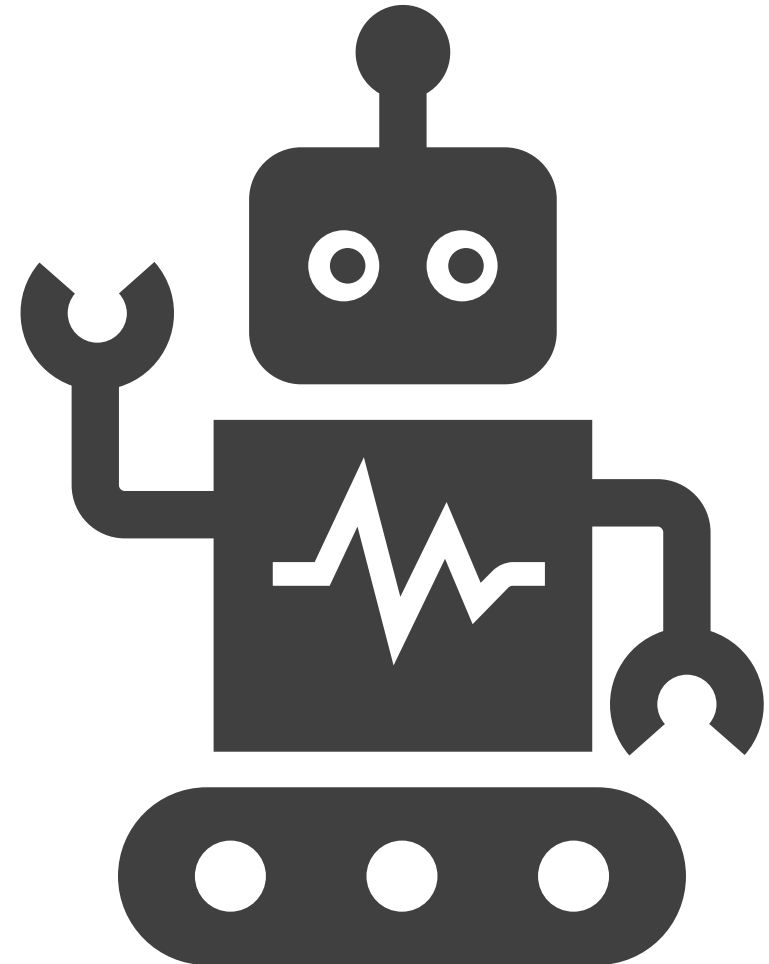


Machine, Machine on the wall, Who's the most Resistant of them all?

Alfred Lee MRCP(UK) DTM&H(RCP) FRCPath FRCPA

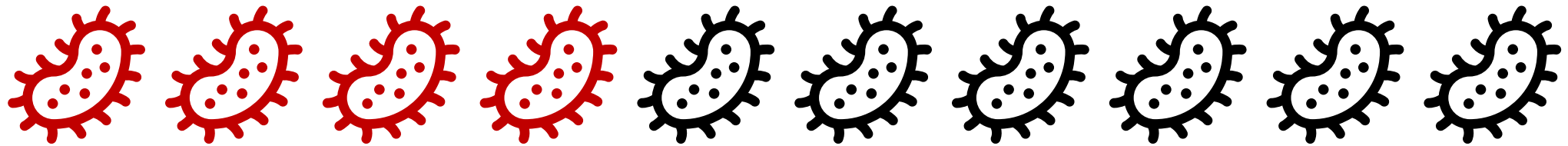
Associate Consultant

PWH Microbiology

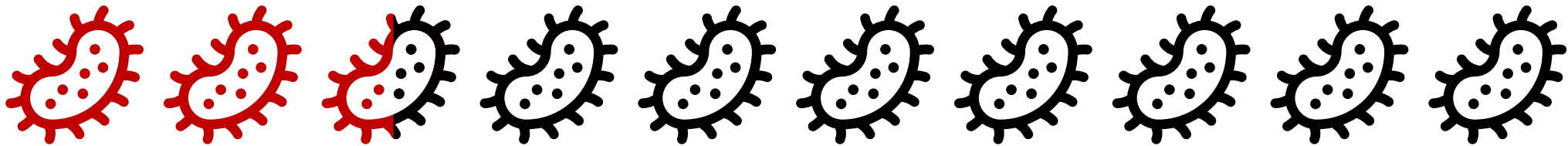


High rate of **antimicrobial resistance**

41% *Staphylococcus aureus* are **MRSA**

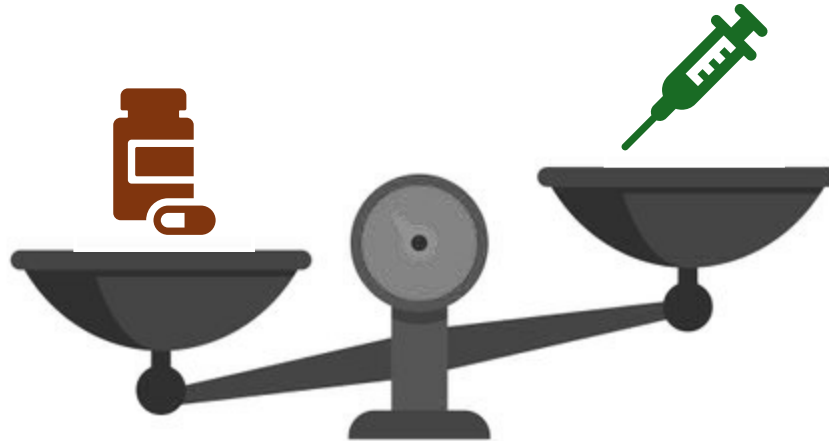


27% *Escherichia coli* are **resistant to 3rd generation cephalosporins**



Every day, we face an **impossible** dilemma.

