.751)	32.0 (259–35.3)	85.4 (83.4–89.3)
LR	.790 (.765–.811)	.715 (.675–.754)	31.9 (26.0–35.2)	85.9 (83.4–89.2)
LR-FP	.785 (.761–.807)	.706 (.665–.746)	33.6 (28.8–40.4)	89.5 (85.7–90.2)
RF	.761 (.736–.785)	.723 (.685–.758)	26.6 (23.4–29.5)	71.8 (66.8–76.7)

[1] At a preset sensitivity of 95%. Note: All models presented in this table used log-transformation and mean imputation. Performance metrics represent performance of the best hyperparameter combination during internal validation, trained on all training data and evaluated in the test set. 95% confidence intervals are based on percentiles of 1,000 bootstrapped samples.

95% CI, 95% confidence interval; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic; E-NET, elastic net; LR, logistic regression; LR-FP, logistic regression with fractional polynomials; NPV, negative predictive value; RF, random forest; XGB, extreme gradient boosting trees.

Figure 5.5). In line with internal validation, the best performance was achieved by an XGB model using all candidate predictors (AUROC .815, 95% CI .794–.836). The only model that did not outperform the performance estimated during internal validation was a RF using the reduced set of predictors (AUROC .761, 95% .736–.785 versus AUROC .767, 95% CI .761–.774 during internal validation). Performance remained comparable if only the first visit of each patient in the test set was used (results not shown). Incremental learning and evaluation of models suggested that internal validation results were not a reliable estimate of future model performance (Figure 5.6). While XGB models generally remained preferable to simpler LR, confidence intervals from internal validation were only able to capture the next-year performance in approximately two out of every six years. Estimated internal validation estimates peaked as early as 2014/15 for both LR and XGB models. External validation estimates exhibited a V-shape and where lowest when evaluated in 2015 —coinciding with the change in laboratory procedure described in Section 5.4.4.

Remark (Incremental learning). During incremental learning and evaluation all internal validation steps described for the main analysis were repeated for each year using only data available *before* that year. The best model chosen during internal validation was then externally evaluated on data from the current year. For example, for the year 2013 in the plot the expected performance (green) was estimated using internal validation with data from November 2011 to December 2012, and an external model performance (purple) was then evaluated on data from 2013. While the main result of interest is the performance of the model in 2018/19, incremental learning can help us understand how indicative external validation is of likely future performance.

Raw model predictions of all models tended to underestimate the probability of bacterial growth, particularly for estimated probabilities of 25–75% (see Figure 5.7 A for calibration curves of LR and XGB models). Corresponding calibration

intercepts ranged from 0.303 (RF using all predictors) to 0.561 (LR-FP using the reduced set of predictors). Re-calibration via Platt scaling on data from January to March 2018 resulted in better calibration of estimated probabilities but re-calibrated probabilities ended up slightly over-estimating the probability of bacterial growth for samples with high estimated probabilities. Calibration slopes on re-calibrated probabilities therefore suggested overfitting for all models, ranging from 0.744 (LR-FP using all predictors) to 0.957 (RF using the reduced set of predictors).

5.6 Discussion

5.6.1 Clinical findings

In this chapter, I demonstrated the use of routinely collected EHR data to develop a model that predicts bacteriuria in patients who attend the ED. I used a large sample of more than 10,000 ED visits at QEHB over a period of almost eight years to train and evaluate the model. Patients included to train the model were more commonly female (although less so than in the primary care analysis presented in Chapter 4) and frequently had a recent history of hospital activity or underlying renal and urological disease. In this patient group, slightly more than a third of urine samples grew a predominant organism during microbiological culture. The best-performing model developed in this chapter — an XGB model using all available variables — was able to predict samples that would later show predominant growth with an AUROC of .815 (95% CI .794–.836) when evaluated on temporally independent test data from 2018/19. This was similar to the AUROC of .808 (95% CI .802–.814) estimated using internal validation during model training. The observed predictive performance implies scope for this approach to be used in clinical practice. Most of the predictive power was based on a small number of predictors — particularly those pertaining to urine flow cytometry. A reduced model based solely on age, sex, history of positive urine culture, and urine flow cytometry measurements performed almost as well as the full model that used all available predictors (AUROC: .807, 95% CI .782–.828). Although bacterial count and WBC measured during urine flow cytometry were already used at QEHB as a decision rule to preclude culture

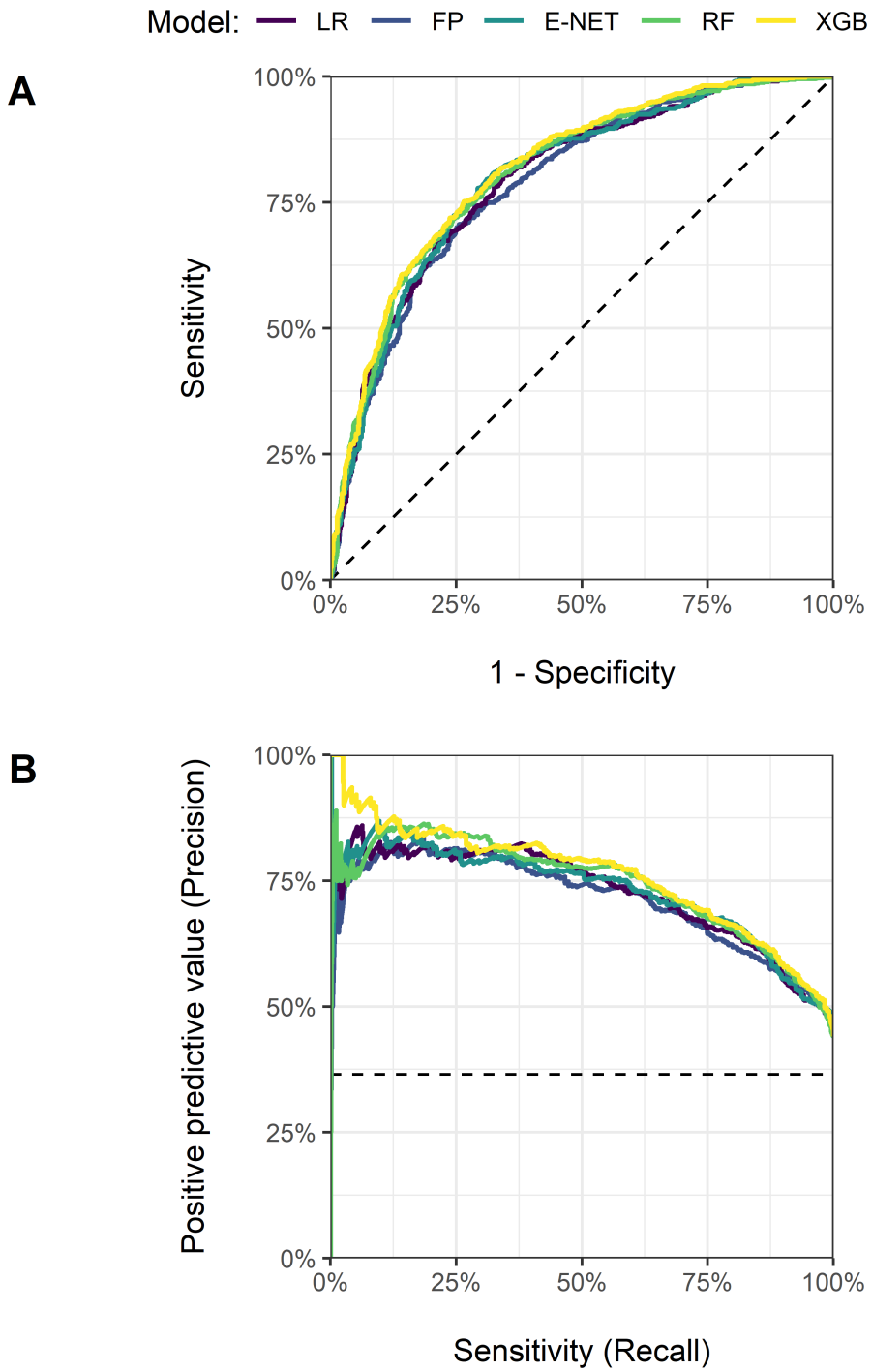

Figure 5.5: **(A)** Receiver operating characteristic and **(B)** precision-recall curves of each considered model class when predicting bacterial growth using all predictors during external validation.

E-NET, elastic net; LR, logistic regression; LR-FP, logistic regression with fractional polynomials; RF, random forest; XGB, extreme gradient boosting trees.

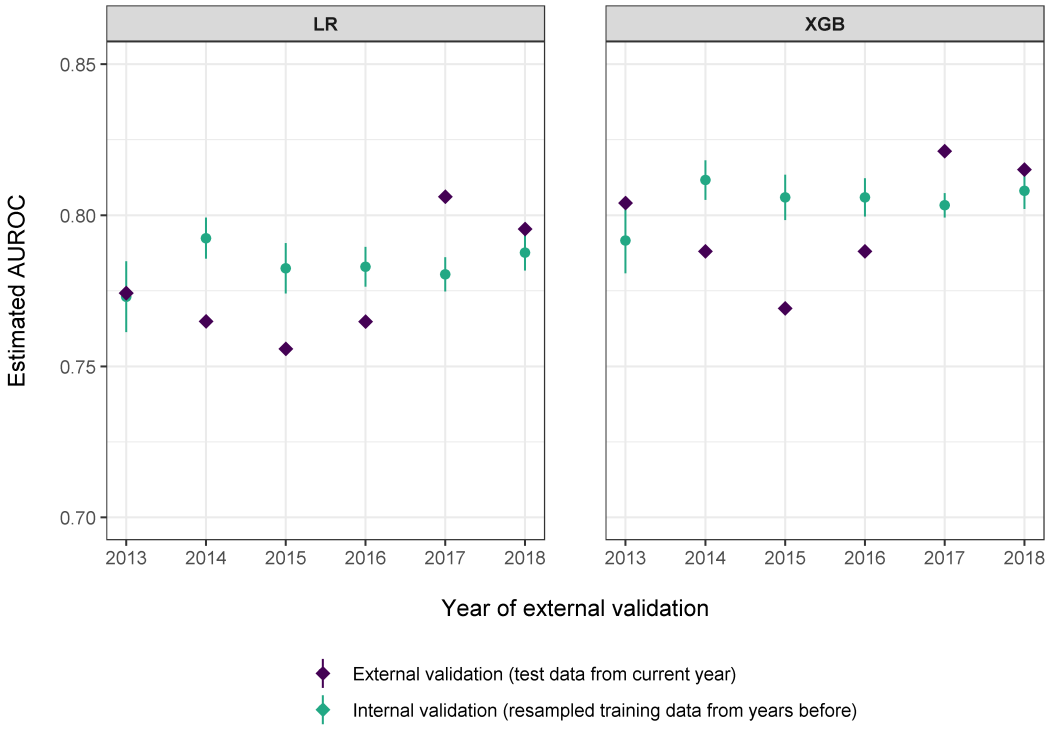

Figure 5.6: [Changes in the estimated AUROC of LR and XGB over time when predicting bacterial growth using incremental learning on all predictors. For each year, data up to that year was used to train the model. Green dots and lines represent estimated performance from internal validation of the training data up to that year. Purple diamonds represent the estimated performance from external validation using data from that year.

AUROC, area under the receiver operating characteristic; LR, logistic regression; XGB, extreme gradient boosting trees.

of urine samples with extremely low probability of growth, the good predictive power of these measurements even in urines that cleared that threshold suggests that the value of these measurements to support prescribing decisions for UTI in clinical practice may currently be underestimated. If the models were used to rule out bacterial growth with a target sensitivity of 95%, the best-performing model would have been able to correctly flag ∼35% of samples without bacterial growth early and achieve an NPV of ∼90%.

5.6.2 Methodological findings

Although the observed predictive performance implied scope for this approach to be used in clinical practice, model performance was found to vary notably over the eight year study period. Particularly around a period of laboratory procedure

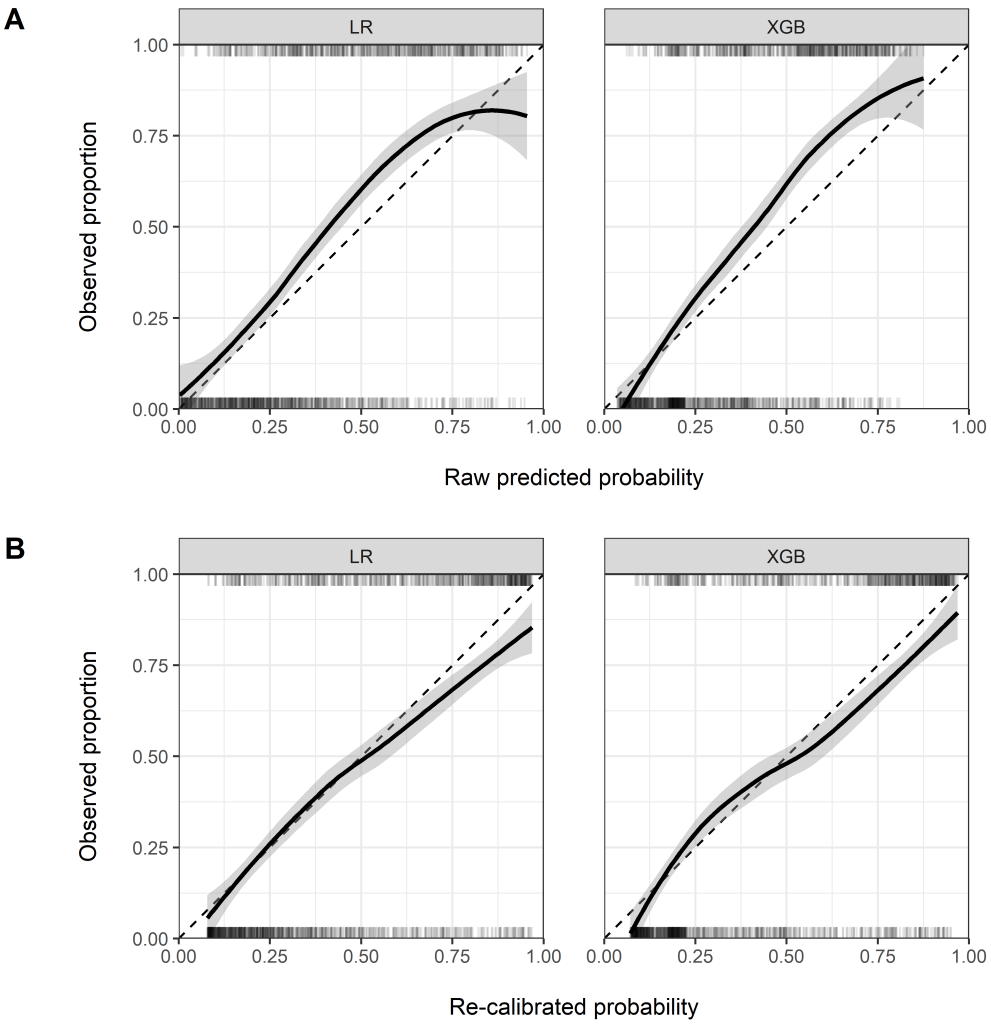

Figure 5.7: Calibration plots of **(A)** raw and **(B)** re-calibrated predictions of LR and XGB when predicting bacterial growth using all predictors. Calibration plots were estimated using LOESS regression and gray areas represent 95% confidence intervals. Rug plots at the top and bottom of each graph show the distribution of positive and negative outcomes respectively.

LOESS, locally estimated scatterplot smoothing; LR, logistic regression; XGB, extreme gradient boosting trees.

changes in 2015, externally evaluated model performance dropped several points of AUROC below the performance suggested by the corresponding internal validation. Reasons for this drop in performance likely relate to the concomitant changes in local practice in relation to clinical diagnoses, laboratory procedures, testing for UTI, and data availability. Model performance increased again towards the end of the study period when models were able to adjust to the new data distributions, equalling or exceeding predicted performance after 2017. Nevertheless, models needed to be re-calibrated in order to maintain acceptable calibration on the temporally external test data from 2018/19. This suggests that — if deployed in clinical practice at QEHB or another hospital — models likely would need to be re-calibrated periodically to assure valid interpretation of model predictions as a probability of bacteriuria.

Except for changes over time, model performance was reasonably robust to the choice of statistical model, pre-processing strategy, and imputation method. Surprisingly, multiple imputation performed considerably worse than simpler alternatives, likely due to the structured way in which data was missing, which might be an indicator that the data was unlikely to be missing at random. Many clinical measurements were missing in almost all patients and/or were missing in groups (e.g. vital signs), and available clinical information may not have been sufficient to account for this missingness. Although data resolution was improved compared to the primary care analysis presented in Chapter 4, important clinical information was therefore still missing. The analysis presented in this chapter demonstrated the promises of predictive modelling in hospital but also highlighted the limitations of EHR data even at digitally mature sites like QEHB. Manual note review in a pilot study preceding this analysis suggested that much of the absent information would already be recorded in the doctors' free-text notes but are not reflected in the structured EHR data usually available to researchers (see Appendix F and Shallcross *et al.* (2020) [24] for a more detailed discussion on this). The implications of missing these key bits of information are discussed below as part of the limitations of this study.

Finally, all information used in this study was obtained retrospectively. Relevant data items were identified one-by-one by a health informatician at QEHB and manually extracted. A large part of the work presented in this chapter was therefore related to combining and cleaning data from several hospital IT systems. If the model were to be used in real-time in the ED, an automated data processing pipeline will be required and continued maintenance necessary to adapt this pipeline to changes in the hospital's IT structure over time. Furthermore, while many IT systems back up data in data warehouses like the one used in this study overnight, live data feeds that allow predictions in real-time are not always routinely supported and may require bespoke software engineering. Substantial technical know-how and resources are therefore required on site to enable the deployment of clinical decision making tools, limiting their spread in often financially strained healthcare settings with highly fragmented IT systems [203].

5.6.3 Strengths and limitations

To the best of my knowledge, this is the first study using EHR data to predict bacteriuria in ED populations in England. A major strength of this analysis is the use of a large sample of high-quality EHR data from the ED of a major teaching hospital. QEHB has a long history of electronic record keeping in English secondary care [152], allowing me to use records collected over a period of multiple years to build a robust model and assess its likely future performance. The use of extensive sensitivity analyses and validation designs ensured the robustness of the primary findings to modelling assumptions.

Like the primary care data used in Chapter 4, the data used in this analysis were recorded as part of routine care rather than with research in mind. Despite its status as a digital leader among English hospitals [151], QEHB did not routinely collect or store data on several key variables relevant to the diagnosis and management of UTI. Perhaps most importantly, prior antibiotic prescribing outside of hospital was not available in the data and antibiotic prescribing during an ED visit was also missing for most of the study period [24]. Urine samples are collected before antibiotics are initiated in the ED, but prior antibiotic treatment that was initiated in the community

may have nevertheless prevented or retarded the growth of microorganisms during culture. The inability to account for recent antibiotic treatment may therefore have decreased the models' power to predict bacteriuria [20], and some of these patients may still require antibiotic treatment for partially treated infections. Data was also notably missing on urine dipstick testing. Dipstick results are commonly used to support the diagnosis of UTI in the ED and have previously been shown to have at least some power in ruling UTI in or out [27, 89]. While cell counts from urine flow cytometry might be viewed as a partial substitute for dipstick results, their absence might have reduced the models performance by omitting key information available to clinicians but not the model. Dipstick results may have influenced clinicians' decision to send urine samples for culture, thus resulting in a pre-selected sample. Absence of dipstick results prevented me from investigating the extent of this issue in my sample, and any impact that it might have on my findings. Finally, vital signs were rarely recorded and almost exclusively confined to patients who were later admitted to hospital. It is unclear whether this was due to procedural differences between patients who do and do not get admitted to hospital, or due to patient severity. In either case, the fact that vital signs were recorded during an ED visit was likely a marker of the patient's state, potentially biasing models that were to use all predictors (including vital signs) in a future setting in which vital signs are more routinely measured and input into the model.

How and when urines were sent for culture at QEHB further limited the reliability and generalisability of the findings. For unknown reasons, a substantial proportion of urine samples included in this analysis were submitted for culture in the absence of any recorded suspicion of UTI (Figure 5.8 A). A prediction of urine culture growth should only guide treatment in the presence of clear symptoms to avoid over-treatment of asymptomatic bacteriuria [31, 32]. How often cultures are requested without symptoms will therefore strongly influence the future performance of the model in clinical practice. This proportion likely depends on local hospital guidelines and best practices, and will vary by hospital. On the other hand, the cohort used in this analysis excluded patients visiting the ED at QEHB

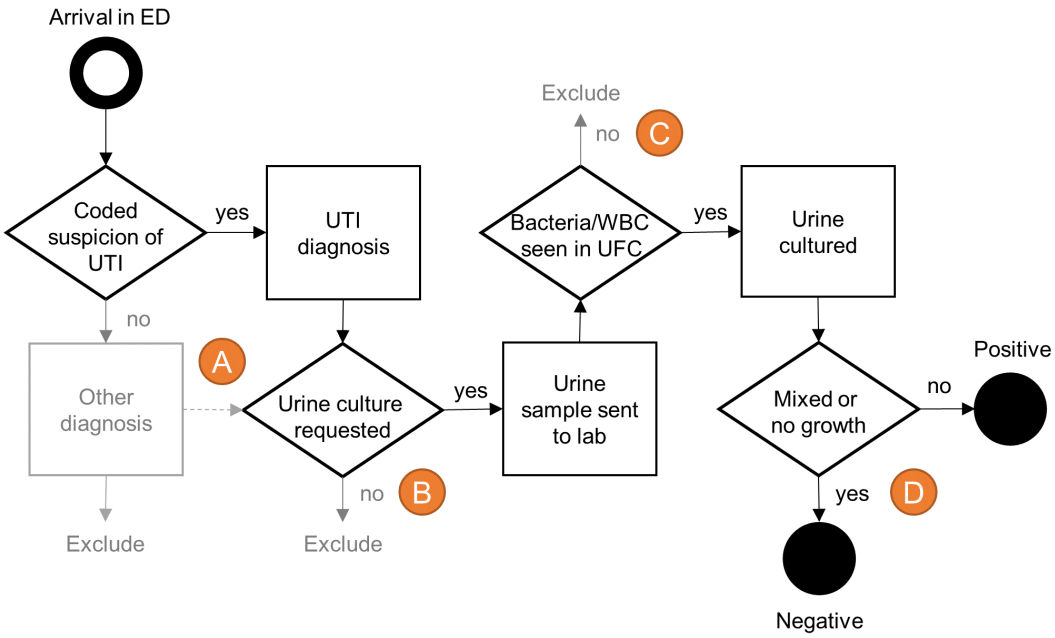

Figure 5.8: Selection of urine samples and opportunities for selection bias when ascertaining bacterial growth in ED urine cultures at QEHB: **(A)** urine cultures requested for patients without a coded suspicion of UTI, **(B)** urine cultures not requested for patients with a coded suspicion of UTI, **(C)** urine samples not cultured due to local standard operating procedure based on UFC measurements, and **(D)** mixed urine cultures treated as contamination by default.

UFC, urine flow cytometry; QEHB, Queen Elizabeth Hospital Birmingham; UTI, urinary tract infection; WBC, white blood cell.

with urinary symptoms but who did not have a urine sample sent for culture (Figure 5.8 B). A substantial number of patients was excluded this way, as more than 60% of all patients with an explicitly recorded suspicion of UTI in the ED did not have a urine sample sent for culture. Whether or not a urine sample is submitted is decided by the clinician. Patients with suspected UTI who were included in this analysis might therefore present a subpopulation with an a priori greater probability of (complicated) UTI, such as patients with underlying renal disease or recurrent UTI. The eventual culture of urine samples — and whether they could be included in this analysis — further depended on locally set thresholds based on bacteria and urinary WBC counts (Figure 5.8 C; see Section 5.4.4). Flow cytometry values observed in this study were thus unlikely to be representative of all urines submitted for culture. Instead, they were positively biased since only samples above a certain

value were cultured. As these thresholds depended entirely on local guidelines — which changed during the study period — model performance might not generalise well to hospitals with different thresholds or no threshold at all (see comparison with results by Müller *et al.* [51] in the next section). Finally, following standard clinical practice, mixed growth was not considered significant growth and treated as contamination of the urine sample (Figure 5.8 D) [42]. Urines with mixed growth might have contained bacteria before the contamination but I was unable to estimate the possible extent of this issue. Since bacteria count was the most important predictor of bacteriuria, it is further likely that samples with mixed growth were commonly misclassified by the model (see Chapter 6 for a detailed discussion of this issue).

5.6.4 Comparison with existing literature

Taylor *et al.* (2018) [109] recently applied machine learning methods to predict bacterial growth in 80,000 ED patients with UTI symptoms consulting at several US hospitals (see Chapter 2 for a summary of the study). Their best model was an XGB model that achieved an AUROC of .904 (95% CI .898–.910). At a pre-set sensitivity of 95%, the model had a specificity of 50% and an approximate NPV of 97%. The reported performance was thus significantly higher than that reported here, but similarly relied on a small number of key variables. Re-analysis performed by me using code and data published by the authors alongside their publication showed that a LR using urinary WBC and bacteria counts alone would have achieved an AUROC of .839 (95% CI .832–.847), close to the published top performance using a more complicated algorithm and all 211 variables available to the authors. Adding the urine dipstick parameters leukocyte esterase, nitrites, and haematuria further increased the achievable LR performance to AUROC .862 (95% CI .855–.869). The model published by Taylor *et al.* therefore similarly relied on information related to urinalysis results, and it is not immediately clear why the performance estimated in my analysis is significantly lower even when comparing variables available in both datasets — i.e., urinary WBC and bacteria counts. Three important differences exist between the data used by Taylor *et al.* and that used in this analysis.

First, the number of patients available to Taylor *et al.* was more than six times larger, possibly enabling the model to fit the data in more detail. It is unlikely, though, that sample size alone accounted for the observed differences in performance. In the incremental learning setting presented earlier, the performance of my models levelled out in 2014 with as little as 3,000 patients. A LR model trained only on a random subset of the data used by Taylor *et al.* of similar size to the data used in this chapter achieved a performance that was virtually identical to the performance of a model trained on all patients, confirming the limited impact of sample size.

Second, the relative frequency with which urine cultures were requested in the ED was much larger in their sample (25.6%) than at QEHB (2.9%), raising questions about the comparability of included patient populations. While the propensity to culture might genuinely be higher in the US than it is in England, data published elsewhere from the same hospital trust found a considerably lower culture rate of 15.2% for approximately the same time period [204]. The US Center for Disease Control and Prevention (CDC) estimated an even lower US nation-wide urine culture rate of 8.1% [205] for 2016. Two further single-center studies from the US [206] and Canada [207] provided estimates that ranged from 2.3% to 6.0%, which was comparable to the numbers observed in this chapter. This suggests that the data used by Taylor *et al.* might have been subject to selection bias, or — if urine cultures were indeed requested for one out of four patients attending the ED — at least was not representative of other hospitals in England *or* the US.

Finally, the higher proportion of patients who had a urine culture requested in the sample used by Taylor *et al.* had implications for the overall proportion of culture-positive samples. At QEHB, 37.2% of cultures ended up positive, which increased to >40% after 2017. In comparison, Taylor *et al.* reported only 22.7% of samples positive. The difference in the prevalence of bacteriuria may be due to a higher proportion of unnecessary cultures in their patient population, or a less stringent urine flow cytometry threshold for culture than that applied at QEHB.

In a much smaller study of 613 patients obtained from laboratory data in

Switzerland, Müller *et al.* (2018) [51] reported similar performance to that reported by Taylor *et al.* when predicting bacterial growth based on urine flow cytometry results alone. Using LR, the authors reported an estimated performance of AUROC .930 (95% CI .900–.940) when predicting bacterial growth in a single train-test split. A comparison of the data used in this chapter and data published by Müller *et al.* provides a compelling explanation for the observed difference in model performance. Samples that were a priori dismissed by the laboratory at QEHB due to low bacteria and urinary WBC counts during urine flow cytometry (Figure 5.9 A; yellow area) *were* cultured in Switzerland (Figure 5.9 B). These samples were highly unlikely to grow bacteria, representing easy wins for a predictive algorithm. As a result, the algorithm developed by Müller *et al.* would have significantly worse performance when transferred to patients at QEHB.

In the only other study I am aware of that attempted to predict bacteriuria in ED patients, Wigton *et al.* (1985) [89] three decades earlier used prospectively collected data to train a discriminant analysis and achieved an AUROC of .780. Their simplified decision rule was based on the presence or absence of five clinical variables: history of UTI, back pain, urinary WBC > 15 per high-power field (HPF), urinary RBC > 5 per HPF, and more than a few observed bacteria. The rule had a specificity of 44% and an NPV of 80% at approximately 93% sensitivity. Little *et al.* (2006) [27] several years later developed a similar decision rule for primary care that achieved an AUROC of .785 using rounded coefficients from logistic regression on dipstick results — i.e., leukocyte esterase, nitrites, and haematuria. Importantly, both studies used prospectively collected data and limited their analysis to female patients with suspected UTI, usually due to dysuria, urinary urgency, or frequency. As a result, the proportion of positive cultures was considerably larger than that observed here or by Taylor *et al.* and Müller *et al.*, with $> 60\%$ of samples positive in both studies.

None of the secondary care studies discussed above have been externally validated at another hospital. Wigton *et al.* evaluated their model on a temporally independent set of patients from the same hospital [89], whereas Taylor *et al.* and

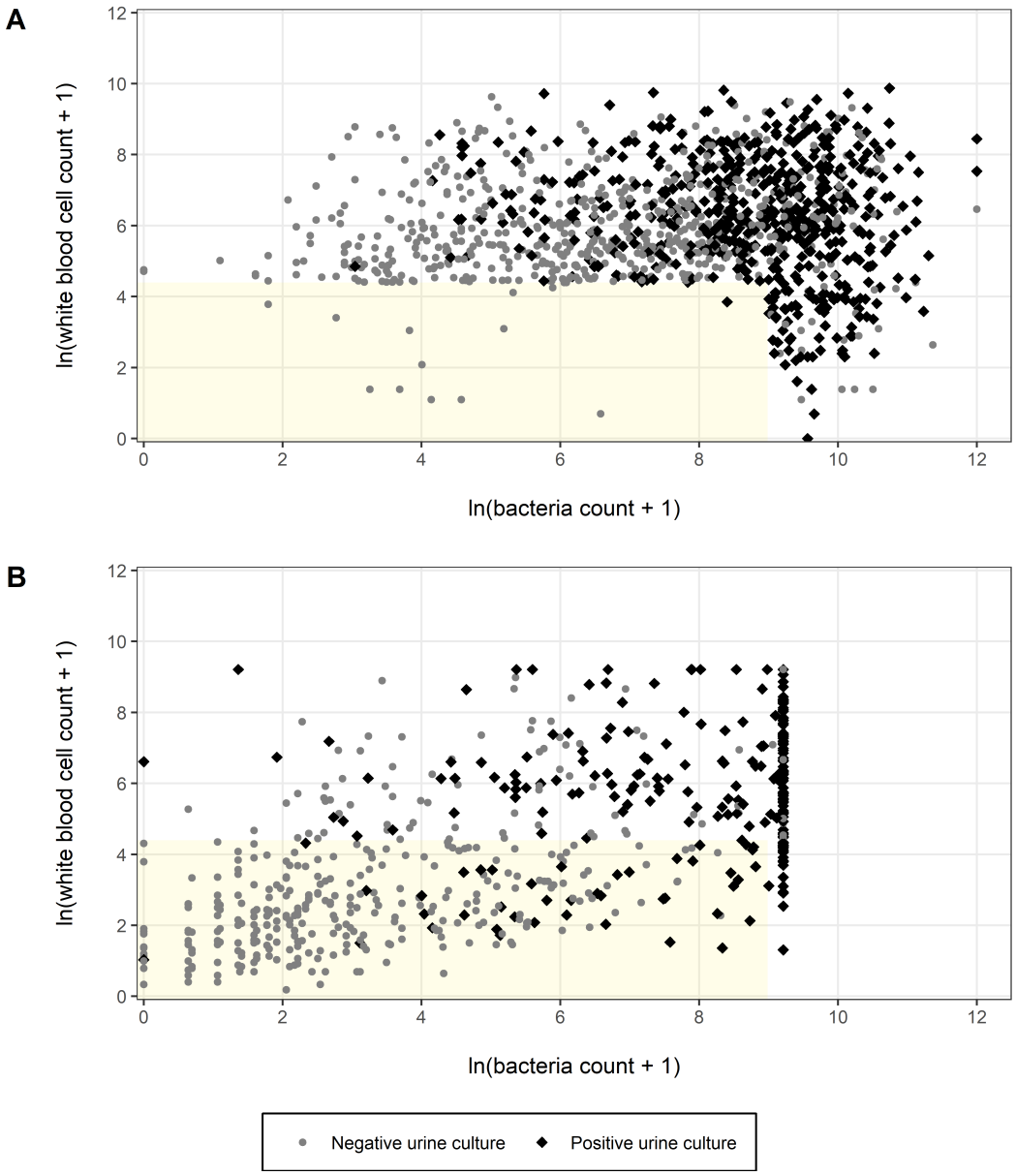

Figure 5.9: Distribution of bacteria and urinary WBC counts by bacterial growth in the **(A)** test set at QEHB compared to **(B)** samples analysed and published by Müller *et al.* (2018) [51].

