Using machine learning in R to understand antibiotic prescribing

Dr Alexander Martin
Dr Christopher Lawrence
South Tees Hospitals NHS Foundation Trust
alexander.martin6@nhs.net

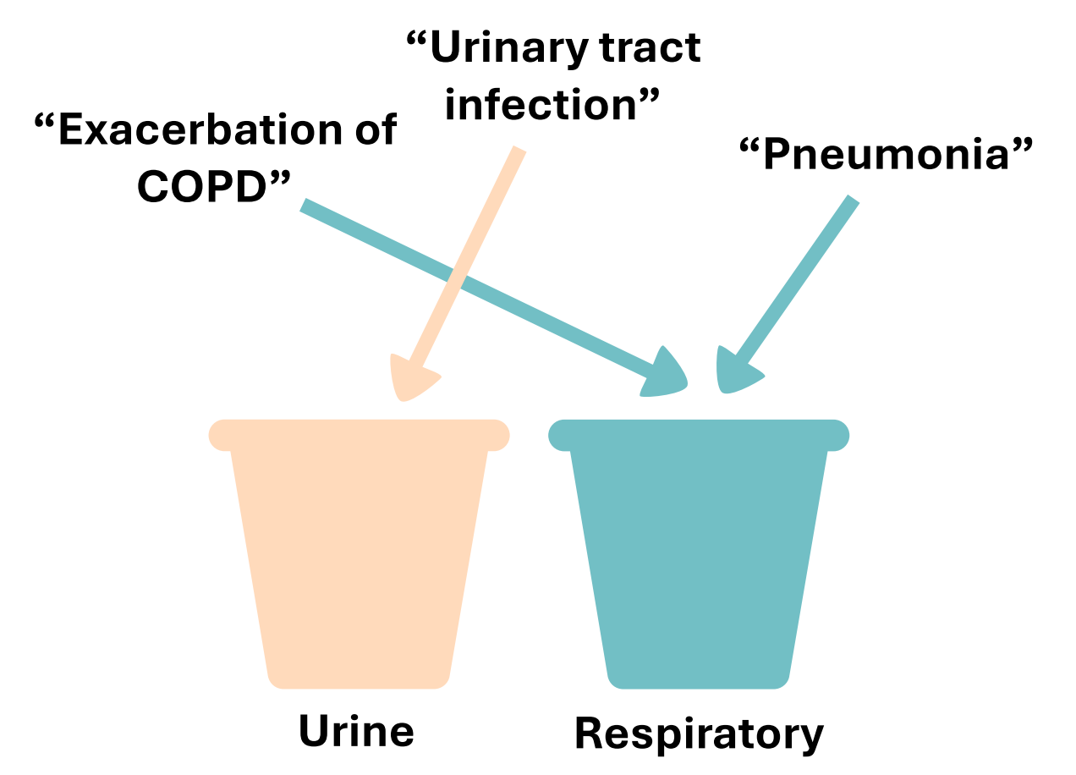
Download the code used in this project

GitHub: bytesizedbugs

Background

- Antimicrobial prescriptions should have a clearly documented indication [1].
- Electronic Prescribing and Medicines Administration (EPMA) systems hold large amounts of data. The indication for an antimicrobial prescription is recorded as free text.
- Categorising antimicrobial indications can help antimicrobial stewardship efforts, but free text data is diverse and difficult to use in its raw form.
- **Natural language processing (NLP)** is a machine learning tool to analyse free text data and extract meaningful information [2].

Why not just search for keywords?