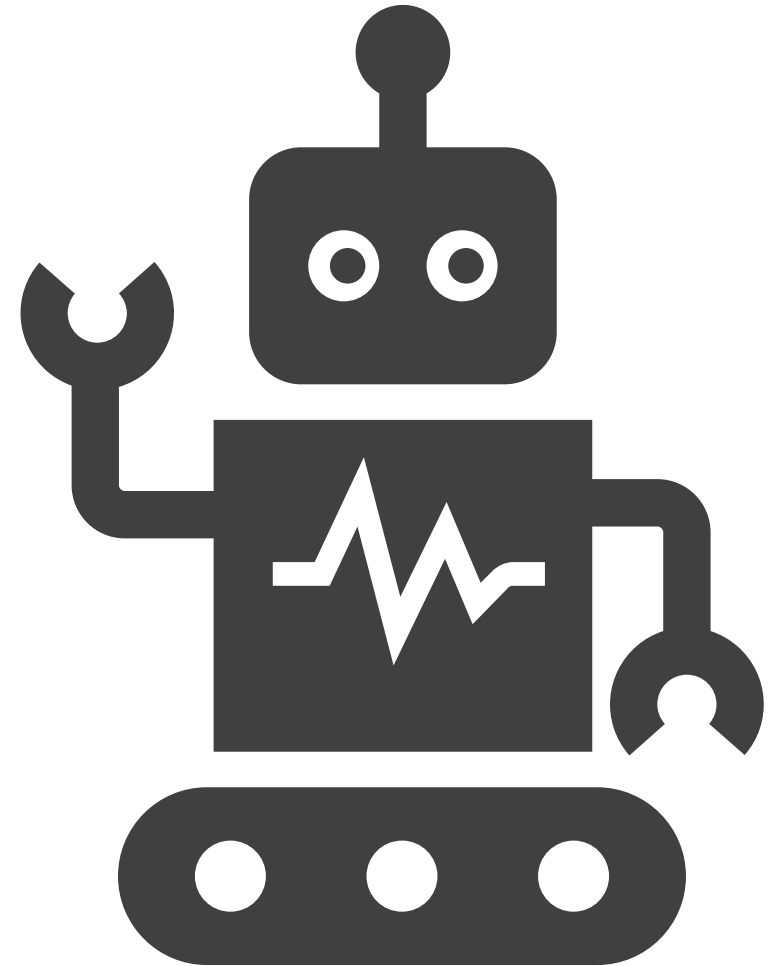
Stringent
antibiotic use to
reduce AMR



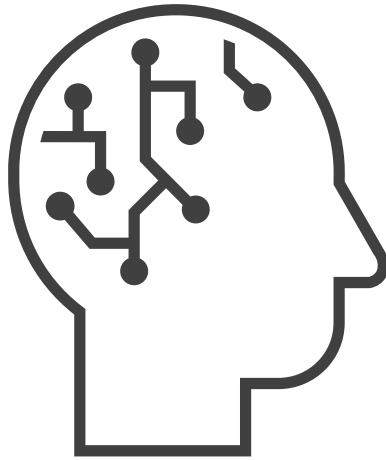
Broad-spectrum
antibiotic use to
ensure efficacy

Machine Learning

Can it help?



What better way to understand this by asking AI itself?



Assistant Poe

Machine learning can significantly enhance the prediction of antimicrobial resistance (AMR) through various approaches:

1. **Data Analysis:** Machine learning algorithms can analyze large datasets from clinical samples, genomic sequences, and antibiotic usage patterns. This helps identify trends and correlations that might not be evident through traditional statistical methods.
2. **Genomic Insights:** By applying machine learning to genomic data, researchers can predict resistance mechanisms based on genetic mutations. Algorithms can classify bacteria and their resistance profiles based on their genetic makeup.
3. **Clinical Decision Support:** Machine learning models can assist healthcare providers by predicting the likelihood of AMR in specific infections. This can guide appropriate antibiotic prescribing, reducing the misuse of antibiotics.
4. **Surveillance Systems:** Machine learning can enhance epidemiological surveillance by processing real-time data from healthcare institutions, enabling early detection of AMR outbreaks.
5. **Feature Selection:** Algorithms can identify and prioritize the most significant features (e.g., patient demographics, prior antibiotic exposure) that contribute to AMR, improving the understanding of risk factors.
6. **Predictive Modeling:** Supervised learning techniques can develop predictive models based on historical data to forecast the emergence of resistance in bacterial populations.
7. **Natural Language Processing (NLP):** NLP can extract relevant information from scientific literature and clinical records, helping to compile comprehensive databases of resistance patterns.
8. **Simulation and Forecasting:** Machine learning can simulate the impact of different antibiotic usage strategies on the evolution of resistance, aiding in public health planning and policy-making.

By leveraging these capabilities, machine learning can provide valuable insights into AMR, ultimately leading to better prevention, treatment, and management strategies.



AMR Prediction

 Assistant

 New chat




3. **Clinical Decision Support:** Machine learning models can assist healthcare providers by predicting the likelihood of AMR in specific infections. This can guide appropriate antibiotic prescribing, reducing the misuse of antibiotics.

Predicting antibiotic susceptibility in urinary tract infection with artificial intelligence—model performance in a multi-centre cohort

Alfred Lok Hang Lee ^{1*}, Curtis Chun Kit To², Ronald Cheong Kin Chan², Janus Siu Him Wong³, Grace Chung Yan Lui⁴, Ingrid Yu Ying Cheung¹, Viola Chi Ying Chow¹, Christopher Koon Chi Lai⁵, Margaret Ip ⁵ and Raymond Wai Man Lai⁶

¹Department of Microbiology, Prince of Wales Hospital, Shatin, Hong Kong SAR, China; ²Department of Anatomical and Cellular Pathology, Faculty of Medicine, Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; ³Department of Orthopaedics and Traumatology, School of Clinical Medicine, LKS Faculty of Medicine, University of Hong Kong, Pokfulam, Hong Kong SAR, China; ⁴Department of Medicine and Therapeutics, Prince of Wales Hospital, Shatin, Hong Kong SAR, China; ⁵Department of Microbiology, Faculty of Medicine, Chinese University of Hong Kong, Shatin, Hong Kong SAR, China; ⁶Chief Infection Control Officer Office, Hospital Authority, Kowloon, Hong Kong SAR, China

*Corresponding author. E-mail: leelh1107@gmail.com
@AlfredL75461620

Received 28 March 2024; accepted 5 July 2024

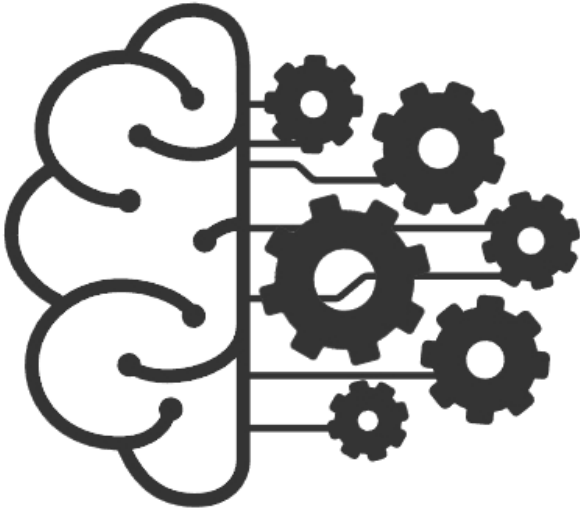
The clinical question of this study:

Can we use existing healthcare data to predict the AST result in patients with UTI?

Data engineering

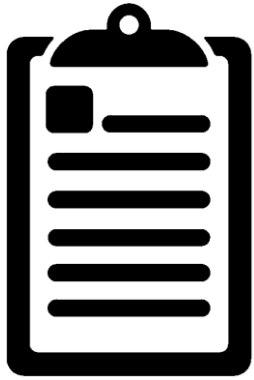


Model development



Model testing





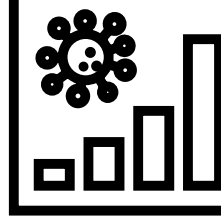
Data collection

Data collection

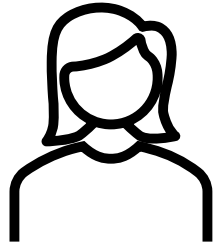
- Inclusion criteria – All adult patients with urinary tract infection and positive urine culture in PWH, NDH, and SH during the period 2015-2020
- Exclusion criteria
 - Asymptomatic bacteriuria
 - Discordance between blood culture and urine culture
 - Common contaminants in urine culture – coagulase-negative staphylococci, diphtheroids, viridans streptococci...

Data collection

- Microbiology data: Laboratory Information System (LIS)
- All other data: Clinical Data Analysis and Reporting System (CDARS)
- The most significant challenge while collecting data is...
 - The output size of CDARS is limited.
 - Larger datasets like blood investigations and medication history require multiple partitions (~50-100) to download.