QEHB, Queen Elizabeth Hospital Birmingham; WBC, white blood cell.

Müller *et al.* only tested their model on a single random hold-out set covering the same time frame as their training data [51, 109]. Nevertheless, all three studies as well as the primary care analysis by Little *et al.* agreed with the results presented here on the importance of a small number of urine flow cytometry and urine dipstick results in predicting likely urine culture growth. While dipstick results are available to clinicians, urine flow cytometry results do not usually get reported back to clinicians in real time and may be an opportunity to provide additional information to clinicians.

5.7 Conclusion

The analysis presented in this chapter suggests that both traditional statistical risk prediction models as well as machine learning may prove useful in diagnosing UTIs in the ED. These models exclusively used data available in the hospital's EHR system and may be directly integrated into electronic patient management systems to provide real-time prediction of a patient's risk of bacteriuria. In low-risk patients, antibiotic prescribing decisions may be delayed until results from such a model are available — most likely as soon as urine flow cytometry was performed. In patients at a higher risk of infectious complication, on the other hand, antibiotics may be initiated as usual and model predictions may later be used to revise empirical prescribing decisions.

While the model showed promising performance in predicting bacteriuria, comparisons to previous research and over time suggested considerable variations in model performance that may be expected if the model is deployed in real-world clinical practice. Possible explanations for these variations include differences in clinical practice or included patient populations. For example, the models developed in this chapter may work well in some patients like those consulting with clear urinary symptoms of UTI but may fail to identify presence or absence of bacteriuria in other patient groups. The model performance presented in this chapter and previously published studies only represent an average performance across all included patients, and does not allow to assess differential model performance

in key patient groups. Yet such differences may have important implication for the use of the model in clinical practice, where the model is likely to encounter differences to the environment it was trained in. Chapter 6 will therefore investigate the robustness of model performance to changes in patient or variable definitions. In order to provide an estimate of possible patient benefit that may be expected from using the model in clinical practice, Chapter 6 further critically compares the observed model performance to proxies of clinical decision making. The results will provide a better understanding of how and when the model works in clinical practice, helping future researchers and clinicians to use the model correctly [208].

Chapter summary

- Local data from the ED of a digitally mature hospital provided more granular data than primary care to identify and study UTI, yet important data items like urine dipsticks or previous antibiotic use in the community were still missing.
- A risk prediction model using only eleven variables (age, sex, history of positive urine culture, and urine flow cytometry results) achieved an AUROC that was very close to the top performance of a model using all 44 available variables, suggesting that most variables had little value in predicting bacterial growth once key variables were accounted for.
- If the model were to be used in clinical practice, robust data processing pipelines will be required within the hospital's IT infrastructure to provide cleaned variables to the model in real-time.
- Even then, observed performance changed over the study period and was lowest in 2015, when changes in laboratory procedures led to changes in the distribution of the data. Model performance thus needs to be monitored regularly during deployment to identify any future deterioration in performance.

- Meaningful comparison of model performance to previous published literature was difficult with notable differences in reported performances despite the use of similar predictors. Re-analysis of some published data suggested that these differences might be due to variations in outcome definitions and underlying patient populations.

My role in the work presented in this chapter

The study presented in this chapter was conceived by Dr Laura Shallcross and Professor Andrew Hayward. The research application was submitted to the Health Research Authority (HRA) for ethical approval by Dr Laura Shallcross. The first study protocol was drafted by Dr Laura Shallcross, Dr Milena Falcaro, Dr Martin Gill, and Prof Nick Freemantle. The study protocol was substantially edited and finalised by me with input from Dr Laura Shallcross, Dr Martin Gill, Dr David McNulty, Ms Orlagh Carroll, and Prof Nick Freemantle. Raw hospital data for this study was extracted from Queen Elizabeth Hospital Birmingham by Dr David McNulty and provided to me unaltered. All further data linkage and pre-processing was performed by me within the UCL Data Safe Haven. I performed all analyses presented in this chapter. All findings were interpreted by me and Dr Laura Shallcross. I wrote this chapter, with feedback from Dr Laura Shallcross, Prof Nick Freemantle, and Prof Andrew Hayward.

Software and code used in this chapter

All analyses in this chapter were performed using R (v3.6.2) and RStudio (v1.2.5033) on Windows 10. Data processing was performed using the *tidyverse* (1.3.0) and *data.table* (1.12.8) packages. All model building was performed using packages from the *tidymodels* (0.1.0) ecosystem, including *rsample* (0.0.5), *recipes* (0.1.9), *parsnip* (0.0.5), *tune* (0.0.1), *yardstick* (0.0.5), and *tidyposterior* (0.0.2). The models themselves were fit using the implementations provided in *mfp* (1.5.2), *glmnet* (3.0-2), *randomForest* (4.6-14), and *xgboost* (0.90.0.2). All code is publicly

available at https://github.com/prockenschaub/phd_code.

Publications resulting from this chapter

Rockenschaub P, Gill MJ, McNulty D, Carroll O, Freemantle N & Shallcross L. Development of risk prediction models to predict urine culture growth for adults with suspected urinary tract infection in the emergency department: protocol for an electronic health record study from a single UK university hospital. *Diagn Progn Res* 2020;4:15

Chapter 6

Variability in model performance when predicting bacteriuria in the ED: sensitivity analyses to inform the likely applicability of EHR models in clinical practice at QEHB

Abstract

Introduction: In the previous chapter, I used EHR data to develop a model which aimed to guide antibiotic prescribing decisions in the emergency department (ED) by predicting the presence/absence of bacteriuria. Although predictive performance in the whole cohort was promising, it is critical to understand how the model might perform in different population subsets, to understand whether it could be of clinical value to decision-making. Few risk prediction models are evaluated in this way and this partly explains why few models are adopted in clinical practice. In this chapter, I evaluate the use of my ED models in different target populations, focusing on factors that influence the presentation and natural history of UTI.

Background: Patients who have a urine sample sent for microbiological culture in the ED are likely to represent a heterogeneous group. While the models developed in Chapter 5 might work well in some of those patients, they might struggle to reliably predict bacterial growth in others. Understanding when the models work and how their performance compares to clinicians is important to judge their likely usefulness in clinical practice. The greatest clinical value of this model would be in ruling out bacterial UTI in low risk patients who have been treated empirically

with antibiotics for suspected UTI but actually have a different condition. To this end, I test my model in patients with and without ED diagnosed UTI, and by age and sex. I further explore the impact of changing the definition of bacteriuria by excluding/including samples classified as mixed growth, and compare model performances to proxies of clinicians' judgement.

Methods: Performance (area under the receiver operating characteristic [AUROC], specificity, and negative predictive value [NPV]) of a logistic regression (LR) model was estimated in key sub-populations (age, sex, and ED diagnosis) visiting the ED at Queen Elizabeth Hospital Birmingham (QEHB). I further investigated the robustness of results to definitions of bacteriuria — i.e., considering mixed growth samples as negative or positive growth, or excluding them from the analysis altogether. Finally, model performance was compared to proxies of clinical judgement based on ED diagnoses and antibiotic prescribing, or discharge diagnoses (admitted patients only).

Results: Compared to overall performance reported in Chapter 5, the model performed significantly worse in patients with an ED diagnosis of UTI, men, and older patients. The classification of mixed growth had a substantial influence on model performance, with AUROCs ranging from .790 (95% CI .767–.815; negative) to .888 (95% CI .870–.905; excluded). Prediction models consistently outperformed proxies of clinical judgement, but the validity of those proxies remained questionable.

Discussion: The results highlight the difficulties of retrospectively analysing UTI in the ED. In order to fully exploit the potential of electronic health records (EHR) for the diagnosis of UTI in the ED, careful intervention design and evaluation at several hospitals with different EHR systems, guidelines, and patient populations will be required.

6.1 Introduction

In Chapter 5, I developed and evaluated a statistical model to predict the risk of bacterial growth in urine samples collected in the emergency department (ED). The

results demonstrated that — in principle — we can predict bacteriuria in the ED with reasonable accuracy, suggesting scope to use these models to support clinicians in the diagnosis of suspected urinary tract infection (UTI). However, differences in model performance over time and compared to previous literature indicated that the developed models may not always be able to achieve the estimated performance when deployed in real-world clinical practice.

Apparent discrepancies between estimated and real-world performance may stem from the fact that estimates presented in Chapter 5 described an average model performance across a retrospective, broadly defined patient population. The study cohort included a wide range of patients who had a urine sample submitted for microbiological culture in the ED at Queen Elizabeth Hospital Birmingham (QEHB), many of whom had a sample cultured despite having no or only very vague evidence of UTI recorded in the electronic health record (EHR) system. Based on EHR data these individuals were unlikely to benefit from antibiotic treatment, but there may be good reasons why these patients were treated with antibiotics which are not evident from EHRs alone. A similar heterogeneity in patient populations was described in previous literature that developed similar risk prediction models for bacteriuria [51, 109]. This has important implications for how study findings need to be interpreted. Both the clinical need for as well as the ability to predict bacteriuria may vary considerably in such heterogeneous populations, and the average achievable model performance is likely to be poor measure of the model's ultimate value to clinical decision making. For example, patients who showed no clear indication of UTI in the ED may have had a very low probability of bacteriuria anyway. While models therefore might find it easier to rule out bacteriuria in those patients, the clinical benefit of predicting — or refuting — bacteriuria in this group may be questionable. There is therefore a discrepancy between developing and evaluating a retrospective predictive model, and creating a model that is truly meaningful for real-world clinical practice and decision making.

In this chapter, I expand on the results presented so far and explore in more detail how robust the estimated model performance presented in Chapter 5 is to

changes in patient mix and variable definitions. The analyses presented here aim to provide insight into whether the created models are likely to provide clinical benefit when deployed in practice. The chapter is divided into four subsections that each address a distinct question about the developed models' applicability to clinical care. First, I assess the performance of the model in patients with different recorded ED diagnoses and investigate variations in the estimated performance when looking at patients with and without recorded suspicion of UTI (Section 6.4). Second, I compare the performance of the model in age and sex stratified populations, which often warrant very different clinical management (Section 6.5). Third, I evaluate the impact of using different definitions of bacteriuria which include or exclude mixed growth during culture (which is usually regarded as contamination but may be treated with antibiotics; Section 6.6). Finally, I compare the estimated model performance to recorded UTI diagnoses and/or antibiotic prescribing as proxies of clinicians' ability to predict the likelihood of bacteriuria (Section 6.7).

6.2 Aims and Objectives

To evaluate the predictive performance of the model in specific patient subgroups and according to different variable definitions, in order to explore the feasibility of using the model to inform antibiotic prescribing decisions for suspected UTI.

Objectives:

6.1 To investigate the performance of the model to guide antibiotic prescribing decisions for patients with and without a recorded UTI diagnosis in the ED.

6.2 To investigate the performance of the model to guide antibiotic prescribing decisions by age and sex.

6.3 To investigate the impact on predictive performance of changing the definition of bacteriuria to include/exclude samples categorised as mixed growth.

6.4 To compare observed model performance to proxies of clinicians' ability to predict bacteriuria, overall and by admission status.

6.3 Data source, patient population, and variables

The data and patient cohort used in this chapter is identical to that described in Chapter 5. As before, the analyses included all non-pregnant adult patients who attended the ED at QEHB between November 1st 2011 and March 31st 2019, who had a urine sample sent for microbiological culture during their ED visit, and who had a valid record of age and sex. In line with the analysis in Chapter 5, patients who attended before January 1st 2018 were used as the training set, patients who attended between January 1st 2018 and March 31st 2018 were used as a calibration set, and all later samples were used as a temporally independent test set. All analyses presented in this chapter used subsets of this full patient cohort for model building and/or testing. Model performance was evaluated on the test set where possible but resorted to internal validation were necessary due to relatively small sample sizes in the subgroups. Definitions for each subgroup are given in the respective analyses sections below.

For simplicity and computational convenience, all models presented in this chapter were limited to logistic regression (LR) on the reduced predictor set (see Chapter 5 for a detailed definition), unless explicitly stated otherwise. In short, this means that models were built using only data on age, sex, history of positive urine culture, and urine flow cytometry parameters. The primary outcome remained urine culture growth of $\geq 10^4$ colony-forming units per millilitre (cfu/mL). Model building was performed as described in Section 5.4 using log transformation of continuous variables and mean imputation with missingness indicators.

6.4 Variations in model performance according to ED diagnosis

Dataset: ED patient cohort from Chapter 5

Evaluation: Internal and partial external validation (due to small sample sizes)

Patients who attend the ED and undergo urine sampling have a wide range of presenting complaints which may be associated with their likelihood of bacteriuria

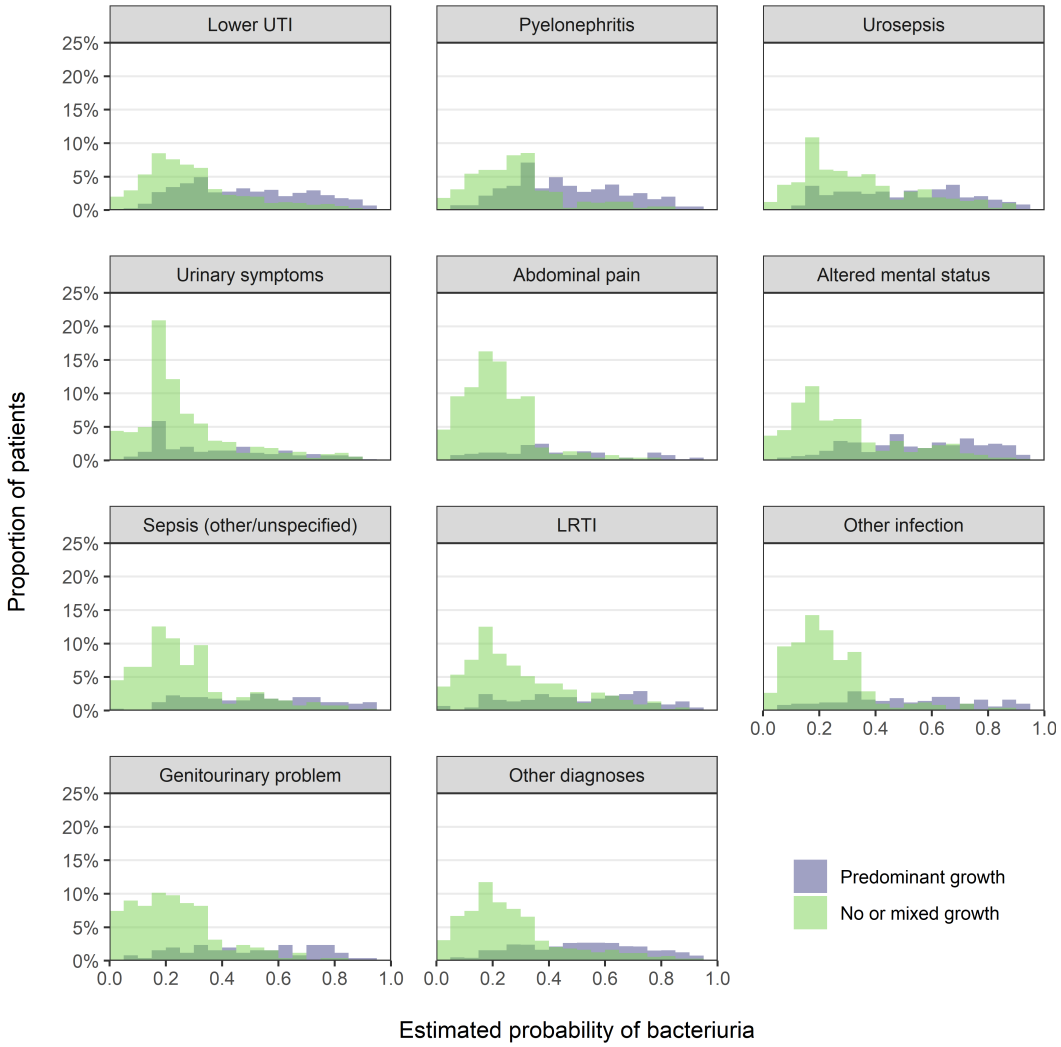

Figure 6.1: Distribution of the probability of bacterial growth in urine cultures predicted via LR in the entire training set, by ED diagnosis.

ED, emergency department; LR, logistic regression.

and whether a urinary pathogen (predominant growth) or a contaminant (mixed growth) is identified during urine culture. Patient presentation in the ED is recorded within structured EHR data using diagnostic codes. More detailed information may be available in document scans or free text but these aren't routinely available to researchers (see Appendix F for a comparison of coded ED diagnosis and free-text data). While codes are not a perfect substitute for a detailed description of symptoms, they do capture information on what the treating clinician thought important — and are thus a proxy for performed clinical investigations interpreted in the context of the clinician's experience. Probability of bacterial growth in

Table 6.1: Multivariable discriminative performance of a LR model using the reduced set of predictors and trained on the entire patient population when using it to predict bacterial growth by ED diagnosis during internal validation.

	AUROC (95% CI)	**Specificity** (95% CI)	**NPV** (95% CI)	P[1]
All patients	.776 (.769–.783)	29.3 (27.7–30.9)	91.1 (90.5–91.6)	
UTI diagnosis	.731 (.721–.740)	22.5 (21.7–23.3)	87.5 (85.9–89.0)	<0.001
Lower UTI	.736 (.726–.747)	23.0 (22.0–23.9)	87.3 (85.4–89.2)	0.003
Pyelonephritis	.746 (.721–.771)	24.5 (21.9–27.1)	85.9 (80.5–91.2)	0.018
Urosepsis	.705 (.685–.726)	19.2 (17.1–21.3)	91.5 (88.3–94.6)	<0.001
Symptoms attributable to UTI	.769 (.754–.784)	29.5 (27.6–31.3)	91.5 (89.7–93.2)	0.314
Urinary symptoms	.663 (.639–.687)	21.4 (19.2–23.7)	85.5 (81.6–89.4)	<0.001
Abdominal pain	.790 (.753–.827)	35.6 (32.8–38.4)	92.5 (90.0–95.0)	0.837
Altered mental status	.810 (.793–.828)	32.0 (28.8–35.2)	95.5 (93.3–97.7)	0.992
Other infections	.802 (.787–.818)	31.9 (30.0–33.7)	94.0 (92.9–95.0)	0.966
Sepsis (other)	.816 (.786–.845)	28.8 (25.7–31.9)	98.8 (97.5–100)	0.996
LRTI	.757 (.730–.783)	29.1 (26.2–32.1)	93.4 (90.8–96.1)	0.091
Other infection	.831 (.807–.856)	36.4 (32.7–40.0)	91.1 (88.6–9.35)	>0.999
Other diagnoses	.809 (.801–.818)	31.8 (30.5–33.1)	94.5 (93.6–95.5)	0.988
Genitourinary problem	.817 (.784–.851)	37.9 (34.6–41.1)	95.6 (92.9–98.2)	0.998
Other reason	.808 (.798–.817)	31.2 (29.8–32.7)	94.5 (93.5–95.5)	0.987

[1] Proportion of posterior samples in which the AUROC in the subgroup was larger than the AUROC in the entire patient population. Obtained by fitting a Bayesian GLMM to the AUROC estimated within each re-sample [192].
Note: The approximate Wald confidence intervals led to the confidence interval for Sepsis (other) to exceed one. To guarantee a sensible interpretation, the interval was capped at one.

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; ED, emergency department; GLMM, generalised linear mixed model; LRTI, lower respiratory tract infection; NPV, negative predictive value; UTI, urinary tract infection.

Chapter 5 was found to differ considerably by ED diagnosis (Table 5.3), ranging from as low as 17.4% in patients with abdominal pain to as high as 49.6% in patients with suspected pyelonephritis. Suspected ED diagnosis further had the third best performance of any single variable to distinguish samples with and without confirmed growth during microbiological culture (Table 5.5), asserting that the diagnosis captures some of the clinician's prior knowledge as to which patients might be at risk of bacteriuria.

Understanding when the model works — i.e., in which patients — and how and why a model might work better or worse for some ED diagnoses (e.g., due to differences in the relationships between key predictors and outcome in patients with and without suspicion of UTI) is important to judge whether the model can be used

Table 6.2: Multivariable discriminative performance of a LR model using the reduced set of predictors and trained separately for each ED diagnosis, when using it to predict bacterial growth during internal validation.

	AUROC (95% CI)	**Specificity** (95% CI)	**NPV** (95% CI)	P[1]
UTI diagnosis	.733 (.723–.742)	25.9 (24.8–27.1)	85.7 (85.0–86.4)	0.553
Lower UTI	.735 (.724–.747)	27.2 (25.7–28.7)	85.7 (84.7–86.6)	0.469
Pyelonephritis	.731 (.709–.753)	24.5 (20.4–28.6)	73.8 (68.8–78.8)	0.169
Urosepsis	.715 (.691–.739)	24.3 (19.9–28.8)	79.5 (75.3–83.7)	0.740
Symptoms attributable to UTI	.765 (.748–.782)	23.9 (19.4–28.4)	88.7 (86.0–91.5)	0.403
Urinary symptoms	.659 (.629–.689)	18.5 (10.7–26.2)	67.5 (56.5–78.4)	0.390
Abdominal pain	.769 (.725–.812)	35.5 (26.2–44.8)	88.7 (85.4–92.0)	0.088
Altered mental status	.798 (.779–.817)	31.1 (24.2–38.1)	80.2 (71.8–88.5)	0.219
Other infections	.799 (.781–.817)	30.5 (24.9–36.1)	91.0 (89.0–92.9)	0.419
Sepsis (other)	.783 (.741–.824)	37.8 (28.0–47.6)	75.8 (64.3–87.4)	0.018
LRTI	.724 (.694–.753)	22.7 (16.5–28.9)	79.2 (72.4–86.0)	0.017
Other infection	.827 (.801–.852)	38.7 (31.8–45.6)	88.8 (82.6–94.9)	0.383
Other diagnoses	.814 (.807–.821)	34.7 (31.8–37.6)	92.7 (92.0–93.3)	0.616
Genitourinary problem	.785 (.748–.822)	38.4 (29.3–47.6)	79.0 (71.6–86.4)	0.023
Other reason	.812 (.805–.819)	34.3 (31.4–37.2)	92.3 (91.6–93.0)	0.601

[1] Proportion of posterior samples in which the AUROC of the model trained only on data from the subgroups (this table) was larger than the AUROC of a model trained on the entire patient population (Table 6.1). Obtained by fitting a Bayesian GLMM to the AUROC estimated within each re-sample [192].

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; ED, emergency department; GLMM, generalised linear mixed model; LRTI, lower respiratory tract infection; NPV, negative predictive value; UTI, urinary tract infection.

in clinical practice [208]. This knowledge may help to anticipate how the model will perform in key populations of interest, and how this performance might change when used at other healthcare sites with different patient case mix [209]. This section therefore estimates differences in model performance and effect sizes when developing and/or evaluating the model in patients with different ED diagnoses[1].

6.4.1 Statistical analysis

A LR model using the reduced set of predictors was fitted to the entire training set (global model), and the distribution of predicted probabilities in the training set

[1]Note that ED diagnoses had already been included as a candidate predictor in models in Chapter 5. This, however, does not fully account for performance differences. Even after including ED diagnoses as predictor, models may perform better in patients with certain diagnoses than in patients with others — for example if a condition makes it less likely that sufficient urine for urine flow cytometry can be sampled. In this case, the model would have less information for patients with this diagnosis and will consequently struggle to predict bacteriuria in those patients.

were plotted by ED diagnosis. The area under the receiver operating characteristic (AUROC), specificity, and negative predictive value (NPV)[2] were then calculated using internal validation as described in Section 5.4.6.4. Separate estimates of AUROC, specificity, and NPV were calculated for each ED diagnosis by applying the global model to only those patients in the hold-out sets that had the respective diagnosis. Specificity and NPV were evaluated at the threshold that achieved 95% sensitivity in the entire patient cohort. Differences in the estimated AUROC when evaluating the model on the entire cohort as opposed to each sub-population were assessed using Bayesian generalised linear mixed models (GLMMs) [192]. Performance was also evaluated on the external test set but due to small numbers only a comparison of explicitly recorded suspicion of UTI (combining urinary symptoms, lower UTI, pyelonephritis, and urosepsis) versus all other diagnoses was performed.

Next, local LR models were fitted for each ED diagnosis separately using only patients with that particular diagnosis. AUROC, specificity, and NPV were re-calculated for all local models and compared to the performance estimated for the same patient group when predicted via the global model. Specificity and NPV were now evaluated at the threshold that achieved 95% sensitivity in the sub-population. The AUROC achieved by each local models was compared to the performance of the global model using Bayesian GLMMs [192].

Finally, effect sizes of the fitted coefficients of all local models were compared graphically to look for ED diagnoses with a notably different relationship between predictors and the outcome. Bayesian GLMMs were used to gradually relax assumptions of fixed coefficients across sub-populations in the global model by adding a random effect for ED diagnosis. Three alternative global models were fitted to the entire training data: a model with only fixed effects, a model with intercepts that varied by ED diagnosis but fixed slopes, and a model with intercepts and slopes that varied by ED diagnosis. Model fit of the alternative global models

[2] No area under the precision-recall curve (AUPRC) was calculated due to differences in the prevalence of bacteriuria between patient groups.

was compared using the Widely Applicable Information Criterion (WAIC)[3] [212]. All Bayesian GLMMs were fitted using MCMC sampling with 4 chains of 2,000 warm-up iterations and 2,000 sampling iterations

6.4.2 Results

The distribution of predicted probabilities of bacterial growth in the training set differed notably between ED diagnoses. Samples in patients *without* a recorded suspicion of UTI were more often — and more often correctly — assigned a low probability of bacterial growth (Figure 6.1). Samples with an explicitly recorded suspicion of UTI, on the other hand, had both a relatively large share of samples with high estimated probability and no bacterial growth as well as low estimated probability despite bacterial growth. This observation was reflected in the estimated performances when stratified by ED diagnosis (Table 6.1). Patients with a suspicion of UTI or urinary symptoms had a statistically significant lower estimated performance than other subgroups. The lowest performance during internal validation was observed in patients with UTI symptoms only (AUROC .663, 95% CI .639–.687) followed by those with suspected urosepsis (AUROC .705, 95% CI .685–.726). The performance in patients without ED diagnoses relating to UTI or urinary symptoms was similar to or above that estimated for the entire cohort (AUROC .776, 95% CI .769—.783). Estimated model performance did not meaningfully improve by fitting a separate model for each ED diagnosis (Table 6.2), and even reduced for patients with suspected lower respiratory tract infection (LRTI) and patients with other suspected genitourinary problems — potentially due to too small sample sizes in those patient groups. A notably smaller difference was observed when evaluating the model on the external test set, resulting in AUROC .780 (95% CI .748–.816) for patients with a UTI diagnosis or urinary symptom and AUROC .799 (95% CI .767–.828) for all other patients. Clinical coding changed in 2017 (see Section 3.3.2 for a detailed discussion) and ED diagnoses did not contain

[3]WAIC was used to compare Bayesian models in favour of the more traditional Deviance Information Criterion (DIC) because the latter is based on a point estimate [210] and thus not fully Bayesian. This may lead to awkward situations in which the DIC may produce a negative estimate of the effective number of model parameters. WAIC avoids these situations by using the entire posterior distribution [211].

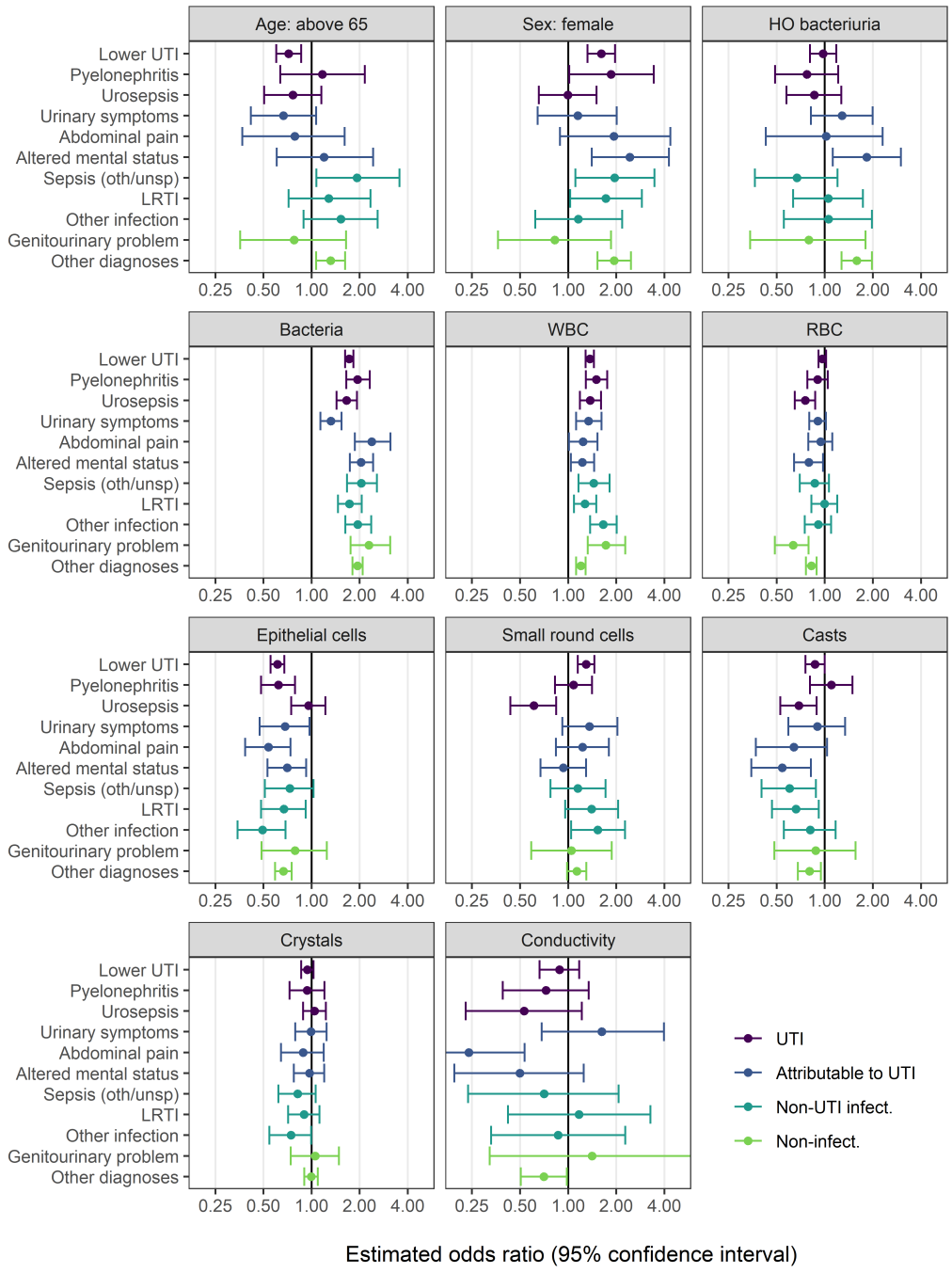

Figure 6.2: Estimated coefficients of an LR model trained separately on patients with each ED diagnosis. Boxes depict the effect of a single predictor (e.g., age above 65 years) estimated via multivariable analysis using only patients with the ED diagnosis listed on the y-axis (e.g. lower UTI). ED diagnosis are grouped by colour into explicit UTI diagnoses, diagnoses attributable to UTI, infectious diagnoses other than UTI, and non-infectious diagnoses.

ED, emergency department; LR, logistic regression; UTI, urinary tract infection.

any patients with urosepsis or abdominal pain after the change, as these codes did not exist in the updated terminology. This suggests that the meaning of ED diagnoses might have shifted over time, perhaps explaining some of the difference in results when looking at internal versus external validation.

Despite the differences in observed performance across ED diagnoses, there was only moderate evidence for variations in the relationship between predictors and the outcome (Figure 6.2). Notably, higher age tended to reduce the chance of bacterial growth in patients with a lower UTI diagnosis, whereas it increased the chance in patients with suspicion of sepsis or other, non-infectious diagnoses. A history of positive urine culture similarly increased the probability of bacterial growth in patients with altered mental status or other, non-infectious diagnoses. Presence of small round cells increased the probability of growth in patients with suspected lower UTI, while decreasing probability in patients with suspected urosepsis. When relaxing assumptions about fixed coefficient effects across all ED diagnoses, a model with varying intercepts and slopes was preferred on the basis of WAIC. The estimated deterioration in WAIC was Δ WAIC$_{intercept} = 45.1$ (standard error 10.9) for a model with varying intercepts and ΔWAIC$_{fixed} = 128.9$ (standard error 17.3) for a model with only fixed effects. However, model performance only marginally improved from .790 (95% CI .769–.811, Table 5.8) to .796 (95% CI .776–817) when a model with varying intercepts and slopes was applied to the external test set.