Analysing free text is a challenge

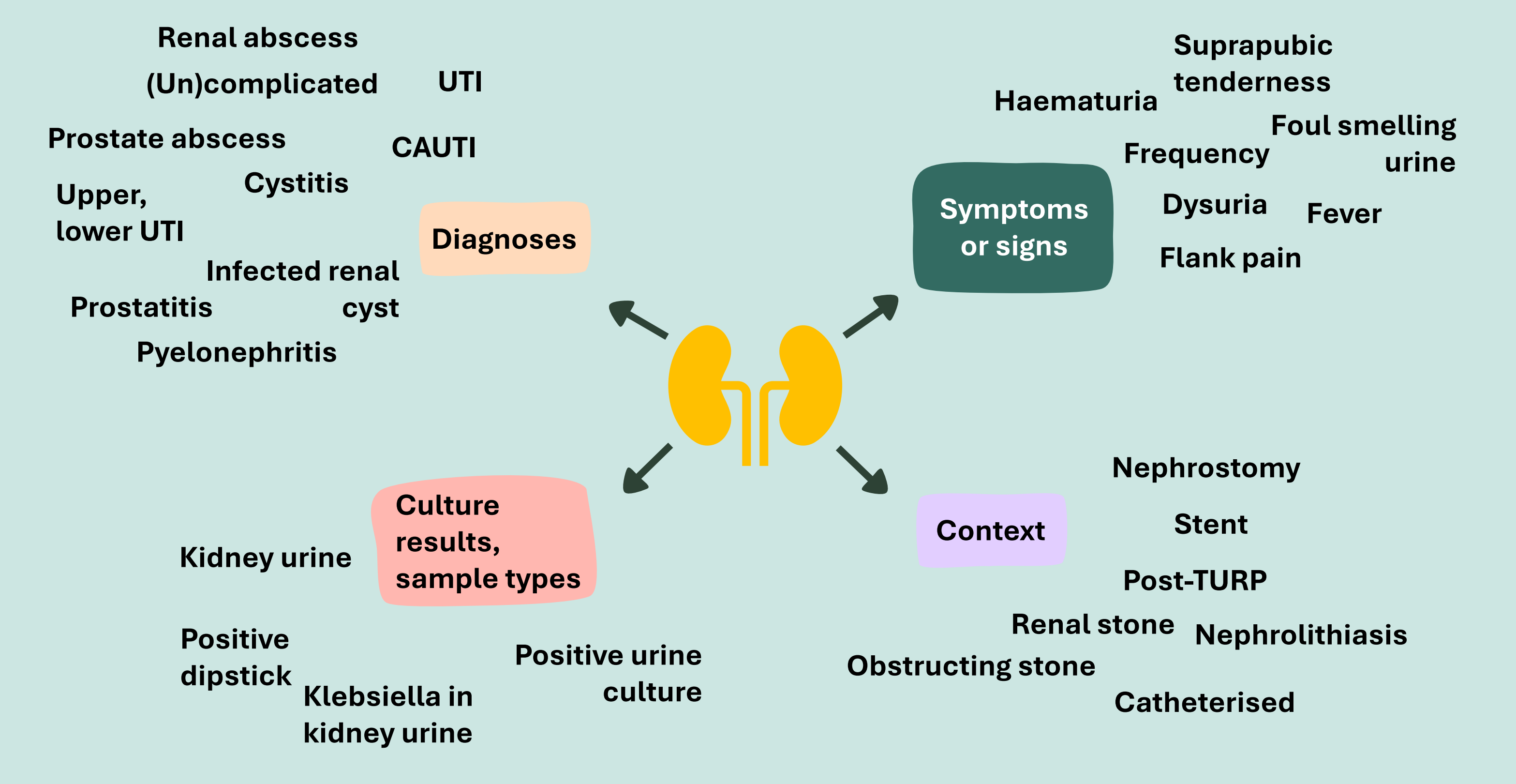
Antibiotic prescription
INDICATION:

UTI, "# prophylaxis"	**Acronyms**
Urine *or* urinary	**Word inflections**
Switch from sol**UTI**on	**False positives**
No sign of UTI	**Negation**
"**Kidney** urine", "...Impaired **kidney** function dose"	**Context**
Chest versus urine	**Uncertain diagnosis**
As per consultant ward round	**Uninformative indications**
Pyelonehprits	**Misspellings**

Aim

- To design an NLP model to categorise large numbers of free-text antibiotic indications.
- Validate and apply this model to understand which antibiotics are being prescribed, assess adherence to guidelines, and ultimately improve patient outcomes through more targeted interventions.

There are many ways to describe one system



Methods

- **74,134** inpatient free-text antimicrobial prescriptions were extracted from a tertiary centre EPMA.
- Indications were manually classified into categories based on specific syndromes or body systems: bacteraemia/line infection, neutropenic and non-neutropenic sepsis, infective endocarditis, diabetic foot infection and urinary tract, respiratory, skin, head and neck, bone and joint, perinatal and genital, plus prophylaxis and prescriptions referencing a specific culture result.
- An indication could occupy multiple categories, e.g. sepsis AND urinary tract infection.
- Notably, **12,967** (18%) indications provided **"no information"**, e.g. "as per microbiology".

- Data analysis was performed in R (version 4.3.1).
- 70% of the dataset (51,894 texts) was used for training and 30% (22,240 texts) for testing. Model discrimination was assessed using sensitivity, specificity, positive and negative predictive value and area under the receiver operating curve (AUC).