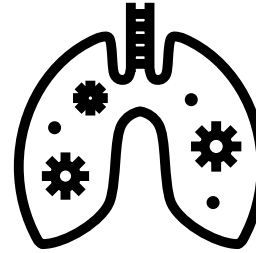


Previous culture &
sensitivity results

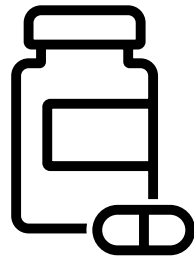


Basic epidemiological
information

**What data
did we
obtain?**



Medical history



Medication history

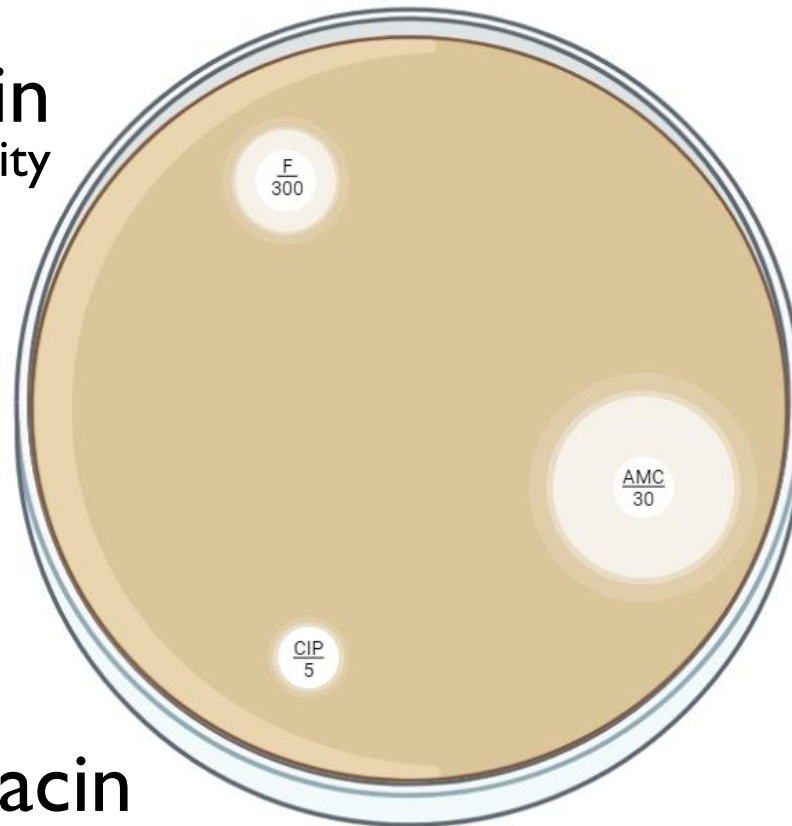


Past investigation results

What are we predicting?

Antibiotic susceptibility in urine culture:

Nitrofurantoin
susceptibility



Amoxicillin-clavulanate
susceptibility

Ciprofloxacin
susceptibility



Data wrangling & cleaning

What does CDARS output look like?

This is one of the 10 spreadsheets containing prior microbiological culture results...



refkey	labnum	date	specialty	specialty2	specimen	specimen2	ID	ID2	Amikacin	Augmentin	Cefotaxim
9310585	19B29970	#####	MED	NEPHRO	Culture, Ur	Midstream	Proteus mi	Proteus mi	Sensitive	Sensitive	
9313070	20B27664	#####	MED	RHEUMA	Culture, Ur	Midstream	Escherichi	Escherichia coli	Sensitive		
9314033	19MB0627	#####	A&E	A&E	Culture, Ur	Midstream	Escherichi	Escherichia coli	Sensitive		
9314033	19B29666	#####	MED	INT_MED	Culture, Ur	Midstream	Escherichi	Escherichia coli	Sensitive		
9314491	18B27037	#####	GYN	GYN	Culture, Ur	Midstream	Pseudomo	Pseudomonas aeruginosa			
9315637	17B25898	#####	MED	INT_MED	Culture, Ur	Midstream	Escherichi	Escherichi	Sensitive	Sensitive	
9318929	17B31531	#####	A&E	A&E	Culture, Ur	Midstream	Escherichi	Escherichia coli	Sensitive		
9320467	16B29153	#####	MED	GER	Culture, Ur	Midstream	Escherichi	Escherichia coli	Sensitive	Resistance	
9322419	20B29214	#####	MED	INT_MED	Culture, Ur	Midstream	Escherichi	Escherichia coli		Resistance	
9323218	18B28562	#####	SUR	SUR	Culture, Ur	Catheteriz	Pseudomo	Pseudomonas aeruginosa			
9323218	18B29701	#####	SUR	SUR	Culture, Ur	Nephrosto	Escherichi	Escherichi	Sensitive	Sensitive	
9326723	18B26921	#####	GYN	GYN	Culture, Ur	Midstream	Escherichi	Escherichia coli	Sensitive	Resistance	
9326723	18B25955	#####			Culture, Ur	Midstream	Escherichi	Escherichia coli	Sensitive	Resistance	
9326852	18B28439	#####	A&E	A&E	Culture, Ur	Midstream	Escherichi	Escherichia coli		Intermediate	
9328182	20B30618	#####	MED	INT_MED	Culture, Ur	Catheteriz	Escherichi	Escherichi	Sensitive	Sensitive	
9328213	16B29300	#####	EM	EM	Culture, Ur	Midstream	Klebsiella	Klebsiella pneumonia		Intermediate	
9328213	16B29300	#####	EM	EM	Culture, Ur	Midstream	Proteus mi	Proteus mirabilis	Sensitive		

Data cleaning

Examination of initial data to look for potential issues

- Outliers
- Missing values
 - Deduction by domain knowledge
 - Imputation by MICE
- Correlation matrix
 - Remove the highly correlated features
 - Dimensionality reduction with Principal Component Analysis

Data wrangling

Using microbiological data as example

- Classification of organisms

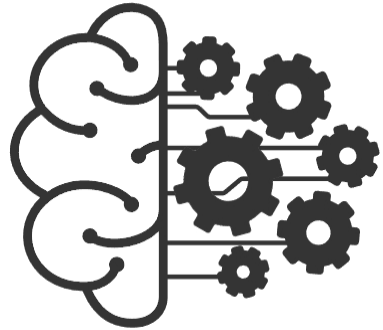
- *Escherichia coli* belongs to the Enterobacterales order.
- *Pseudomonas aeruginosa* belongs to Non-fermenters.
- *Trichosporon asahii* belongs to Yeasts.

- Classification of AMR

- *Providencia stuartii* is intrinsically resistant to Nitrofurantoin.
- *Stenotrophomonas maltophilia* is intrinsically resistant to Amoxicillin-clavulanate.

Data wrangling

- Combine/Merge multiple datasheets into one single dataframe.
- Each row represents one patient.
- Each column represents one feature.
- Combination and validation
 - Excel: *vlookup*
 - R: *cbind*
 - No matter the function you use, **always** check after each combination.



Data analysis and Model development

Data analysis

- Software
 - Microsoft 365 Excel: data cleaning and wrangling
 - R and the randomForest package: statistical analysis, machine learning
- Hardware
 - Razor Blade 15 (my laptop)

Exploratory data analysis: Summary of key features

Table 1. Baseline characteristics of the study cohort ($n = 26\,087$)

Variable	Count (%) / mean (SD)
(a) Antibiotic susceptibility of causative agent in the cohort	
Nitrofurantoin	20 794 (79.7%)
Amoxicillin-clavulanate	18 658 (71.5%)
Ciprofloxacin	17 410 (66.8%)
(b) Basic epidemiological factors	
Male gender	8888 (34.1%)
Use of immunosuppressive within the past 5 years	2283 (8.75%)
Age in years	67.9 (21.0)
Patient source	
Emergency department	3277 (12.6%)
Hospital in-patient	14 303 (54.8%)
Clinic	8507 (32.6%)

(g) Mean percentage of urinary tract bacterial pathogen susceptible to antibiotics in the past 5 years^a

Nitrofurantoin	76.54
Amoxicillin-clavulanate	68.55
Meropenem	91.03
Ciprofloxacin/Levofloxacin	62.82
Amikacin	99.53
Cefuroxime	65.99%
Cefotaxime/ceftriaxone	70.31%

Other pre-ML processing

- Train-test split
 - Data split into **training set (2015-2019)** and **test set (2020)**
 - **Training set**: model training & development
 - **Test set**: model testing & evaluation
- Scaling of features
 - Subtraction of mean and division by standard deviation
 - Mean = 0, SD = 1
- Addressing class imbalance with resampling techniques

Fitting various ML models to the data...