6.4.3 Discussion

6.4.3.1 Clinical findings

Evaluating model performance by ED diagnosis showed that the model performed significantly worse in patients with a recorded suspicion of UTI or urinary symptoms. There are several possible explanations for these findings. A large proportion of patients with non-UTI diagnoses may never have been at risk of UTI and may have had a urine sample sent in the absence of any urinary symptoms [207]. Patients with recorded suspicion of UTI may further have been more likely to have had antibiotics prior to arrival in hospital. For example, some patients who

consulted for UTI in the ED might have first consulted their General Practitioner and already received initial treatment in the community, thwarting bacterial growth (see Chapters 2 and 5 for a detailed discussion of the possible effect of prior antibiotic treatment on urine culture growth). If ED clinicians were more likely to label patients who were previously diagnosed with UTI in the community as "having UTI", the corresponding higher prevalence of prior antibiotic treatment in the community might render the prediction of bacteriuria more difficult in this patient group. Surprisingly, the model performed worst in patients with urinary symptoms alone. This might have been due to the inclusion of patients with catheter problems in this subgroup. Samples from patients with catheters might differ fundamentally from those from other patients with suspected UTI[4]. A model trained on patients with and without catheters might therefore not be appropriate in patients with catheters, or predicting bacterial growth in catheter urine might be a more difficult prediction task. Patients with a diagnosis of urinary symptoms also contained a high proportion of patients with missing urine flow cytometry values, leading to a notable peak of the predicted probabilities around $\sim$0.2 (Figure 6.1). As discussed in Section 4.4.6, urine flow cytometry values could not be obtained if the urine sample was either too viscous or if the quantity of urine wasn't sufficient. The most commonly recorded urinary symptom was urinary retention, which may have made it less likely that sufficient urine could be collected for those patients.

6.4.3.2 Methodological findings

Examination of the predicted probabilities suggested that the observed reduced performance in patients with suspicion of UTI may be driven by the fact that fewer patients in that group had a low predicted probability of bacterial growth. Patients without suspicion of UTI, on the other hand, commonly showed a clear cluster of patients with low probabilities (Figure 6.1). These patients would have few bacteria or urinary white blood cells (WBCs) found during urine flow cytometry, making it very unlikely that the sample would show predominant bacterial growth. This has important implications for the apparent model performance. AUROC — the main

[4]For example, they warrant their own national guidelines [213].

metric used in this analysis — represents the probability that a patient whose urine sample eventually exhibits bacterial growth will be assigned a higher score by the model than a patient whose urine sample did not grow bacteria (see Section 5.4.6.3). The inclusion of "easy targets" that had a very low probability may therefore artificially boost the estimated model performance. While the model struggles to distinguish bacterial growth from no growth in a group of patients with suspicion of UTI (a primary patient group of interest), it does so more successfully when applied to all patients. In the latter case, the model is also rewarded for correctly distinguishing bacterial growth in pairs of patient where one has a suspicion of UTI while the other does not. This may be an easier task but one that is unlikely to be of clinical relevance.

Surprisingly, estimated strengths and directions of model coefficients were comparable between ED diagnoses and fitting separate models to each group did not improve performance. This suggests that the differences in performance were not due to different predictor-outcome relationships. Instead, differences are more likely to be due to differences in covariates between ED diagnoses or due to unmeasured predictors not currently identifiable from EHR data.

6.5 Variations in model performance according to age and sex

Dataset: ED patient cohort from Chapter 5

Evaluation: Internal and external validation

Expected model performance might not only depend on suspected diagnosis in the ED, but may also be influenced by a patient's age and sex. In particular, the prevalence of asymptomatic bacteriuria may vary significantly between these patient groups, with possible implications for model performance and interpretation. Prevalence of asymptomatic bacteriuria was previously found to increase with age, and to be more prevalent in women (see Chapter 1). It is unclear how an increased prevalence of asymptomatic bacteriuria — or other patient

Table 6.3: Multivariable discriminative performance of a LR model trained on the entire patient population when using it to predict bacterial growth by age and sex during internal validation.

	AUROC (95% CI)	**Specificity** (95% CI)	**NPV** (95% CI)	P[1]
Sex				
Male	.733 (.719–.748)	33.0 (31.3–34.7)	90.0 (88.5–91.5)	<0.001
Female	.797 (.787–.807)	26.3 (25.5–27.2)	92.7 (91.6–93.7)	>0.999
Age				
$<$ 65 years	.791 (.781–.801)	30.5 (29.6–31.4)	92.3 (91.1–93.5)	0.998
$\geq$ 65 years	.760 (.751–.770)	27.1 (25.9–28.2)	90.8 (89.4–92.1)	0.003

[1] Proportion of posterior samples in which the AUROC in the subgroup was larger than the AUROC in the entire patient population. Obtained by fitting a Bayesian GLMM to the AUROC estimated within each re-sample [192].

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; GLMM, generalised linear mixed model; LRTI, lower respiratory tract infection; NPV, negative predictive value.

characteristics linked to age and sex — may influence the capability of the model developed in Chapter 5 to predict bacterial growth in ED urine samples. This section therefore investigates differences in model performance when applying the model to samples collected from men and women and from patients aged <65 years and $\geq$65 years.

6.5.1 Statistical analysis

Similar to Section 6.4, a LR model using the reduced set of predictors was fitted to the full dataset and evaluated in subgroups defined by age ($<$ 65 years, $\geq$ 65 years) and sex. Evaluation was performed via both internal and external validation. Discriminatory performance was assessed via AUROC, specificity, and NPV, and differences in AUROC were tested via Bayesian GLMMs [192]. Since the previous analysis showed a difference in performance for patients with suspected UTI (combining urinary symptoms, lower UTI, pyelonephritis, urosepsis), the stratification by age and sex was repeated in the subset of patients with suspected UTI. Model calibration was assessed graphically in each subgroup using calibration plots. As model probabilities showed some miscalibration in Chapter 5, probabilities were re-calibrated before plotting using Platt scaling (see Section 5.4.6.4).

6.5.2 Results

Performance was statistically significantly better in women (AUROC .797, 95% CI .787–.807) compared to men (AUROC .733, 95% CI .719–.748) and in patients under the age of 65 years (AUROC .791, 95% CI .781–.801) compared to those above the age of 65 years (AUROC .760, 95% CI .751–.770; Table 6.3). Performance further improved minimally when evaluation was limited to women under the age of 65 years (AUROC .798, 95% CI .787–.809). The observed changes in model performance were similar when limiting the analysis to only patients with a recorded suspicion of UTI (Table 6.4). The model consistently performed worst in male patients regardless of age or suspected UTI diagnosis. Differences in AUROC were further exacerbated during external validation (AUROC .824, 95% .798–.848 in women; and AUROC .822, 95% .790–.853 in patients <65 years). Despite considerably lower performance in terms of AUROC, the model achieved higher specificity in men (Tables 6.3 and 6.4) and in older patients with suspected UTI (Table 6.4). Model calibration remained reasonable in women, but the model showed considerable miscalibration in men even after re-calibration (Figure 6.3). Especially in men aged <65 years, the model severely underestimated the probability of bacterial growth during urine culture.

6.5.3 Discussion

6.5.3.1 Clinical findings

Stratification of model predictions by age and sex showed a better discriminatory performance in patients <65 years, and better discriminatory performance and calibration in women, irrespective of ED diagnosis. Observed differences in performance did not corresponded directly with previously estimated prevalence of asymptomatic bacteriuria [33, 34, 35]. While performance was lower in elderly patients, which are more likely to have asymptomatic bacteriuria, performance was also lower in men, which are less likely to have asymptomatic bacteriuria (see Chapter 1). This suggests other underlying factors that impact performance in patients with different age and sex. UTIs in men are usually rare and — if they

occur — are considered as a complicated UTI by default [17], warranting separate guidance [13]. Men have previously been reported to have different aetiology than women [214], which was also observed in the data used here. While four out of five isolated pathogens (81.5%) in women were *Escherichia coli*, they made up only about half (54.8%) of all isolates in men included in this analysis. A potential reason for this discrepancy may be an increased prevalence of catheter use in included male patients, but the available data did not contain enough detail to ascertain the presence of a catheter at arrival.

6.5.3.2 Methodological findings

The observed differences in performance suggest that men and women might require separate prediction models when used in clinical practice, as prediction of bacterial growth was generally easier and more reliable in women both in terms of model discrimination and calibration. Age, on the other hand, was a less important factor. Performance for example hardly changed when comparing the predictions in all women to predictions in younger women only. Age was unevenly distributed between the sexes — with only 44% of women aged $\geq$65 years compared to 69% of men. The observed variation by age could therefore be partially explained by an underlying correlation with sex, suggesting that a stratification by sex may be sufficient to ensure adequate performance in clinical practice. Finally, differences in performances by age and sex widened when looking only at data from 2018/19. This suggests that the increased performance observed during external validation in Chapter 5 may have been driven primarily by an increased performance in younger patients and women in those years.

6.6 Ambiguity introduced by mixed growth

Dataset: ED patient cohort from Chapter 5

Evaluation: Internal and external validation

Following common clinical practice in England [43] and at QEHB, samples with mixed growth in the absence of a predominant isolate were considered negative —

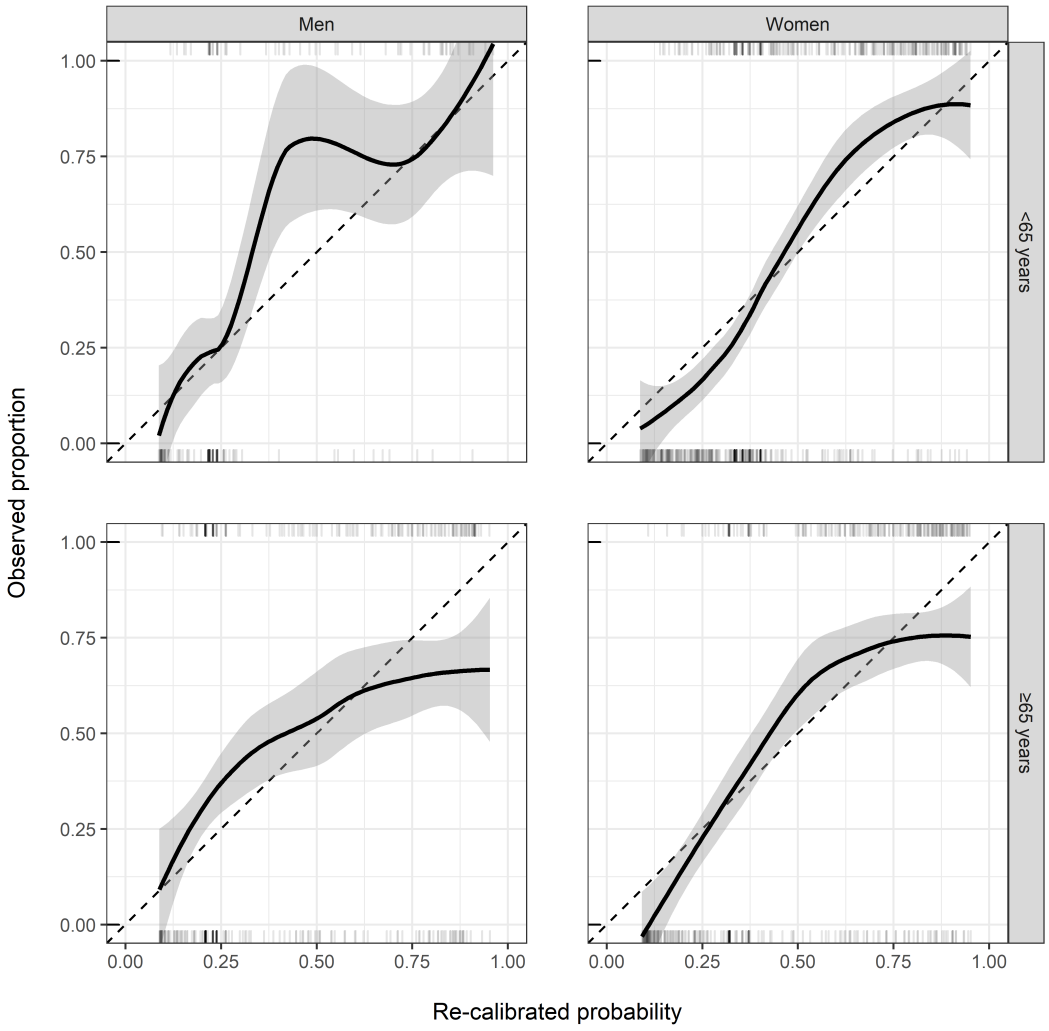

Figure 6.3: Calibration of LR model predictions when predicting bacterial growth by age and sex.

ED, emergency department; LR, logistic regression.

i.e., equivalent to no growth. Treating mixed growth as no growth is justified by the observation that mixed growth often does not represent bacteria in the patient's urine but instead indicates contamination of the urine during sampling (see Section 1.3.2 for a more detailed discussion) [42]. However, this view has been occasionally challenged. A study from Israel found that 30% of patients hospitalised with urosepsis had mixed growth in their urine samples. In 50% of those cases, an isolate in the mixed growth could be matched with bacteria grown from blood samples taken from the same patient, or mixed growth remained present in multiple sequential urine cultures of that patient [215]. Another study found that *E. coli* may

Table 6.4: Multivariable discriminative performance of a LR model trained on UTI patients when using it to predict bacterial growth by age and sex during internal validation.

	AUROC (95% CI)	**Specificity** (95% CI)	**NPV** (95% CI)	Prob[1]
Sex				
Male	.702 (.682–.722)	30.8 (28.8–32.8)	86.5 (84.1–88.9)	<0.001
Female	.744 (.729–.759)	19.0 (17.4–20.6)	88.1 (85.7–90.4)	0.965
Age				
$<$ 65 years	.747 (.731–.763)	19.8 (18.3–21.3)	89.5 (86.8–92.1)	0.983
$\geq$ 65 years	.716 (.701–.731)	28.0 (26.5–29.4)	85.8 (83.8–87.8)	0.029

[1] Proportion of posterior samples in which the AUROC in the subgroup was larger than the AUROC in the entire patient population. Obtained by fitting a Bayesian GLMM to the AUROC estimated within each re-sample [192].

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; GLMM, generalised linear mixed model; LRTI, lower respiratory tract infection; NPV, negative predictive value.

potentially be more invasive in the presence of other pathogens [216], and a study comparing microbiological culture and 16S rRNA metagenomic sequencing on mid-stream urine from symptomatic patients and asymptomatic controls found that mixed growth of symptomatic patients more often contained *Enterobacteriaceae*, which are uropathogens. Caution is therefore warranted when overly relying on the assumption that mixed growth always equals no growth. Unless the urine sample is repeated, we do not obtain a definitive result for a patient whose initial sample showed mixed growth. Instead, the underlying sample is simply labelled as negative and could have either contained or not contained pathogens before any contamination happened, or may not have been contaminated at all.

Mixed growth was common in the cohort used in this analysis, accounting for 23.5% of all samples and 37.4% of negative samples. Due to its frequency, mixed growth might have had a large impact on model training and evaluation. Just like urine sampled from a patient with true bacteriuria, samples which later show mixed growth will most likely have had high bacteria counts during urine flow cytometry. While these bacteria might have entered the urine through imperfect sampling, urine flow cytometry cannot distinguish between bacteria that was present in the patient's urinary tract prior to sampling and bacteria that entered the urine for example from the skin. This might be problematic, since bacteria counts were the most powerful predictor of culture growth (see Table 5.5). Prediction models

may therefore struggle to separate predominant growth (which were labelled as positive) from mixed growth (which were considered negative in Chapter 5). By universally classifying mixed growth as no growth, the inclusion of mixed growths in the negative population might therefore have confounded the predictive models and biased the estimated relationship between bacteriuria counts during urine flow cytometry and probability of culture growth.

This section assesses the distribution of predicted probabilities in samples with predominant growth, mixed growth, and no growth, and investigates differences in the estimated model performance if mixed growth is excluded from the analysis or considered positive growth.

Remark: (Discriminatory performance in multinomial settings). Metrics of discriminatory performance quantify how well model predictions can separate one outcome category from another. When the outcome is binary, this comparison is well defined. For example, the AUROC in a binary problem measures the probability that a randomly chosen patient with the outcome is assigned a higher score than a randomly chosen patient without the outcome [190]. How performance should be quantified in multinomial problems is less clear-cut and performance might depend on which outcome categories are compared. While some outcome categories might be well separated, others might not be. One might use a one-vs-all comparison — i.e., a separate AUROC is calculated for each category, estimating how well that category is separated from a single category made up of all other categories. However, this results in as many performance metrics as there are categories. If a single metric is desired, it may be calculated as a simple average or a weighted average of all one-vs-all comparisons. Alternatively, Hand and Till (2001) [217] proposed a natural extension of the AUROC that directly calculates the average probability that a patient with outcome i is assigned a higher score for i than a patient from another outcome j.

Table 6.5: Proportion of urine samples classified as positive bacterial growth and changes in discriminative performance of a LR model under different classifications of mixed growth during internal validation.

Classification of mixed growth	All patients		ED diagnosis of UTI	
	% positive	AUROC (95% CI)	% positive	AUROC (95% CI)
Considered negative[1]	35.1	.776 (.769–.783)	42.1	.731 (.721–.740)
Excluded	46.1	.870 (.864–.875)	55.9	.836 (.828–.843)
Considered positive	58.9	.842 (.838–.847)	66.8	.823 (.815–.830)

[1] This definition was used in the main analysis in Chapter 5.

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; ED, emergency department; LR, logistic regression; UTI, urinary tract infection.

6.6.1 Statistical analysis

The distribution of the two primary urine flow cytometry parameters — bacteria and urinary WBC counts — was stratified by culture result (predominant growth, mixed growth, no growth) and compared graphically via box plots. The distribution of probabilities predicted by a LR model using the reduced set of predictors fitted on all training data was further compared via histograms stratified by culture result.

Separate LR models were fitted using two alternative classifications of mixed growth. First, mixed growth was excluded entirely from the analysis and the model was fitted only on samples with predominant growth or no growth. Second, mixed growth was reclassified as positive growth, thus predicting samples with any growth (predominant or mixed) versus samples with no growth. Model performance was estimated via AUROC in both internal and external validation, and the entire procedure was repeated using only patients with an ED diagnosis of UTI (combining urinary symptoms, lower UTI, pyelonephritis, and urosepsis).

The possibility of estimating mixed growth separately from predominant growth and no growth was investigated using multinomial logistic regression. Instead of fitting one set of coefficients, separate coefficients were fitted for predicting predominant growth and mixed growth, with no growth as the reference category. Model performance was evaluated using binary one-vs-all AUROC as well as multinomial adaptions of AUROC (simple average, weighted average, and the Hand-Till method [217]) in both internal and external validation sets.

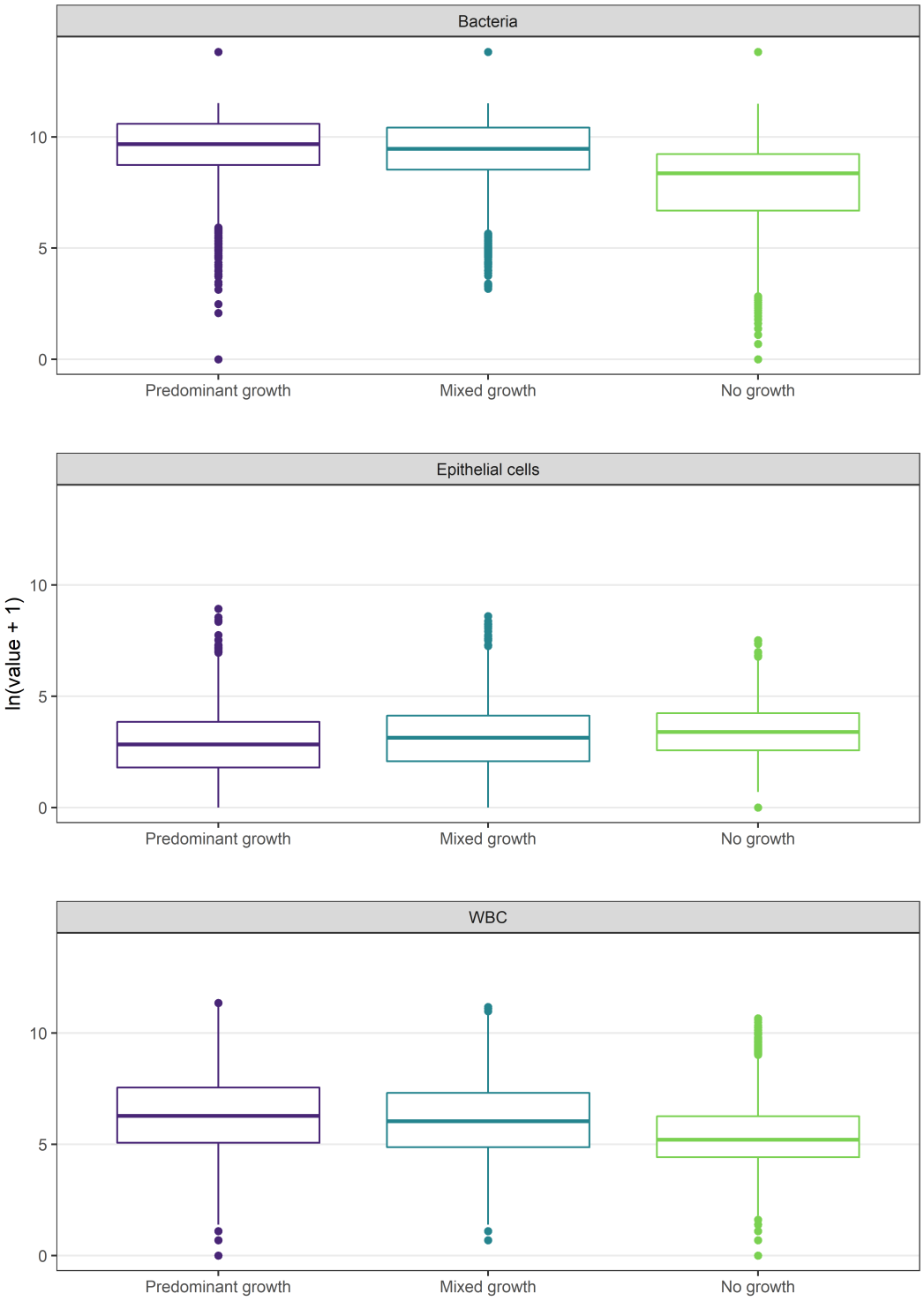

Figure 6.4: Distribution of bacteria counts, epithelial cells, and urinary WBC, stratified by urine culture results.

WBC, white blood cell counts.

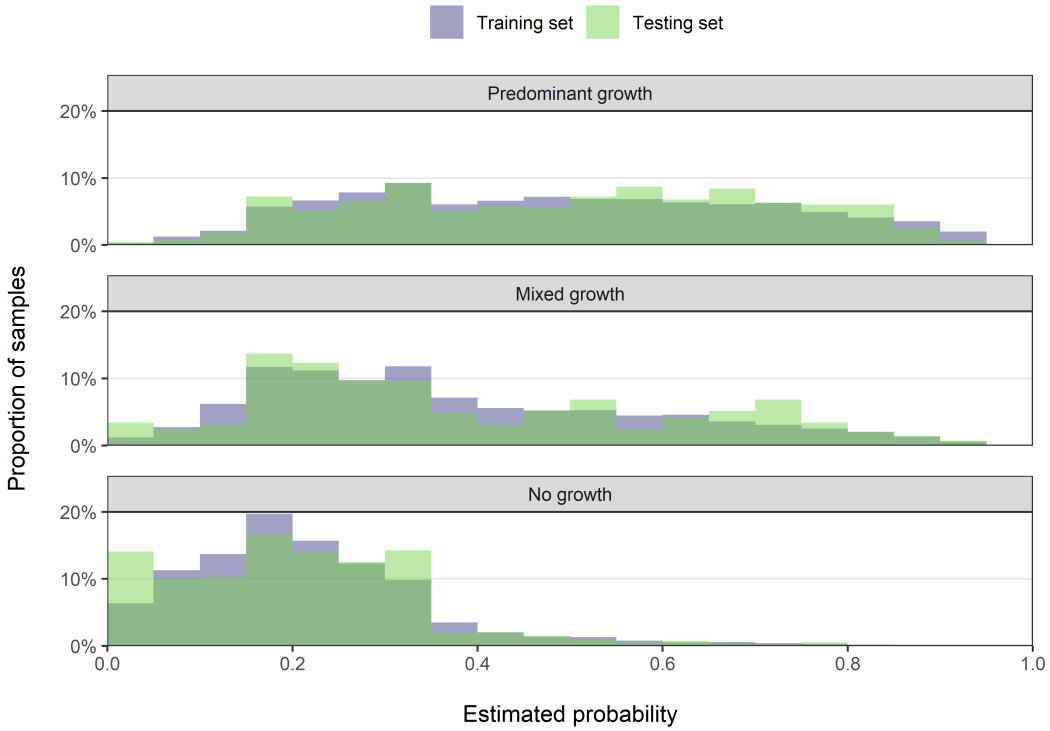

Figure 6.5: Distribution of predicted probability of bacterial growth in urine samples predicted by LR, stratified by ED urine culture results.

LR, logistic regression.

6.6.2 Results

The distribution of both bacteria counts and urinary WBC counts in mixed growth samples lay between those of samples with predominant growth and samples with no growth, although they were more similar to the former (Figure 6.4). A similar pattern could be observed for model predictions (Figure 6.5). While the mode of the predicted probabilities of mixed growth samples was close to the mode of samples without growth, predictions were skewed to the right and almost half of all samples with mixed growth had a predicted probability of 0.4 or higher. This was not observed in samples without growth, for which a probability of ≥ 0.4 was observed in only around 9% of samples.

The difficulty of correctly assigning mixed growth was also reflected in model performances. While a LR model that considered mixed growth to be negative achieved an AUROC of .790 (95% CI .767–.815) in the external test set, performance increased to AUROC .888 (95% CI .870—.905) when excluding

Table 6.6: Proportion of urine samples classified as positive bacterial growth and changes in discriminative performance of a LR model under different classifications of mixed growth during external validation.

Classification of mixed growth	All patients		ED diagnosis of UTI	
	% positive	AUROC (95% CI)	% positive	AUROC (95% CI)
Considered negative[1]	44.1	.790 (.767–.815)	50.6	.769 (.734–.802)
Excluded	54.4	.888 (.870–.905)	61.8	.858 (.825–.886)
Considered positive	63.1	.866 (.847–.885)	68.7	.846 (.817–.875)

[1] This definition was used in the main analysis in Chapter 5.

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; ED, emergency department; LR, logistic regression; UTI, urinary tract infection.

mixed growth altogether and AUROC .866 (95% CI .847-–.885) when considering it positive growth (Table 6.6). Results using only patients with an ED diagnosis of UTI (Table 6.6) and from internal validation were highly comparable (Table 6.5).

Although mixed growth was associated with patient characteristics and medical history (Table 6.7), a multinomial regression model using the reduced predictor set had difficulties distinguishing mixed growth from predominant growth or no growth. While the predominant growth and no growth were well separated with one-vs-all AUROCs of .798 (95% .776–.819) and .861 (95% .843–.878), the model only achieved an AUROC of .629 (95% CI .594–.666) when trying to distinguish mixed growth from no mixed growth — i.e., a combined class of predominant growth and no growth. Similar results were found when using all predictors instead of just the reduced set of predictors. The one-vs-all AUROCs of the multinomial regression where therefore close to the results obtained by considering mixed growth as either positive or negative in binary logistic regression (Tables 6.5 and 6.6), with some but not much extra information gained from separately estimating mixed growth. The relatively bad performance when trying to identify mixed growth samples was reflected differently in the summary AUROC measures. With .749, the Hand-Till method yielded the lowest summary AUROC, emphasising the low performance that would be obtained by comparing mixed growth with either predominant or no growth. The simple macro average was higher at .762, since now only one of the three comparisons considers mixed growth on

Table 6.7: Characteristics and medical histories of patients with ED urine cultures that resulted in mixed growth.

	Overall	Bacterial growth		p-value[1]
		No or predominant	Mixed	
Total number of visits (%)	12,680 (100.0)	9,742 (76.8)	2,938 (23.2)	
Age ≥65 years (%)	6,584 (51.9)	4,635 (70.4)	1,949 (29.6)	<0.001
Female (%)[2]	8,368 (66.0)	6,694 (80.0)	1,674 (20.0)	<0.001
Ethnicity (%)				
White	9,256 (73.0)	6,959 (75.2)	2,297 (24.8)	<0.001
Asian	1,671 (13.2)	1,367 (81.8)	304 (18.2)	
Black	552 (4.4)	421 (76.3)	131 (23.7)	
Mixed or other	526 (4.1)	443 (84.2)	83 (15.8)	
Unknown	675 (5.3)	552 (81.8)	123 (18.2)	
CCI (%)				
0	7,386 (58.2)	6,050 (81.9)	1,336 (18.1)	<0.001
1-2	3,023 (23.8)	2,186 (72.3)	837 (27.7)	
≥3	2,271 (17.9)	1,506 (66.3)	765 (33.7)	
Individual comorbidities (%)[2]				
Cancer	915 (7.2)	658 (71.9)	257 (28.1)	<0.001
Underlying renal condition	2,733 (66.4)	1,815 (66.4)	918 (33.6)	<0.001
Underlying urological condition	3,614 (28.5)	2,347 (64.9)	1,267 (35.1)	<0.001
Renal or urological surgery	2,484 (19.6)	1,543 (62.1)	941 (37.9)	<0.001
Hospital activity in prior year (%)[2]				
Any hospitalisation	6,067 (47.8)	4,319 (71.2)	1,748 (28.8)	<0.001
Urine sample taken	6,195 (48.9)	4,364 (70.4)	1,831 (29.6)	<0.001
Urine sample positive	3,062 (24.1)	2,137 (69.8)	925 (30.2)	<0.001
Antibiotics in hospital	3,194 (25.2)	2,131 (66.7)	1,063 (33.3)	<0.001

[1] Obtained via χ^2 tests. [2] These were binary yes/no variables. For legibility, only the positive (yes) category is shown. p-values represent comparisons with the negative/no category. Note: Percentages in the overall column represent column-%, whereas percentages in the bacterial growth columns represent row-%.

CCI, Charlson Comorbidity Index.

its own. Finally, the highest AUROC was estimated by weighted macro averaging, whose AUROC of .789 reflects the relatively higher frequency of predominant or no growth in the dataset.

6.6.3 Discussion

6.6.3.1 Clinical findings

Due to the frequency of mixed growth in my sample, the proportion of culture-positive samples varied considerably depending on whether mixed growth was considered positive, negative, or was excluded. The findings presented here are

in line with results reported by Müller *et al.* (2018) [51]. Using a LR model based on flow cytometry measurements, they report an AUROC of .660 (95% .610–.700) when trying to predict mixed growth from non-mixed growth, compared to the performance of AUROC .629 (95% CI .594–.666) found here. They concluded that standard urine flow cytometry results do not reliably identify mixed growth.

The inspected model probabilities might further suggest that the mixed growth group might have contained both samples that should be considered positive — i.e., genuine polymicrobial growth — as well as samples that should be considered negative due to contamination. Other authors have argued that strict microbiological protocols might miss important bacteriuria [215, 218], and urine flow cytometry measurements together with predictive models such as the one presented here might be used to distinguish between them. Future research may determine the prevalence and importance of true polymicrobial growth and whether the distribution of model predictions observed here indeed reflects a mixture of contaminated samples and genuine polymicrobial growth (see Chapter 7 for further discussion).

6.6.3.2 Methodological findings

Given that urine flow cytometry results of mixed growth looked very similar to those of predominant growth and no prediction model could reliably identify mixed growth, it remains unclear how mixed growth should be labelled when the model would be used in clinical practice. While Taylor *et al.* (2018) [109] considered mixed growth as contamination in line with standard clinical practice , Müller *et al.* considered them as positive growth. In the absence of further information that can reliably predict mixed growth, classification of mixed growth becomes a trade off between apparent sensitivity and specificity, since samples with mixed growth were frequently misclassified either way. The most meaningful performance measure may therefore be the one that excludes mixed growth altogether. While this technically only estimates the model performance in an unrealistic situation in which no mixed growth — and thus no sample contamination or genuine polymicrobial growth — exists, it avoids arbitrarily classifying mixed growth and provides a (sensitive) upper boundary on achievable performance. If used, however,

it is important to acknowledge that such a model will nevertheless frequently misclassify mixed growth, further emphasising the importance of additional UTI symptoms to diagnose suspected bacterial UTI.