An **XGBoost** model [3] was trained to categorise antibiotic prescriptions.

1 Pre-process text strings (e.g. change to lower case, remove numbers) to create a corpus of texts.

2 A bag-of-words represents words in the corpus, ignoring word order and syntax.

3 Create a Document-Term Matrix, with a row representing each indication and each column representing a word in the corpus. Count the frequency with which each word appears in each row.

4 Make a guess whether a string belongs to a given category. Calculate the residuals (errors) and split the data using a decision tree to reduce errors. For example, if 'sputum' or 'cough' are common words in respiratory infection prescriptions, one tree might split on the presence or absence of these words.

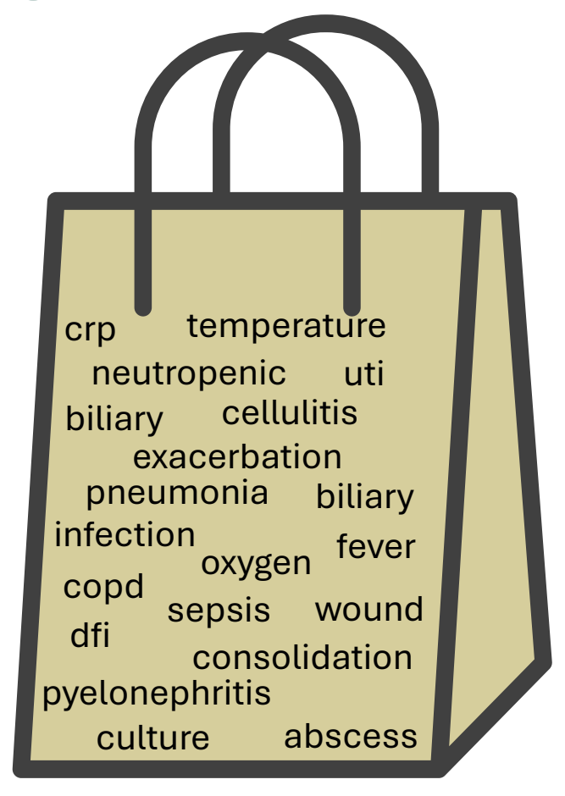
5 A single tree will have poor accuracy on its own. New trees are trained to correct previous errors by focusing on misclassified cases, adjusting the prediction gradually.

6 With more iterations, the predictions improve and the error reduces.

1 Pre-processing

Spell check
Remove numbers
Transform to lowercase
Remove stopwords
Stemming
Create a Corpus

2 Bag-of-words



3 Document-Term Matrix

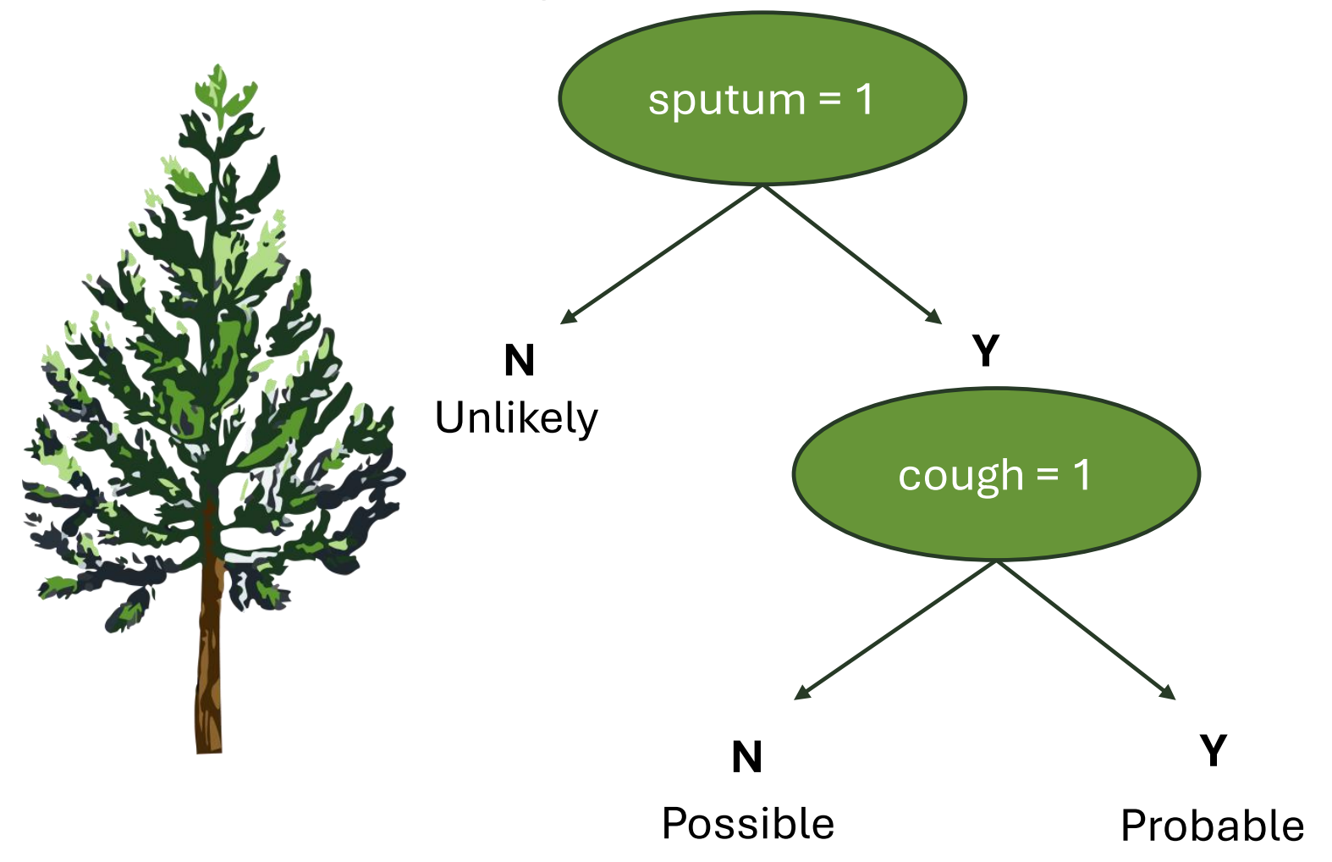
string	word1	word2	word3	...
string1	1	0	0	
string2	0	1	0	
string3	0	0	1	
string4	1	0	1	
string5	0	0	0	
...				