Random Forest

Support Vector Machine

Decision Tree

Logistic regression

Artificial neural network



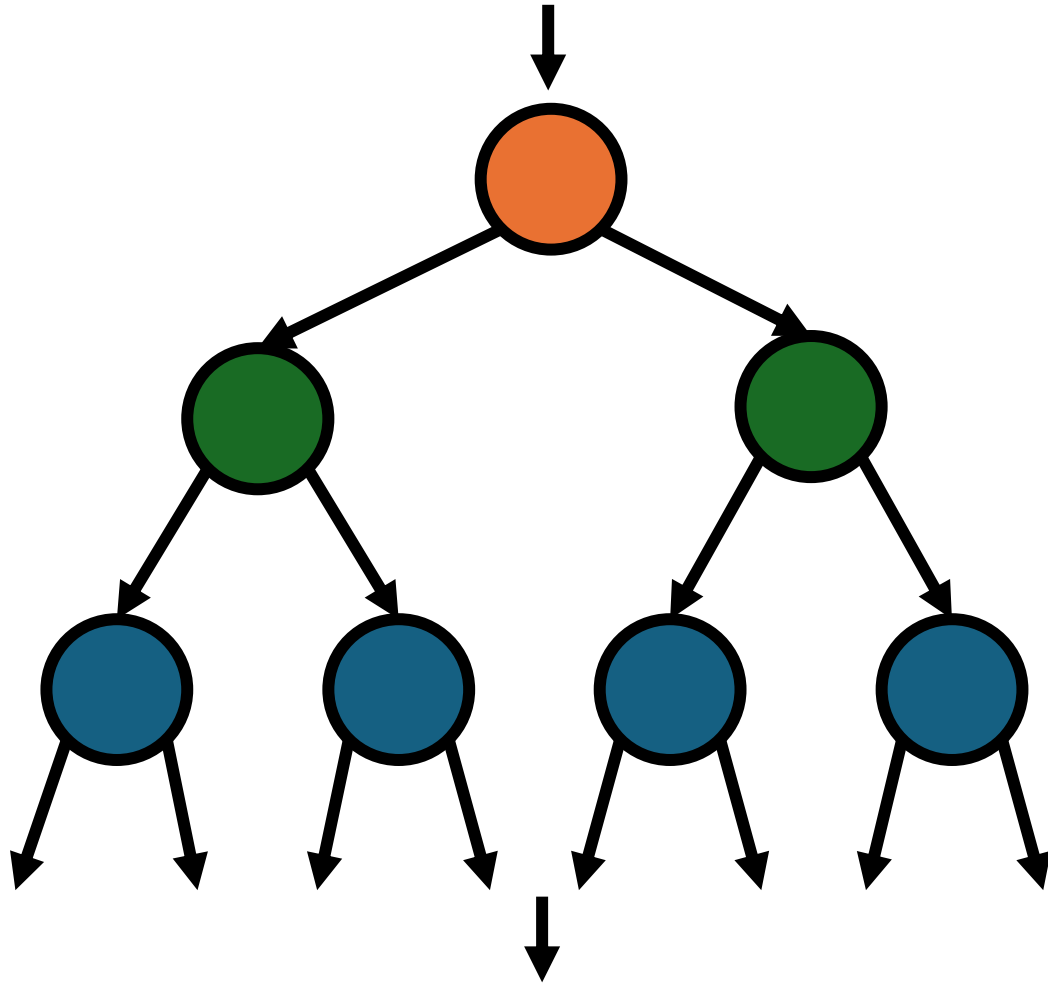
Random Forest was the
best performing model
across all three antibiotics.



What is Random Forest?

(Image generated by StableDiffusion XL[®] using the prompt “random forest”)

Training data



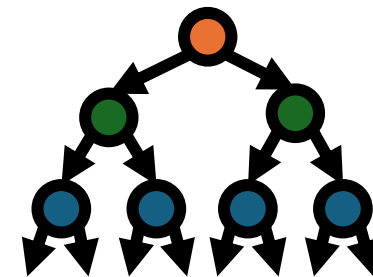
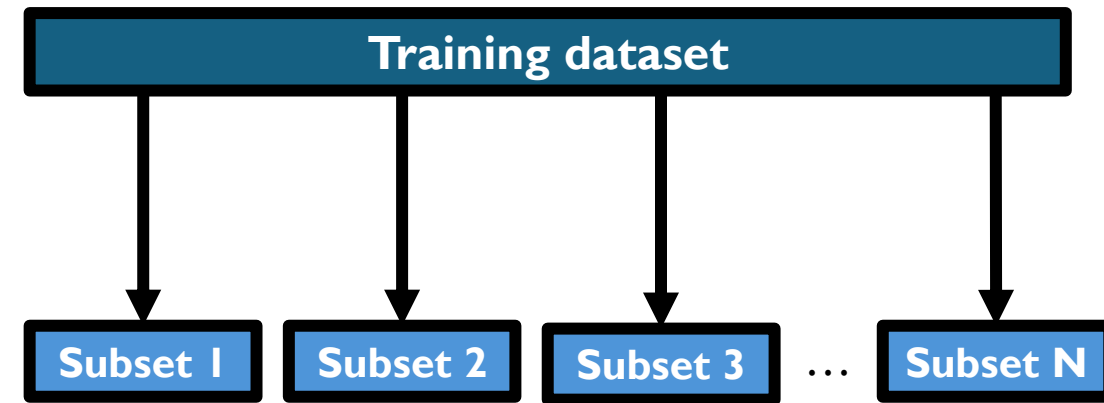
Classification result

Decision Tree

- Each node represents a decision rule based on a feature.
- Subjects are divided into subgroups based on a series of rules (nodes).
- As more rules are applied (training), each subgroup becomes more homogenous (Gini index).
- Training stops when Gini index is 0 or there are no features left.



Planting trees into a forest...

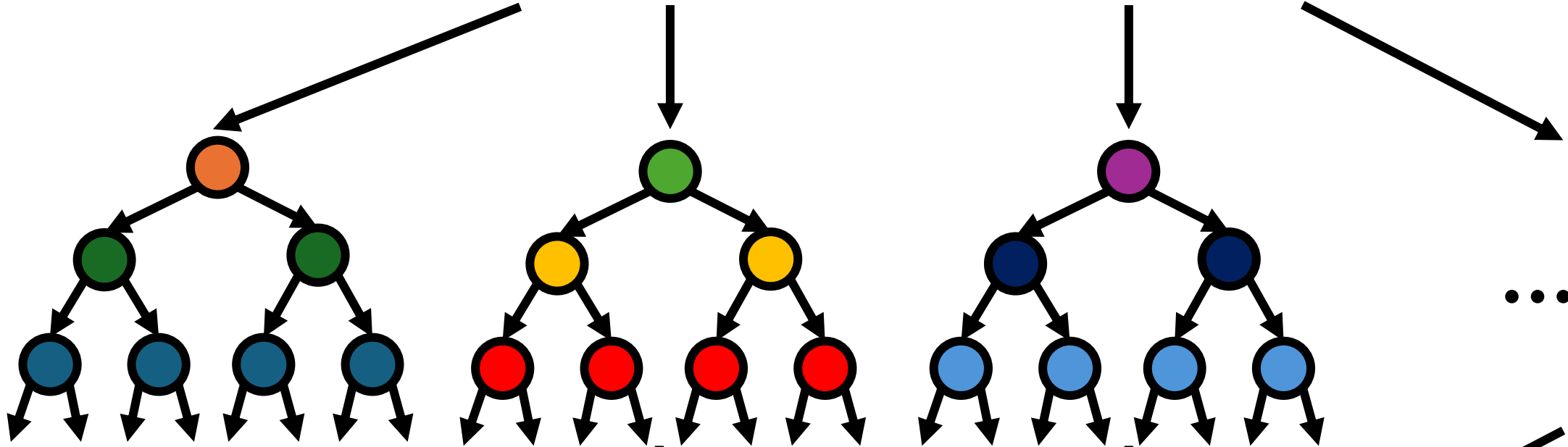


From each **subset** of data, a **decision tree** is created.