In order to predict mixed growth early and allow for the collection of a second — hopefully uncontaminated — sample, advances in laboratory technology may provide additional information that may allow for more reliable prediction. For example, future routine inclusion of advanced urine flow cytometry measurements like laser flow cytometry have been suggested as a potential way to identify mixed growth early [219] and could be used to extend the model developed in Chapter 5.

6.7 Comparison with clinicians' performance

Dataset: ED patient cohort from Chapter 5

Evaluation: External validation only

All evaluations so far have assessed the model's ability to identify patients at risk of bacteriuria in comparison to a "random" model without information. The value of such an evaluation to judge a model's usability in clinical practice is limited, however. An attempt at a more meaningful evaluation of clinical model performance may instead judge the model against the real-world diagnostic abilities of clinicians, assessing whether the model can separate patients as well or better than is currently the case in routine care. Ideally, this comparison is performed prospectively in a randomized controlled trial, but running a trial is expensive and time-consuming [220]. As a result, many evaluations of clinical risk prediction models are performed retrospectively on routinely collected data or large cohorts. Defining a valid and meaningful comparison on retrospective data, however, can prove challenging. Inferring clinical reasoning from routine data is limited by incomplete and coarse data recording. We could see in Chapter 4 and Appendix F that information on clinician's decision making is often limited to broad diagnosis codes and prescribing records, without further detail on the diagnostic process or rationale for prescribing. It is therefore difficult to unambiguously categorise

clinician behaviour from EHR data.

In the only study that I am aware of that attempted to retrospectively compare model performance in predicting bacteriuria to the performance of clinicians in the same healthcare setting, Taylor *et al.* used ED diagnoses and empirical antibiotic prescribing in the ED as proxies for clinical judgement [109]. In the last section of this chapter, I therefore apply similar criteria to data from QEHB and critically assess the evidence that can be derived from such comparison.

6.7.1 Statistical analysis

Clinician's "prediction" of bacteriuria was approximated using diagnostic codes and antibiotic prescribing in the ED, constructing two proxies of clinical judgement with different sensitivities and specificities. A sensitive estimate was defined using either presence of an ED diagnosis of UTI (lower UTI, pyelonephritis, urosepsis)[5] and/or a prescription of systemic antibiotics included in QEHB's 2018 prescribing guidelines for UTI (amoxicillin, cefalexin, cefuroxime, ciprofloxacin, co-amoxiclav, ertapenem, gentamicin, nitrofurantoin, trimethoprim, vancomycin) or sepsis (ceftriaxone, ciprofloxacin, co-amoxiclav, gentamicin, meropenem, piperacillin / tazobactam, vancomycin) in the absence of a recorded diagnosis of another infection. A more specific estimate of clinical judgement was based solely on the presence of an ED diagnosis of UTI. For patients who were admitted to hospital as a result of their ED visit, recorded 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) diagnoses of UTI on discharge from hospital (either as primary diagnosis only or at any position) were used as a third proxy of clinical judgement[6]. See Appendix H for a detailed list of all codes used.

Accuracy, sensitivity, and specificity of each estimate of clinical judgement were calculated in patients consulting after March 31st 2018 (test set). No AUROC was estimated, since the definitions of clinical judgement did only provide binary

[5] Note that for this analysis, a record of urinary symptoms was not considered to be an ED diagnosis of UTI.

[6] No ICD-10 discharge codes are recorded for patients who are directly discharged from the ED, see Section 3.3.2 for a more detailed discussion.

A

Antibiotics	UTI diagnosis: Yes	UTI diagnosis: No	
Yes	268 (85.9 col-%)	56 (26.3 col-%)	324 (61.7%)
No	44 (14.1 col-%)	157 (73.7 col-%)	201 (38.3%)
	312 (59.7%)	213 (40.3%)	525 (100%)

B

Antibiotics	UTI diagnosis: Yes	UTI diagnosis: No	
Yes	277 (84.7 col-%)	281 (41.0 col-%)	558 (55.1%)
No	50 (15.3 col-%)	405 (59.0 col-%)	455 (44.9%)
	327 (32.3%)	686 (67.7%)	1,013 (100%)

Figure 6.6: Distribution of UTI diagnoses and antibiotics in the ED for **A)** patients discharged directly from the ED and **B)** patients admitted to hospital.

ED, emergency department; UTI, urinary tract infection.

predictions. Predictions of a LR model using the reduced set of predictors were pre-set to the same sensitivity (in the case of ED diagnosis of UTI and/or antibiotic prescription) or specificity (in the case of ED diagnosis of UTI alone) as achieved by the above proxies of clinical judgement, and the performance on the other, non-fixed metrics was compared. The 2.5% and 97.5% percentiles of 1,000 bootstrapped resamples were used to calculate approximate 95% CIs for each metric. The same analysis was performed separately in patients who were admitted to hospital after their ED visit and in patients who were discharged directly from the ED.

6.7.2 Results

Out of the 1,538 visits in the test set, 1,013 (65.9%) resulted in an admission to hospital and 525 (34.1%) were discharged directly from the ED (Figure 6.6 A and B). An ED diagnosis of UTI was more common among discharged patients (59.4% versus 32.6% of patients admitted to hospital). In both groups, approximately 85% of patients with an ED diagnosis of UTI were prescribed an antibiotic for UTI in the ED. Since patients without a diagnosis of UTI included patients with other suspected infections and recommended antibiotics overlapped, the proportion of antibiotic prescribing was still high in this patient group. Antibiotic prescribing was

less common in patients discharged from the ED without a UTI diagnosis (26.3% versus 41.0% of patients admitted to hospital).

A LR model using the reduced set of predictors universally outperformed clinical judgement as estimated by ED diagnosis of UTI and/or prescription of antibiotics (Table 6.8) or ED diagnosis of UTI alone (Table 6.9). At the same or higher sensitivity as ED diagnosis of UTI and/or prescription of systemic antibiotic in the ED (60.8, 95% CI 57.3–64.3), the LR model achieved a specificity of 83.4 (95% CI 81.0—85.9; Table 6.8). In comparison, ED diagnosis of UTI and/or prescription of systemic antibiotic only achieved an estimated specificity of 51.0 (95% CI 47.8—54.3). When fixed at the same or higher specificity as ED diagnosis of UTI alone (63.7, 95% CI 60.5—66.7; Table 6.9), the LR model achieved a sensitivity of 79.5 (95% CI 76.6—82.5) compared to that sensitivity of 48.2 (95% CI 44.3—52.1) when relying on an ED diagnosis of UTI. Results were comparable in admitted and discharged patients, although estimates became less precise due to diminished sample sizes (Tables 6.8 and 6.9).

Clinical performance as estimated by recorded ICD-10 code on discharge (admitted patients only) was similar to the performance of ED diagnosis in admitted patients. Using diagnosis of UTI at any position among a patient's discharge diagnoses achieved a sensitivity of 37.6 (95% CI 34.1–41.3) and a specificity of 74.0 (95% CI 71.1–76.9; table not shown). Limiting the diagnosis to primary diagnoses — i.e., the recorded reason for admission — decreased sensitivity to 27.0 (95% CI 23.5–30.5) but increased specificity to 83.1 (95% CI 80.7–85.6).

6.7.3 Discussion

6.7.3.1 Clinical findings

When estimating clinicians' performance of predicting a patient's risk of bacteriuria using diagnosis codes of UTI and/or prescriptions of systemic antibiotics, a simple LR model based on age, sex, history of positive urine culture, and urine flow cytometry measurements consistently outperformed clinicians. The comparison of clinician and model performance shown here closely mirrors the approach taken by Taylor *et al.* [109]. In line with their results, I find a consistently

Table 6.8: Comparison of discriminative model performance to clinical judgement as defined by ED diagnosis of UTI and/or prescription of systemic antibiotics in the ED.

Decision rule	Accuracy (95% CI)	Sensitivity (95% CI)[1]	Specificity (95% CI)
All patients			
UTI diagnosis or antibiotics	55.3 (52.9–57.8)	60.8 (57.3–64.4)	51.0 (47.8–54.3)
LR (reduced predictor set)	73.4 (71.3–75.5)	60.8 (57.3–64.3)	83.4 (81.0–85.9)
Admitted patients			
UTI diagnosis or antibiotics	54.6 (51.4–57.4)	51.3 (46.8–55.7)	57.0 (52.5–61.0)
LR (reduced predictor set)	72.0 (69.3–74.7)	51.3 (47.0–56.2)	87.2 (84.5–89.9)
Discharged patients			
UTI diagnosis or antibiotics	56.8 (52.4–60.8)	77.1 (72.2–82.5)	38.4 (32.5–44.4)
LR (reduced predictor set)	70.9 (67.0–74.5)	78.7 (73.9–84.0)	63.8 (57.9–69.1)

[1] LR sensitivity was fixed at the same sensitivity as that achieved by UTI diagnosis or antibiotic prescribing. Due to the relative small sample size actually observed sensitivity might differ slightly from that achieved by UTI diagnosis or antibiotics. In these cases, the closest sensitivity higher than that achieved by UTI diagnosis or antibiotics was chosen.

CI, confidence interval; LR, logistic regression; UTI, urinary tract infection.

Table 6.9: Comparison of discriminative model performance to clinical judgement as defined by ED diagnosis of UTI.

Decision rule	Accuracy (95% CI)	Sensitivity (95% CI)[1]	Specificity (95% CI)
All patients			
UTI diagnosis	56.9 (54.3–59.3)	48.2 (44.3–52.1)	63.7 (60.5–66.7)
LR (reduced predictor set)	70.9 (68.7–73.1)	79.5 (76.6–82.5)	64.1 (61.2–67.2)
Admitted patients			
UTI diagnosis	57.2 (54.3–60.0)	37.5 (32.9–42.0)	71.6 (67.7–75.2)
LR (reduced predictor set)	73.3 (70.8–76.1)	75.3 (71.5–79.1)	71.9 (68.4–75.5)
Discharged patients			
UTI diagnosis	56.4 (52.2–60.6)	66.7 (60.6–72.4)	47.1 (41.8–53.0)
LR (reduced predictor set)	66.5 (62.3–70.5)	87.6 (83.2–91.4)	47.5 (41.8–53.4)

[1] LR specificity was fixed at the same specificity as that achieved by UTI diagnosis. Due to the relative small sample size actually observed specificity might differ slightly from that achieved by UTI diagnosis. In these cases, the closest specificity higher than that achieved by UTI diagnosis was chosen.

CI, confidence interval; LR, logistic regression; UTI, urinary tract infection.

higher performance of the prediction model in classifying bacterial growth in urine samples. They concluded that the implementation of their model in clinical practice has the potential to "reduce the number of false positives and false negatives for UTI diagnosis" [109]. The validity of this claim, however, depends on two factors: the extent to which bacteriuria is indicative of clinically confirmed UTI in the patient cohort and the extent to which these proxies for clinical judgement can be interpreted as predictions of bacteriuria — or whether they represent a summary of wider clinical judgement and context. The patient cohort used in this study included patients regardless of presenting complaints, and bacteriuria alone should therefore not be interpreted as evidence of UTI in this cohort [17]. While Taylor *et al.* only included "symptoms potentially attributable to a UTI", these symptoms were defined broadly. For example, less than 20% of their cohort had recorded urinary symptoms while the vast majority of patients was included due to vague symptoms such as abdominal pain, fatigue, or fever. This was also the case in data from QEHB. These symptoms may or may not be indicative of UTI even if the patient has bacteriuria [31, 32]. Consequently, it is unclear whether urine samples which were categorised as false negatives — i.e., samples that showed bacterial growth during culture but had no diagnosis of UTI or systemic antibiotic prescribing — did indeed fulfil the clinical criteria for UTI, or whether they were mostly asymptomatic and thus did not require treatment for UTI despite bacterial growth. Similarly, diagnosing UTI in patients without microbiologically confirmed bacteriuria (false positives) and treating them with antibiotics might still be the correct course of action in the presence of clear urinary symptoms strongly indicative of UTI [56], particularly if patients received prior antibiotic treatment in the community that might have retarded microbiological growth during incubation [20]. The apparent conclusion reached in this chapter and earlier by Taylor *et al.* may therefore have been exaggerated by an imprecise definition of clinic judgement. Rather than showing the superiority of prediction models over clinicians' performance, these results instead primarily highlight the difficulties of reliably defining the presence of UTI from routine EHR data.

6.7.3.2 Methodological findings

A tendency to overstate the amount and quality of evidence for increased clinical performance of (machine learning based) risk prediction models has also been noted elsewhere [220]. This remains in stark contrast to the number of models successfully translated into clinical practice [100]. Instead of overstating model performance during retrospective evaluation, conclusions should appropriately reflect the level of evidence that can be created from retrospective data and statements should be tempered accordingly. Ultimately, performance of any model will have to be judged based on prospective clinical outcomes [221]. For example, the use of a model for the prediction of bacteriuria may be considered advantageous — and acceptable for clinical use — if it not only correctly identifies bacteriuria early but also reduces antibiotic prescribing without leading to an increase in complications of infection. While retrospective evaluation may or may not give some evidence as to how the performance of the model might compare to that of clinicians, this has to be done and interpreted cautiously. Prospective evaluation will be required unless clinical reasoning can be unambiguously inferred from the EHR records.

6.8 Conclusion

In the detailed sensitivity analyses performed in this chapter, I was able to show that apparent model performances may vary widely according to the patient populations and variable definitions used in model training and evaluation. The performance of risk prediction models developed in Chapter 5 was lower for patients with recorded suspicion of UTI as well as in older patients and men. When judged solely on retrospectively estimated performance, the model might therefore be deemed acceptable for clinical practice in some patient groups while unsuitable for others. Mixed culture growth presented a particular challenge for models predicting the risk of bacteriuria. Stratified reporting of estimated performance is therefore desirable, although stratification even on a single variable like ED diagnosis may already significantly reduced the available sample size, potentially

leading to unstable models and/or evaluations. This exemplifies a previously noted difficulty in comparing performances between published retrospective prediction modelling studies [100]. If UTI cases can't be identified with certainty and clinical coding depends on local hospital and laboratory guidelines, superficially similar cohorts might contain very different case mixes. Even identical models may show very different performance in these circumstances, rendering the comparison of simple performance summaries like an AUROC for the entire patient population meaningless. Better quantitative and qualitative methods will be required to understand and communicate differences between patient cohorts derived from EHR systems at different hospitals or even different countries.

A lack of detail in structured EHR diagnoses further made it difficult to fully reconstruct clinical decision making for the purpose of model validation, preventing reliable comparison of model performance to standard care. Whereas the statistical models were evaluated exactly on the information that was available to them during training, clinicians' decision making was ascribed after the fact based on incomplete information. Decisions that might appear sub-optimal retrospectively given only the incomplete information captured in EHR data might instead have been considered appropriate had I access to the same information as the clinician in the ED. By stripping away contextual information not recorded in the EHR data used here and in Chapter 5 — e.g., information on previous antibiotic use in the community enquired during anamnesis but not recorded in structured data fields — any comparison was inevitably skewed against clinicians. It is therefore difficult to judge the likely improvement in performance — if any — that may be achieved by deploying the model in clinical practice. Prospective evaluation alongside clinical care will be necessary to understand if and how the models developed here may indeed provide clinical benefit to patients.

Chapter summary

- Patients who have an ED urine sample submitted for culture at QEHB represent a heterogeneous population, with important implications for model performance and applicability.
- An inability to unambiguously define patients with suspected UTI in the ED raises questions about how model performance in EHR studies should be reported, and how inevitable variation in patient mix can be accounted for and clearly communicated during publication.
- Reconstructing clinical decision making for UTI from EHR data remains difficult, and comparisons of clinician and model performance based on routinely collected EHR data should be interpreted cautiously. Prospective evaluation will be necessary to compare performances with reasonable certainty.

My role in the work presented in this chapter

The analysis presented in this chapter was conceived by me. The research application for the wider project was submitted to the Health Research Authority (HRA) for ethical approval by Dr Laura Shallcross. Raw hospital data for this study was extracted from Queen Elizabeth Hospital Birmingham by Dr David McNulty and processed by me as described in Chapter 5. I performed all analyses presented in this chapter. All findings were interpreted by me. I wrote this chapter, with feedback from Dr Laura Shallcross, Prof Nick Freemantle, and Prof Andrew Hayward.

Software and code used in this chapter

All analyses in this chapter were performed using R (v3.6.2) and RStudio (v1.2.5033) on Windows 10. Data processing was performed using the *tidyverse* (1.3.0) and *data.table* (1.12.8) packages. All model building was performed using packages from the *tidymodels* (0.1.0) ecosystem, including *rsample* (0.0.5), *recipes* (0.1.9), *parsnip* (0.0.5), *tune* (0.0.1), *yardstick* (0.0.5), and *tidyposterior* (0.0.2). All

code is publicly available at https://github.com/prockenschaub/phd_code.

Publications resulting from this chapter

No publications have resulted from this chapter yet.

Chapter 7

Conclusions, limitations, and future research

Abstract

In this chapter, I summarise the main findings of this thesis and discuss their implications for patient care in England. I highlight the major strengths and limitations of my approach and outline areas in need of further research.

Using data from major national and local electronic health record (EHR) databases in England, I demonstrated that previous studies assessing the diagnosis and management of UTI using EHR data may have been at high risk of bias. This bias mainly originated from a difficulty to identify patients with (suspected) UTI from EHR data, and to retrospectively infer the rationale behind clinical decisions such as diagnostic testing and antibiotic prescribing. I showed that estimated treatment effects and prediction performances were sensitive to variable definitions and heterogeneity in the included patient population, making it often difficult to interpret the obtained findings.

This thesis analysed several research databases — including data from over 600 primary care practices and one of England's leading digital hospital — and performed comprehensive sensitivity analyses to assess the reliability of findings on the diagnosis and management of UTI obtained from EHR data. Findings were primarily limited by the exclusive use of routinely collected data. Incentive schemes that promote more detailed recording of UTI or easier access to free-text notes may be required to reliably identify patients with UTI from EHR data. While several large national databases already exist for primary care, similar infrastructure will be required in secondary care to generalise findings beyond a small number of digitally

advanced hospitals. A combination of retrospective and prospective analysis as well as qualitative research may be necessary to fully understand the diagnosis and management of UTI, and to design systems and interventions that can meaningfully support clinicians in routine practice.

Future research should include a review of case notes to fully understand clinical coding for UTI, retrospective and prospective evaluation of predictive models at multiple sites, and the use of recent innovations like metagenomic sequencing to assess the reliability of microbiological culture as the gold standard for UTI diagnosis.

7.1 Overview of key research findings

7.1.1 Identification of UTI cases in EHR data relies primarily on coarse diagnostic codes

Whether EHR data can be effectively used to support the diagnosis and management of disease is contingent on the ability to identify cases of the disease reliably [222]. As discussed in Chapter 1, clinical diagnosis of UTI is complex. This complexity is mirrored in retrospective EHR data. A range of clinical information may be used to identify (probable) cases of UTI in EHRs, including recorded symptoms and diagnoses, diagnostic tests (urine dipsticks and urine flow cytometry), antibiotic prescribing, and microbiological culture. In this thesis, I showed that the identification of cases of UTI in major EHR databases relied primarily on diagnostic codes captured in structured data fields (Chapters 2, 4, and 6).

While these codes in principle allow to differentiate between lower UTI, pyelonephritis, and urosepsis, little additional information was available to assess the overlap between these clinical infectious syndromes, or the severity of infection. Data on specific urinary symptoms was rarely recorded in structured fields in primary care (Chapter 4), and was limited even at digitally mature hospital sites (Chapters 5 and 6). The additional lack of urine dipstick test results meant that I was unable to distinguish between patients with genuine urinary symptoms and patients in whom suspicion of UTI might have been driven by a positive dipstick

result in the absence of symptoms. Culture results were overwhelmingly missing in primary care databases, since routine culture of urine is not routinely recommended in this setting in the absence of known risk factors [13]. While cultures were more commonly available in secondary care — allowing me to confirm the presence of bacteriuria — they were also found to be frequently performed in patients without a recorded symptom or suspicion of UTI, questioning their usability to identify cases of symptomatic UTI.

The predominant use of diagnostic codes to identify UTI in EHR databases has implications for the reliability of the obtained results. In primary care, a previous study suggested that definitions based solely on diagnostic codes may miss a large proportion of patients in primary care who were prescribed antibiotics for UTI but do not have a corresponding diagnosis recorded [6]. As a result, a large proportion of patients who did suffer from UTI may have been missed in Chapter 4, introducing risk of bias if these patients systematically differed from those with a recorded diagnosis of UTI. In data captured in the emergency department (ED) — where diagnostic codes until recently were limited to national codes or bespoke local terminologies (Chapter 3) — diagnostic coding was similarly imprecise, with little information on urinary symptoms recorded in the structured EHR data. As a result, the patient population in Chapter 5 included all patients who had a urine sample submitted for microbiological culture, many of whom may not have had urine symptoms.

Conclusion

Clinical coding of suspicion of UTI in structured EHR data is currently limited. More detailed recording, particularly regarding the presence or absence of urinary symptoms will be required to account for the large variations in the clinical presentation of UTIs and allow for the estimation of reliable and comparable results from EHR data. Better recording of UTI symptoms may be achieved for example by incentivising more detailed coding in structured data via national payment programmes (see for example

the Quality Outcomes Framework (QOF) [223] or Commissioning for Quality and Innovation (CQUIN) [54]), or by providing researchers with easier access to full-text information recorded in clinical notes.

7.1.2 Findings on UTI obtained from EHR are highly dependent on included patient case mix

The above described inability to precisely define inclusion criteria for community-onset UTI likely affected the findings reported here and elsewhere. I repeatedly encountered difficulties when comparing the results of models developed here to findings that had previously been published. Through careful sensitivity analyses, I was able to show that the results of the employed statistical models as well as their clinical interpretation strongly depended on the included patient case mix. This finding is not unique to UTI nor EHR data, and has also been reported for prospective studies of other diseases [224]. However, prospective data collection allows researchers to define precisely which patients should be included in the study. EHR data, on the other hand, represents real-world clinical practice and captures only information that is routinely recorded as part of it. While this can be an advantage, it also means that the patient populations captured within EHR data reflect the full heterogeneity encountered in clinical practice. It is therefore paramount to consider a priori how the data analysis — and any results thereof — may be used prospectively to inform clinical decision making, and to clearly define the intended target population with this goal in mind. The interpretation of the obtained findings then needs to consider how closely the analysis was able to mirror that target population and how possible deviations from it may have affected the results.

In Chapter 4, more thoroughly defining the date of UTI episode start and excluding patients with evidence of hospital-acquired infection may partially explain the differences between the results presented here and those reported by Gharbi *et al.* [112]. Delaying antibiotic prescribing will be most viable in a population of low-risk patients with new community-onset UTI. The more

stringent exclusion of cases that originated in hospital may have removed spurious associations between disease severity, treatment decisions in primary care, and risk of complications. By more closely reflecting the target population, study results therefore become increasingly interpretable and applicable.

The impacts of patient case mix were particularly apparent in the ED analyses in Chapters 5 and 6. I was able to show that there was considerable variation in model performance depending on the characteristics of patients included in the study population. Estimated discriminative performance and calibration of the predictive models ranged widely by age, sex, ED diagnosis, and over time (Chapters 5 and 6). Results were further impacted by local procedures such as reflex cultures[1] and rules about the interpretation of mixed growth. These variations impeded the comparison of the model developed here with findings published previously in the United States [109] and in Switzerland [51]. Interpreting differences in model performance without a clear reference to the patient population in which the model was evaluated may therefore lead to arbitrary conclusions of model superiority. Cohort differences were not immediately obvious from published findings alone, and obtaining analysis code and patient-level information for these studies was central to enable a more meaningful comparison between study findings.

Conclusion

The results obtained from EHR analyses may vary considerably depending on characteristics of included patients. In the context of UTI, such variation may be caused by unintentional heterogeneity in the study population due to coarse diagnostic coding. Routine subgroup analyses based on important confounding factors such as age, sex, and initial UTI diagnosis may be necessary to evaluate the sensitivity of obtained results to such heterogeneity.

[1] A set of criteria that need to be met in order for a urine culture to be incubated, e.g. bacteria seen during urine flow cytometry. See Chapter 1 for a discussion of reflex cultures.

7.1.3 The context of diagnosing and managing UTI is often insufficiently captured in EHR data

Decisions on how to manage suspected or confirmed UTIs are also influenced by the wider clinical and social context. For example, clinicians may be more likely to prescribe an antibiotic to sicker patients or if the social circumstances of the patient do not provide an adequate safety net. While the fact that an antibiotic was prescribed may be automatically recorded in EHR data [121], this wider clinical and social context is often difficult to capture. Even purely clinical histories such as antibiotic prescribing in the community prior to a patient's ED visit weren't always available in EHR data due to lack of linkage and/or recording of information. An inability to account for this context, however, may lead to biased results.

The perils of insufficient contextual information were particularly apparent in Chapter 4. Delayed antibiotic prescribing — the main exposure of interest — was not generally well recorded in national primary care databases [139]. I was therefore forced to infer delayed prescribing indirectly through the absence of an antibiotic prescribing record. This led to a paradox finding in which patients that tended to be at higher risk of complications were more — rather than less — likely to be treated with delayed prescribing strategies [72]. It is therefore unlikely that the definition of delayed or withheld antibiotic prescribing used in this thesis represented a purposeful decision to delay antibiotics in anticipation of self-limiting infection. As a result, the findings obtained from the EHR analysis may not reflect the effect of interest, but instead estimate the influence of underlying patient characteristics that warranted the differences in prescribing, or recording of prescribing. A similar issue could be observed in Chapters 5 and 6, in which missing information on antibiotic exposure prior to arrival in the ED may have confounded the prediction of bacterial growth in ED urine samples.

More advanced methodological approaches may be used to try and limit the impact of this problem. Balancing techniques like propensity score analysis may allow to create a balanced patient population in which the influence of treatment context has been accounted for. However, due to the likely unobserved

nature of many confounding factors, additional methodological approaches did not meaningfully change point estimates when applied in Chapter 4. The biggest change was instead observed when additional contextual information was included in the model. I was able to demonstrate that the inclusion of home visits accounted for more than half of the difference between crude and adjusted estimates. It is likely that further unobserved factors — and therefore residual confounding — remained. The reliability of the obtained estimates is therefore questionable, and it is unclear if and how they should be interpreted.

Conclusion

Reasons for why a particular diagnostic test (e.g., urine culture) was performed or why a particular treatment (e.g., delayed prescribing) was chosen are often missing from EHR data, which may prevent researchers from ascertaining the exposure of interest. Unless the context of clinical decision making for UTI is better captured in EHR data (e.g., through more detailed recording in structured data fields or free-text analysis), treatment effects or model performances obtained from retrospective EHR analyses remain questionable.

7.2 Strengths and limitations of this thesis

Strengths of this thesis are its use of state-of-the-art English EHR databases — both on a national level and locally — together with comprehensive sensitivity analyses that allowed me to critically evaluate the reliability of the results obtained here and in recent key publications [109, 112]. I was able to show that findings may change considerably depending on cohort and variable definitions, and that careful interpretation is necessary to obtain meaningful, generalisable insights on the diagnosis and management of UTI from EHR data. I was further able to identify areas in which improved recording of UTI is essential to fully exploit routine data on UTIs in English clinical practice.

Limitations of the clinical findings reported in this thesis mainly related

to the way that clinical information on UTIs was recorded in retrospective records. Identification of UTI in primary and secondary care in this thesis relied predominantly on diagnosis of UTI using structured diagnostic codes. As discussed earlier, limited resolution of these codes likely introduced unintended heterogeneity into the study population, erroneously including (excluding) patients without (with) symptoms of UTI and leading to imprecise or biased results. The reliability of clinical coding may also have changed over time, or may have varied between included healthcare sites. Further lack of recording of several key variables — including urinary symptoms, urine dipstick tests, prior antibiotic use, and reasons for antibiotic prescribing — prevented me from fully assessing the reliability of diagnostic codes. Taking a pragmatic approach, I instead built on previously published code lists to define and identify UTI within the included databases. In the subset of patients for whom I was able to compare recorded diagnostic codes to information extracted manually from ED free-text notes (Appendix F), my findings showed a limited agreement between UTI symptoms recorded in free-text and coded UTI diagnoses in the ED. Retrospective EHR data therefore only allowed for an incomplete reconstruction of the real-world clinical decision making involved in diagnosing and treating UTIs.

Methodological findings presented in this thesis were limited by the choice of data sources and case studies. While I was able to access a widely-used national database for all primary care analyses, no national dataset currently exists in secondary care that includes prescriptions or laboratory values. I therefore used data from a single tertiary hospital. Although this hospital is one of England's digital centres of excellence, other hospitals might have better integration of some of the key information crucially missing in this thesis. Using only a single hospital further has implications for the generalisability of the results. Many NHS trusts still lack fully-integrated EHR systems, potentially limiting the applicability of the models developed in this thesis. Patient populations and/or clinical procedures (such as urine flow cytometry thresholds for urine culture) may further differ at other hospitals, potentially affecting the transferability of my findings to other hospitals.

While key research findings presented above proved to be important in the case studies chosen in this thesis, they might not play as large a role for other research questions. For example, if other exposures are at a lesser risk of confounding by indication — which may be the case for example for antibiotic treatment duration — EHR data may contain sufficient contextual information to account for remaining differences between treatment groups.

Either way, the results presented in this thesis highlight that the analysis of EHR records can only be one piece in improving the management of UTI. In order to fully understand healthcare processes and be able to intervene at the right place, at the right time, and in the right way, a deeper understanding of the wider healthcare context is essential. Working closely with clinical colleagues at the UCL Institute of Health Informatics (Dr Anna Aryee and Dr Arnoupe Jhass) and at Queen Elizabeth Hospital Birmingham (Dr Martin Gill) was invaluable to understand the expected clinical behaviours and interpret deviations from them. In parallel to the work on this thesis, I also had the pleasure to be involved in the *Preserving Antibiotics through Safe Stewardship (PASS)* project [160], which acknowledges the complexity of real-life infectious disease management and antibiotic stewardship. Bringing together clinicians, statisticians, epidemiologists, behavioural scientists, and designers in an interdisciplinary mixed-methods approach, PASS attempted to provide a more comprehensive picture of the underlying healthcare processes, allowing it to go beyond the work presented in this thesis and work towards designing effective clinical interventions.

7.3 Translating EHR research into clinical practice

Digital transformation of the National Health Service (NHS) has become a major national policy goal, accruing more than £4 billion in national investment between 2016 and 2019 [99]. The commitment to a digital NHS is reflected in the NHS Long Term Plan, which aims to enable the comprehensive use and exchange of healthcare data across clinical settings by 2024 [98]. While there is strong focus on building capacity and digital maturity, the NHS Long Term Plan also ascertains the intention

to use the increasingly abundant data to enable medical research, new population health management approaches, and the development of decision support systems and artificial intelligence. Below, I discuss the implications of the NHS' plans for the findings of this thesis and discuss several issues that need to be considered when attempting to apply my findings within NHS England.

7.3.1 Technological considerations

Many of the findings described in this thesis relate to the technological challenge of accurately inferring clinical information from EHR data and providing clinicians with helpful information in real-time.

Data collection and linkage The NHS Long Term Plan includes a pledge to push all healthcare providers towards a minimum level of digital maturity, and to support healthcare personnel in recording detailed, high-quality information as a "by-product of care" [98]. The NHS is also accelerating the integration of care records from primary practices, hospitals, community services, and social care via so-called Local Health and Care Record (LHCR) programmes. Better capture of rich contextual information like urinary symptoms or medical history may alleviate many of the issues reported in this thesis (e.g., identification of antibiotic exposure in the community prior to ED visit) and allow for a robust definition of patient populations, exposures, and outcomes (Chapters 4–6). Increased documentation of care processes, however, also usually requires additional resources [225]. Careful planning will therefore be needed to ensure relevant information is captured without imposing an undue burden on an already stretched healthcare personnel.

Population health management systems Access to detailed data across healthcare settings will be fundamental to realising the NHS' ambition to develop population health management systems [98]. Clinical coding for UTI currently contains substantial uncertainty about the patient's health status and management plan, and richer information capture will be necessary to gain actionable insights on the management of UTI (Chapter 4). Similar caveats will apply to other acute (infectious) conditions, and a holistic approach will be required that improves the capture of necessary clinical detail in often very short consultations.

Clinical decision support systems If they are to be used across the NHS, clinical decision support systems will require standardised data in real-time (Chapter 5). The NHS has committed to increase the recruitment of technical expertise to develop and maintain the infrastructure necessary to effectively deploy decision support systems [98], which will be needed to enable the use of prediction models like the one developed in this thesis beyond the currently still small number of digitally mature hospitals.