Example

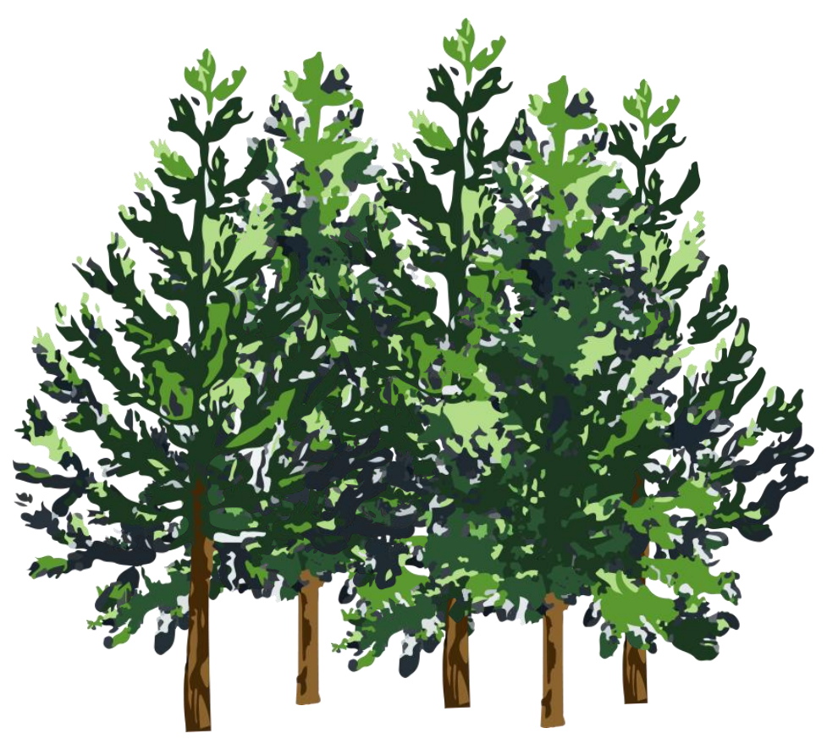
string	SOB	LRTI	dysuria
tachypnoea, SOB, ?LRTI	1	1	0
dysuria, frequency	0	0	1