With **N subsets** of data, there are **N decision trees**.

N smaller datasets from random sampling

Model training



Model testing

Majority voting of the N decision trees



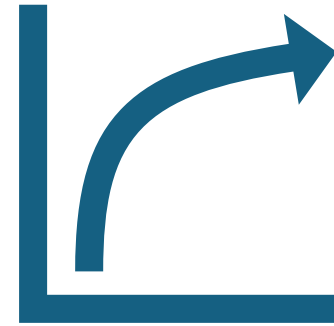
Model testing

What makes a good model?

Accuracy
Sensitivity
Specificity
NPV
PPV



Harmonic F_1 score
(harmonic mean of precision and recall)



AUC-ROC

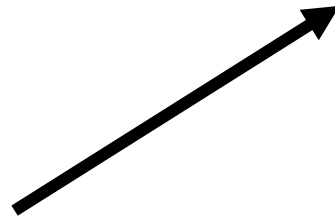
Table 5. Performance of random forest in predicting antibiotic susceptibility

	Nitrofurantoin	Amoxicillin-clavulanate	Ciprofloxacin
Accuracy	0.902	0.894	0.882
Sensitivity	0.840	0.765	0.860
Specificity	0.916	0.943	0.894
F1	0.938	0.928	0.906
AUC-ROC	0.941	0.939	0.937
PPV	0.960	0.913	0.919
NPV	0.705	0.837	0.820

Model explainability

ML models used to be **black boxes**.

We know how to train them, but we do not know how they make predictions.

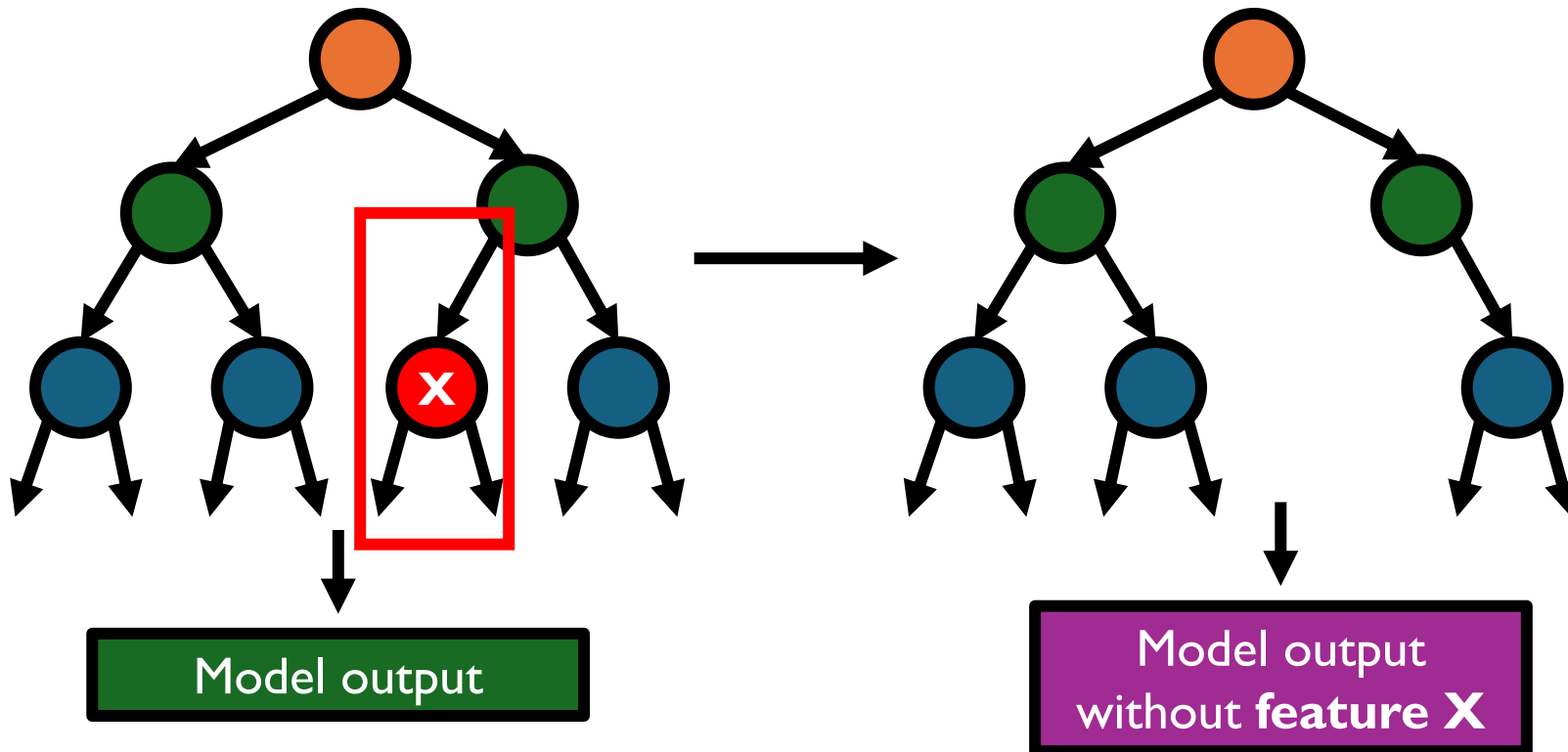


For humans to trust and use ML models, we need to understand **how they work**.

Fortunately, there are methods to **crack open** these black boxes.

SHapley Additive exPlanations (SHAP) values

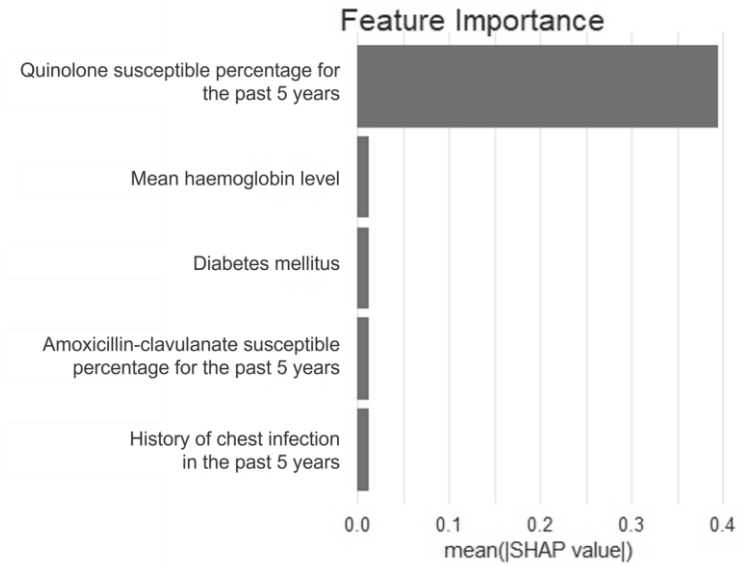
You cannot realise the value of a person until they are gone.



To determine the significance of a feature, we look at the difference in model output by **removing** the feature.