Regulatory approvals Finally, decision support systems — and perhaps also population health management systems — based on EHR data will be required to provide similar quality assurance as other healthcare products. In Chapter 6, I showed that predictive models might perform very differently in different patient populations. The UK Medicines and Healthcare products Regulatory Agency (MHRA) together with industry stakeholders has recognised this issues and is currently developing guidance for validation of artificial intelligence in healthcare [226]. Performance of EHR models may also change over time (Chapter 5), introducing further regulatory complications. The US Food and Drug Administration (FDA) has recently published a white paper discussing the additional challenges of continued learning in a fluid healthcare system but further work is required to ensure the continued safety of EHR-based systems once deployed in clinical practice [227].

7.3.2 Barriers to implementing learning from this thesis into clinical practice

The use of EHR-driven technology does not only pose technological challenges, but doctor-patient interactions are complex behavioural situations by themselves and are influenced by a multitude of contextual factors and motivations [228]. Evidence and models derived from EHR data — even when carefully obtained and technologically feasible — therefore depend on being accepted by both patients and clinicians in order to create meaningful clinical impact, and need to consider how doctors interact with their patients, where undesired behaviours originate, and when and how they can best support doctors [100].

Asymptomatic bacteriuria Asymptomatic bacteriuria is a chief concern for the usability of statistical models that aim to guide diagnosis and management of UTI [25, 30, 31]. Throughout this thesis I have made the assumption that patients undergoing treatment for suspected UTI in primary care and the ED have symptomatic infection, but this is unlikely to be the case and highlights the complexity of diagnosing and managing UTI. Unless combined with effective training and clear guidelines, the naive use of statistical models faces the same limitations as existing diagnostic tests and guidelines, and might similarly lead to over-treatment of suspected UTI if clinicians rely on them to guide empirical prescribing in the absence of clear clinical symptoms.

Point of intervention Thought must be given to decide at which point clinicians ought to be supported in their decision making. Possible points of intervention include the empirical prescribing process during initial presentation[2] or later during regular medication reviews. Close collaboration with clinicians as well as prospective studies will be necessary to explore the respective merits of each option.

Presentation of evidence Similarly, it remains unclear how evidence derived from statistical models should be provided to clinicians. Possible options include binary yes/no predictions at "optimal" cut-offs, expected probabilities of the outcome for a particular patient, a list of variables that generally indicate likely presence of the outcome, or a combination of the above. In any case, clinicians must be trained to adequately use the evidence provided by the models and appropriately account for it during their clinical decision making. Rather than making the decision for the clinician, models should provide additional evidence that can be interpreted by the clinician in the appropriate context [208].

Clinician-model feedback Once a model is deployed into clinical practice, the use of the model may itself impact clinical behaviour, which in turn may change the distribution — and perhaps even underlying meaning — of the variables used to build the model [208]. Models therefore need to be monitored for changes in their performance over time, and re-calibrated or re-trained as necessary (with

[2] Assuming that all inputs like urine flow cytometry are already available at this point.

implications for regulatory approval, see Section 7.3.1).

7.4 Future research

EHR databases provide a potentially valuable resource that may be leveraged to improve the diagnosis and management of UTI in England, but more research is required to fully understand the validity of the captured information and its impact on clinical practice.

Case note reviews in primary and secondary care More research is required to understand the recording of UTI in EHR databases, particularly in the absence of evidence of urinary symptoms in structured data fields. Case note review in a random subset of UTI patients identified from CPRD and at several other hospital sites may be used to partially answer this question, assessing the expected reliability of results obtained from current EHR sources. These findings may be used to create guidance and incentives that improve recording of key diagnostic information for UTI in local and national EHR databases.

Multi-centre validation studies In this thesis, I was only able to investigate the performance of a risk prediction model for bacteriuria at the same hospital at which it was developed. In order to investigate the generalisability of the estimated models, retrospective and prospective external validation at multiple hospital sites across England needs to be performed. I secured seed funding from the Department of Health and Social Care to validate the model on retrospective data from University College London Hospital (UCLH). If the model shows promise, the intention is to evaluate its use prospectively to inform early antibiotic cessation in a low risk group of patients who have been treated empirically for suspected UTI in the ED and who have subsequently been admitted to hospital.

Methodological research on evaluation of prediction models Variance in model performance suggested differences in the applicability of prediction models for bacteriuria in key patient populations. However, estimates quickly became imprecise due to a considerable reduction in sample sizes when stratifying by patient characteristics. While some initial work has been performed to isolate

case-mix effects in clinical prediction models [224, 229], further research is required to propose and promote methods to estimate and clearly report model performance in different key populations, identify changes in the meaning of variables derived from EHR data, and distinguish issues of model fit from changes in patient case-mix.

Investigating the reliability of urine culture as the gold standard for diagnosis of UTI This thesis has treated culture results as the ground truth for confirming bacterial UTI in the presence of urinary symptoms. This view is increasingly challenged [56], and future research using for example metagenomic sequencing may investigate the reliability of urine culture results in confirming the presence of uropathogens in routine practice. To this end, I have received funding for a pilot project that prospectively compares model predictions, urine culture results, and results from 16S metagenomic sequencing. The proposed analysis will undertake metagenomic sequencing in a small number of urine samples to investigate the presence of uropathogens in individuals with negative culture results but who were predicted to have likely bacteriuria by the statistical model.

In conclusion, this thesis has shown that there may be scope to use routinely-collected EHR data to investigate community-onset UTI. However, deriving insights on the diagnosis and treatment of UTIs from EHR data is currently curtailed by a difficulty to ascertain patient state and treatment decisions. Results from EHR studies to date need to be interpreted carefully and critically, especially when assessing treatment effects. A combination of improved data collection, more comprehensive linkage of data sources, tailored statistical methodologies, and further validation studies will be needed in order to use EHR data to derive reliable and reproducible conclusions on the management of community-onset UTI.

Appendix A

Review data extraction form

The following data extraction template was used to chart data from each full-text article included in the scoping review presented in Chapter 2, as per Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist (Appendix G).

Table A.1: Review data extraction form

Study characteristics	
Author(s)	
Year of publication	
Healthcare setting	◯ Community ◯ Primary care ◯ Hospital (emergency department) ◯ Hospital (outpatient clinic) ◯ Hospital (inpatient ward)
Geographical location	◯ US ◯ UK ◯ Europe: _____ ◯ Other: _____
Participants	
Definition of community-onset UTI	◯ Recorded symptoms of UTI ◯ Recorded diagnosis of UTI ◯ Urinalysis requested ◯ Urine culture requested ◯ Prescription of antibiotics

Other inclusion/exclusion criteria	
Number of included patients (or samples)	
Methodology	
Exposures (if any)	
Covariates	
Outcomes	
Statistical methods	
Results	
Summary of key findings	

Appendix B

Studies excluded during full-text review

The following 58 studies were excluded from the scoping review presented in Chapter 2 after full-text eligibility screening. References are grouped by reason for exclusion.

Manual note review or non-EHR data

Briongos-Figuero L.S., Gomez-Traveso T., Bachiller-Luque P. et al. Epidemiology, risk factors and comorbidity for urinary tract infections caused by extended-spectrum beta-lactamase (ESBL)-producing enterobacteria. *Int. J. Clin. Pract.* 2012;66:891–896

Brosh-Nissimov T., Navon-Venezia S., Keller N. & Amit S. Risk analysis of antimicrobial resistance in outpatient urinary tract infections of young healthy adults. *J. Antimicrob. Chemother.* 2019;74:499–502

Calbo E., Romani V., Xercavins M. et al. Risk factors for community-onset urinary tract infections due to Escherichia coli harbouring extended-spectrum beta-lactamases. *J. Antimicrob. Chemother.* 2006;57:780–783

de La Blanchardiere A., Dargere S., Guerin F. et al. Non-carbapenem therapy of urinary tract infections caused by extended-spectrum beta-lactamase-producing Enterobacteriaceae. *Med. Mal. Infect.* 2015;45:169–172

Drekonja D.M., Okoye N.C., Kuskowski M.A. & Johnson J.R. Appropriateness of urinary tract infection diagnosis and treatment duration. *Arch. Intern. Med.* 2010;170:489–490

Ena J., Arjona F., Martinez-Peinado C., del mar Lopez-Perezagua M. & Amador C. Epidemiology of urinary tract infections caused by extended-spectrum beta-lactamase-producing Escherichia coli. *Urology* 2006;68:1169–1174

Faine BA, Harland KK, Porter B, Liang SY & Mohr N. A clinical decision rule identifies risk factors associated with antimicrobial-resistant urinary pathogens in the emergency department: a retrospective validation study. *Ann. Pharmacother.* 2015;49:649–655

Forster CS, Powell EA, DeBurger B, Courter J, Haslam DB & Mortensen JE. Association of systemic antimicrobials with the expression of beta-lactamases in bacteria cultured from urological patients. *Diagn. Microbiol. Infect. Dis.* 2019;94:391–394

Foudraine DE, Bauer MP, Russcher A et al. Use of Automated Urine Microscopy Analysis in Clinical Diagnosis of Urinary Tract Infection: Defining an Optimal Diagnostic Score in an Academic Medical Center Population. *J. Clin. Microbiol.* 2018;56

Hay A.D., Thomas M., Montgomery A. et al. The relationship between primary care antibiotic prescribing and bacterial resistance in adults in the community: A controlled observational study using individual patient data. *J. Antimicrob. Chemother.* 2005;56:146–153

Hitzenbichler F, Simon M, Holzmann T et al. Antibiotic resistance in E. coli isolates from patients with urinary tract infections presenting to the emergency department. *Infection* 2018;46:325–331

Hsu CY, Fang HC, Chou KJ, Chen CL, Lee PT & Chung HM. The clinical impact of bacteremia in complicated acute pyelonephritis. *Am. J. Med. Sci.* 2006;332:175–180

Hyun M., Lee J.Y., Kim H.A. & Ryu S.Y. Comparison of Escherichia coli and Klebsiella pneumoniae acute pyelonephritis in Korean patients. *Infect. Chemother.* 2019;51:130–141

Ikram R, Psutka R, Carter A & Priest P. An outbreak of multi-drug resistant Escherichia coli urinary tract infection in an elderly population: A case-control study of risk factors. *BMC Infect. Dis.* 2015;15:224

Kang SC, Hsu NW, Tang GJ & Hwang SJ. Impact of urinary catheterization on geriatric inpatients with community-acquired urinary tract infections. *J. Chin.*

Med. Assoc. 2007;70:236–240

Karlovic K, Nikolic J & Arapovic J. Ceftriaxone treatment of complicated urinary tract infections as a risk factor for enterococcal re-infection and prolonged hospitalization: A 6-year retrospective study. *Bosn. J. Basic Med. Sci.* 2018;18:361–366

Khawcharoenporn T, Vasoo S, Ward E & Singh K. High rates of quinolone resistance among urinary tract infections in the ED. *Am. J. Emerg. Med.* 2012;30:68–74

Killgore K.M., March K.L. & Guglielmo B.J. Risk factors for community-acquired ciprofloxacin-resistant Escherichia coli urinary tract infection. *Ann. Pharmacother.* 2004;38:1148–1152

Linsenmeyer K., Strymish J. & Gupta K. Two simple rules for improving the accuracy of empiric treatment of multidrug-resistant urinary tract infections. *Antimicrob. Agents Chemother.* 2015;59:7593–7596

Al Majid F. & Buba F. The Predictive and discriminant values of urine nitrites in urinary tract infection. *Biomed. Res.* 2010;21:297–299

Meier S., Weber R., Zbinden R., Ruef C. & Hasse B. Extended-spectrum beta-lactamase-producing Gram-negative pathogens in community-acquired urinary tract infections: An increasing challenge for antimicrobial therapy. *Infection* 2011;39:333–340

Al Mohajer M., Musher D.M., Minard C.G. & Darouiche R.O. Clinical significance of Staphylococcus aureus bacteriuria at a tertiary care hospital. *Scand. J. Infect. Dis.* 2013;45:688–695

Mohr NM, Harland KK, Crabb V et al. Urinary Squamous Epithelial Cells Do Not Accurately Predict Urine Culture Contamination, but May Predict Urinalysis Performance in Predicting Bacteriuria. *Acad. Emerg. Med.* 2016;23:323–330

Osthoff M., McGuinness S.L., Wagen A.Z. & Eisen D.P. Urinary tract infections due to extended-spectrum beta-lactamase-producing Gram-negative bacteria: Identification of risk factors and outcome predictors in an Australian tertiary referral hospital. *Int. J. Infect. Dis.* 2015;34:79–83

Ryanto S., Wong M., Czarniak P. et al. The use of initial dosing of gentamicin in the management of pyelonephritis/urosepsis: A retrospective study. *PLoS One* 2019;14:e0211094

Safrin S, Siegel D & Black D. Pyelonephritis in adult women: inpatient versus outpatient therapy. *Am. J. Med.* 1988;85:793–798

Spoorenberg V., Hulscher M.E.J.L., Akkermans R.P., Prins J.M. & Geerlings S.E. Appropriate antibiotic use for patients with urinary tract infections reduces length of hospital stay. *Clin. Infect. Dis.* 2014;58:164–169

Tal S., Guller V., Levi S. et al. Profile and prognosis of febrile elderly patients with bacteremic urinary tract infection. *J. Infect.* 2005;50:296–305

Tan CK, Ulett KB, Steele M, Benjamin Jr WH & Ulett GC. Prognostic value of semi-quantitative bacteruria counts in the diagnosis of group B streptococcus urinary tract infection: a 4-year retrospective study in adult patients. *BMC Infect. Dis.* 2012;12:273

Tasbakan MI, Pullukcu H, Sipahi OR, Yamazhan T & Ulusoy S. Nitrofurantoin in the treatment of extended-spectrum beta-lactamase- producing Escherichia coli-related lower urinary tract infection. *Int. J. Antimicrob. Agents* 2012;40:554–556

Vardi M., Kochavi T., Denekamp Y. & Bitterman H. Risk factors for urinary tract infection caused by enterobacteriaceae with extended-spectrum beta-lactamase resistance in patients admitted to internal medicine departments. *Isr. Med. Assoc. J.* 2012;14:115–118

Vellinga A, Tansey S, Hanahoe B, Bennett K, Murphy AW & Cormican M. Trimethoprim and ciprofloxacin resistance and prescribing in urinary tract infection associated with Escherichia coli: a multilevel model. *J. Antimicrob. Chemother.* 2012;67:2523–2530

Patients with already microbiologically confirmed UTI only

Cohen-Nahum K, Saidel-Odes L, Riesenberg K, Schlaeffer F & Borer A. Urinary tract infections caused by multi-drug resistant Proteus mirabilis: Risk factors and clinical outcomes. *Infection* 2010;38:41–46

Hebert C., Gao Y., Rahman P. et al. Prediction of Antibiotic Susceptibility for Urinary Tract Infection in a Hospital Setting. *Antimicrob. Agents Chemother.* 2020;64:e02236–19

Jackson HA, Cashy J, Frieder O & Schaeffer AJ. Data mining derived treatment algorithms from the electronic medical record improve theoretical empirical therapy for outpatient urinary tract infections. *J. Urol.* 2011;186:2257–2262

Malcolm W, Fletcher E, Kavanagh K et al. Risk factors for resistance and MDR in community urine isolates: population-level analysis using the NHS Scotland Infection Intelligence Platform. *J. Antimicrob. Chemother.* 2018;73:223–230

Montelin H, Forsman KJ & Tangden T. Retrospective evaluation of nitrofurantoin and pivmecillinam for the treatment of lower urinary tract infections in men. *PLoS One* 2019;14:e0211098

Rattanaumpawan P, Nachamkin I, Bilker WB et al. Risk factors for ambulatory urinary tract infections caused by high-MIC fluoroquinolone-susceptible Escherichia coli in women: results from a large case-control study. *J. Antimicrob. Chemother.* 2015;70:1547–1551

Steinke D.T., Seaton R.A., Phillips G., MacDonald T.M. & Davey P.G. Prior trimethoprim use and trimethoprim-resistant urinary tract infection: A nested case-control study with multivariate analysis for other risk factors. *J. Antimicrob. Chemother.* 2001;47:781–787

Yelin I, Snitser O, Novich G et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nat. Med.* 2019;25:1143–1152

Laboratory or pharmacy data only

Christensen R.L., Creekmore F.M., Strong M.B. & Lugo R.A. The predictability of urinary pathogens based on the urinalysis nitrite test in hospitalized patients. *Hosp. Pharm.* 2007;42:52–56

Kayalp D., Dogan K., Ceylan G., Senes M. & Yucel D. Can routine automated urinalysis reduce culture requests? *Clin. Biochem.* 2013;46:1285–1289

Kim H, Kim HR, Kim TH & Lee MK. Age-Specific Cutoffs of the Sysmex

UF-1000i Automated Urine Analyzer for Rapid Screening of Urinary Tract Infections in Outpatients. *Ann. Lab. Med.* 2019;39:322–326

Metlay JP, Strom BL & Asch DA. Prior antimicrobial drug exposure: a risk factor for trimethoprim-sulfamethoxazole-resistant urinary tract infections. *J. Antimicrob. Chemother.* 2003;51:963–970

Müller M, Seidenberg R, Schuh SK et al. The development and validation of different decision-making tools to predict urine culture growth out of urine flow cytometry parameter. *PLoS One* 2018;13:e0193255

Wolterink I., Verheij T., Platteel T., Bruel Van Den A., Stam A. & Pol Van De A. Nitrofurantoin failure in elderly men: A retrospective observational study. *Antibiotics* 2020;9:211

Conference abstract

Asahata S., Ainoda Y., Fujita T. et al. Investigation of Haemophilus influenzae and H. parainfluenzae bacteriuria. *Int. J. Antimicrob. Agents* 2013;42:S151

Gerasimovska V. & Gerasimovska-Kitanovska B. Extended spectrum beta-lactamase (ESBL) strains of E. coli as a cause of urinary tract infections in hospitalized patients. *Antimicrob. Resist. Infect. Control* 2015;4

Lombardi K.M., Pourmand A. & Mazer-Amirshahi M. Pattern of antibiotic resistance of urinary tract infections in the emergency department. *Acad. Emerg. Med.* 2018;25:S241

Masel J., Kamau E., Brooks D. & Gleeson T. Clinical significance of staphylococcus aureus bacteriuria. *Open Forum Infect. Dis.* 2018;5:S467

Mitrani-Gold F., Suppapanya N. & Mundy L.M. Decision rules for microbiologically-evaluable complicated urinary tract infection in an electronic medical record-linked administrative database. *Pharmacoepidemiol. Drug Saf.* 2013;22:300–301

Petty L., Conlon A., Vaughn V. et al. Patient- and hospital-level factors and outcomes associated with treatment of asymptomatic bacteriuria in hospitalized patients: A multi-hospital cohort study. *Open Forum Infect. Dis.* 2018;5:S536–S537

Vivian G.Y.S., Gabriel L.W.T., Cheng A.L. et al. The epidemiology of urinary

tract infection in a Singapore teaching hospital: Results of a pilot audit. *Int. J. Antimicrob. Agents* 2017;50:S211–S212

Full-text not available

Holloway J, Joshi N & O'Bryan T. Positive urine nitrite test: an accurate predictor of absence of pure enterococcal bacteriuria. *South. Med. J.* 2000;93:681–682

Lee J.W., Oh K.J., Park S.C. & Rim J.S. The clinical features of complicated urinary tract infections by Pseudomonas aeruginosa. *Korean J. Urol.* 2008;49:1149–1154

Samli M.M., Dincel C., Karalar M., Sargin R., Aktepe O.C. & Altindis M. Evaluation of urinary tract infections with respect of clinical and laboratory findings. *Turk Urol. Derg.* 2003;29:87–94

Not available in English language

Diaz-Granados L.E.S., Mendoza O.E.S. & Nunez J.F.G. Clinical characteristics and risk factors for urinary tract infection with extended spectrum betalactamase infections in the emergency service of the Central Military Hospital. *Infectio* 2018;22:147–152

Oshida Y., Hirashima O., Tanaka T. & Fujimoto T. [The characteristics of urinary tract infection with urosepsis]. *Kansenshogaku Zasshi* 2014;88:678–684

Appendix C

Risk of bias appraisal tools

Studies included in the scoping review in Chapter 2 were assessed for risk of bias using two commonly used risk assessment tools: an adaption of the well-known Newcastle - Ottawa Quality Assessment Scale (NOS) for cohort studies of exposures, and the recently developed Prediction model Risk of Bias Assessment Tool (PROBAST) for studies that developed or evaluated clinical prediction models. A template for both tools is presented below.

C.1 Risk of bias in cohort studies of exposures

The following tool was proposed by Busse and Guyatt [103] and is based on the popular Newcastle - Ottowa Quality Assessment Scale (NOS). This tool was chosen over the traditional NOS due to its broader assessment of the quality of all predictors (and not only the exposure) and because it explicitly distinguishes between risk of bias — i.e., risk to the internal validity of the study – and applicability of the study results to the review question. The tool used in this thesis was restructured to align with the four categories of the Prediction model Risk of Bias Assessment Tool (PROBAST) — participants, predictors, outcomes, analysis (see Section C.2 below) — and additionally included an appraisal regarding the applicability of study results to the review question.

Table C.1: Adapted Newcastle - Ottowa Quality Assessment Scale

DOMAIN 1: Participants		
A. Risk of Bias		
Describe the sources of data and criteria for participant selection:		
1.1 Was selection of exposed and non-exposed cohorts drawn from the same population?		
1.2 Were co-interventions similar between groups?		
Risk of bias introduced by selection of participants	**RISK:** (low/high/unclear)	
Rationale of bias rating:		
B. Applicability		
Describe included participants, setting and dates:		
1.3 Was the exposed cohort representative of the average patient?		
Concern that the included participants and setting do not match the review question	**CONCERN:** (low/high/unclear)	
Rationale of applicability rating:		
DOMAIN 2: Predictors		
A. Risk of Bias		
List and describe the exposure and other prognostic factors included in the final model, e.g. definition and timing of assessment:		
2.1 Can we be confident in the assessment of exposure?		
2.2 Can we be confident in the assessment of the presence or absence of prognostic factors?		

Risk of bias introduced by predictors or their assessment	**RISK:** (low/high/unclear)	
Rationale of bias rating:		
B. Applicability		
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** (low/high/unclear)	
Rationale of applicability rating:		
DOMAIN 3: Outcomes		
A. Risk of Bias		
Describe the outcome(s), how it was defined and determined, and the time interval between predictor assessment and outcome determination:		
3.1 Can we be confident that the outcome of interest was not present at start of study?		
3.2 Can we be confident in the assessment of outcome?		
Risk of bias introduced by the outcome or its determination	**RISK:** (low/high/unclear)	
Rationale of bias rating:		
B. Applicability		
At what time point was the outcome determined: If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:		
3.3 Was the follow up of cohorts adequate?		
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** (low/high/unclear)	
Rationale of applicability rating:		

DOMAIN 4: Analysis		
A. Risk of Bias		
Describe numbers of participants, number of predictors, outcome events and events per predictor:		
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition):		
Describe missing data on predictors and outcomes as well as methods used for missing data:		
4.1 Did the study match exposed and unexposed for all variables that are associated with the outcome of interest or did the statistical analysis adjust for these prognostic variables?		
Risk of bias introduced by the analysis	**RISK:** (low/high/unclear)	
Rationale of bias rating:		

C.2 Risk of bias in risk prediction studies

Wolff *et al.* (2019) [104] recently proposed a novel tool to assess the risk of bias in studies developing (and/or evaluating) clinical risk prediction models: the Prediction model Risk of Bias Assessment Tool (PROBAST). The tool was used without further adaption to assess the risk of bias in risk prediction studies reviewed in Chapter 2.

Table C.2: Prediction model Risk of Bias Assessment Tool (PROBAST)

DOMAIN 1: Participants
A. Risk of Bias
Describe the sources of data and criteria for participant selection:

		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?			
1.2 Were all inclusions and exclusions of participants appropriate?			
Risk of bias introduced by selection of participants	**RISK:** (low/high/unclear)		
Rationale of bias rating:			
B. Applicability			
Describe included participants, setting and dates:			
Concern that the included participants and setting do not match the review question	**CONCERN:** (low/high/unclear)		
Rationale of applicability rating:			
DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment:			
		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?			
2.2 Were predictor assessments made without knowledge of outcome data?			
2.3 Are all predictors available at the time the model is intended to be used?			
Risk of bias introduced by predictors or their assessment	**RISK:** (low/high/unclear)		
Rationale of bias rating:			
B. Applicability			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** (low/high/unclear)		

Rationale of applicability rating:			
DOMAIN 3: Outcomes			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:			
		Dev	Val
3.1 Was the outcome determined appropriately?			
3.2 Was a pre-specified or standard outcome definition used?			
3.3 Were predictors excluded from the outcome definition?			
3.4 Was the outcome defined and determined in a similar way for all participants?			
3.5 Was the outcome determined without knowledge of predictor information?			
3.6 Was the time interval between predictor assessment and outcome determination appropriate?			
Risk of bias introduced by the outcome or its determination	**RISK:** (low/high/unclear)		
Rationale of bias rating:			
B. Applicability			
At what time point was the outcome determined: If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** (low/high/unclear)		
Rationale of applicability rating:			
DOMAIN 4: Analysis			

A. Risk of Bias		
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor:		
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition):		
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):		
Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism:		
Describe any participants who were excluded from the analysis:		
Describe missing data on predictors and outcomes as well as methods used for missing data:		
	Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		
4.2 Were continuous and categorical predictors handled appropriately?		
4.3 Were all enrolled participants included in the analysis?		
4.4 Were participants with missing data handled appropriately?		
4.5 Was selection of predictors based on univariable analysis avoided?		
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		
4.7 Were relevant model performance measures evaluated appropriately?		
4.8 Were model overfitting and optimism in model performance accounted for?		

4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?			
Risk of bias introduced by the analysis	**RISK:** (low/high/unclear)		
Rationale of bias rating:			

Appendix D

Propensity score analysis and coarsened exact matching

Retrospective observational studies of treatment effects are at high risk of confounding by indication due to the non-random assignment of medications. Clinicians' decisions to prescribe are likely influenced by clinical and social contexts. Applied to the case study in Chapter 4, this might lead to biases of uncertain direction. If delayed or withheld prescribing was systematically chosen for patients at the lowest risk of complications, the protective effect of antibiotics would be underestimated. In contrast, if a failure to record antibiotics in certain situations like home visits leads to patients at the highest risks to be wrongly classified as having received no antibiotics, the protective effect of antibiotics would instead be overestimated. These biases are caused by an imbalance in patient characteristics, where certain types of patients — e.g., the very sick — are mostly found in one treatment group and not the other.

In Chapter 4, I used matching and weighting techniques (propensity score [PS] matching or weighting and coarsened exact matching [CEM]) to try and account for these imbalances [168]. The following tables and figures present in-depth results from PS and CEM relating to the analyses in Chapter 4. Table D.1 presents univariate differences in covariates before (raw data) and after matching procedures were applied. Figure D.1 shows the distribution of the multivariate imbalance measure $\mathscr{L}_1$ across 250 random draws of bins, as recommended in Iacus, King & Porro (2011) [166]. Creating such a figure is helpful, since value of $\mathscr{L}_1$ depends on the (arbitrary) choice of bins [166]. The $\mathscr{L}_1$ values reported in Table 4.5

Table D.1: Standardised differences (continuous variables) and χ^2 distances (categorical variables) in the raw data, after one-to-one PS matching, and after CEM.

	Raw data	PS matching	CEM
Continuous variables			
Age (per 5 years)	-0.354	0.289	-1.088
CCI	-0.097	0.003	-0.230
Number of hospital stays (per 5)[1]	-0.015	0.000	-0.007
Number of nights in hospital (per 7)[1]	-0.114	0.004	-0.045
Number of ED visits (per 5)	-0.026	0.001	-0.016
Categorical variables			
IMD	65.442	2.355	155.120
Geographical region	14.265	2.579	92.428
Financial year	40.476	3.780	25.318
Smoking status	53.123	59.369	73.423
Recurrent UTI	2241.674	0.117	3821.223
Recent antibiotic[2]	6014.463	21.515	7196.474
Index event was home visit	3734.163	40.046	1992.171
Recent hospitalisation[2]	233.385	0.041	1136.282
Recent ED visit[2]	1466.154	0.949	1802.357

[1] Within 12 months prior to episode start. [2] Within 30 days prior to episode start.

CCI, Charlson Comorbidity Index; CEM, coarsened exact matching; ED, emergency department; IMD, Index of Multiple Deprivation; PS, propensity score; UTI, urinary tract infection.

correspond to the median value in the raw data (dotted red line in Figure D.1). Using these rebalanced data, Tables D.2–D.5 show generalised estimating equations results for the analysis of an association between delayed or withheld antibiotic prescribing and risk of progression to severe urinary tract infection (Table D.2), death (Table D.3), hospitalisation for lower respiratory tract infection (Table D.4), and hospitalisations for other reasons (Table D.5).

Table D.2: Multivariable associations between delayed or withheld antibiotic prescribing for UTI and progression to severe UTI within 30 days, adjusting for covariates using GEEs and Huber-White sandwich estimators and re-balancing via PS or CEM.

	PS matching	**PS weighting**	**CEM**[3]
	aOR (95%-CI)	aOR (95%-CI)	aOR (95%-CI)
Delayed / withheld prescribing	1.57 (1.44–1.71)	1.49 (1.41–1.58)	1.87 (1.72–2.04)
Age (per 5 years)	1.11 (1.10–1.13)	1.08 (1.08–1.09)	1.09 (1.08–1.10)
IMD quintiles			
Q1	1	1	1
Q2	1.07 (0.94–1.21)	1.11 (1.04–1.17)	1.14 (1.04–1.26)
Q3	1.09 (0.96–1.24)	1.15 (1.09–1.22)	1.20 (1.09–1.33)
Q4	1.14 (0.99–1.31)	1.29 (1.21–1.37)	1.26 (1.13–1.41)
Q5	1.25 (1.09–1.45)	1.40 (1.31–1.49)	1.14 (1.01–1.29)
Region			
South	1	1	1
London	0.92 (0.80–1.08)	1.00 (0.94–1.07)	0.95 (0.83–1.08)
Midlands and East	1.04 (0.93–1.15)	1.07 (1.02–1.12)	1.17 (1.08–1.26)
North and Yorkshire	0.94 (0.84–1.06)	1.06 (1.01–1.11)	1.03 (0.93–1.13)
Financial year			
2007	0.82 (0.69–0.98)	0.99 (0.92–1.07)	0.88 (0.77–1.00)
2008	0.98 (0.84–1.16)	1.08 (1.01–1.16)	0.93 (0.82–1.07)
2009	1.00 (0.85–1.17)	1.05 (0.98–1.12)	1.04 (0.92–1.18)
2010	1	1	1
2011	1.00 (0.85–1.17)	0.99 (0.92–1.06)	0.97 (0.85–1.10)
2012	1.00 (0.85–1.18)	0.91 (0.85–0.98)	1.03 (0.91–1.18)
2013	0.89 (0.75–1.05)	1.02 (0.95–1.10)	1.03 (0.91–1.18)
2014	1.07 (0.90–1.28)	0.96 (0.89–1.04)	1.11 (0.96–1.28)
CCI	1.11 (1.09–1.14)	1.14 (1.13–1.15)	1.16 (1.13–1.18)
Smoking status			
Non-smoker	1	1	1
Ex-smoker	0.97 (0.88–1.07)	1.05 (1.00–1.10)	0.94 (0.86–1.03)
Smoker	1.20 (1.04–1.37)	1.35 (1.28–1.43)	1.17 (1.05–1.31)
Recurrent UTI	1.11 (1.01–1.22)	0.76 (0.72–0.80)	1.12 (1.02–1.23)
Recent antibiotic[1]	1.18 (1.08–1.30)	1.17 (1.12–1.22)	1.36 (1.24–1.49)
Index event was home visit	1.98 (1.75–2.25)	2.04 (1.89–2.20)	3.31 (2.75–3.99)
Hospital stays			
Recent hospitalisation[1]	1.24 (1.07–1.45)	1.81 (1.70–1.92)	1.92 (1.37–2.70)
Number of stays (per 5)[2]	2.46 (1.98–3.06)	2.87 (2.54–3.24)	-
Number of nights (per 7)[2]	1.04 (1.02–1.06)	1.02 (1.02–1.03)	1.09 (1.03–1.15)
ED visits			
Recent visit[1]	1.45 (1.23–1.70)	1.36 (1.23–1.50)	-
Number of visits (per 5)[2]	0.98 (0.87–1.11)	1.25 (1.18–1.33)	3.20 (2.37–4.33)

[1] Within 30 days prior to episode start.
[2] Within 12 months prior to episode start.
[3] ED visit in the prior 30 days and number of hospitalisations in the prior year where excluded as covariates in the CEM analysis due to small remaining numbers after matching.

95% CI, 95% confidence interval; aOR, adjusted odds ratio; CCI, Charlson Comorbidity Index; CEM, coarsened exact matching; ED, emergency department; GEE, generalised estimating equations; IMD, Index of Multiple Deprivation; PS, propensity score; UTI, urinary tract infection.