4 Single Decision Tree

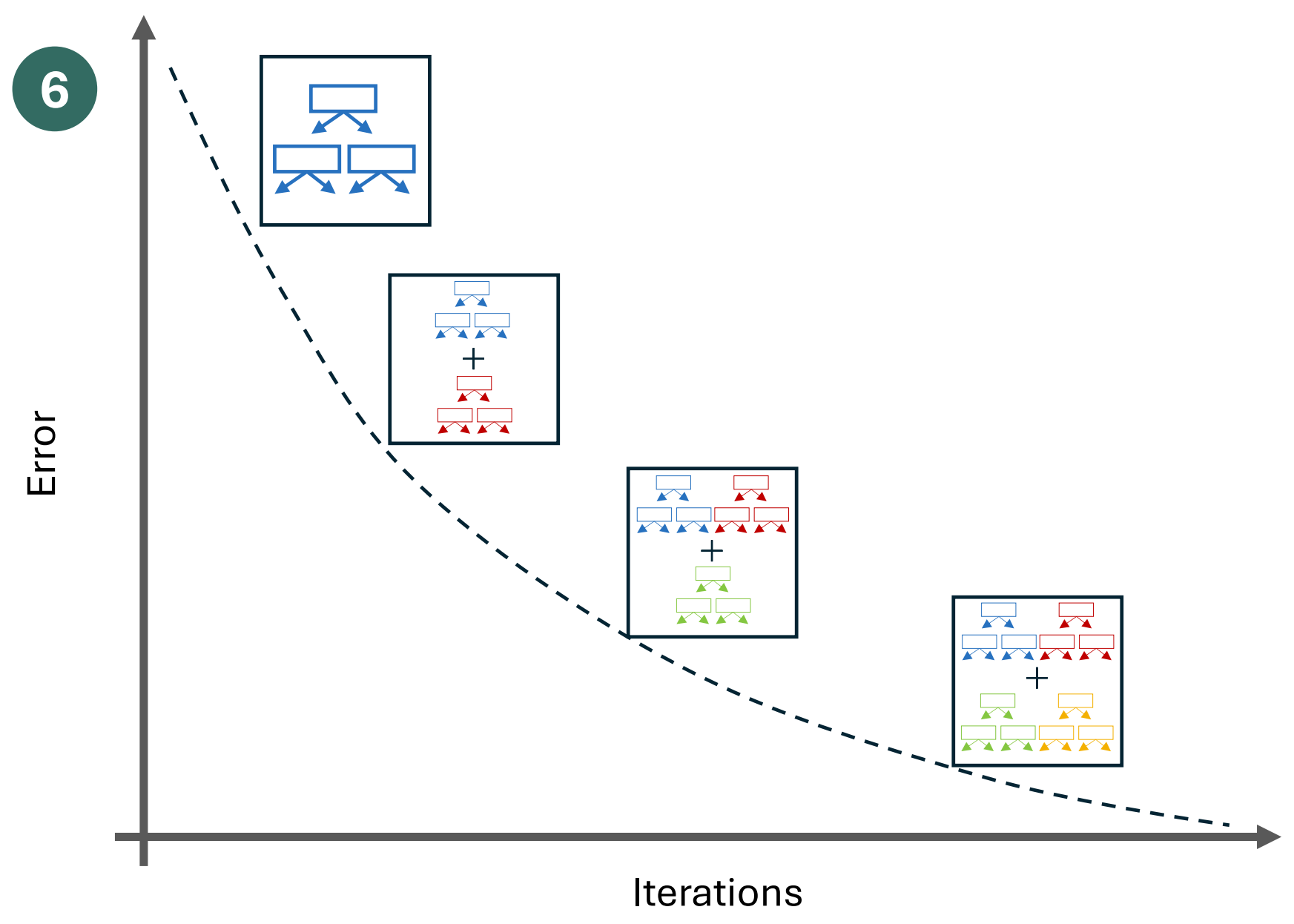
Probability of chest infection



5 Multiple trees



6



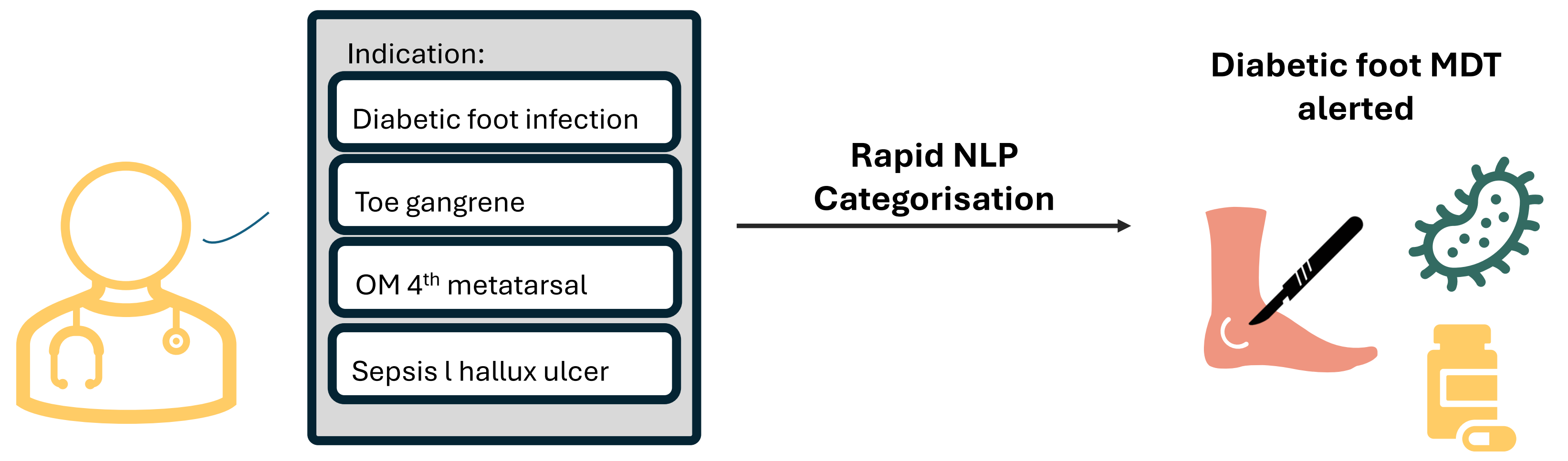
Results & Discussion

Model discrimination

- Individual models were applied to the test dataset. Overall **discrimination was excellent**, with AUCs for binary discrimination of each category ranging from **0.95 – 1.00**.
- The model with the poorest discrimination was for bacteraemia, perhaps because it is often associated with other infections e.g. abdominal or urinary sepsis. Sensitivity was 0.83 and specificity was 0.90.
- The model with the best discrimination was for diabetic foot infection, with **sensitivity of 0.995 and specificity of 0.999**.

Rapid, regular reports

- The models were applied to a new dataset of prescriptions for internal validation.
- **7,747** prescriptions were categorised in **11 seconds** of runtime.
- Discrimination was again excellent, with AUC ranging from 0.90 to 0.99.
- This rapid categorisation allows reports to be produced regularly. Clinical teams can stay informed about antimicrobial prescribing for conditions of interest.



Indication:

Diabetic foot infection
Toe gangrene
OM 4th metatarsal
Sepsis l hallux ulcer

Understanding prescribing and microbiological sampling behaviour

- In 2023, of the 6,526 unique inpatient episodes with an antimicrobial prescription, **5,129** (79%) had at least one "informative" documented indication, and **1,397** (21%) had only "no information" prescription indications.
- People in hospital who are prescribed an antimicrobial have a microbiological sample taken [1].
- Laboratory records showed that a **microbiological sample** was statistically more likely to have been received for those patients with "informative" antibiotic indications **(80%)** than "no information" indications **(62%), *p*<0.001** (Chi-square).