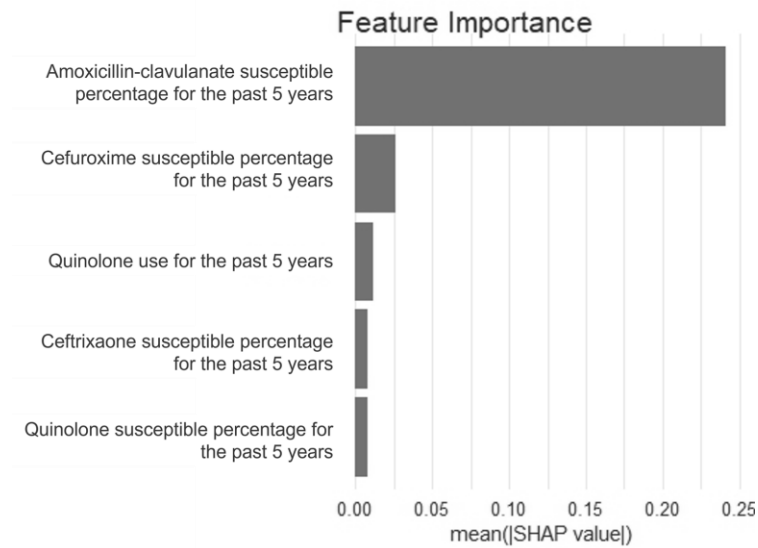
SHAP value is the **difference** in model output after the feature is omitted.

Ciprofloxacin



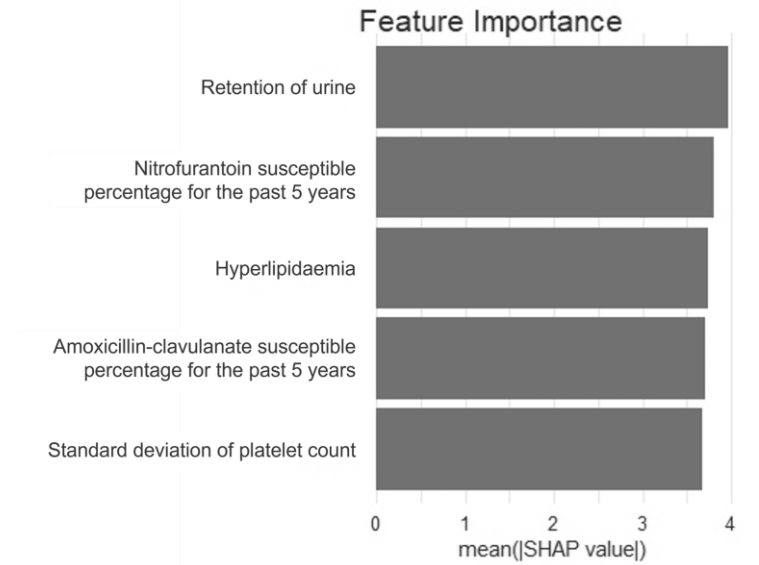
Most important feature:
Previous Ciprofloxacin susceptibility

Amoxicillin-clavulanate



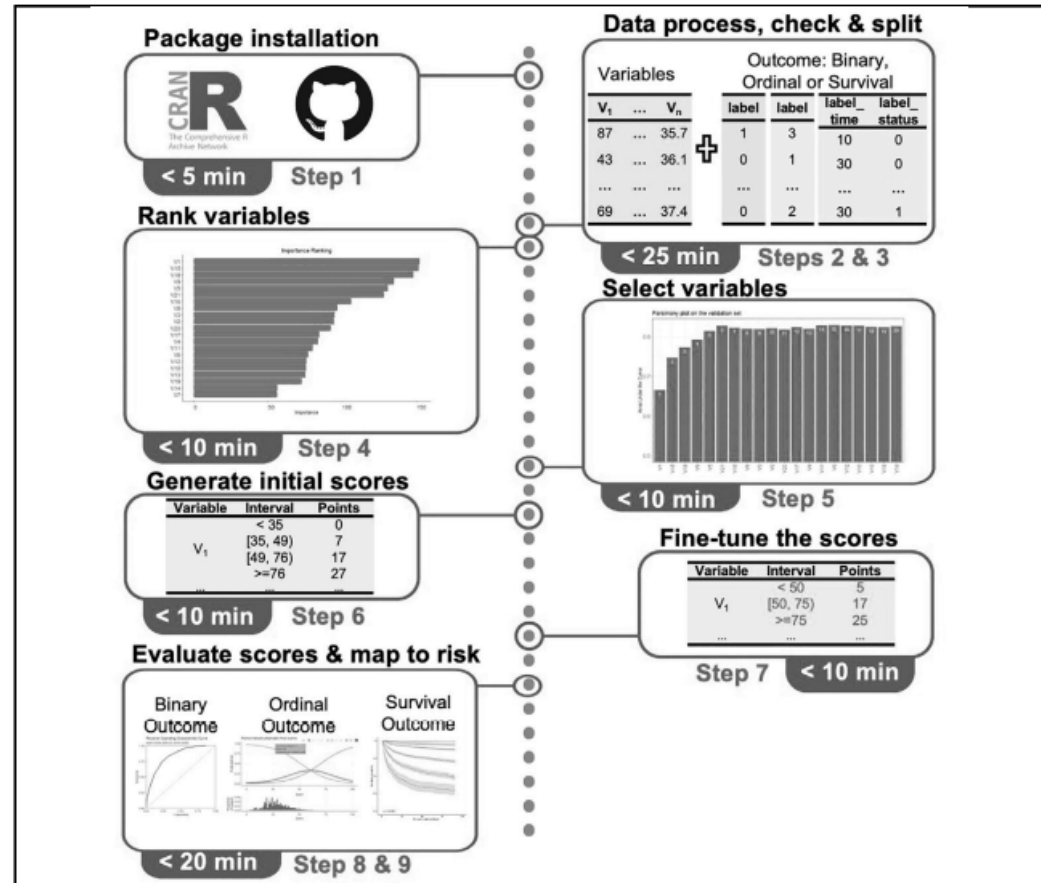
Most important feature:
Previous Augmentin susceptibility

Nitrofurantoin



Various features are important.

A universal AutoScore framework to develop interpretable scoring systems for predicting common types of clinical outcomes



Feng Xie, Yilin Ning,
 Mingxuan Liu, ...,
 Roger Vaughan,
 Bibhas Chakraborty,
 Nan Liu

liu.nan@duke-nus.edu.sg

Highlights
 A machine learning
 framework for
 automated
 development of
 clinical risk scores

A common workflow
 to handle binary,
 survival, and ordinal
 outcomes

Detailed
 demonstration of R
 package usage using
 publicly shared
 clinical data

Conclusion

- Random forest model aids judicious empirical antibiotics use in UTI.
- Given the reasonable performance and accuracy, these accurate models may aid clinicians in choosing between different first-line antibiotics for UTI.

How about other sites of infection? This pipeline also works on bloodstream infection!

European Journal of Clinical Microbiology & Infectious Diseases (2021) 40:1049–1061
<https://doi.org/10.1007/s10096-020-04120-2>

ORIGINAL ARTICLE



Deep learning model for prediction of extended-spectrum beta-lactamase (ESBL) production in community-onset Enterobacteriaceae bacteraemia from a high ESBL prevalence multi-centre cohort

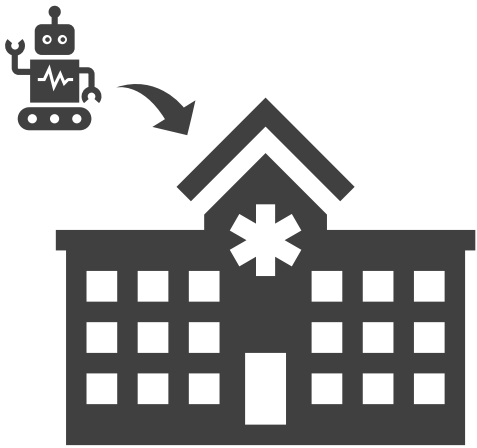
Alfred Lok Hang Lee¹  · Curtis Chun Kit To² · Angus Lang Sun Lee² · Ronald Cheong Kin Chan² · Janus Siu Him Wong³ · Chun Wai Wong¹ · Viola Chi Ying Chow¹ · Raymond Wai Man Lai¹

Received: 3 September 2020 / Accepted: 2 December 2020 / Published online: 5 January 2021
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Table 1. Performance of machine learning across different studies in predicting antibiotic susceptibility patterns.