Table D.3: Multivariable associations between delayed or withheld antibiotic prescribing for UTI and death within 30 days, adjusting for covariates using GEEs and Huber-White sandwich estimators and re-balancing via PS or CEM.

	PS matching	**PS weighting**	**CEM**[3]
	aOR (95%-CI)	aOR (95%-CI)	aOR (95%-CI)
Delayed / withheld prescribing	1.31 (1.13–1.53)	1.49 (1.41–1.58)	1.45 (1.21–1.74)
Age (per 5 years)	1.52 (1.46–1.57)	1.08 (1.08–1.09)	1.95 (1.87–2.04)
IMD quintiles			
Q1	1	1	1
Q2	1.07 (0.85–1.33)	1.11 (1.04–1.17)	1.18 (0.96–1.46)
Q3	1.08 (0.86–1.36)	1.15 (1.09–1.22)	1.43 (1.16–1.76)
Q4	1.35 (1.07–1.71)	1.29 (1.21–1.37)	1.30 (1.02–1.67)
Q5	1.22 (0.94–1.58)	1.40 (1.31–1.49)	1.36 (1.01–1.82)
Region			
South	1	1	1
London	0.57 (0.41–0.78)	1.00 (0.94–1.07)	0.49 (0.32–0.75)
Midlands and East	0.95 (0.80–1.14)	1.07 (1.02–1.12)	0.91 (0.77–1.08)
North and Yorkshire	0.96 (0.78–1.17)	1.06 (1.01–1.11)	0.79 (0.63–0.99)
Financial year			
2007	0.95 (0.71–1.28)	0.99 (0.92–1.07)	1.28 (0.99–1.67)
2008	0.88 (0.66–1.19)	1.08 (1.01–1.16)	0.90 (0.68–1.20)
2009	1.05 (0.79–1.39)	1.05 (0.98–1.12)	1.02 (0.77–1.35)
2010	1	1	1
2011	1.03 (0.77–1.37)	0.99 (0.92–1.06)	0.80 (0.59–1.07)
2012	1.12 (0.84–1.48)	0.91 (0.85–0.98)	1.03 (0.77–1.36)
2013	0.81 (0.60–1.11)	1.02 (0.95–1.10)	0.79 (0.57–1.07)
2014	1.05 (0.77–1.43)	0.96 (0.89–1.04)	0.99 (0.72–1.35)
CCI	1.19 (1.16–1.23)	1.14 (1.13–1.15)	1.14 (1.09–1.19)
Smoking status			
Non-smoker	1	1	1
Ex-smoker	0.99 (0.83–1.17)	1.05 (1.00–1.10)	0.99 (0.82–1.20)
Smoker	1.36 (1.00–1.86)	1.35 (1.28–1.43)	2.23 (1.48–3.35)
Recurrent UTI	0.77 (0.65–0.92)	0.76 (0.72–0.80)	0.75 (0.60–0.94)
Recent antibiotic[1]	1.10 (0.94–1.29)	1.17 (1.12–1.22)	1.28 (1.05–1.55)
Index event was home visit	2.13 (1.80–2.53)	2.04 (1.89–2.20)	2.22 (1.74–2.83)
Hospital stays			
Recent hospitalisation[1]	1.89 (1.48–2.40)	1.81 (1.70–1.92)	1.82 (0.66–4.99)
Number of stays (per 5)[2]	2.55 (1.65–3.94)	2.02 (1.58–2.60)	-
Number of nights (per 7)[2]	1.05 (1.02–1.07)	1.02 (1.02–1.03)	1.09 (1.00–1.19)
ED visits			
Recent visit[1]	1.15 (0.86–1.55)	1.36 (1.14–1.62)	-
Number of visits (per 5)[2]	0.65 (0.44–0.97)	1.25 (1.18–1.33)	0.91 (0.36–2.31)

[1] Within 30 days prior to episode start.
[2] Within 12 months prior to episode start.
[3] ED visit in the prior 30 days and number of hospitalisations in the prior year where excluded as covariates in the CEM analysis due to small remaining numbers after matching.

95% CI, 95% confidence interval; aOR, adjusted odds ratio; CCI, Charlson Comorbidity Index; CEM, coarsened exact matching; ED, emergency department; IMD, Index of Multiple Deprivation; PS, propensity score; UTI, urinary tract infection.

Table D.4: Multivariable associations between delayed or withheld antibiotic prescribing for UTI and hospitalisation for LRTI within 30 days, adjusting for covariates using GEEs and Huber-White sandwich estimators and re-balancing via PS or CEM.

	PS matching	**PS weighting**	**CEM**[3]
	aOR (95%-CI)	aOR (95%-CI)	aOR (95%-CI)
Delayed / withheld prescribing	1.34 (1.02–1.76)	1.49 (1.41–1.58)	1.62 (1.20– 2.18)
Age (per 5 years)	1.32 (1.25–1.38)	1.08 (1.08–1.09)	1.37 (1.31– 1.43)
IMD quintiles			
Q1	1	1	1
Q2	1.06 (0.69–1.62)	1.11 (1.04–1.17)	1.01 (0.70– 1.47)
Q3	1.44 (0.96–2.16)	1.15 (1.09–1.22)	1.39 (0.98– 1.97)
Q4	1.37 (0.89–2.12)	1.29 (1.21–1.37)	1.74 (1.20– 2.53)
Q5	1.11 (0.68–1.79)	1.40 (1.31–1.49)	1.75 (1.14– 2.69)
Region			
South	1	1	1
London	0.75 (0.45–1.25)	1.00 (0.94–1.07)	0.69 (0.39– 1.20)
Midlands and East	1.03 (0.74–1.43)	1.07 (1.02–1.12)	0.99 (0.75– 1.31)
North and Yorkshire	1.05 (0.73–1.51)	1.06 (1.01–1.11)	0.97 (0.70– 1.35)
Financial year			
2007	0.67 (0.38–1.21)	0.99 (0.92–1.07)	0.68 (0.43– 1.07)
2008	0.92 (0.54–1.54)	1.08 (1.01–1.16)	0.80 (0.52– 1.24)
2009	0.95 (0.57–1.60)	1.05 (0.98–1.12)	0.93 (0.61– 1.41)
2010	1	1	1
2011	1.13 (0.69–1.87)	0.99 (0.92–1.06)	0.85 (0.55– 1.30)
2012	1.20 (0.73–1.97)	0.91 (0.85–0.98)	0.88 (0.57– 1.35)
2013	1.08 (0.64–1.82)	1.02 (0.95–1.10)	0.60 (0.36– 0.99)
2014	0.92 (0.52–1.66)	0.96 (0.89–1.04)	0.77 (0.46– 1.29)
CCI	1.13 (1.06–1.20)	1.14 (1.13–1.15)	1.14 (1.06– 1.22)
Smoking status			
Non-smoker	1	1	1
Ex-smoker	1.08 (0.79–1.48)	1.05 (1.00–1.10)	1.07 (0.80– 1.43)
Smoker	2.36 (1.56–3.58)	1.35 (1.28–1.43)	1.94 (1.25– 3.01)
Recurrent UTI	0.66 (0.47–0.91)	0.76 (0.72–0.80)	0.74 (0.52– 1.05)
Recent antibiotic[1]	1.23 (0.93–1.64)	1.17 (1.12–1.22)	1.57 (1.15– 2.12)
Index event was home visit	1.34 (0.93–1.94)	2.04 (1.89–2.20)	2.70 (1.67– 4.39)
Hospital stays			
Recent hospitalisation[1]	1.17 (0.74–1.85)	1.81 (1.70–1.92)	4.02 (1.48–10.96)
Number of stays (per 5)[2]	2.98 (1.75–5.07)	2.12 (1.52–2.96)	-
Number of nights (per 7)[2]	1.05 (1.01–1.09)	1.02 (1.02–1.03)	0.99 (0.78– 1.26)
ED visits			
Recent visit[1]	1.40 (0.87–2.27)	1.30 (0.97–1.73)	-
Number of visits (per 5)[2]	0.92 (0.73–1.15)	1.25 (1.18–1.33)	1.53 (0.37– 6.22)

[1] Within 30 days prior to episode start.
[2] Within 12 months prior to episode start.
[3] ED visit in the prior 30 days and number of hospitalisations in the prior year where excluded as covariates in the CEM analysis due to small remaining numbers after matching.

95% CI, 95% confidence interval; aOR, adjusted odds ratio; CCI, Charlson Comorbidity Index; CEM, coarsened exact matching; ED, emergency department; GEE, generalised estimating equations; IMD, Index of Multiple Deprivation; LRTI, lower respiratory tract infection; PS, propensity score; UTI, urinary tract infection.

Table D.5: Multivariable associations between delayed or withheld antibiotic prescribing for UTI and hospitalisation for reasons unrelated to UTI, LRTI, or bloodstream infection within 30 days, adjusting for covariates using GEEs and Huber-White sandwich estimators and re-balancing via PS or CEM.

	PS matching	**PS weighting**	**CEM**[3]
	aOR (95%-CI)	aOR (95%-CI)	aOR (95%-CI)
Delayed / withheld prescribing	1.39 (1.29–1.50)	1.49 (1.41–1.58)	1.58 (1.46–1.70)
Age (per 5 years)	1.09 (1.08–1.10)	1.08 (1.08–1.09)	1.09 (1.08–1.10)
IMD quintiles			
Q1	1	1	1
Q2	1.10 (0.98–1.23)	1.11 (1.04–1.17)	1.06 (0.97–1.15)
Q3	1.17 (1.04–1.31)	1.15 (1.09–1.22)	1.16 (1.06–1.26)
Q4	1.40 (1.25–1.58)	1.29 (1.21–1.37)	1.27 (1.16–1.39)
Q5	1.46 (1.29–1.65)	1.40 (1.31–1.49)	1.53 (1.39–1.68)
Region			
South	1	1	1
London	0.89 (0.79–1.02)	1.00 (0.94–1.07)	0.96 (0.86–1.07)
Midlands and East	1.07 (0.98–1.17)	1.07 (1.02–1.12)	1.09 (1.02–1.16)
North and Yorkshire	1.00 (0.90–1.10)	1.06 (1.01–1.11)	1.06 (0.99–1.15)
Financial year			
2007	1.03 (0.89–1.19)	0.99 (0.92–1.07)	1.00 (0.90–1.12)
2008	1.13 (0.98–1.30)	1.08 (1.01–1.16)	1.04 (0.93–1.15)
2009	1.10 (0.96–1.26)	1.05 (0.98–1.12)	1.06 (0.95–1.17)
2010	1	1	1
2011	1.01 (0.88–1.17)	0.99 (0.92–1.06)	0.99 (0.89–1.10)
2012	0.91 (0.79–1.06)	0.91 (0.85–0.98)	0.85 (0.76–0.95)
2013	1.11 (0.96–1.28)	1.02 (0.95–1.10)	0.93 (0.83–1.04)
2014	0.93 (0.79–1.09)	0.96 (0.89–1.04)	0.98 (0.87–1.11)
CCI	1.13 (1.11–1.15)	1.14 (1.13–1.15)	1.19 (1.16–1.21)
Smoking status			
Non-smoker	1	1	1
Ex-smoker	1.08 (0.99–1.18)	1.05 (1.00–1.10)	1.00 (0.93–1.08)
Smoker	1.39 (1.24–1.54)	1.35 (1.28–1.43)	1.31 (1.21–1.43)
Recurrent UTI	0.71 (0.65–0.78)	0.76 (0.72–0.80)	0.67 (0.61–0.73)
Recent antibiotic[1]	1.08 (0.99–1.17)	1.17 (1.12–1.22)	1.18 (1.08–1.28)
Index event was home visit	1.92 (1.72–2.15)	2.04 (1.89–2.20)	2.41 (2.01–2.89)
Hospital stays			
Recent hospitalisation[1]	1.71 (1.52–1.91)	1.81 (1.70–1.92)	2.43 (1.88–3.14)
Number of stays (per 5)[2]	4.10 (3.43–4.91)	4.64 (4.19–5.14)	-
Number of nights (per 7)[2]	1.01 (1.00–1.03)	1.02 (1.02–1.03)	1.07 (1.02–1.13)
ED visits			
Recent visit[1]	1.55 (1.37–1.76)	1.58 (1.47–1.70)	-
Number of visits (per 5)[2]	1.13 (1.03–1.24)	1.25 (1.18–1.33)	5.16 (4.00–6.67)

[1] Within 30 days prior to episode start.
[2] Within 12 months prior to episode start.
[3] ED visit in the prior 30 days and number of hospitalisations in the prior year where excluded as covariates in the CEM analysis due to small remaining numbers after matching.

95% CI, 95% confidence interval; aOR, adjusted odds ratio; CCI, Charlson Comorbidity Index; CEM, coarsened exact matching; ED, emergency department; GEE, generalised estimating equations; IMD, Index of Multiple Deprivation; LRTI, lower respiratory tract infection; PS, propensity score; UTI, urinary tract infection.

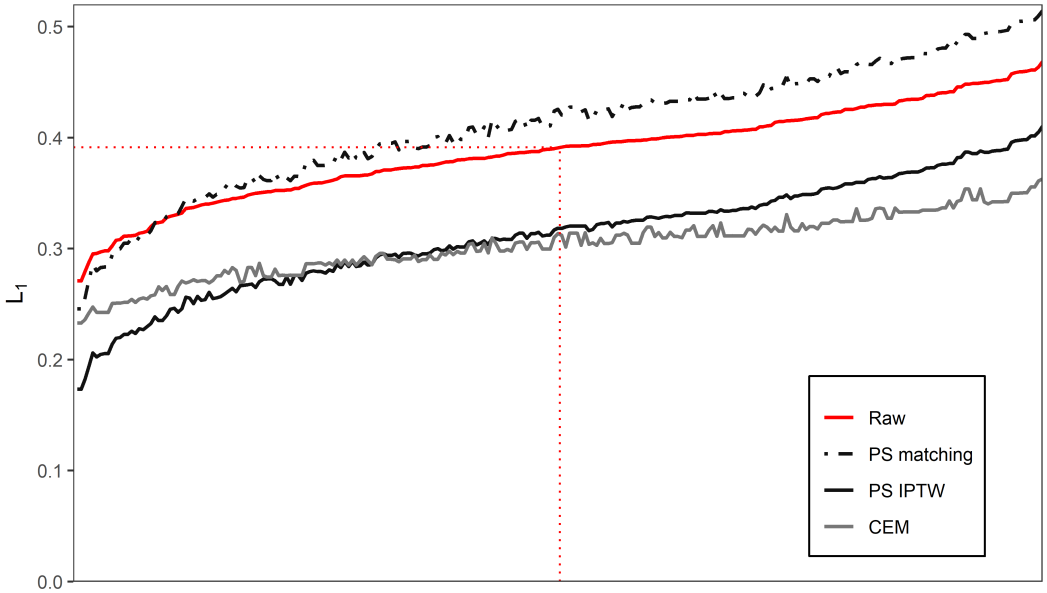

Figure D.1: $\mathscr{L}_1$ profile of the raw data from Chapter 4 (red line) compared to balanced data after PS matching, PS weighting, and CEM, evaluated in 250 randomly chosen bins H and ordered by the value of $\mathscr{L}_1$ in the raw data.

CEM, coarsened exact matching; IPTW, inverse probability of treatment weighting; PS, propensity score.

Appendix E

Comparison of pre-processing and imputation methods

Many different data pre-processing and imputation methods could have been chosen to prepare the data in Chapter 5 for analysis. I chose to apply a selection of commonly used pre-processing steps (identity, log transformation, Yeo-Johnson transformation) and imputation approaches (mean imputation, k-nearest neighbour imputation, kmeans imputation, multiple imputation), and compare their relative performances. A summary of my results were reported in Chapter 5. More detailed results of these sensitivity analyses are presented below.

E.1 Pre-processing steps

The choice of transformations prior to model fitting in Chapter 5 had a strong impact on the performance of linear models. Logistic regression (LR) that received raw, — i.e., not transformed — data had much lower estimated performance of in terms of its area under the receiver operating characteristic (AUROC): .661 (95% CI .654–.668; Table E.1). Tree-based models performed similar irrespective of which transformation was used. A log transformation of all urinalysis and lab data except those representing percentages achieved the best discriminatory performance across models and was therefore presented in Tables 5.6 and 5.8. A more flexible Yeo-Johnson transformation, which automatically chooses the best power transformations, performed very similar to the simpler log transformation. A gradient boosting tree following a Yeo-Johnson transformation achieved an identical performance of AUROC .808 (95% CI .802–.814), but the transformation led to slightly worse performance with other algorithms like LR.

Table E.1: Discriminative performance during internal validation for LR and XGB models by pre-processing method

	LR	XGB
Transformation	AUROC (95% CI)	AUROC (95% CI)
Raw / identity	.661 (.654–.668)	.808 (.802–.814)
Log transformation	.788 (.782–.793)	.808 (.802–.814)
Yeo-Johnson transformation	.785 (.779–.791)	.808 (.802–.814)

Note: All models presented in this table used mean imputation. Performance metrics represent performance of the best hyperparameter combination during internal validation.

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; LR, logistic regression; XGB, extreme gradient boosting tree.

E.2 Imputation methods

The addition of more complex imputation methods did not improve the estimated performance of the model (Table E.2). Models trained on data imputed with a KNN approach with a preset size of 5 neighbours achieved a peak performance of AUROC of .805 (95% CI .799–.810) which was comparable but slightly lower than the performance estimated for simple mean imputation with missing indicator variables. Similarly, the k-means binning strategy used previously by Taylor *et al.* (2018) [109] did not improve performance and resulted in a peak AUROC of .779 (95% CI .774–.785). Due to its increased computational burden, multiple imputation was only applied for LR which doesn't require hyperparameter tuning. Using multiple imputation led to significantly lower performance estimates of AUROC .643 (95% CI .634–.652) for LR.

Since the sub-par performance of multiple imputation was quite surprising, a careful re-examination of the source code was performed to exclude implementation errors as a possible source of discrepancy but no errors could be identified. A possible explanation for these findings might be the fact that almost all of the model performance appeared to be derived from urinalysis parameters. Predictions for patients with missing data on these variables were therefore generally poor, particularly if all urinalysis parameters were jointly missing as was usually the case (Figure 5.3). As briefly discussed in the methods section, the main reason for missing information on urinalysis parameters were high urine viscosity or a low urine sample quantity. No other variable measured in the ED was strongly

Table E.2: Discriminative performance during internal validation for LR and XGB by imputation method (after log-transformation).

	LR	**XGB**
Imputation method	AUROC (95% CI)	AUROC (95% CI)
Mean imputation	.788 (.782–.793)	.808 (.802–.814)
k-nearest neighbour	.786 (.781–.792)	.805 (.799–.810)
k-means	.777 (.771–.783)	.779 (.774–.785)
Multiple imputation	.643 (.634–.652)	-

Note: All models presented in this table used log transformation. Performance metrics represent performance of the best hyperparameter combination during internal validation.

95% CI, 95% confidence interval; AUROC, area under the receiver operating characteristic; LR, logistic regression; XGB, extreme gradient boosting tree.

associated with the probability of urinalysis parameters being missing or the values of urinalysis parameters, preventing more sophisticated imputation procedures from capitalising on correlation structures between predictors. Consequently, it is possible that a combination of mean imputation and missing indicators was the best — or at least most efficient — representation of the missing pattern in this dataset. The even larger drop in model performance when using multiple imputation on all predictors might be explained by the extremely high proportion of missing values among vital signs and laboratory measurements. While these values would be largely ignored by classification algorithms using mean imputation, multiple imputation captures the large uncertainty surrounding those values and propagates them to the classification algorithm. This might cause a substantial increase of random noise without any increase in true signals, leading to the drop in observed performance.

Appendix F

Agreement between coded diagnoses and case notes in the ED

In Chapters 5 and 6, I relied on the correctness of clinical coding in the emergency department (ED). Patients who had a recorded diagnosis of lower urinary tract infection (UTI), pyelonephritis, or urosepsis were considered to have a suspected UTI. However, as briefly touched upon throughout the analysis chapters, a recorded diagnosis of UTI might represent a variety of cases, ranging from elderly patients presenting with confusion to young women with urinary symptoms such as dysuria and haematuria. These characteristics, in turn, might influence the a priori probability of positive bacterial growth during culture and change the interpretation of model predictions when using it in clinical practice — e.g., in the case of likely asymptomatic bacteriuria (see Chapter 1). In this additional analysis, I therefore compared recorded ED diagnoses to manually curated information documented in patients' case notes, evaluating whether ED diagnosis of UTI at Queen Elizabeth Hospital Birmingham (QEHB) can be interpreted as clear suspicion of UTI or whether it describes a heterogeneous patient population.

Data: In a pilot study prior to the analysis presented in Chapter 5, I previously selected a random subset of 1,000 patients (300 patients discharged from the ED and 700 patients admitted to hospital) who visited the ED at QEHB between January 2014 and May 2017 and who had a urine sample submitted for microbiological culture within 48 hours of arrival in hospital [24]. After excluding children, patients directly referred to specialist care, and patients without a valid hospital

record or identifier, 943 patients were included in this pilot study[1]. Case notes recorded in the ED were manually reviewed by Dr Laura Shallcross, extracting all recorded free-text information on urinary symptoms (dysuria, haematuria, frequency, urgency, hesitancy, and difficulty), relevant pain (abdominal, back, flank, loin, and suprapubic), fever, and vomiting. A detailed description of the data can be found in Shallcross *et al.* (2020) [24].

Methods: Re-analysing the above data, I assessed the sensitivity, specificity, and positive predictive value (PPV) of ED diagnoses for UTI (lower UTI, pyelonephritis, and urosepsis) at QEHB in reflecting recorded evidence of UTI in ED case notes, defined as presence of at least two urinary symptoms, or a single urine symptom and relevant pain. Coding reliability was assessed for a combined category of UTI (lower UTI, pyelonephritis, or urosepsis), as well as for every condition individually. Approximate 95% confidence intervals (95% CI) were calculated by bootstrapping the original data 1,000 times and taking the 2.5% and 97.5% percentiles.

Results: Out of the random selection of 943 ED patients who had a urine sample submitted for microbiological culture, 307 (32.6%) patients had a coded ED diagnosis of UTI and 162 (17.2%) had evidence of UTI recorded in their case notes (Figure F.1 A). The most commonly recorded urinary symptom in the case notes was dysuria (151; 16.0%) followed by urinary frequency (98; 10.4%) and haematuria (54; 5.7%; Table F.1). Most urinary symptoms as well as pain and fever were positively associated with an ED diagnosis of UTI. ED diagnosis of UTI had a sensitivity of 61.7% (95% CI 54.4–69.1) and a PPV of 32.6% (95% CI 27.3–37.9) when using it to identify patients with evidence of UTI in their case notes (Figure F.1 A). Specificity of diagnosis codes was higher but still limited at 73.5% (95% CI 70.5–76.8). The reliability of diagnosis codes varied by condition. While 45.6% of pyelonephritis cases had evidence of UTI according to case note review, this was true for only 29.3% of lower UTI and 31.1% of urosepsis cases

[1] The patients included in the pilot study overlapped with the cohort used in Chapters 5 and 6. However, the exact extent of overlap could not be determined since different patient identifiers were provided for this data extract due to concerns about patient confidentiality.

Table F.1: Number and proportion of urinary symptoms recorded in the case notes of 943 randomly selected ED patients who had a urine sample submitted for microbiological culture at QEHB.

	All	ED diagnosis of UTI		p-value[1]
		Yes	No	
Total number of patients (row-%)	943 (100)	307 (32.6)	636 (67.4)	
Urinary symptoms (%)				
Dysuria	151 (16.0)	96 (31.5)	55 (8.6)	<0.001
Urinary frequency	98 (10.4)	59 (19.2)	39 (6.1)	<0.001
Haematuria	54 (5.7)	26 (8.5)	28 (4.4)	0.018
Difficulty urinating	29 (3.1)	11 (3.6)	18 (2.8)	0.670
Urinary urgency	15 (1.6)	<10 (<3.3)	<10 (<1.6)	0.369
Other urinary symptoms[2]	32 (3.4)	18 (5.9)	14 (2.2)	0. 007
Pain (%)	223 (23.6)	99 (32.2)	124 (19.5)	<0.001
Fever (%)	158 (16.8)	83 (27.0)	75 (11.8)	<0.001
Vomiting (%)	154 (16.3)	54 (17.6)	100 (15.7)	0.527

[1] Obtained via χ^2 tests.
[2] Including hesitancy, urinary retention, malodorous urine, and catheter related symptoms.

ED, emergency department; UTI, urinary tract infection.

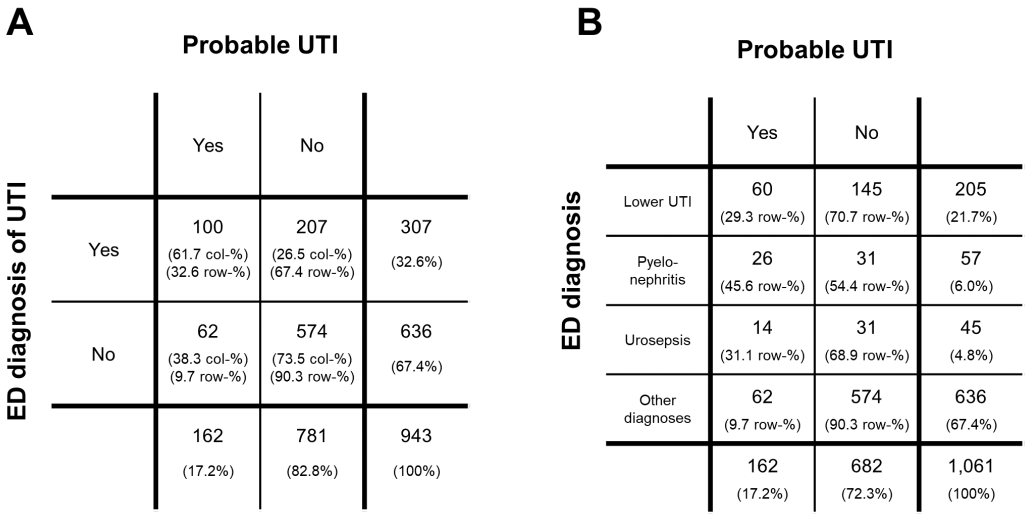

Figure F.1: Agreement between evidence of UTI recorded in case notes and coded ED diagnosis for **A)** UTI / No UTI and **B)** Lower UTI / Pyelonephritis / Urosepsis / No UTI.

Evidence of UTI in the case notes was defined as the presence of at least two urinary symptoms, or a single urine symptom and relevant pain. Note that the total number of UTI cases differs between the analysis shown here and that published in Shallcross *et al.* (2020) [24], since the published results included additional information collected from the case notes to determine ED diagnosis.

ED, emergency department; UTI, urinary tract infection.

(Figure F.1 B). More overall cases (27.7%) would have had evidence of UTI if fever and vomiting recorded in the case notes were also considered symptoms, but sensitivity of coded ED diagnoses would have been reduced from 61.7% to 54.4%, likely due to the large number of alternative causes of these symptoms. Notably, by including just fever 71.9% of pyelonephritis cases would end up being classified as having evidence of UTI, highlighting the importance of this symptom for coding of pyelonephritis in the ED at QEHB. Sensitivity and specificity remained similar but PPV also improved to 42.3% (95% CI 36.8–47.5) when including fever.

Discussion: There was limited agreement between coded ED diagnosis for suspected lower UTI, pyelonephritis, and urosepsis in the ED and evidence of UTI status derived manually from case notes. Only a minority of patients with a diagnosis of UTI in the ED had clear urinary symptoms documented in their case notes, whereas many more had diffuse symptoms of local pain, fever, or vomiting. These findings demonstrate the difficulty of diagnosing UTI in the ED more generally, and highlight the limitations of determining patient status and suspected diagnosis from clinical codes alone.

Appendix G

Reporting guidelines

The quality of reporting of medical research in peer-reviewed articles has repeatedly been found to be poor [105, 169, 170, 200]. In order to promote a more comprehensive and standardised way of describing research findings in published literature, reporting checklists have been devised. These checklist aid researchers by specifying the minimum information that should be included in a manuscript to allow for a sufficient appraisal of the performed analyses. Over time, separate checklists have been developed for most major study designs (e.g., randomized controlled trials, retrospective cohort studies, etc.).

In each chapter of this thesis that described original primary research, I used the appropriate reporting checklist to ensure all relevant information was included. For Chapter 2, I used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Extension for Scoping Reviews (ScR) statement [105]. For Chapter 4, I used the the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [169] and the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) [170], which extends STROBE for electronic health records research. For Chapter 5, I used the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [200]. As Chapter 6 primarily presented sensitivity analyses for Chapter 5, no reporting statement was used. Filled in statements for each chapter are presented below.