- Causation cannot be inferred, but this finding highlights that taking samples for microbiology samples is key for diagnosis and good prescribing.
- NLP embedded into the EPMA could categorise the treatment indication and prompt clinicians to ensure appropriate microbiological samples are taken.

Limitations

- The "Bag of words" approach does not understand word context, such as negatives ("**not** neutropenic") or qualifiers ("**previous** C. diff").
- More complex NLP models, such as **ClinicalBERT**, perform better as they can understand the context of each word in relation to the rest of the clinical record [4].
- Finally, seamlessly incorporating NLP into workflows will require deep integration with the electronic patient record.

Conclusions

- Machine learning and natural language processing can accurately classify antimicrobial indications.
- Fast runtimes will allow regular reports to be produced to support a proactive approach to stewardship.
- 'Indications' which provide no useful information represent a significant group of antimicrobial prescriptions and a challenge for stewardship. Further investigation into the link between uninformative indications and failure to take microbiological samples is warranted.
- Newer models have the potential to provide deeper insights into clinical records, support antimicrobial stewardship and improve patient outcomes.

[1] National Institute for Health and Care Excellence (2016) Antimicrobial stewardship quality standard.
[2] Hirschberg J, Manning CD (2015) Advances in natural language processing. Science. PMID: 26185244.
[3] Tianqi C, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System.
[4] Huang K, Altosaar J, Ranganath R (2019). ClinicalBERT: Modeling clinical notes and predicting hospital readmission.