Authors	Year of Publication	Medical Setting	Geographical Setting	Input Data	ML Algorithms	Performance Evaluation	Bacterial Species
Goodman et al. [27]	2016	Hospital admissions	USA	Blood cultures/AST	Recursive partitioning, DT	PPV 0.908-NPV 0.919	<i>Escherichia coli</i> , <i>Klebsiella pneumoniae</i> , <i>Klebsiella oxytoca</i>
Vazquez-Guillamet et al. [29]	2017	Hospital admissions	USA	EHR data/Blood cultures/AST	Recursive partitioning, DT	AUC 0.61–0.80	GNB
Sousa et al. [28]	2019	Hospital admissions	Spain	Clinical/demographic data/Blood cultures/AST	DT	AUC 0.76	BL-GNB
Moran et al. [20]	2020	Hospital admissions and primary care	UK	Blood/urine cultures	XGBoost	AUC 0.70	<i>Escherichia coli</i> , <i>Klebsiella pneumoniae</i> and <i>Pseudomonas aeruginosa</i>
Feretzakis et al. [33]	2020	Medical wards	Greece	Demographics/Cultures/AST/Bacterial Gram stain/Type of sample	MLR	AUC 0.758	All isolated bacterial species
Feretzakis et al. [34]	2020	Intensive Care Unit	Greece	Demographics/Cultures/AST/Bacterial Gram stain/Type of sample	LR, RF, k-NN, J48, MLP	AUC 0.726	All isolated bacterial species
Feretzakis et al. [35]	2021	Intensive Care Unit	Greece	Demographics/Cultures/AST/Bacterial Gram stain/Type of sample	JRip, RF, MLP, Class. Regr, REPTree	F-measure 0.884, AUC 0.933	<i>Pseudomonas aeruginosa</i> , <i>Acinetobacter baumannii</i> , <i>Klebsiella pneumoniae</i>
Martínez-Agüero et al. [36]	2019	Intensive Care Unit	Spain	Demographics/Clinical data/Type of sample/Cultures/AST	LR, k-NN, DT, RF, MLP	Accuracy for quinolone resistance 88.1 ± 1.6	<i>Pseudomonas</i> , <i>Stenotrophomonas</i> , <i>Enterococcus</i>
McGuire et al. [5]	2021	Hospital admissions	USA	Demographic, medication, vital sign, laboratory, billing code, procedure, culture, and sensitivity data (67 features)	XGBoost	AUC 0.846	Bacterial isolates with CR
Pascual-Sánchez et al. [32]	2021	Intensive Care Unit	Spain	EHR data	LR, DT, RF, XGBoost, MLP	AUC 0.76	MDR bacteria
García-Vidal et al. [31]	2021	FN Hematological Patients	Spain	EHR data	RF, GBM, XGBoost, GLM	AUC 0.79	MDR- <i>Pseudomonas aeruginosa</i> /ESBL-E
Henderson et al. [30]	2022	HIV patients	USA	EHR data	PLR, naïve Bayes, gradient boosting, SVM, RF	AUC 0.70	MDR-E

EHR: electronic health record, AST: antibiotic susceptibility testing, DT: decision tree, PPV: positive predictive value, NPV: negative predictive value, RF: random forest, XGBoost: eXtreme Gradient Boosting, MLR: multinomial logistic regression, MLP: multilayer perceptron, JRip (RIPPER): repeated incremental pruning to produce error reduction, Class. Regr.: a classifier using regression, k-NN: k-Nearest Neighbors, GBM: gradient boosting machine, SVM: support vector machines, GLM: generalized linear model, AUROC: area under receiver operating curve, CR: carbapenem resistance, ESBL: extended-spectrum beta-lactamase, BL: beta-lactamase, GNB: Gram-negative bacteria, MDR-E: multi-drug-resistant Enterobacterales, PLR: penalized logistic regression, FN: febrile neutropenic, HIV: human immunodeficiency virus.



Where do we go from here?

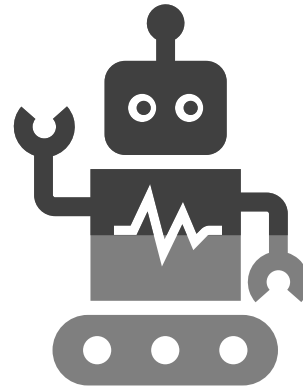
In HA, we are sitting on a territory-wide dataset that goes back to the 1990s.

We should be doing our absolute best to utilise this data to improve our care.