G.1 PRISMA-ScR checklist (Chapter 2)

Table G.1: PRISMA-ScR reporting checklist for Chapter 2: Use of EHR data to guide diagnosis and management of suspected community-acquired UTI in adults

Section/Topic	PRISMA-ScR item	Location
Title		
Title	Identify the report as a scoping review.	Chapter 2
Abstract		
Structured summary	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	Chapter 2
Introduction		
Rationale	Describe the rationale for the review in the context of what is already known. Explain why the review questions/objectives lend themselves to a scoping review approach.	Sections 2.1 and 2.2
Objectives	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	Section 2.3
Methods		
Protocol and registration	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	Section 2.3
Eligibility criteria	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	Sections 2.4.1 and 2.4.2

Table G.1 Continued: PRISMA-ScR reporting checklist for Chapter 2

Section/Topic	PRISMA-ScR item	Location
Information sources	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	Section 2.4.3
Search	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	Table 2.1
Selection of sources of evidence	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	Section 2.4.4
Data charting process	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	Section 2.4.5
Data items	List and define all variables for which data were sought and any assumptions and simplifications made.	Section 2.4.5
Critical appraisal of individual sources of evidence	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	Section 2.4.6
Synthesis of results	Describe the methods of handling and summarizing the data that were charted.	Section 2.4.7
Results		
Selection of sources of evidence	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	Figure 2.1

Table G.1 Continued: PRISMA-ScR reporting checklist for Chapter 2

Section/Topic	PRISMA-ScR item	Location
Characteristics of sources of evidence	For each source of evidence, present characteristics for which data were charted and provide the citations.	Table 2.2
Critical appraisal within sources of evidence	If done, present data on critical appraisal of included sources of evidence.	Table 2.3
Results of individual sources of evidence	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	Section 2.5.2
Synthesis of results	Summarize and/or present the charting results as they relate to the review questions and objectives.	Section 2.5.2
Discussion		
Summary of evidence	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	Section 2.6
Limitations	Discuss the limitations of the scoping review process.	Section 2.6.1
Conclusions	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	Section 2.6.2
Funding		
Funding	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	Not applicable

G.2 STROBE and RECORD checklists (Chapter 4)

Table G.2: STROBE reporting checklist for Chapter 4: Antibiotic prescribing for lower UTI in primary care and subsequent risk of infectious complication

Section/Topic	STROBE item	Location
Title and abstract		
Title and abstract	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	Chapter 4 Chapter 4
Introduction		
Background rationale	Explain the scientific background and rationale for the investigation being reported	Section 4.2
Objectives	State specific objectives, including any prespecified hypotheses	Section 4.3
Methods		
Study Design	Present key elements of study design early in the paper	Section 4.4
Setting	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Sections 4.4.1–4.4.6
Participants	(a) Cohort study - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up (b) Cohort study - For matched studies, give matching criteria and number of exposed and unexposed	Section 4.4.1 Matching was only performed in secondary analysis (Section 4.4.8)
Variables	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	Sections 4.4.4–4.4.6

Table G.2 Continued: STROBE reporting checklist for Chapter 4

Section/Topic	STROBE item	Location
Data sources/ measurement	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Sections 4.4.4–4.4.6
Bias	Describe any efforts to address potential sources of bias	Sections 4.4.4 and 4.4.8
Study size	Explain how the study size was arrived at	Figure 4.2
Quantitative variables	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why	Section 4.4.7
Statistical methods	(a) Describe all statistical methods, including those used to control for confounding	Section 4.4.8
	(b) Describe any methods used to examine subgroups and interactions	Section 4.4.8
	(c) Explain how missing data were addressed	Section 4.4.7
	(d) Cohort study - If applicable, explain how loss to follow-up was addressed	Section 4.4.1
	(e) Describe any sensitivity analyses	Section 4.4.8
Other information		
Participants	(a) Report the numbers of individuals at each stage of the study (e.g., numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed)	Figure 4.2
	(b) Give reasons for non-participation at each stage.	Figure 4.2
	(c) Consider use of a flow diagram	Figure 4.2
Descriptive data	(a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders	Table 4.2
	(b) Indicate the number of participants with missing data for each variable of interest	Section 4.4.7

Table G.2 Continued: STROBE reporting checklist for Chapter 4

Section/Topic	STROBE item	Location
	(c) Cohort study - summarise follow-up time (e.g., average and total amount)	Section 4.5
Outcome data	Cohort study - Report numbers of outcome events or summary measures over time	Table 4.2
Main results	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included	Table 4.3 and Section 4.4.7
	(b) Report category boundaries when continuous variables were categorized	Only applicable for coarse matching, see Section 4.4.8
	(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	Section 4.5
Other analyses	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	Sections 4.5.2–4.5.4 and Appendix D
Discussion		
Key results	Summarise key results with reference to study objectives	Section 4.6
Limitations	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Section 4.6.3
Interpretation	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Section 4.6.4
Generalisability	Discuss the generalisability (external validity) of the study results	Sections 4.7
Other information		

Table G.2 Continued: STROBE reporting checklist for Chapter 4

Section/Topic	STROBE item	Location
Funding	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Not applicable

Table G.3: RECORD reporting checklist for Chapter 4: Antibiotic prescribing for lower UTI in primary care and subsequent risk of infectious complication

Section/Topic	RECORD item	Location
Title and abstract		
Title and abstract	a) The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included.	Chapter 4
	b) If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract.	Chapter 4
	c) If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.	Chapter 4
Methods		
Participants	a) The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided.	Appendix H
	b) Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided.	No direct validation studies exist.
	c) If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.	Section 3.2.1

Table G.3 Continued: RECORD reporting checklist for Chapter 4

Section/Topic	RECORD item	Location
Variables	A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Appendix H
Data access and cleaning methods	a) Authors should describe the extent to which the investigators had access to the database population used to create the study population. b) Authors should provide information on the data cleaning methods used in the study.	Section 4.4.1 Sections 4.4.4–4.4.7
Linkage	State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	Section 3.2.1
Results		
Participants	Describe in detail the selection of the persons included in the study (i.e. study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.	Section 4.4.1 and Figure 4.2
Discussion		
Limitations	Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Section 4.6.3
Other information		

Table G.3 Continued: RECORD reporting checklist for Chapter 4

Section/Topic	RECORD item	Location
Accessibility of protocol, raw data, and programming code	Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	End of chapter

G.3 TRIPOD checklist (Chapter 5)

Table G.4: TRIPOD reporting checklist for Chapter 5: Using EHR data to predict bacteriuria in the ED: a case study using data from QEHB

Section/Topic	TRIPOD item	Location
Title and abstract		
Title	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Not applicable
Abstract	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	Chapter 5
Introduction		
Background and objectives	a) Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models. b) Specify the objectives, including whether the study describes the development or validation of the model or both.	Section 5.2 Section 5.3
Methods		
Source of data	a) Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable. b) Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	Section 5.4.1, identical for development and validation datasets Sections 5.4.1 and 5.4.6.4
Participants	a) Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres. b) Describe eligibility criteria for participants.	Section 5.4 Section 5.4.1

Table G.4 Continued: TRIPOD reporting checklist for Chapter 5

Section/Topic	TRIPOD item	Location
	c) Give details of treatments received, if relevant.	Not applicable
Outcome	a) Clearly define the outcome that is predicted by the prediction model, including how and when assessed. b) Report any actions to blind assessment of the outcome to be predicted.	Section 5.4.4 Not applicable (retrospective data). Outcome was predefined in study protocol [180].
Predictors	a) Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured. b) Report any actions to blind assessment of predictors for the outcome and other predictors.	Section 5.4.5 Not applicable (retrospective data).
Sample size	Explain how the study size was arrived at.	Figure 5.1
Missing data	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Section 5.4.6.2
Statistical analysis methods	a) Describe how predictors were handled in the analyses. b) Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation. c) For validation, describe how the predictions were calculated. d) Specify all measures used to assess model performance and, if relevant, to compare multiple models. e) Describe any model updating (e.g., recalibration) arising from the validation, if done.	Section 5.4.6.2 Sections 5.4.6.3 and 5.4.6.4 Sections 5.4.6.3 and 5.4.6.4 Section 5.4.6.4 Section 5.4.6.4

Table G.4 Continued: TRIPOD reporting checklist for Chapter 5

Section/Topic	TRIPOD item	Location
Risk groups	Provide details on how risk groups were created, if done.	Not performed.
Development vs. validation	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	Section 5.4.6.4
Results		
Participants	a) Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. b) Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome. c) For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Figure 5.1 and Tables 5.2–5.4 Tables 5.2–5.4 and Figure 5.3 Figure 5.2 (for outcomes) and Appendix E
Model development	a) Specify the number of participants and outcome events in each analysis. b) If done, report the unadjusted association between each candidate predictor and outcome.	Section 5.5 and Figure 5.2 Tables 5.2–5.5
Model specification	a) Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point). b) Explain how to use the prediction model.	Appendix E Appendix E
Model performance	Report performance measures (with CIs) for the prediction model.	Tables 5.6–5.8 and Figures 5.4–5.7
Model-updating	If done, report the results from any model updating (i.e., model specification, model performance).	Figures 5.6–5.7

Table G.4 Continued: TRIPOD reporting checklist for Chapter 5

Section/Topic	TRIPOD item	Location
Discussion		
Limitations	Discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data).	Section 5.6.3
Interpretation	a) For validation, discuss the results with reference to performance in the development data, and any other validation data. b) Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	Sections 5.6 and 5.6.4 Sections 5.6.4 and 5.7
Implications	Discuss the potential clinical use of the model and implications for future research.	Section 5.7
Other information		
Supplementary information	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	End of chapter
Funding	Give the source of funding and the role of the funders for the present study.	Not applicable

Appendix H

Codelists

H.1 Primary care

Diagnoses in primary care were identified using Read codes. Read codes used in this thesis were based on those used by Gharbi *et al.* (2019).

H.1.1 Urinary tract infection

Table H.1: Read codes for lower UTI in primary care.

Read code	Description
1AG..00	Recurrent urinary tract infections
1J4..00	Suspected UTI
K15..00	Cystitis
K150.00	Acute cystitis
K15z.00	Cystitis NOS
K190.00	Urinary tract infection, site not specified
K190.11	Recurrent urinary tract infection
K190100	Pyuria, site not specified
K190200	Post operative urinary tract infection
K190300	Recurrent urinary tract infection
K190311	Recurrent UTI
K190400	Chronic urinary tract infection
K190500	Urinary tract infection
K190z00	Urinary tract infection, site not specified NOS
SP07700	Infect+inflam react due pros dev, implt+graft in urinary syst
SP07Q00	Catheter associated urinary tract infection
SP07Q11	CAUTI - catheter associated urinary tract infection
K15y.00	Other specified cystitis
K152z00	Other chronic cystitis NOS
K152.00	Other chronic cystitis
K152y00	Chronic cystitis unspecified
K155.00	Recurrent cystitis
K15yz00	Other cystitis NOS
1AZ6000	Mild lower urinary tract symptoms
7N51.00	[SO]Lower urinary tract

Table H.1 Continued: Read codes for lower UTI in primary care.

Read code	Description
1AZ6100	Moderate lower urinary tract symptoms
1AZ6.00	Lower urinary tract symptoms
Kyu5100	[X]Other cystitis
14D4.00	H/O: recurrent cystitis

Table H.2: Read codes for pyelonephritis in primary care.

Read code	Description
K100.00	Chronic pyelonephritis
K100000	Chronic pyelonephritis without medullary necrosis
K100100	Chronic pyelonephritis with medullary necrosis
K100400	Nonobstructive reflux-associated chronic pyelonephritis
K100500	Chronic obstructive pyelonephritis
K100600	Calculous pyelonephritis
K100z00	Chronic pyelonephritis NOS
K101.00	Acute pyelonephritis
K101000	Acute pyelonephritis without medullary necrosis
K101z00	Acute pyelonephritis NOS
K104.00	Xanthogranulomatous pyelonephritis
K10y.00	Pyelonephritis and pyonephrosis unspecified
K10y000	Pyelonephritis unspecified
K10y300	Pyelonephritis in diseases EC
K10yz00	Unspecified pyelonephritis NOS
K100200	Chronic pyelitis
K10y400	Pyelitis in diseases EC
K101200	Acute pyelitis
K10y100	Pyelitis unspecified
K102000	Renal abscess
K102.00	Renal and perinephric abscess
K102100	Perinephric abscess
K102z00	Renal and perinephric abscess NOS
K10..00	Infections of kidney
K10z.00	Infection of kidney NOS
K10..11	Renal infections
K10..00	Infections of kidney
K10z.00	Infection of kidney NOS
K21..11	Prostatitis and other inflammatory diseases of prostate
K210.00	Acute prostatitis
K211.00	Chronic prostatitis
K214.00	Prostatitis in diseases EC
K214z00	Prostatitis in diseases EC NOS
K21z.00	Prostatitis NOS
K213.00	Prostatocystitis
K212.00	Abscess of prostate
K10y200	Pyonephrosis unspecified

Table H.2 Continued: Read codes for pyelonephritis in primary care.

Read code	Description
K105.00	Chronic infective interstitial nephritis
A160200	Tuberculous pyelonephritis
A160100	Tuberculous pyelitis

Table H.3: Read codes for recurrent UTI in primary care.

Read code	Description
1AG..00	Recurrent urinary tract infections
K190.11	Recurrent urinary tract infection
K190300	Recurrent urinary tract infection
K190311	Recurrent UTI
K190400	Chronic urinary tract infection
K152z00	Other chronic cystitis NOS
K152.00	Other chronic cystitis
K152y00	Chronic cystitis unspecified
K155.00	Recurrent cystitis
14D4.00	H/O: recurrent cystitis
K100.00	Chronic pyelonephritis
K100000	Chronic pyelonephritis without medullary necrosis
K100100	Chronic pyelonephritis with medullary necrosis
K100400	Nonobstructive reflux-associated chronic pyelonephritis
K100500	Chronic obstructive pyelonephritis
K100600	Calculous pyelonephritis
K100z00	Chronic pyelonephritis NOS
K100200	Chronic pyelitis
K211.00	Chronic prostatitis
K105.00	Chronic infective interstitial nephritis

H.1.2 Bloodstream infection

Table H.4: Read codes for bloodstream infection in primary care.

Read code	Description
A38z.11	Sepsis
A38..00	Septicaemia
A3C..00	Sepsis
K190600	Urosepsis
A38z.00	Septicaemia NOS
A381.00	Staphylococcal septicaemia
R106.00	[D]Unspecified bacteraemia
A380.00	Streptococcal septicaemia
A382.00	Pneumococcal septicaemia

Table H.4 Continued: Read codes for bloodstream infection in primary care.

Read code	Description
A384200	Escherichia coli septicaemia
A384211	E.coli septicaemia
A384.00	Septicaemia due to other gram negative organisms
A3Cz.00	Sepsis NOS
A3Cy.00	Other specified sepsis
A381000	Septicaemia due to Staphylococcus aureus
A38y.00	Other specified septicaemias
A380100	Septicaemia due to streptococcus, group B
A384300	Pseudomonas septicaemia
A384000	Gram negative septicaemia NOS
A380300	Septicaemia due to streptococcus pneumoniae
A384100	Haemophilus influenzae septicaemia
A380400	Septicaemia due to enterococcus
A380000	Septicaemia due to streptococcus, group A
Ayu3J00	[X]Septicaemia, unspecified
AB2y300	Candidal septicaemia
A270100	Listeria septicaemia
A383.00	Septicaemia due to anaerobes
A3C3.00	Sepsis due to Gram negative bacteria
A381100	Septicaemia due to coagulase negative staphylococcus
A3C0100	Sepsis due to Streptococcus group B
A380500	Vancomycin resistant enterococcal septicaemia
A3C1.00	Sepsis due to Staphylococcus
A384400	Serratia septicaemia
A3C2.11	Sepsis due to anaerobes
AB2y500	Candidal sepsis
A3C1000	Sepsis due to Staphylococcus aureus
A3C0300	Sepsis due to Streptococcus pneumoniae
A3C0000	Sepsis due to Streptococcus group A
A3C0.00	Sepsis due to Streptococcus
A270611	Listerial sepsis
A384z00	Other gram negative septicaemia NOS
A3C0z00	Streptococcal sepsis, unspecified
A3C0y00	Other streptococcal sepsis
Ayu3F00	[X]Streptococcal septicaemia, unspecified
A396.00	Sepsis due to Actinomyces
A3C2.00	Sepsis due to anaerobic bacteria
Ayu3E00	[X]Other streptococcal septicaemia
A271100	Erysipelothrix septicaemia
Ayu3G00	[X]Septicaemia due to other gram-negative organisms
A3C3.11	Sepsis due to Gram negative organisms
AB2y511	Sepsis due to Candida
A3C3y00	Sepsis due to other Gram negative organisms
A270600	Sepsis due to Listeria monocytogenes
Ayu3H00	[X]Other specified septicaemia

H.1.3 Charlson Comorbidity Index

The list to calculate the Charlson Comorbidity Index in primary care contains more than 3,000 Read codes. Due to its volume, the list was not included here. A full list of all codes used in this analysis can be found in the online repository for Shallcross *et al.* (2020) [133].

H.2 Secondary care

Diagnoses in secondary care were identified using 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) codes (inpatient admissions) and Emergency Care Data Set (ECDS)/bespoke codes (emergency department visits). ICD-10 codes used in this thesis were partially based on those used by Gharbi *et al.* (2019).

H.2.1 Urinary tract infection

Table H.5: ICD-10 codes for lower UTI in hospital.

ICD-10 code	Description
N30.0	Acute cystitis
N30.9	Cystitis, unspecified
N30.8	Other cystitis
N39.0	Urinary tract infection, site not specified

Table H.6: ICD-10 codes for pyelonephritis in hospital.

ICD-10 code	Description
N10	Acute tubulo-interstitial nephritis
N12	Tubulo-interstitial nephritis, not specified as acute or chronic
N13.6	Pyonephrosis
N15.1	Renal and perinephric abscess
N15.8	Other specified renal tubulo-interstitial diseases
N15.9	Renal tubulo-interstitial disease, unspecified
N16.0	Renal tubulo-interstitial disorders in infectious and parasitic diseases classified elsewhere
N28.8	Other specified disorders of kidney and ureter
N34.0	Urethral abscess
N34.1	Nonspecific urethritis
N34.2	Other urethritis
N34.3	Urethral syndrome, unspecified

Table H.6 Continued: ICD-10 codes for pyelonephritis in hospital.

ICD-10 code	Description
N41.0	Acute prostatitis
N41.1	Chronic prostatitis
N41.2	Abscess of prostate
N41.3	Prostatocystitis
N41.8	Other inflammatory diseases of prostate
N41.9	Inflammatory disease of prostate, unspecified
N11.0	Nonobstructive reflux-associated chronic pyelonephritis
N11.1	Chronic obstructive pyelonephritis
N11.8	Other chronic tubulo-interstitial nephritis
N11.9	Chronic tubulo-interstitial nephritis, unspecified

Table H.7: ECDS / bespoke codes for suspected UTI in the emergency department.

ECDS / bespoke	Description
195	Pyelonephritis
203	Urinary Tract Infection / UTI
204	Urosepsis
1418111000	Pyelonephritis
1513111000	Urinary tract infection

H.2.2 Bloodstream infection

Table H.8: ICD-10 codes for bloodstream infection in hospital.

ICD-10 code	Description
R57.2	Septic shock
R65.1	Systemic Inflammatory Response Syndrome of infectious origin with organ failure
R65.0	Systemic Inflammatory Response Syndrome of infectious origin without organ failure
A40.0	Sepsis due to streptococcus, group A
A40.1	Sepsis due to streptococcus, group B
A40.2	Sepsis due to streptococcus, group D
A40.3	Sepsis due to Streptococcus pneumoniae
A40.8	Other streptococcal sepsis
A40.9	Streptococcal sepsis, unspecified
A41.0	Sepsis due to Staphylococcus aureus
A41.1	Sepsis due to other specified staphylococcus
A41.2	Sepsis due to unspecified staphylococcus
A41.3	Sepsis due to Haemophilus influenzae
A41.4	Sepsis due to anaerobes
A41.5	Sepsis due to other Gram-negative organisms

Table H.8 Continued: ICD-10 codes for bloodstream infection in hospital.

ICD-10 code	Description
A41.8	Other specified sepsis
A41.9	Sepsis, unspecified
R65.0	Severe sepsis without septic shock
R65.1	Severe sepsis with septic shock

H.2.3 Lower respiratory tract infection

Table H.9: ICD-10 codes for lower respiratory tract infection in hospital.

ICD-10 code	Description
J09	Influenza due to identified zoonotic or pandemic influenza virus
J10	Influenza due to identified seasonal influenza virus
J10.0	Influenza with pneumonia, seasonal influenza virus identified
J10.1	Influenza with other respiratory manifestations, seasonal influenza virus identified
J10.8	Influenza with other manifestations, seasonal influenza virus identified
J11	Influenza, virus not identified
J11.0	Influenza with pneumonia, virus not identified
J11.1	Influenza with other respiratory manifestations, virus not identified
J11.8	Influenza with other manifestations, virus not identified
J12	Viral pneumonia, not elsewhere classified
J12.0	Adenoviral pneumonia
J12.1	Respiratory syncytial virus pneumonia
J12.2	Parainfluenza virus pneumonia
J12.3	Human metapneumovirus pneumonia
J12.8	Other viral pneumonia
J12.9	Viral pneumonia, unspecified
J13	Pneumonia due to Streptococcus pneumoniae
J14	Pneumonia due to Haemophilus influenzae
J15	Bacterial pneumonia, not elsewhere classified
J15.0	Pneumonia due to Klebsiella pneumoniae
J15.1	Pneumonia due to Pseudomonas
J15.2	Pneumonia due to staphylococcus
J15.3	Pneumonia due to streptococcus, group B
J15.4	Pneumonia due to other streptococci
J15.5	Pneumonia due to Escherichia coli
J15.6	Pneumonia due to other Gram-negative bacteria
J15.7	Pneumonia due to Mycoplasma pneumoniae
J15.8	Other bacterial pneumonia
J15.9	Bacterial pneumonia, unspecified
J16	Pneumonia due to other infectious organisms, not elsewhere classified
J16.0	Chlamydial pneumonia
J16.8	Pneumonia due to other specified infectious organisms
J17	Pneumonia in diseases classified elsewhere

Table H.9 Continued: ICD-10 codes for lower respiratory tract infection in hospital.

ICD-10 code	Description
J17.0	Pneumonia in bacterial diseases classified elsewhere
J17.1	Pneumonia in viral diseases classified elsewhere
J17.2	Pneumonia in mycoses
J17.3	Pneumonia in parasitic diseases
J17.8	Pneumonia in other diseases classified elsewhere
J18	Pneumonia, organism unspecified
J18.0	Bronchopneumonia, unspecified
J18.1	Lobar pneumonia, unspecified
J18.2	Hypostatic pneumonia, unspecified
J18.8	Other pneumonia, organism unspecified
J18.9	Pneumonia, unspecified
J20	Acute bronchitis
J20.0	Acute bronchitis due to Mycoplasma pneumoniae
J20.1	Acute bronchitis due to Haemophilus influenzae
J20.2	Acute bronchitis due to streptococcus
J20.3	Acute bronchitis due to coxsackievirus
J20.4	Acute bronchitis due to parainfluenza virus
J20.5	Acute bronchitis due to respiratory syncytial virus
J20.6	Acute bronchitis due to rhinovirus
J20.7	Acute bronchitis due to echovirus
J20.8	Acute bronchitis due to other specified organisms
J20.9	Acute bronchitis, unspecified
J21	Acute bronchiolitis
J21.0	Acute bronchiolitis due to respiratory syncytial virus
J21.1	Acute bronchiolitis due to human metapneumovirus
J21.8	Acute bronchiolitis due to other specified organisms
J21.9	Acute bronchiolitis, unspecified
J22	Unspecified acute lower respiratory infection
J40	Bronchitis, not specified as acute or chronic
J41	Simple and mucopurulent chronic bronchitis
J41.0	Simple chronic bronchitis
J41.1	Mucopurulent chronic bronchitis
J41.8	Mixed simple and mucopurulent chronic bronchitis
J42	Unspecified chronic bronchitis
J43.0	MacLeod syndrome
J43.1	Panlobular emphysema
J43.2	Centrilobular emphysema
J43.8	Other emphysema
J43.9	Emphysema, unspecified
J44.0	Chronic obstructive pulmonary disease with acute lower respiratory infection
J44.1	Chronic obstructive pulmonary disease with acute exacerbation, unspecified
J44.8	Other specified chronic obstructive pulmonary disease
J44.9	Chronic obstructive pulmonary disease, unspecified
J85.1	Abscess of lung with pneumonia

H.2.4 Comorbidities

All lists for comorbidities only list 3-character codes and include all 4-character subcodes of the codes included.

Table H.10: ICD-10 codes for renal disease in hospital.

ICD-10 code	Description
N00	Acute nephritic syndrome
N01	Rapidly progressive nephritic syndrome
N02	Recurrent and persistent haematuria
N03	Chronic nephritic syndrome
N04	Nephrotic syndrome
N05	Unspecified nephritic syndrome
N06	Isolated proteinuria with specified morphological lesion
N07	Hereditary nephropathy, not elsewhere classified
N08	Glomerular disorders in diseases classified elsewhere
N13	Obstructive and reflux uropathy
N14	Drug- and heavy-metal-induced tubulo-interstitial and tubular conditions
N15	Other renal tubulo-interstitial diseases
N16	Renal tubulo-interstitial disorders in diseases classified elsewhere
N17	Acute renal failure
N18	Chronic kidney disease
N19	Unspecified kidney failure
N25	Disorders resulting from impaired renal tubular function
N26	Unspecified contracted kidney
N27	Small kidney of unknown cause
N28	Other disorders of kidney and ureter, not elsewhere classified
N29	Other disorders of kidney and ureter in diseases classified elsewhere

Table H.11: ICD-10 codes for urological disease in hospital.

ICD-10 code	Description
N20	Calculus of kidney and ureter
N21	Calculus of lower urinary tract
N22	Calculus of urinary tract in diseases classified elsewhere
N23	Unspecified renal colic
N30	Cystitis
N31	Neuromuscular dysfunction of bladder, not elsewhere classified
N32	Other disorders of bladder
N33	Bladder disorders in diseases classified elsewhere
N34	Urethritis and urethral syndrome
N35	Urethral stricture
N36	Other disorders of urethra
N37	Urethral disorders in diseases classified elsewhere
N39	Other disorders of urinary system
N40	Hyperplasia of prostate

Table H.11 Continued: ICD-10 codes for urological disease in hospital.

ICD-10 code	Description
N41	Inflammatory diseases of prostate
N42	Other disorders of prostate

Table H.12: ICD-10 codes for cancer in hospital.

ICD-10 code	Description
C00	Malignant neoplasm of lip
C01	Malignant neoplasm of base of tongue
C02	Malignant neoplasm of other and unspecified parts of tongue
C03	Malignant neoplasm of gum
C04	Malignant neoplasm of floor of mouth
C05	Malignant neoplasm of palate
C06	Malignant neoplasm of other and unspecified parts of mouth
C07	Malignant neoplasm of parotid gland
C08	Malignant neoplasm of other and unspecified major salivary glands
C09	Malignant neoplasm of tonsil
C10	Malignant neoplasm of oropharynx
C11	Malignant neoplasm of nasopharynx
C12	Malignant neoplasm of piriform sinus
C13	Malignant neoplasm of hypopharynx
C14	Malignant neoplasm of other and ill-defined sites in the lip, oral cavity and pharynx
C15	Malignant neoplasm of oesophagus
C16	Malignant neoplasm of stomach
C17	Malignant neoplasm of small intestine
C18	Malignant neoplasm of colon
C19	Malignant neoplasm of rectosigmoid junction
C20	Malignant neoplasm of rectum
C21	Malignant neoplasm of anus and anal canal
C22	Malignant neoplasm of liver and intrahepatic bile ducts
C23	Malignant neoplasm of gallbladder
C24	Malignant neoplasm of other and unspecified parts of biliary tract
C25	Malignant neoplasm of pancreas
C26	Malignant neoplasm of other and ill-defined digestive organs
C30	Malignant neoplasm of nasal cavity and middle ear
C31	Malignant neoplasm of accessory sinuses
C32	Malignant neoplasm of larynx
C33	Malignant neoplasm of trachea
C34	Malignant neoplasm of bronchus and lung
C37	Malignant neoplasm of thymus
C38	Malignant neoplasm of heart, mediastinum and pleura
C39	Malignant neoplasm of other and ill-defined sites in the respiratory system and intrathoracic organs
C40	Malignant neoplasm of bone and articular cartilage of limbs

Table H.12 Continued: ICD-10 codes for cancer in hospital.

ICD-10 code	Description
C41	Malignant neoplasm of bone and articular cartilage of other and unspecified sites
C43	Malignant melanoma of skin
C44	Other malignant neoplasms of skin
C45	Mesothelioma
C46	Kaposi sarcoma
C47	Malignant neoplasm of peripheral nerves and autonomic nervous system
C48	Malignant neoplasm of retroperitoneum and peritoneum
C49	Malignant neoplasm of other connective and soft tissue
C50	Malignant neoplasm of breast
C51	Malignant neoplasm of vulva
C52	Malignant neoplasm of vagina
C53	Malignant neoplasm of cervix uteri
C54	Malignant neoplasm of corpus uteri
C55	Malignant neoplasm of uterus, part unspecified
C56	Malignant neoplasm of ovary
C57	Malignant neoplasm of other and unspecified female genital organs
C58	Malignant neoplasm of placenta
C60	Malignant neoplasm of penis
C61	Malignant neoplasm of prostate
C62	Malignant neoplasm of testis
C63	Malignant neoplasm of other and unspecified male genital organs
C64	Malignant neoplasm of kidney, except renal pelvis
C65	Malignant neoplasm of renal pelvis
C66	Malignant neoplasm of ureter
C67	Malignant neoplasm of bladder
C68	Malignant neoplasm of other and unspecified urinary organs
C69	Malignant neoplasm of eye and adnexa
C70	Malignant neoplasm of meninges
C71	Malignant neoplasm of brain
C72	Malignant neoplasm of spinal cord, cranial nerves and other parts of central nervous system
C73	Malignant neoplasm of thyroid gland
C74	Malignant neoplasm of adrenal gland
C75	Malignant neoplasm of other endocrine glands and related structures
C76	Malignant neoplasm of other and ill-defined sites
C77	Secondary and unspecified malignant neoplasm of lymph nodes
C78	Secondary malignant neoplasm of respiratory and digestive organs
C79	Secondary malignant neoplasm of other and unspecified sites
C80	Malignant neoplasm without specification of site
C81	Hodgkin lymphoma
C82	Follicular lymphoma
C83	Non-follicular lymphoma
C84	Mature T/NK-cell lymphomas
C85	Other and unspecified types of non-Hodgkin lymphoma
C86	Other specified types of T/NK-cell lymphoma

Table H.12 Continued: ICD-10 codes for cancer in hospital.

ICD-10 code	Description
C88	Malignant immunoproliferative diseases
C90	Multiple myeloma and malignant plasma cell neoplasms
C91	Lymphoid leukaemia
C92	Myeloid leukaemia
C93	Monocytic leukaemia
C94	Other leukaemias of specified cell type
C95	Leukaemia of unspecified cell type
C96	Other and unspecified malignant neoplasms of lymphoid, haematopoietic and related tissue
C97	Malignant neoplasms of independent (primary) multiple sites

Table H.13: ICD-10 codes for immunosuppression in hospital.

ICD-10 code	Description
D80	Immunodeficiency with predominantly antibody defects
D81	Combined immunodeficiencies
D82	Immunodeficiency associated with other major defects
D83	Common variable immunodeficiency
D84	Other immunodeficiencies
D86	Sarcoidosis
D89	Other disorders involving the immune mechanism, not elsewhere classified

H.2.5 Pregnancy

Pregnancy was defined as any code of ICD-10 Chapter XV: Pregnancy, childbirth and the puerperium (O00-O99). Pregnancy tests were identified via the laboratory system's bespoke labels.

H.2.6 Charlson Comorbidity Index

As listed in Table 1 of Quan, Hude, Vijaya Sundararajan, Patricia Halfon, Andrew Fong, Bernard Burnand, Jean-Christophe Luthi, L. Duncan Saunders, Cynthia A. Beck, Thomas E. Feasby, and William A. Ghali. 2005. "Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data." Medical Care 43 (11): 1130–39.

Appendix I

Colophon

This document was created using LaTeX and BibLaTeX and edited on the Overleaf platform (https://www.overleaf.com/). The thesis layout is based on the *UCL Thesis LaTeX Template* written by Ian Kirker (© 2014). The document was set in Times Roman typeface. Tables were created using the *booktabs*, *tabularx*, and *longtable* packages. Images and plots were included using the *graphicx* package. The raw LaTeX code used to create this thesis can be found at https://github.com/prockenschaub/phd_thesis.

Bibliography

1. Stamm WE, Norrby SR. Urinary tract infections: disease panorama and challenges. *J. Infect. Dis.* 2001;183 Suppl 1:S1–4.
2. Foxman B. Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. *Dis. Mon.* 2003;49:53–70.
3. Butler CC, Hawking MKD, Quigley A, McNulty CAM. Incidence, severity, help seeking, and management of uncomplicated urinary tract infection: a population-based survey. *Br. J. Gen. Pract.* 2015;65:e702–7.
4. François M, Hanslik T, Dervaux B et al. The economic burden of urinary tract infections in women visiting general practices in France: a cross-sectional survey. *BMC Health Serv. Res.* 2016;16.
5. Mulder M, Baan E, Verbon A, Stricker B, Verhamme K. Trends of prescribing antimicrobial drugs for urinary tract infections in primary care in the Netherlands: a population-based cohort study. *BMJ Open* 2019;9:e027221.
6. Dolk FCK, Pouwels KB, Smith DRM, Robotham JV, Smieszek T. Antibiotics in primary care in England: which antibiotics are prescribed and for which conditions? *J. Antimicrob. Chemother.* 2018;73:ii2–ii10.
7. NHS Digital. Hospital Accident & Emergency Activity 2018-19. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident--emergency-activity/2018-19. Accessed: 2020-7-6.
8. Simmering JE, Tang F, Cavanaugh JE, Polgreen LA, Polgreen PM. The Increase in Hospitalizations for Urinary Tract Infections and the Associated Costs in the United States, 1998-2011. *Open Forum Infect Dis* 2017;4.

9. Ahmed H, Farewell D, Jones HM et al. Incidence and antibiotic prescribing for clinically diagnosed urinary tract infection in older adults in UK primary care, 2004-2014. *PLoS One* 2018;13:e0190521.

10. Foxman B, Barlow R, D'Arcy H, Gillespie B, Sobel JD. Urinary tract infection: self-reported incidence and associated costs. *Ann. Epidemiol.* 2000;10:509–515.

11. Bonkat G, Bartoletti RR, Bruyère F et al. Urological Infections. Tech. rep. 2019.

12. Rowe TA, Juthani-Mehta M. Urinary tract infection in older adults. *Aging health* 2013;9.

13. National Institute for Health and Care Excellence. Urinary tract infection (lower): antimicrobial prescribing [NG109]. https://www.nice.org.uk/guidance/ng109. Accessed: 2020-1-10.

14. Farrell DJ, Morrissey I, De Rubeis D, Robbins M, Felmingham D. A UK multicentre study of the antimicrobial susceptibility of bacterial pathogens causing urinary tract infection. *J. Infect.* 2003;46:94–100.

15. Naber KG, Schito G, Botto H, Palou J, Mazzei T. Surveillance study in Europe and Brazil on clinical aspects and Antimicrobial Resistance Epidemiology in Females with Cystitis (ARESC): implications for empiric therapy. *Eur. Urol.* 2008;54:1164–1175.

16. National Institute for Health and Care Excellence. Pyelonephritis (acute): antimicrobial prescribing [NG111]. https://www.nice.org.uk/guidance/ng111/chapter/Recommendations. Accessed: 2020-3-6.

17. Schmiemann G, Kniehl E, Gebhardt K, Matejczyk MM, Hummers-Pradier E. The diagnosis of urinary tract infection: a systematic review. *Dtsch. Arztebl. Int.* 2010;107:361–367.

18. Tsiga E, Panagopoulou E, Sevdalis N, Montgomery A, Benos A. The influence of time pressure on adherence to guidelines in primary care: an experimental study. *BMJ Open* 2013;3:e002700.

19. Foxman B, Ki M, Brown P. Antibiotic Resistance and Pyelonephritis. *Clin. Infect. Dis.* 2007;45:281–283.

20. Murray P, Traynor P, Hopson D. Evaluation of microbiological processing of urine specimens: comparison of overnight versus two-day incubation. *J. Clin. Microbiol.* 1992;30:1600–1601.

21. Saukko PM, Oppenheim BA, Cooper M, Rousham EK. Gaps in communication between different staff groups and older adult patients foster unnecessary antibiotic prescribing for urinary tract infections in hospitals: a qualitative translation approach. *Antimicrob. Resist. Infect. Control* 2019;8:130.

22. O'Kelly K, Phelps K, Regen EL et al. Why are we misdiagnosing urinary tract infection in older patients? A qualitative inquiry and roadmap for staff behaviour change in the emergency department. *Eur. Geriatr. Med.* 2019;10:585–593.