Clinical microbiologists



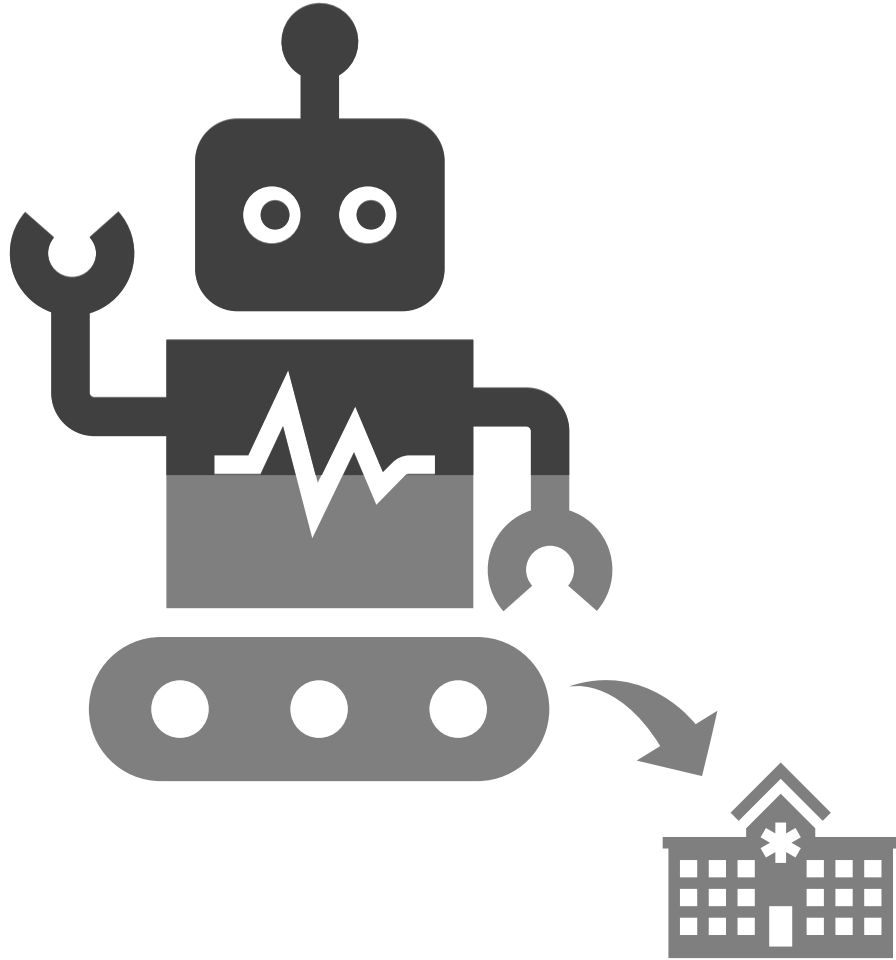
Clinical representatives



Data scientist



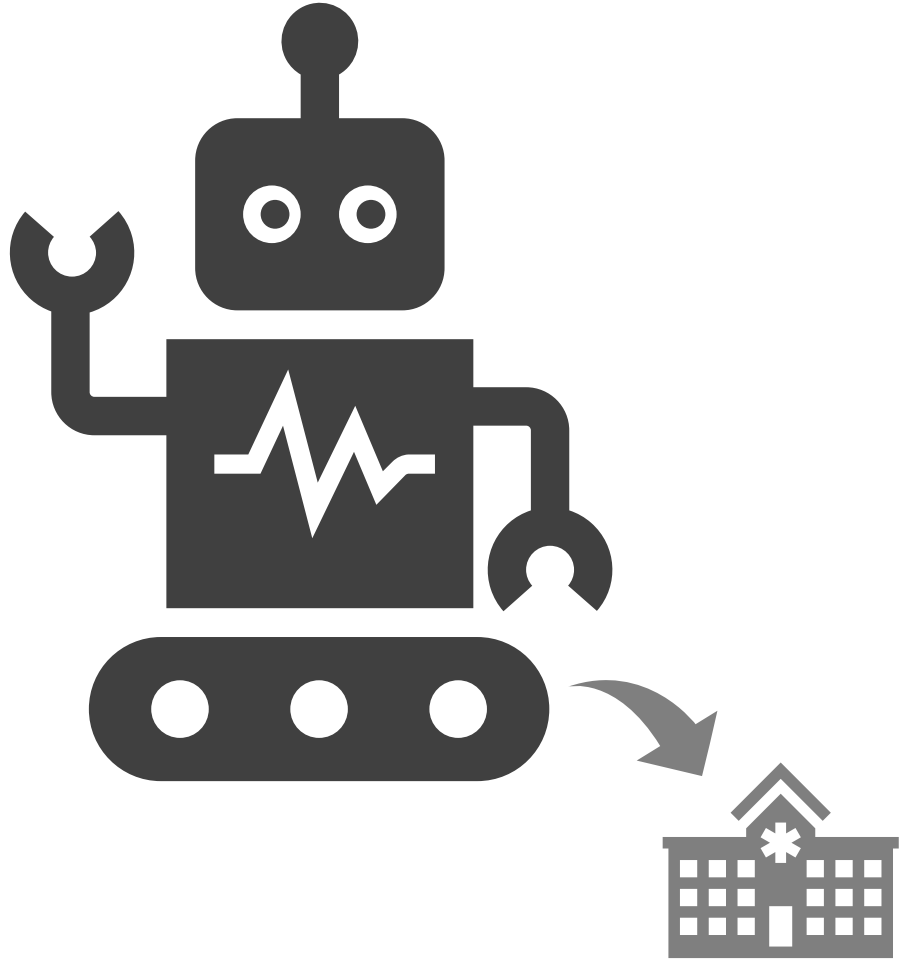
IT support



How do we make it work?

BEFORE the model

- Identify key infections that drive antibiotic use for ML prediction
 - Surveys or point prevalence study on antimicrobial use for common infections.
 - Examples – urinary tract infection, pneumonia, or undifferentiated sepsis.
- Pipelines for data engineering
 - Consistent methodology
 - Domain knowledge for encoding & interpretation of electronic data
 - Validation of data acquisition
 - Dilemmas with diagnostic codes

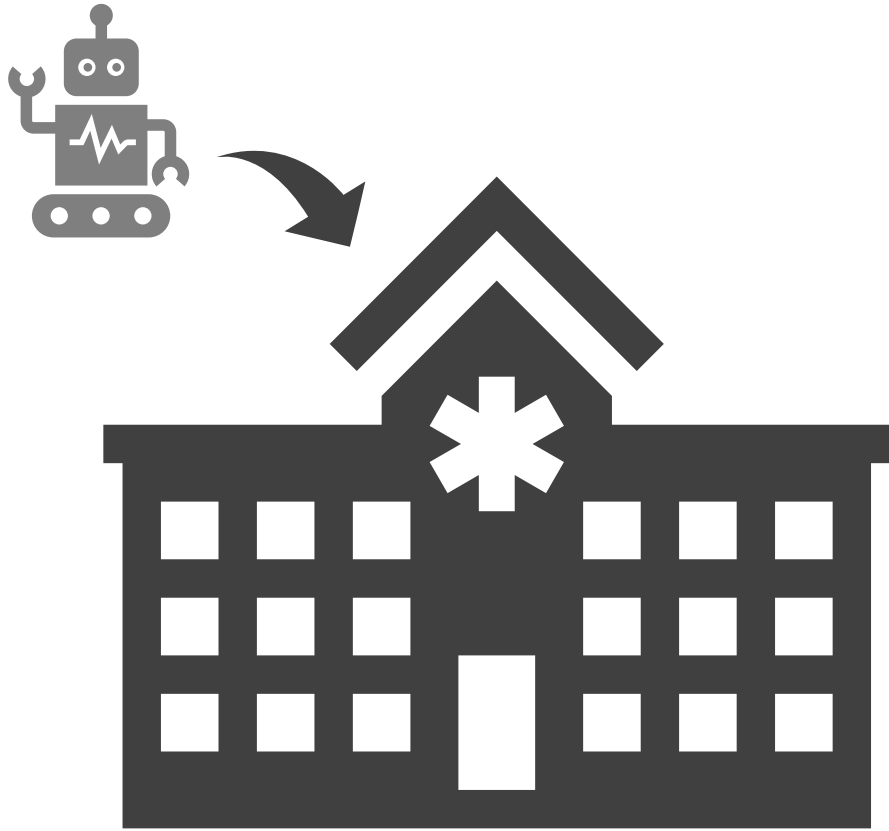


How do we make it work?

CREATING the model

- Model training and testing
 - If a single feature predominates, perhaps we don't need an AI model.
 - Simple scoring system
 - Complex integrated AI system
 - Interpretability issues

- Prospective evaluation



How do we make it work?

IMPLEMENTING the model

- Buy-in from clinicians
 - Model interpretability
 - Principles and pitfalls
- Integration to ePR/CMS
- **Ongoing monitoring of the data pipeline and model performance**

Change is constant

Monitoring data and model performance

- An example is antibiotic breakpoint.
 - CLSI changed the breakpoint for Piperacillin-tazobactam (Tazocin) in 2022.
 - Susceptible rate of *Escherichia coli* dropped from **94%** in 2021 to **82%** in 2022.
- All things change with time, and so should our predictions.
 - Laboratory reporting practice, prescription pattern, patient population, community antimicrobial resistance burden...
- To ensure patient safety, we need to regularly monitor the data and the model performance.

Acknowledgements

Hospital Authority

- Dr Raymond Lai (CICO Office)
- Dr Viola Chow (PWH Microbiology)
- Dr Ingrid Cheung (PWH Microbiology)
- Dr Ronald Chan (NTEC Pathology)
- Dr Grace Lui (PWH Infectious Disease)

Universities

- Professor Margaret Ip (CUHK Microbiology)
- Professor Christopher Lai (CUHK Microbiology)
- Mr Curtis To (CUHK Anatomical and Cellular Pathology)
- Professor Janus Wong (HKU Orthopaedics)

Graphics acknowledgement

- Icons from Microsoft 365 Powerpoint and BioRender
- Graphics from StableDiffusion XL

Machine, Machine on the wall, Who's the most Resistant of them all?

Alfred Lee MRCP(UK) DTM&H(RCP) FRCPath FRCPA

Associate Consultant

PWH Microbiology