23. McNulty C, Joseph A, Cooper E, Jones L. Diagnosis of urinary tract infections: Quick reference tool for primary care for consultation and local adaption. Tech. rep. Public Health England, 2020.

24. Shallcross LJ, Rockenschaub P, McNulty D et al. Diagnostic uncertainty and urinary tract infection in the emergency department: a cohort study from a UK hospital. *BMC Emerg. Med.* 2020;20:40.

25. Rousham E, Cooper M, Petherick E, Saukko P, Oppenheim B. Overprescribing antibiotics for asymptomatic bacteriuria in older adults: a case series review of admissions in two UK hospitals. *Antimicrob. Resist. Infect. Control* 2019;8:71.

26. Hustey FM, Meldon SW, Smith MD, Lex CK. The effect of mental status screening on the care of elderly emergency department patients. *Ann. Emerg. Med.* 2003;41:678–684.

27. Little P, Turner S, Rumsby K et al. Developing clinical rules to predict urinary tract infection in primary care settings: sensitivity and specificity

of near patient tests (dipsticks) and clinical scores. *Br. J. Gen. Pract.* 2006;56:606–612.

28. Colgan R, Williams M, Johnson JR. Diagnosis and treatment of acute pyelonephritis in women. *Am. Fam. Physician* 2011;84:519–526.

29. Pietrow PK, Karellas ME. Medical management of common urinary calculi. *Am. Fam. Physician* 2006;74:86–94.

30. Wagenlehner FM, Naber KG, Weidner W. Asymptomatic bacteriuria in elderly patients: significance and implications for treatment. *Drugs Aging* 2005;22:801–807.

31. Nicolle LE, Bradley S, Colgan R et al. Infectious Diseases Society of America guidelines for the diagnosis and treatment of asymptomatic bacteriuria in adults. *Clin. Infect. Dis.* 2005;40:643–654.

32. Mayne S, Bowden A, Sundvall PD, Gunnarsson R. The scientific evidence for a potential link between confusion and urinary tract infection in the elderly is still confusing - a systematic literature review. *BMC Geriatr.* 2019;19:32.

33. Bengtsson C, Bengtsson U, Björkelund C, Lincoln K, Sigurdsson JA. Bacteriuria in a population sample of women: 24-year follow-up study. Results from the prospective population-based study of women in Gothenburg, Sweden. *Scand. J. Urol. Nephrol.* 1998;32:284–289.

34. Boscia JA, Kobasa WD, Knight RA et al. Epidemiology of bacteriuria in an elderly ambulatory population. *Am. J. Med.* 1986;80:208–214.

35. Baldassarre JS, Kaye D. Special problems of urinary tract infection in the elderly. *Med. Clin. North Am.* 1991;75:375–390.

36. Mody L, Juthani-Mehta M. Urinary tract infections in older women: a clinical review. *JAMA* 2014;311:844–854.

37. Coulthard MG, Kalra M, Lambert HJ et al. Redefining urinary tract infections by bacterial colony counts. *Pediatrics* 2010;125:335–341.

38. Stamm WE, Counts GW, Running KR et al. Diagnosis of coliform infection in acutely dysuric women. *N. Engl. J. Med.* 1982;307:463–468.

39. National Institute for Health and Care Excellence. Urinary tract infection (lower) - women. https://cks.nice.org.uk/urinary-tract-infection-lower-women. Accessed: 2020-1-13.

40. Hindman R, Tronic B, Bartlett R. Effect of delay on culture of urine. *J. Clin. Microbiol.* 1976;4:102–103.

41. LaRocco MT, Franek J, Leibach EK et al. Effectiveness of Preanalytic Practices on Contamination and Diagnostic Accuracy of Urine Cultures: a Laboratory Medicine Best Practices Systematic Review and Meta-analysis. *Clin. Microbiol. Rev.* 2016;29:105–147.

42. Valenstein P, Meier F. Urine culture contamination: a College of American Pathologists Q-Probes study of contaminated urine cultures in 906 institutions. *Arch. Pathol. Lab. Med.* 1998;122:123–129.

43. Public Health England. Standards for microbiology investigations (UK SMI). https://www.gov.uk/government/collections/standards-for-microbiology-investigations-smi. Accessed: 2019-7-29. 2014.

44. Burton RJ, Albur M, Eberl M, Cuff SM. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC Med. Inform. Decis. Mak.* 2019;19:171.

45. Urinalysis. In: *Concise Medical Dictionary*. 8th ed. Oxford University Press, 2010.

46. Kouri T, Fogazzi G, Gant V et al. European Urinalysis Guidelines. *Scand. J. Clin. Lab. Invest.* 2000;60:1–96.

47. Young JL, Soper DE. Urinalysis and urinary tract infection: update for clinicians. *Infect. Dis. Obstet. Gynecol.* 2001;9:249–255.

48. McKinnon KM. Flow Cytometry: An Overview. *Curr. Protoc. Immunol.* 2018;120:5.1.1–5.1.11.

49. Giesen CD, Greeno AM, Thompson KA et al. Performance of flow cytometry to screen urine for bacteria and white blood cells prior to urine culture. *Clin. Biochem.* 2013;46:810–813.

50. Boonen KJM, Koldewijn EL, Arents NLA, Raaymakers PAM, Scharnhorst V. Urine flow cytometry as a primary screening method to exclude urinary tract infections. *World J. Urol.* 2013;31:547–551.

52. Simerville JA, Maxted WC, Pahira JJ. Urinalysis: a comprehensive review. *Am. Fam. Physician* 2005;71:1153–1162.

53. Tiso M, Schechter AN. Nitrate reduction to nitrite, nitric oxide and ammonia by gut bacteria under physiological conditions. *PLoS One* 2015;10:e0119712.

54. Commissioning for Quality and Innovation (CQUIN). CCG indicator specifications for 2020-2021. Tech. rep. NHS England, 2020.

55. Felt JR, Yurkovich C, Garshott DM et al. The Utility of Real-Time Quantitative Polymerase Chain Reaction Genotype Detection in the Diagnosis of Urinary Tract Infections in Children. *Clin. Pediatr.* 2017;56:912–919.

56. Heytens S, De Sutter A, Coorevits L et al. Women with symptoms of a urinary tract infection but a negative urine culture: PCR-based quantification of Escherichia coli suggests infection in most cases. *Clin. Microbiol. Infect.* 2017;23:647–652.

57. Knight GM, Dyakova E, Mookerjee S et al. Fast and expensive (PCR) or cheap and slow (culture)? A mathematical modelling study to explore screening for carbapenem resistance in UK hospitals. *BMC Med.* 2018;16:141.

58. Padmavathy B, Vinoth Kumar R, Patel A et al. Rapid and sensitive detection of major uropathogens in a single-pot multiplex PCR assay. *Curr. Microbiol.* 2012;65:44–53.

59. Wagenlehner FM, Naber KG. Understanding clinical variables to improve empirical antibiotic therapy for UTI. *Nat. Rev. Urol.* 2019;16:695–696.

60. Melander RJ, Zurawski DV, Melander C. Narrow-Spectrum Antibacterial Agents. *Medchemcomm* 2018;9:12–21.

61. Howard B, Furman B. Nitrofurantoin. In: *Reference Module in Biomedical Sciences*. Elsevier, 2018.

62. Krockow EM, Colman AM, Chattoe-Brown E et al. Balancing the risks to individual and society: a systematic review and synthesis of qualitative research on antibiotic prescribing behaviour in hospitals. *J. Hosp. Infect.* 2019;101:428–439.

63. Knottnerus BJ, Geerlings SE, Moll van Charante EP, Riet G ter. Women with symptoms of uncomplicated urinary tract infection are often willing to delay antibiotic treatment: a prospective cohort study. *BMC Fam. Pract.* 2013;14:71.

64. Ryves R, Eyles C, Moore M et al. Understanding the delayed prescribing of antibiotics for respiratory tract infection in primary care: a qualitative analysis. *BMJ Open* 2016;6:e011882.

65. Pujades-Rodriguez M, West RM, Wilcox MH, Sandoe J. Lower Urinary Tract Infections: Management, Outcomes and Risk Factors for Antibiotic Re-prescription in Primary Care. *EClinicalMedicine* 2019;14:23–31.

66. Pouwels KB, Hopkins S, Llewelyn MJ et al. Duration of antibiotic treatment for common infections in English primary care: cross sectional analysis and comparison with guidelines. *BMJ* 2019;364:l440.

67. Imison C, Castle-Clarke S, Watson R, Edwards N. Delivering the benefits of digital health care. Tech. rep. The Nuffield Trust, 2016.

68. Smith DRM, Dolk FCK, Pouwels KB et al. Defining the appropriateness and inappropriateness of antibiotic prescribing in primary care. *J. Antimicrob. Chemother.* 2018;73:ii11–ii18.

69. Falagas ME, Kotsantis IK, Vouloumanou EK, Rafailidis PI. Antibiotics versus placebo in the treatment of women with uncomplicated cystitis: a meta-analysis of randomized controlled trials. *J. Infect.* 2009;58:91–102.

70. Bleidorn J, Gágyor I, Kochen MM, Wegscheider K, Hummers-Pradier E. Symptomatic treatment (ibuprofen) or antibiotics (ciprofloxacin) for uncomplicated urinary tract infection?–results of a randomized controlled pilot trial. *BMC Med.* 2010;8:30.

71. Gágyor I, Bleidorn J, Kochen MM et al. Ibuprofen versus fosfomycin for uncomplicated urinary tract infection in women: randomised controlled trial. *BMJ* 2015;351:h6544.

72. Rapid responses: Gharbi et al. (2019) Antibiotic management of urinary tract infection in elderly patients in primary care and its association with bloodstream infections and all cause mortality: population based cohort study. https://www.bmj.com/content/364/bmj.l525/rapid-responses. Accessed: 2019-6-12.

73. Tomas ME, Getman D, Donskey CJ, Hecker MT. Overdiagnosis of Urinary Tract Infection and Underdiagnosis of Sexually Transmitted Infection in Adult Women Presenting to an Emergency Department. *J. Clin. Microbiol.* 2015;53:2686–2692.

74. Goossens H, Ferech M, Vander Stichele R, Elseviers M, ESAC Project Group. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *Lancet* 2005;365:579–587.

75. Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. *BMJ* 2010;340:c2096.

76. Pouwels KB, Dolk FCK, Smith DRM, Robotham JV, Smieszek T. Actual versus 'ideal' antibiotic prescribing for common conditions in English primary care. *J. Antimicrob. Chemother.* 2018;73:19–26.

77. Yelin I, Snitser O, Novich G et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. *Nat. Med.* 2019;25:1143–1152.

78. Antibiotics: Side effects. https://www.nhs.uk/conditions/antibiotics/side-effects/. Accessed: 2020-10-20.

79. Blumenthal KG, Peter JG, Trubiano JA, Phillips EJ. Antibiotic allergy. *Lancet* 2019;393:183–198.

80. Knoop FC, Owens M, Crocker IC. Clostridium difficile: clinical disease and diagnosis. *Clin. Microbiol. Rev.* 1993;6:251–265.

81. Bignardi GE. Risk factors for Clostridium difficile infection. *J. Hosp. Infect.* 1998;40:1–15.

82. Bosch CMA van den, Hulscher MEJL, Akkermans RP et al. Appropriate antibiotic use reduces length of hospital stay. *J. Antimicrob. Chemother.* 2017;72:923–932.

83. Staa TP van, Palin V, Li Y et al. The effectiveness of frequent antibiotic use in reducing the risk of infection-related hospital admissions: results from two large population-based cohorts. *BMC Med.* 2020;18:40.

84. Vogel T, Verreault R, Gourdeau M et al. Optimal duration of antibiotic therapy for uncomplicated urinary tract infection in older women: a double-blind randomized controlled trial. *CMAJ* 2004;170:469–473.

85. Huttner A, Verhaegh EM, Harbarth S et al. Nitrofurantoin revisited: a systematic review and meta-analysis of controlled trials. *J. Antimicrob. Chemother.* 2015;70:2456–2464.

86. Carey MR, Vaughn VM, Mann J et al. Is Non-Steroidal Anti-Inflammatory Therapy Non-Inferior to Antibiotic Therapy in Uncomplicated Urinary Tract Infections: a Systematic Review. *J. Gen. Intern. Med.* 2020;35:1821–1829.

87. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. *Trials* 2009;10:37.

88. Wojno KJ, Baunoch D, Luke N et al. Multiplex PCR Based Urinary Tract Infection (UTI) Analysis Compared to Traditional Urine Culture in Identifying Significant Pathogens in Symptomatic Patients. *Urology* 2020;136:119–126.

89. Wigton RS, Hoellerich VL, Ornato JP et al. Use of clinical findings in the diagnosis of urinary tract infection in women. *Arch. Intern. Med.* 1985;145:2222–2227.

90. Little P, Turner S, Rumsby K et al. Validating the prediction of lower urinary tract infection in primary care: sensitivity and specificity of urinary dipsticks and clinical scores in women. *Br. J. Gen. Pract.* 2010;60:495–500.

91. Munn Z, Peters MDJ, Stern C et al. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med. Res. Methodol.* 2018;18:143.

92. Haeusler GM, Thursky KA. Electronic health record data for antimicrobial prescribing. *Lancet Infect. Dis.* 2020.

93. Hansell A, Hollowell J, Nichols T, McNiece R, Strachan D. Use of the General Practice Research Database (GPRD) for respiratory epidemiology: a comparison with the 4th Morbidity Survey in General Practice (MSGP4). *Thorax* 1999;54:413–419.

94. Majeed A, Moser K. Age- and sex-specific antibiotic prescribing patterns in general practice in England and Wales in 1996. *Br. J. Gen. Pract.* 1999;49:735–736.

95. Nichols GA, Brodovicz KG, Kimes TM, Déruaz-Luyet A, Bartels DB. Prevalence and incidence of urinary tract and genital infections among patients with and without type 2 diabetes. *J. Diabetes Complications* 2017;31:1587–1591.

96. Hawker JI, Smith S, Smith GE et al. Trends in antibiotic prescribing in primary care for clinical syndromes subject to national recommendations to reduce antibiotic resistance, UK 1995-2011: analysis of a large database of primary care consultations. *J. Antimicrob. Chemother.* 2014;69:3423–3430.

97. Public Health England. English surveillance programme for antimicrobial utilisation and resistance (ESPAUR) report 2018-2019. Tech. rep. Public Health England, 2019.

98. National Health Service. The NHS Long Term Plan. Tech. rep. 2019.

99. Castle-Clarke S, Hutchings R. Achieving a digital NHS. Tech. rep. Nuffield Trust, 2019.

100. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 2019;17:195.

101. Peters MDJ, Godfrey C, McInerney P et al. JBI Manual for Evidence Synthesis. Ed. by Aromataris E, Munn Z. 2020. Chap. Scoping Reviews.

102. Gunter TD, Terry NP. The emergence of national electronic health record architectures in the United States and Australia: models, costs, and questions. *J. Med. Internet Res.* 2005;7:e3.

103. Busse JW, Guyatt GH. Tool to Assess Risk of Bias in Cohort Studies. Evidence Partners Inc.

104. Wolff RF, Moons KGM, Riley RD et al. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* 2019;170:51–58.

105. Tricco AC, Lillie E, Zarin W et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* 2018;169:467–473.

106. Drekonja DM, Rector TS, Cutting A, Johnson JR. Urinary tract infection in male veterans: treatment patterns and outcomes. *JAMA Intern. Med.* 2013;173:62–68.

107. Grigoryan L, Zoorob R, Wang H et al. Less workup, longer treatment, but no clinical benefit observed in women with diabetes and acute cystitis. *Diabetes Res. Clin. Pract.* 2017;129:197–202.

108. Ahmed H, Farewell D, Francis NA, Paranjothy S, Butler CC. Risk of adverse outcomes following urinary tract infection in older people with renal impairment: Retrospective cohort study using linked health record data. *PLoS Med.* 2018;15:e1002652.

109. Taylor RA, Moore CL, Cheung KH, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. *PLoS One* 2018;13:e0194085.

110. Ahmed H., Farewell D., Francis N.A., Paranjothy S., Butler C.C. Choice of empirical antibiotic therapy and adverse outcomes in older adults with suspected urinary tract infection: Cohort study. *Open Forum Infect. Dis.* 2019;6.

111. Ahmed H., Farewell D., Francis N.A., Paranjothy S., Butler C.C. Impact of antibiotic treatment duration on outcomes in older men with suspected urinary tract infection: Retrospective cohort study. *Pharmacoepidemiol. Drug Saf.* 2019.

112. Gharbi M, Drysdale JH, Lishman H et al. Antibiotic management of urinary tract infection in elderly patients in primary care and its association with bloodstream infections and all cause mortality: population based cohort study. *BMJ* 2019;364:l525.

113. Mistry C., Palin V., Li Y. et al. Development and validation of a multivariable prediction model for infection-related complications in patients with common infections in UK primary care and the extent of risk-based prescribing of antibiotics. *BMC Med.* 2020;18:118.

114. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* 2014;6:10.

115. Kyriacou DN, Lewis RJ. Confounding by Indication in Clinical Research. *JAMA* 2016;316:1818–1819.

116. Roland M, Guthrie B, Thomé DC. Primary medical care in the United kingdom. *J. Am. Board Fam. Med.* 2012;25 Suppl 1:S6–11.

117. Grosios K, Gahan PB, Burbidge J. Overview of healthcare in the UK. *EPMA J.* 2010;1:529–534.

118. Powell T. Briefing Paper: The structure of the NHS in England. Tech. rep. CBP 07206. House of Commons, 2020.

119. Adlington K, Finn R, Ghafur S, Smith CR, On behalf of the National Medical Director's Clinical Fellows 2014-15. Comissioning. What's the big deal. Tech. rep. Faculty of Medical Leadership and Management, 2015.

120. 2019-20 Annual Report and Accounts. Tech. rep. NHS Digital, 2020.

121. Herrett E, Gallagher AM, Bhaskaran K et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int. J. Epidemiol.* 2015;44:827–836.

122. Wolf A, Dedman D, Campbell J et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int. J. Epidemiol.* 2019;48:1740–1740g.

123. Mathur R, Bhaskaran K, Chaturvedi N et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J. Public Health* 2014;36:684–692.

124. Bhaskaran K, Forbes HJ, Douglas I, Leon DA, Smeeth L. Representativeness and optimal use of body mass index (BMI) in the UK Clinical Practice Research Datalink (CPRD). *BMJ Open* 2013;3:e003389.

125. Padmanabhan S. CPRD GOLD Data Specification. Tech. rep. CPRD, 2017.

126. Gill B. The English Indices of Deprivation 2015. Tech. rep. Department for Communities and Local Government, 2015.

127. Price SJ, Stapley SA, Shephard E, Barraclough K, Hamilton WT. Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case-control study. *BMJ Open* 2016;6:e011664.

128. Chisholm J. The Read clinical classification. *BMJ* 1990;300:1092.

129. Denaxas SC, George J, Herrett E et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int. J. Epidemiol.* 2012;41:1625–1638.

130. National Information Board. Personalised Health and Care 2020. Tech. rep. National Health Service, 2014.

131. Robinson D, Schulz E, Brown P, Price C. Updating the Read Codes: user-interactive maintenance of a dynamic clinical vocabulary. *J. Am. Med. Inform. Assoc.* 1997;4:465–472.

132. Anderson RN, Miniño AM, Hoyert DL, Rosenberg HM. Comparability of cause of death between ICD-9 and ICD-10: preliminary estimates. *Natl. Vital Stat. Rep.* 2001;49:1–32.

133. Shallcross L, Rockenschaub P, Blackburn R et al. Antibiotic prescribing for lower UTI in elderly patients in primary care and risk of bloodstream infection: A cohort study using electronic health records in England. *PLoS Med.* 2020;17:e1003336.

134. Herrett E, Shah AD, Boggon R et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ* 2013;346:f2350.

135. Gallagher AM, Dedman D, Padmanabhan S, Leufkens HGM, Vries F de. The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations. *Pharmacoepidemiol. Drug Saf.* 2019;28:563–569.

136. Rockenschaub P, Ansell D, Shallcross L. Linking individual-level data on diagnoses and dispensing for research on antibiotic use: Evaluation of a novel data source from English secondary care. *Pharmacoepidemiol. Drug Saf.* 2018;27:206–212.

137. Dregan A, Moller H, Murray-Thomas T, Gulliford MC. Validity of cancer diagnosis in a primary care database compared with linked cancer registrations in England. Population-based cohort study. *Cancer Epidemiol.* 2012;36:425–429.

138. Gulliford MC, Sun X, Anjuman T, Yelland E, Murray-Thomas T. Comparison of antibiotic prescribing records in two UK primary care electronic health record systems: cohort study using CPRD GOLD and CPRD Aurum databases. *BMJ Open* 2020;10:e038767.

139. Smieszek T, Pouwels KB, Dolk FCK et al. Potential for reducing inappropriate antibiotic prescribing in English primary care. *J. Antimicrob. Chemother.* 2018;73:ii36–ii43.

140. Bottle A, Cohen C, Lucas A et al. How an electronic health record became a real-world research resource: comparison between London's Whole Systems Integrated Care database and the Clinical Practice Research Datalink. *BMC Med. Inform. Decis. Mak.* 2020;20:71.

141. Lewis JD, Schinnar R, Bilker WB, Wang X, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidemiol. Drug Saf.* 2007;16:393–401.

142. Bradley SH, Lawrence NR, Carder P. Using primary care data for health research in England - an overview. *Future Healthc J* 2018;5:207–212.

143. Clegg A, Bates C, Young J et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing* 2016;45:353–360.

144. Lewer D, Bourne T, George A et al. Data Resource: the Kent Integrated Dataset (KID). *Int J Popul Data Sci* 2018;3:427.

145. Public Health England. Fingertips. An Introduction to PHE's Data Visualisation Platform. Tech. rep. 2020.

146. EBM DataLab. OpenPrescribing. https://openprescribing.net/. Accessed: 2020-10-6.

147. Croker R, Walker AJ, Goldacre B. Why did some practices not implement new antibiotic prescribing guidelines on urinary tract infection? A cohort study and survey in NHS England primary care. *J. Antimicrob. Chemother.* 2019;74:1125–1132.

148. Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data Resource Profile: Hospital Episode Statistics Admitted Patient Care (HES APC). *Int. J. Epidemiol.* 2017;46:1093–1093i.

149. NHS Digital. HES Data Dictionary. Outpatients. Tech. rep. 2018.

150. NHS Digital. HES Data Dictionary. Accident and Emergency. Tech. rep. 2018.

151. Department of Health and Social Care. New plans to expand the use of digital technology across the NHS. https://www.gov.uk/government/news/new-plans-to-expand-the-use-of-digital-technology-across-the-nhs. Accessed: 2020-10-21. 2016.

152. Freemantle N, Ray D, Falcaro M et al. BMI upon discharge from hospital and its relationship with survival: an observational study utilising linked patient records. *J. R. Soc. Med.* 2016;109:230–238.

153. NHS Digital. Emergency Care Data Set. https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dcb0092-2062-commissioning-data-sets-emergency-care-data-set. Accessed: 2020-10-21. 2020.

154. Baird B, Charles A, Honeyman M, Maguire D, Das P. Understanding pressures in general practice. Tech. rep. The King's Fund, 2016.

155. Little P, Moore MV, Turner S et al. Effectiveness of five different approaches in management of urinary tract infection: randomised controlled trial. *BMJ* 2010;340:c199.

156. Kronenberg A, Bütikofer L, Odutayo A et al. Symptomatic treatment of uncomplicated lower urinary tract infections in the ambulatory setting: randomised, double blind trial. *BMJ* 2017;359:j4784.

157. Vik I, Bollestad M, Grude N et al. Ibuprofen versus pivmecillinam for uncomplicated urinary tract infection in women-A double-blind, randomized non-inferiority trial. *PLoS Med.* 2018;15:e1002569.

158. Spurling GK, Del Mar CB, Dooley L, Foxlee R, Farley R. Delayed antibiotic prescriptions for respiratory infections. *Cochrane Database Syst. Rev.* 2017;9:CD004417.

159. R Core Team. R: A language and environment for statistical computing. Vienna, Austria, 2018.

160. Shallcross L, Lorencatto F, Fuller C et al. An interdisciplinary mixed-methods approach to developing antimicrobial stewardship interventions: Protocol for the Preserving Antibiotics through Safe Stewardship (PASS) Research Programme. *Wellcome Open Res* 2020;5:8.

161. Shallcross L, Beckley N, Rait G, Hayward A, Petersen I. Antibiotic prescribing frequency amongst patients in primary care: a cohort study using electronic health records. *J. Antimicrob. Chemother.* 2017;72:1818–1824.

162. Quan H, Li B, Couris CM et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am. J. Epidemiol.* 2011;173:676–682.

163. Hernan MA, Robins JM. Causal Inference: What If. CRC Press, 2020.

164. Bender R, Kuss O, Hildebrandt M, Gehrmann U. Estimating adjusted NNT measures in logistic regression analysis. *Stat. Med.* 2007;26:5586–5595.

165. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001;57:120–125.

166. Iacus SM, King G, Porro G. Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *J. Am. Stat. Assoc.* 2011;106:345–361.

167. Lee J, Little TD. A practical guide to propensity score analysis for applied clinical research. *Behav. Res. Ther.* 2017;98:76–90.

168. King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. *Polit. Anal.* 2019;27:435–454.

169. Elm E von, Altman DG, Egger M et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann. Intern. Med.* 2007;147:573–577.

170. Benchimol EI, Smeeth L, Guttmann A et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med.* 2015;12:e1001885.

171. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, 2006.

172. Agresti A. Categorical Data Analysis. Somerset, UNITED STATES: John Wiley & Sons, Incorporated, 2012.

173. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.* 2013;20:144–151.

174. Ferry SA, Holm SE, Stenlund H, Lundholm R, Monsen TJ. Clinical and bacteriological outcome of different doses and duration of pivmecillinam compared with placebo therapy of uncomplicated lower urinary tract infection in women: the LUTIW project. *Scand. J. Prim. Health Care* 2007;25:49–57.

175. Blunt I. Focus on preventable admissions: trends in emergency admissions for ambulatory care sensitive conditions, 2001 to 2013. Tech. rep. The Health Foundation and The Nuffield Trust, 2013.

176. Lachs MS, Nachamkin I, Edelstein PH et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann. Intern. Med.* 1992;117:135–140.

177. Devillé WLJM, Yzermans JC, Duijn NP van et al. The urine dipstick test useful to rule out infections. A meta-analysis of the accuracy. *BMC Urol.* 2004;4:4.

178. Honeyman M, Dunn P, McKenna H. A digital NHS? An introduction to the digital agenda and plans for implementation. Tech. rep. The King's Fund, 2016.

179. Fitzpatrick F, Tarrant C, Hamilton V et al. Sepsis and antimicrobial stewardship: two sides of the same coin. *BMJ Qual. Saf.* 2019;28:758–761.

180. Rockenschaub P, Gill MJ, McNulty D et al. Development of risk prediction models to predict urine culture growth for adults with suspected urinary tract infection in the emergency department: protocol for an electronic health record study from a single UK university hospital. *Diagn Progn Res* 2020;4:15.

181. Kouri TT, Kähkönen U, Malminiemi K, Vuento R, Rowan RM. Evaluation of Sysmex UF-100 urine flow cytometer vs chamber counting of supravitally stained specimens and conventional bacterial cultures. *Am. J. Clin. Pathol.* 1999;112:25–35.

182. Chiquet J, Grandvalet Y, Rigaill G. On coding effects in regularized categorical regression. *Stat. Modelling* 2016;16:228–237.

183. Gower JC. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 1971;27:857–871.

184. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* 2011;30:377–399.

185. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, Articles* 2011;45:1–67.

186. Moons KGM, Donders RART, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. *J. Clin. Epidemiol.* 2006;59:1092–1101.

187. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biom. J.* 2015;57:614–632.

188. Ambler G, Royston P. Fractional polynomial model selection procedures: investigation of type i error rate. *J. Stat. Comput. Simul.* 2001;69:89–108.

189. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* 2012;13:281–305.

190. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.

191. Van Calster B, Nieboer D, Vergouwe Y et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.* 2016;74:167–176.

192. Benavoli A, Corani G, Demšar J, Zaffalon M. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *J. Mach. Learn. Res.* 2017;18:2653–2688.

193. Steyerberg EW, Harrell Jr FE, Borsboom GJ et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* 2001;54:774–781.

194. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY, 2009.

195. Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* 2016;69:245–247.

196. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. Springer, Cham, 2019.

197. Efron B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* 1983;78:316–331.

198. Harrell F. Comparison of Strategies for Validating Binary Logistic Regression Models. http : / / hbiostat . org / doc / simval . html. Accessed: 2020-5-11. 2018.

199. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* 2020;27:621–633.

200. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med.* 2015;13:1.

201. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur. Heart J.* 2014;35:1925–1931.

202. Van Calster B, McLernon DJ, Smeden M van et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230.

203. Sutton RT, Pincock D, Baumgart DC et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3:17.

204. Coughlin RF, Peaper D, Rothenberg C et al. Electronic Health Record-Assisted Reflex Urine Culture Testing Improves Emergency Department Diagnostic Efficiency. *Am. J. Med. Qual.* 2020;35:252–257.

205. Rui P, Kang K, Ashman JJ. National Hospital Ambulatory Medical Care Survey: 2016 Emergency Department Summary Tables. Tech. rep. Centers for Disease Control and Prevention, 2016.

206. Munigala S, Poirier R, Liang S et al. Location, Location, Location: A Change in Urine Testing Order Sets on Culturing Practices at an Academic Medical Center Emergency Department. *Open Forum Infect Dis* 2016;3.

207. Stagg A, Lutz H, Kirpalaney S et al. Impact of two-step urine culture ordering in the emergency department: a time series analysis. *BMJ Qual. Saf.* 2018;27:140–147.

208. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020;2:e489–e492.

209. Debray TPA, Vergouwe Y, Koffijberg H et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J. Clin. Epidemiol.* 2015;68:279–289.

210. Spiegelhalter DJ, Best NG, Carlin BP, Linde A van der. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol.* 2002;64:583–639.

211. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 2017;27:1413–1432.

212. Watanabe S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.* 2010;11:3571–3594.

213. National Institute for Health and Care Excellence. Urinary tract infection (catheter-associated): antimicrobial prescribing [NG113]. https://www.nice.org.uk/guidance/ng113/. Accessed: 2020-10-12.

214. Magliano E, Grazioli V, Deflorio L et al. Gender and age-dependent etiology of community-acquired urinary tract infections. *ScientificWorldJournal* 2012;2012:349597.

215. Siegman-Igra Y, Kulka T, Schwartz D, Konforti N. The significance of polymicrobial growth in urine: contamination or true infection. *Scand. J. Infect. Dis.* 1993;25:85–91.

216. Croxall G, Weston V, Joseph S et al. Increased human pathogenic potential of Escherichia coli from polymicrobial urinary tract infections in comparison to isolates from monomicrobial culture samples. *J. Med. Microbiol.* 2011;60:102–109.

217. Hand DJ, Till RJ. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* 2001;45:171–186.

218. Sathiananthamoorthy S, Malone-Lee J, Gill K et al. Reassessment of Routine Midstream Culture in Diagnosis of Urinary Tract Infection. *J. Clin. Microbiol.* 2019;57.

219. Yang CC, Yang SSD, Hung HC et al. Rapid differentiation of cocci/mixed bacteria from rods in voided urine culture of women with uncomplicated urinary tract infections. *J. Clin. Lab. Anal.* 2017;31.

220. Nagendran M, Chen Y, Lovejoy CA et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.

221. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* 2018;286:800–809.

222. Denaxas S, Gonzalez-Izquierdo A, Direk K et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J. Am. Med. Inform. Assoc.* 2019.

223. Roland M. Linking physicians' pay to the quality of care - a major experiment in the United kingdom. *N. Engl. J. Med.* 2004;351:1448–1454.

224. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am. J. Epidemiol.* 2010;172:971–980.

225. McMillan B, Eastham R, Brown B, Fitton R, Dickinson D. Primary Care Patient Records in the United Kingdom: Past, Present, and Future Research Priorities. *J. Med. Internet Res.* 2018;20:e11293.

226. Rowley A, Turpin R, Walton S. The emergence of artificial intelligence and machine learning algorithms in healthcare: Recommendations to support governance and regulation. Tech. rep. BSI, 2019.

227. US Food and Drug Administration. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD): discussion paper and request for feedback. Tech. rep. 2019.

228. McKay R, Mah A, Law MR, McGrail K, Patrick DM. Systematic Review of Factors Associated with Antibiotic Prescribing for Respiratory Tract Infections. *Antimicrob. Agents Chemother.* 2016;60:4106–4118.

229. Klaveren D van, Gönen M, Steyerberg EW, Vergouwe Y. A new concordance measure for risk prediction models in external validation settings. *Stat. Med.* 2016;35:4136–4152.